Josef Dick · Frances Y. Kuo
Henryk Woźniakowski *Editors*

Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan

Springer

Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan

Josef Dick • Frances Y. Kuo • Henryk Woźniakowski
Editors

Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan

Springer

Editors

Josef Dick
School of Mathematics and Statistics
University of New South Wales
Sydney, Australia

Frances Y. Kuo
School of Mathematics and Statistics
University of New South Wales
Sydney, Australia

Henryk Woźniakowski
Institute of Applied Mathematics
and Mechanics
University of Warsaw
Warsaw, Poland

Department of Computer Science
Columbia University
New York, USA

ISBN 978-3-319-72455-3 ISBN 978-3-319-72456-0 (eBook)
https://doi.org/10.1007/978-3-319-72456-0

Library of Congress Control Number: 2018936907

Mathematics Subject Classification (2010): 11K38, 41A55, 41A63, 65C05, 65D30, 65D32, 65N30, 65N35, 65R20, 65R30

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Preface

On June 17, 2018, Professor Ian Hugh Sloan will celebrate his 80th birthday. We are delighted in wishing him well on this happy occasion. Ian has been a teacher, a mentor, a research collaborator, and a very dear friend to many of us.

We decided to give Ian a special birthday present in the form of this book as a tribute to his many important contributions in various areas of computational mathematics. At this point, we wish to thank the colleagues who contributed to this book as authors and/or referees. In fact, the response of sending papers to celebrate the 80th birthday of Ian was so great and from so many colleagues that it was indeed a difficult job for us to limit the number of pages of the book. We are very grateful to Springer Verlag that they agreed from the very beginning that the number of pages of the book will not be an issue.

The book consists of nearly 60 articles written by international leaders in a diverse range of areas in contemporary computational mathematics. These papers highlight impact and many achievements of Ian in his distinguished academic career. The papers also present the current state of knowledge in such areas as quasi-Monte Carlo and Monte Carlo methods for multivariate integration, multi-level methods, finite element methods, uncertainty quantification, spherical designs and integration on the sphere, approximation and interpolation of multivariate functions, and oscillatory integrals and in general in information-based complexity and tractability, as well as in a range of other topics.

This book tells an important part of the life story of the renowned mathematician, family man, colleague, and friend who has been an inspiration to so many of us.

We believe that the best way to begin this book is by presenting a few words about Ian. We are also sure that the reader will enjoy reading the family perspectives on Ian by his wife Jan, his children Jenni and Tony, and his grandchildren Sam, Gus, Mack, Corrie, and Kiara. (Granddaughter Eliza missed the opportunity to contribute due to travelling.)

Ian Hugh Sloan was born on June 17, 1938, in Melbourne, Australia. He did his schooling at Scotch College, Melbourne, and Ballarat College. The father of Ian was a senior mathematics master at Scotch College and later Principal at Ballarat and apparently took good care of the background on mathematical education of his son.

Then Ian was educated at the University of Melbourne, where he obtained BSc in 1958 in physics and BA (hons) in 1960 in pure and applied mathematics. Ian met his future wife Jan at the University of Melbourne, and they were married in 1961. He obtained MSc at the University of Adelaide in 1961 in mathematical physics for a thesis entitled "*Ionization in Nebulae*". Ian was supervised in Adelaide by Professor Herbert Green who was one of Australia's first professors of mathematical physics. It is worth mentioning that Ian completed his master's degree in record time of 7 months. Ian received his PhD in 1964 at the University of London in theoretical physics based on the thesis "*Electron Collisions by Neutral and Ionized Helium*" and was supervised by renowned mathematical physicist Professor Sir Harrie Massey who worked on the Manhattan Project and at the Australian Woomera Rocket Range. Ian completed his PhD again in record time of just 30 months. Part of his PhD work involved computations on an early mainframe machine in Manchester.

Ian Sloan started his professional career as a research scientist for the Colonial Sugar Refining Company, during 1964–1965, in Melbourne. Since 1965 Ian Sloan has been at the University of New South Wales as a member of the School of Mathematics. He was appointed Lecturer in 1965 and became involved in research in theoretical nuclear physics. Ian had a very good start at UNSW and published 10 single authored papers in the first 5 years. He was promoted to Senior Lecturer in 1968. His research focus was shifting from theoretical physics to applied mathematics, especially towards numerical analysis, first for integral equations relevant to scattering theory and then to computational mathematics mostly of multivariate integration and approximation. He was promoted to Associate Professor in 1973 and was appointed to a personal Chair in Mathematics in 1983 and then to Scientia Professor in 1999. He served as Head of the School of Mathematics of UNSW from 1986 to 1990 and from 1992 to 1993.

Ian had many visiting positions during his career. He was associated, in particular, with (in alphabetical order) Cornell University, ESI in Vienna, Hong Kong Polytechnic University, IBM Paris, ICERM in Providence in the USA, King Fahd University of Petroleum and Minerals in Saudi Arabia, Mittag-Leffler Institute in Stockholm, Newton Institute at Cambridge, Politecnico di Torino, Technical University of Vienna, University of Bath, University of Maryland (two sabbaticals 1971–1972 and 1979–1980), University of Stuttgart, and Weierstrass Institute in Berlin.

Ian Sloan received many honours and awards during his academic career. In 1993 he was elected a Fellow of the Australian Academy of Science; in 1997 he was awarded the ANZIAM Medal of the Australian Mathematical Society; during 1998–2000 he was the President of the Australian Mathematical Society; in 2001 he received the Australian Academy of Science's Thomas Ranken Lyle Medal; in 2001 he was awarded the Centenary Medal; in 2002 he shared the inaugural George Szekeres Medal of the Australian Mathematical Society with Alf van der Poorten of Macquarie University; during 2003–2007 he was the President of the International Council for Industrial and Applied Mathematics (ICIAM); in 2005 he received the Information-Based Complexity (IBC) Prize; in the June 2008 Queen's Birthday Honours, he was appointed an Officer of the Order of Australia (AO); in 2009 he became a Fellow of the Society for Industrial and Applied Mathematics (SIAM); in 2012 he became a Fellow of the American Mathematical Society; and in 2014 he was elected a Fellow of the Royal Society of New South Wales (FRSN).

Ian Sloan has been serving on editorial boards of many international computational mathematical journals. These include *Journal of Integral Equations and Applications* (1987–2012), *SIAM Journal on Numerical Analysis* (1991–1997 and 2003–2012), *Journal of Complexity* (1999–2009 as Associate Editor and since 2009 as Senior Editor), *Numerische Mathematik* (2004–2014), *Advances in Computational Mathematics* (2000–2015), *Computational Methods in Applied Mathematics* (2000–2015), *Chinese Journal of Engineering Mathematics* (2007–), *International Journal of Geomathematics* (2011–), *International Journal for Mathematics in Industry* (2013–), and *Foundations of Computational Mathematics* (2015–).

Ian Sloan loves to work with other people. The list of his collaborators is very impressive, and many of them contributed papers to this book. He was a PhD advisor of Reginald Cahill (1971), John Aarons (1972), Ivan Graham (1980), Stephen Joe (1985), Sunil Kumar (1987), Yi Yan (1989), Thanh Tran (1994), Timothy Langtry (1995), Thang Cao (1995), Yi Zeng (1998, co-supervisor), Josef Dick (2004), Kassem Mustapha (2004, co-supervisor), Benjamin Waterhouse (2007), Paul Leopardi (2007), Jan Baldeaux (2010), Cong Pei An (2011, co-supervisor), James Nichols (2013), Andrew Chernih (2013), Yu Guang Wang (2015), Alexander Gilbert (current, co-supervisor), and Yoshihito Kazashi (current).

Ian Sloan has so far published more than 280 peer-reviewed papers in leading journals of theoretical physics and computational mathematics, book chapters, and refereed conference proceedings, as well as one book with Stephen Joe entitled *Lattice Methods for Multiple Integration* published by Oxford University Press in 1994. His papers cover various areas such as the numerical solution of integral equations, boundary integral equations, numerical integration, interpolation and approximation of multivariate functions, partial differential equations with random coefficients, and information-based complexity and tractability. The list of Ian's publications is included later in this book. He is one of a select few on the 2001 Thompson ISI list of highly cited authors.

Professor Ian Sloan has made outstanding contributions to mathematical research over the last 50 years. Ian's impact is felt widely today; the Bencze-Redish-Sloan equation and the Sloan iteration for integral equations (see the article by Thanh Tran in this monograph) have been named after him. Further key contributions were the introduction of weighted spaces and the study of tractability and inventing the component-by-component construction of lattice rules.

As many of us know, Ian loves to travel. He almost always travels with his wife Jan, and it is sometimes easier to meet Ian and Jan abroad than in Sydney. He has travelled to all parts of the world giving invited talks on his work really almost everywhere. He is making many friends during his trips and has many interesting, not always mathematical, stories to tell about his travels. More importantly, Ian does not slow down. He is as energetic and active today as he was years ago. Ian Sloan is a role model and inspiration to his friends and colleagues.

In addition to most authors of this book who also served as referees, the following people also served as referees for the book: Michael Feischl, Alexander Gilbert, Juan Gonzalez Criado del Rey, Michael Griebel, Thomas Hou, James Hyman, Stephen Joe, Pierre LÉcuyer, Klaus Ritter, Robert Schaback, Frank Stenger, Kosuke Suzuki, Mario Ullrich, Clayton Webster, Takehito Yoshiki, and Penchuan Zhang. We sincerely thank all authors and referees for their contributions.

We are also grateful to Martin Peters of Springer Verlag for his strong support of this book from the very beginning and for making it possible that every contributor receives a free copy of the book.

Josef Dick
Frances Y. Kuo
Henryk Woźniakowski

Family Perspectives

Jan Sloan (Wife)

The Ian Sloan I know is always in a hurry and never has enough time. There is always a paper to finish, a deadline to meet, someone to see, or something to do. I understand that this started early: he began school a few days after turning four because he grew impatient watching other children through the school fence. (His father was a mathematics teacher and at that time was a house master of a boarding house which happened to be next to the primary school.) A few years later, we spent our honeymoon in Adelaide so that he could put in an appearance at morning tea in the Math Physics Department, to allow him to complete a master's degree in two terms—morning tea satisfied the "minimum three-term" requirement! Then he completed a PhD in theoretical physics at the University College London in two and a half years, with his final oral exam taken in a taxi. The oral in a taxi was because his supervisor, Professor Sir Harrie Massey, was in even more of a hurry: he was an important person in the space programme at Cape Canaveral (now Cape Kennedy), so was rarely in London—indeed Ian saw him only six times during his PhD, or seven if you count the taxi. But there was another reason for urgency, namely, that Ian wanted to finish his PhD in superfast time.

I believe his PhD experience made him exceptionally self-reliant and independent. He never needs others to tell him what to do.

After a memorable two and a half years in London, we returned to Australia for Ian to take up a research position in an industrial lab (to which he owed some loyalty since they had paid him to do the PhD in London). Ian says this period of industrial research was the most miserable period of his working life, since in reality there was nothing for him to do. (His industry boss allowed him to go gracefully, saying "We have to accept that some people are not cut out for research.") After a gloomy 10 months in industrial research, he was rescued by an advertisement for a casual teaching position at the University of New South Wales. And after more than 50 years, he is still there!

The person I know is courageous and of strong character. If something needs to be said, he will say it. A choice example came early in his career at UNSW when he told his Head of Department that he should either do his job properly or resign. I suggested, but to no avail, that this was not the right way to advance his career. He is still stirring trouble, recently organising his colleagues in a collective letter of complaint to the Vice Chancellor. I can tell him that this is unwise, but it makes no difference.

He is also uncompromisingly honest. Our children will remember that they were never allowed to use his university-supplied paper and pencils.

He is always checking facts, correcting spelling and pronunciation, and finding errors—no doubt a desirable characteristic professionally, but one not always appreciated at home. Another challenging habit is that he is always reading the fine print, advertisements, and plaques on the wall, anything at all. It is a bad idea to take him food shopping, as he reads all the labels.

Surprisingly, he has a terrible memory. He often tells me of introducing himself in friendly fashion, only to get the response “I know who you are, we met yesterday”.

Many collaborators over the years, first in physics, then later in mathematics, have commented that he has great energy. That is a side that I rarely see, but it is true that he will often be at his desk sending emails in the early hours of the morning. And during our many overseas travels, I am often woken by the clicking of the computer keys when all sensible people are asleep.

Travel has always been an important part of our lives. This is especially true now that children are grown up, and I am not working and Ian is not teaching, but even in the early days, the two year-long sabbaticals we spent at the University of Maryland were important to us—and to our two children who spent formative years in American schools. Through our many travels, we have made the most beautiful and long-lasting friendships.

He is an enthusiastic reader, with a wide range of interests, and he often has five books on the go at one time. (One grandson inherited the interest, and when small always walked with a book in hand.) Both of us are keen on music. For Ian, this used to take the form, when the children were small, of playing Chopin etudes and Beethoven sonatas on the piano in the evening. We attend many concerts, and both love theatre. For us this started all those years ago in London, where for two and a half years we attended one stellar performance or another almost every week.

Ian often talks to me about his various projects, in spite of my complete lack of mathematical training and aptitude. Such talk is not about mathematical details, but rather about strategy, and about the human and intellectual struggles. And I like to know who among our friends is involved in each project.

He seems to me to change fields more often than most people. Privately, he tells me that he is something of a butterfly: he sips the nectar then moves on.

Over the years I have watched from the sidelines the evolution of many projects and papers, first in physics then later in mathematics. Often there is drama, through the struggle to get ideas worked out and papers written and accepted, and sometimes (though rarely) the pain of rejection. But I have also shared in the pleasure of the many awards and signs of recognition he has received over the years.

And I have watched him prepare many talks. It is clear to me that he takes great effort to communicate his ideas and results clearly, and I believe that his colleagues agree.

In spite of his passion for research, he has always been willing to take on administrative burdens, often against my advice. He has been Head of this and Chair of that, always with apparent success. An extreme example early in his mathematical life was taking on the Editorship of the journal that later became the *ANZIAM Journal*, at a time when the journal was in disarray, with the previous Editor having resigned, recommending that the journal be closed down. The journal still exists and is apparently doing well.

Along the way we have found time to have two children and by now six grandchildren, all of whom are just fine, and a source of great pride to both of us. Ian claims to be a good family man, in spite of never having enough time, and says we should judge by the results. I am not so sure, but will ask the family to pass their own judgement.

Jenni Johnson (Daughter)

…and judge we do…somewhat mercilessly at times!

Ian (or dad as I like to call him, despite his best attempts to level the playing field by getting us to call him Ian) is a conundrum. The man at home is a constant source of amusement, often leading the way by laughing at himself, a very endearing characteristic which I have decided to emulate, as it helps get one through otherwise difficult situations. I recall him telling us about his first attempt to use a lapel microphone for one of his lectures. As the lecture progressed and he became more and more animated, he found himself wrapped up in the cord, struggling to stay upright and keep control of the students, who were vastly amused.

He can burst with energy if there is something stimulating to argue about or discuss; arms will fly, the eyes become beady, the veins on his rather proud dome pulse with excitement, and the voice volume rises. Just as easily, he will be off to sleep in a flash if there is too much "small talk"—so it is a constant challenge for us all to keep things interesting. Some might say he has a short attention span, but I am not so sure. I think it's an efficiency measure to keep his brain agile for things that really matter.

While his brain is agile, he has very poor body awareness, and this sometimes gets him into trouble both at home and in far-flung places. He recently injured his knee playing tennis, but was only really aware of the problem the next day when he described "struggling to get out of the bed sheets". Most lesser mortals would have described pain or swelling as the first symptom and usually at the time of the accident. He has been known to walk into walls leaving a permanent nose imprint (to the delight of children and grandchildren) or walk around with socks filled with blood oblivious to a recent insult.

He manages to move around the world mixing with the best. He is adventurous and curious, keeping us well entertained with his stories of international largesse. Yes of course there is always a talk or a workshop to prepare, a paper to review, a student to meet, a keynote address, an important conference to prepare for, and the endless awards and accolades he receives for his work. However the best stories come from his adventures overseas. You may find him trapped in the mines of Chile, losing his navel in a hospital in Poland, on the shores of Lake Baikal in Russia drinking vodka out of a glass rifle, or racing uncontrollably across the sand dunes of Saudi Arabia in a four wheel drive. There is always a funny story to be told and some more friends to be collected along the way.

He adapts to his environment like a chameleon, feeling just as at home in China as in Italy, Korea, or the USA. He learns the essentials of the language and picks up a few colleagues along the way to assist with the immersion process. He is on and off planes at the drop of a hat, yet seems to bounce back to be back at work on the same day. He manages a black tie affair at the Academy of Science, a family "muck in", or is equally comfortable in his "gardening gear and matching glasses". I think this is because he is humble, extremely tolerant, and as I have heard from others "a true gentleman".

One of his most treasured abilities is multitasking. He will often be reading five books at one time, arranging travel for his next adventure, writing several papers, arranging the next season's tickets to the opera, and watching the football all at the same time. In order to get through all this, the 24 h of the day he has available must be maximally utilised; he is well known to be stomping around at 3:00 am writing down the skeletal thoughts of his next paper. He will always have paper and a pencil available readily just in case a thought needs to be captured.

I believe this extreme behaviour comes from a desire to keep learning and continues to be relevant. He strives to make a difference. I think this characteristic has been handed down to the next generations, something for which we are very grateful. The long hours and high productivity are really not a chore for him. Work and life seem to be like sweet treats in a cake shop—all to be sampled and enjoyed or passed off as an "interesting experience". (For him "interesting" is a deceptive adjective: it can well mean he didn't like it!)

So maths, science, and family are his passions, but he also enjoys music, theatre, food, wine, and sport even though this is not his personal forte. He enjoys all the children's sports and is a proud grandparent. He has recently taken up tennis again and has joined a whole series of clubs, really quite strange, as he seemed to have an intolerance of this kind of activity as a younger man. There is always a surprise in store for us!

I find it difficult to reconcile the man we know, somewhat distracted, warm hearted, generous, and humble, with the man and his stellar academic career. We are very proud to have been part of it all!

Tony Sloan (Son)

My father has an eclectic sense of humour. Rather than saying “time flies”, he’ll rush around like a mad professor screaming “tempus fugit”. In my field of accountancy, we say that small amounts of money are “immaterial”. Ian’s mathematical translation of this term is “epsilon”. So I learned about “epsilon” (very small) and “infinity” (more than very large) early in my life. When Ian received his Lyle Medal a decade or so ago, he roared with laughter, almost uncontrollably, when someone said that something was getting “ever closer to infinity”. When George Bush was talking about weapons of mass destruction, Ian was proudly holding his own weapon of math instruction—his calculator.

He gave me a tee shirt, channelling Maxwell’s equations, which he thought was hilarious. It read: “And God Said

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \cdot \mathbf{E} = \mathbf{0},$$
$$\nabla \times \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}, \qquad \nabla \cdot \mathbf{B} = \mathbf{0}.$$

And Then There Was Light”. Once he borrowed it when he was teaching Maxwell’s equations, and at the appropriate moment, in the style of Superman, he stripped off his outer shirt to reveal the message.

While most people have graduated and have proper jobs in the real world, he sticks with maths because he says it’s the only job which really counts. But I took a different point of view, and often used to ask him: “When are you going to get a real job?”

Age shall not weary him, technology shall not defeat him, and retirement is just moving ever closer to infinity for Ian. As he progresses around the bell curve of life, and slowly transforms from a vertical bar into a more cuddly version, higher honours no doubt await him, Professor to the math gods perhaps or maybe “Sir Cumference”, with the Queen’s blessing.

Sam Johnson (Grandson)

For much of my early childhood, I was confused as to what my grandfather (or, as he prefers to sign his emails to the family, Ian/Dad/Papa) actually did for his job. I had been told that he was a “very smart man who likes maths”, but much of my experiences with him seemed to involve English. For example, Papa always had a strong compulsion to drill into the heads of his grandchildren the correct spelling of “raspberry”. He would always make sure to identify the silent “p”, and would clap in delight the first time one of us spelled it correctly. (Jan adds: the word “raspberry” is special to Ian because, as I discovered early and to my great delight, he himself didn't know it contained a silent "p", and only believed me after checking the dictionary.)

Another strong image of mine as a child is going to Ian's office. As I took after my grandfather in my bookishness, he seemed to take great amusement in taking out various papers that he had been working on and seeing if I could read and pronounce the titles. While at the time (and to this day) the words I was saying appeared to be nonsense, nonetheless he made sure that by the end of the session I had come away being able to pronounce a new word.

Later in life, I came to the belief that rather than being a professor, my grandfather was actually a professional traveller, as he seemed to spend the majority of his time in a different country. You can imagine my shock when I eventually found out that he was journeying to all these exotic locales to sit with colleagues and discuss mathematical principles. I did find, however, that the trips were not quite as boring as I had imagined, as I discovered when I was invited to take a holiday with both my grandparents to South America in early 2013: there we travelled in zodiacs, gazed at a glacier calving, and saw amazing mountains.

Ian has always been an enigma to me and the other grandchildren, somehow managing to be an incredibly interesting person, despite his work making sense to nobody at all except for those he works with. One always finds out new things when having a discussion with him, as he seems to have a vast array of knowledge on a colossal number of topics.

Upon reflection, the initial description of my grandfather seems to be very apt: he is indeed a "very smart man who likes maths". However, Papa always tempers a brilliant mind and a sharp wit with a caring and loving heart, who never fails to make time for his family. I have always seen my grandfather as one who has never stopped writing, reading, loving, or pondering, and I hope he never will.

Gus Sloan (Grandson)

It was not until early 2016 that I realised how revered Ian is among his colleagues, when one of them, a publican at my part-time workplace, read the surname on my security licence. "Sloan is a very famous name at UNSW", he said to me—"but I am sure you've no interest". Don't be so quick to judge sir, as I explained I was his grandson. From then on he frequently reminded me of how intelligent Ian was, with an ability to think critically and inform others clearly.

That reminds me of when I was 13 and struggled with maths at school, and Ian would spend hours teaching me different tricks to do calculations faster, despite it being far too simple he would still give me his time. I owe much of my success in school to these early lessons, without such I probably wouldn't have succeeded in the Higher School Certificate. I still use these skills today.

He is always keen to hear of what is happening in our lives despite having a more interesting life than all of us combined. In the past year I have made a conscious effort to educate myself on difficult mathematical concepts so that I would be able to ask him questions at family meetings, such as the Banach-Tarski paradox to which I was able to provide a measly amount of information and Ian was able to clarify and make it interesting for me.

Because my grandparents are constantly travelling around the world, we only see each other a few times a year at family gatherings. But Ian never changes, always with a smile on his dial and a glass of red in his hand and those hauntingly beady eyes. He always wants to chat about his latest adventure or some political issue, except when he falls asleep in the middle of a party. His days of watching me play rugby have now ended since my playing days are behind me, but his interest in watching the Australian rugby Wallabies is still there in spades.

I look forward to the future as we both grow older and wiser, perhaps there will come a day where I am the one answering questions from him!

Mack Sloan (Grandson)

Not even the human calculator himself could have predicted how his life would turn out all these years later. Although, from a young age, he had an undoubted mathematical ability and passion, it would have been impossible for him to think that he would be travelling the world, teaching and researching what he loves, with whom he loves, only to arrive home again to be with the family he loves.

It is interesting to ponder how the small, "epsilon" actions in the past can have big implications for the present. One of these moments was when Jan and Ian started dating. Jan wanted to party on her birthday, but Ian had a very important exam the next day: he needed to be fresh, alert, and prepared. And so, he was faced with the choice: does he rest so he can perform well on the looming exam, or does he instead have a night out with Jan? He bravely declined the party invitation, survived the relationship crisis, and unsurprisingly performed the best in the class for the exam. But what would have happened if he had gone to the party? Would he still be a mathematician? Would he have met someone else other than Jan at the party? Would he have had the privilege to travel the world? Would he come to live a happy life with a beautiful wife, two children, and six grandchildren?

Much like Ian, the butterfly effect is complex, yet very interesting. Not until I stepped back and looked into Ian's accomplishments did I begin to understand how much Ian has impacted my life, and our families. The decisions he chose to make, and the ones he chose not to make, have turned out to be good ones and to have had a positive impact on each of our lives.

I don't know where Ian is travelling half the time, or what area of maths he is in, but it seems to me that he is living a life that many could only dream of. Through his commitment and dedication to both academia and Jan, he is truly a gentleman. As I grow older and wiser, and I reflect on my work and study habits, I realise that my ability to focus and my modus operandi are directly due to Ian. While I may never have his brain power, I can see more clearly every day the enormous impact Ian, my

Papa, has had on my life in terms of problem solving and deep thinking. We are all very proud to be his grandchildren and to enjoy his tales and good humour.

Corrie Sloan (Grandson, Aged 12)

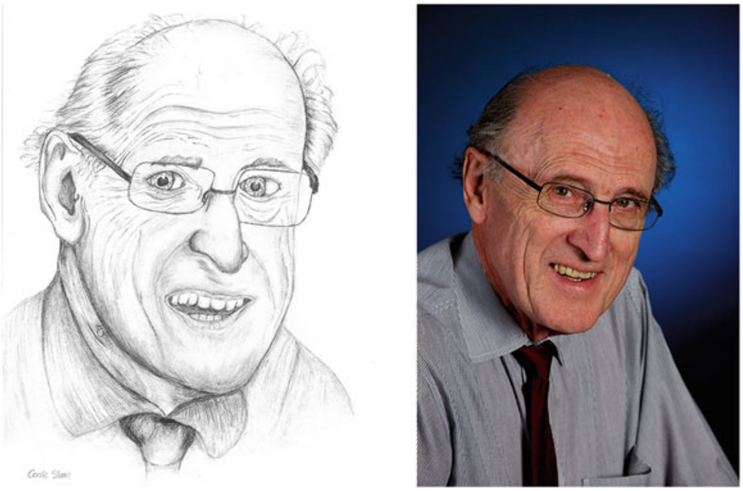
When I was young, I wondered what job my grandfather, Ian, did. I always thought it was strange that my Dad graduated before him, but I was continually reminded that Ian works at the university, so doesn't need to leave. I remember that everyone told me that my grandfather (Papa) was so smart and that he was a scientist, so I pulled out my iPad and googled Ian Sloan. I saw hundreds of pictures of him, but the strangest one of all was of him in his well-cut garden drinking his famous red wine. I still have no idea how he got that picture onto the Internet. Whenever I go to Papa's for a sleepover, the first thing I see is him on his hands and knees working in the garden, picking up some sticks or digging up a plant. But whenever I see him, he's always got a jolly smile and rosy cheeks ready to greet me. Papa makes the best home-made creamy and delicious porridge I have ever tasted. Whenever I am sleeping over, his porridge is the one thing to get me up in the morning.

Whenever Papa has the chance to come and watch me play sport or come to grandparents' day at school, he is there and never in a hurry. I remember the time when Papa and Janny came to one of my representative rugby games, and I scored two tries in the corner where they were standing. I looked up at Papa and Janny and they both clapped and smiled right at me.

When my sister, Kiara, and I go over to Papa's place, he always takes us on an adventure into the bushland. My Dad tells me that when he was little he used to always go into the bush where we go with Papa. He said one time he went to his favourite rock to rest on, slept for about 10 min, and then woke up but realised he was looking right into the eyes of a brown snake. My Dad stayed still for a few minutes and the snake went away. I always ask Papa if he's seen a brown snake but he says no. When we are in the bush, we always see lots of animals rustling in the bushes and birds chirping in the trees. I always enjoy walking through the bush with my grandfather, Papa.

When we have family gatherings at Papa's house, it is always a day to remember. With Papa's great preaching and Janny's great cooking, it always turns out fun. All the grandchildren go outside and play touch footy or kick tennis balls, while the adults stay inside and Papa talks about some political problem. I don't think any of us has seen Papa without a glass of red wine in his grasp or a smile on his face. He always makes us laugh.

Papa is a loving grandfather, a wordsmith, a preacher, a funny clown, and a math God. I look forward to the time when I learn all of my grandfather's tricks and math skills so I too can take part in the mathematical and political discussions that happen inside Papa's house.



Corrie’s sketch of Papa and a photo taken in 2011

Kiara (Button) Sloan (Granddaughter, Aged 9)

I’m always asking Papa math questions, and he always answers them correctly, I think! I’m not sure because he says things in completely different ways—since he’s a University Professor, he says them in a university way instead of a Year 4 context. So I ask, what does that mean? what does that mean? Over and over.

Since Papa is such a genius, he always needs to go to meetings about math stuff, and those meetings are ALWAYS in another state or country. But it makes it all the merrier when Papa and Janny come back and tell us all about their trip.

Papa always takes us on awesome bush walks like the last time Corrie and I were at Papa’s house, Papa took us on a bush walk around a lake. He said if we kept walking we could get to Newcastle, hundreds of kilometres away. I’m so pleased we didn’t end up walking to Newcastle!

Papa comes to nearly all my dance and singing concerts and performances. The only time he doesn’t come is when he is overseas on a work trip. But I love it when Papa and Janny get back and can come to my performances and see how much I have improved.

Papa is the best grandfather any girl could ask for. He is kind and funny, always has a huge smile, and is jolly good at making porridge. I think if he opened up a porridge shop he would be famous. He’s the best Papa anyone could be, and I wouldn’t want to change one single thing about him.

Family

We want to thank sincerely all of those who have contributed to this memorable celebration of Ian's 80th. We especially thank Henryk, Frances, and Josef and all of those who have contributed to this volume.

Jan Sloan and Family

A Fortunate Scientific Life

In the course of a long career, my scientific directions have seen many changes. I am often asked: why so many changes? To me there is often no great change: there is always a connecting thread. But what is also true is that I have grasped every opportunity to learn about something new and interesting. And over the years, I have been helped to move in new directions by many wonderful friends and collaborators. In large part this essay is a homage to those friends and collaborators.

But in truth my early research years were much more solitary. My PhD research at the University College London was concerned with the theory and computation of the scattering of electrons by atoms. There were experts about in the Department, but there was little tradition of collaboration or communication. Indeed I recall being less than amused to find out, well into the PhD research, that another student had been given a project overlapping very significantly with my own. That gave extra incentive, if any were needed, to finish my PhD quickly.

My first significant publication, from those PhD years, was [157], a paper appearing in the Proceedings of the Royal Society (something for a young researcher to be proud of) concerned with an improved method for computing scattering cross sections for electrons hitting upon simple atoms.

One very fortunate aspect of my early career was its timing, in that the first general purpose electronic computer available to university researchers in the UK (a big beast of a machine at the University of Manchester) had just become available. (This was fortunate because until that time a PhD would typically include 6 months of laborious hand calculations to solve numerically one simple integro-differential equation.) In that new era, students were able to write Fortran programmes on paper tape, to be transferred to Manchester overnight. We learned to be experts at patching paper tape.

After returning to Australia, and an unhappy year in an industrial research laboratory, I joined the Applied Mathematics Department at the University of New South Wales. At that time it was in everything but name a theoretical physics department, under the leadership of John M. Blatt, who after a very distinguished career in nuclear physics was beginning to dabble in other areas, including control theory and economics. Probably it was under his influence that I started to work on scattering problems in nuclear physics, especially quantum mechanical problems involving a small number of protons or neutrons. At first the going was tough, because when I began I knew nobody working in the field, either in Australia or anywhere else in the world.

Nevertheless, I managed to make some progress and attract some attention in low-energy scattering problems in nuclear physics. An important milestone for me came through an invitation in the early 1970s to spend a sabbatical year at the University of Maryland. I remain grateful to my host, Gerry Stephenson (later at Los Alamos National Laboratory). In the nuclear theory group there, I learned for the first time about the stimulating effect of a strong group environment, a lesson I have since taken to heart. My friend E.F. (Joe) Redish was an enthusiastic member of that group.

For me that year at Maryland was highly productive. In those years the Faddeev equations (devised by the great Russian physicist L.D. Faddeev) were attracting great interest. They are a beautiful set of integral equations which allow the nuclear three-body problem (e.g., the problem of two neutrons and one proton, or equivalently of a deuteron, i.e., heavy hydrogen, nucleus and an incident neutron) to be solved on a computer essentially exactly. One significant fruit of my time at Maryland was the paper [170], in which I developed analogous equations for four particles, rather than the three particles of the Faddeev equations. While an increase from three to four might not seem much, the problem does become harder, not only mathematically but also computationally—it is still hard to obtain computationally exact solutions for more than three particles.

Later the four-particle equations were generalised to any number of particles, independently by both D. Bencze and Joe Redish, to make what are still recognised as the Bencze-Redish-Sloan equations. But by then my own interests had moved elsewhere.

In those physics years, the principal mathematical tools for those of us working on scattering problems were integral equations. There were many good idea floating around, but I often felt that those good ideas were accompanied by a cavalier attitude to the question of proof. I liked the fact that in physics one can take space to explain the ideas behind an approximation scheme, but sometimes this was at the cost of a simple proof. Still, I enjoyed my time in this area.

For me a turning point came in the mid-1970s when, for reasons I no longer remember, I decided to develop and write up some of the ideas for solving integral equations for publication in a numerical analysis journal. It was something of a shock to me (knowing as I did the more relaxed publication standards in physics) to find my paper in trouble with a referee. (The referee later turned out to be L.M. (Mike) Delves, who before my time had been a staff member at the University of New South Wales, and indeed whose old golf clubs were left behind in my first office at UNSW.) The referee thought there were some good ideas, but noted (very appropriately, as I now think) that there were no proofs and, more importantly from my point of view, wanted the paper rewritten in a way that would (in my view) have buried the intuition behind the method. That rewriting was something I was not willing to do, so in the end the paper appeared as [265] in the *Journal of Computational Physics*.

That experience made me determined to prove the merit of my ideas, but to do that I had first to master the theory of the numerical solution of integral equations, which I did through the books of my later friends Philip Anselone and Ken Atkinson.

In the end I was able to prove the merit of what is now often called the "Sloan iteration" for integral equations of the second kind. In brief, the essence of a second-kind integral equation is that the unknown solution appears on both the left and right sides of the defining equation. (On the left the solution is on its own; on the right it appears under an integral sign.) It is therefore natural, if one already has some approximation to the solution, to substitute that approximation into the right-hand side of the equation, and so generate another approximation, hopefully a better one. I was eventually able to prove, under more or less natural conditions, that if one starts with a so-called Galerkin approximation (which I will not explain!), then the iterated Galerkin approximation generated in that way always converges faster than the original Galerkin approximation as the dimension of the approximating space increases. A dramatic description is that the new iterated approximation is "superconvergent".

The original superconvergence work appeared principally in the references [174, 176]. To me there was special pleasure in the fact that the second of those papers appeared in the journal *Mathematics of Computation*, the very journal that had (deservedly!) given my first venture such a hard time.

By the time that particular interest had been worked through, I discovered that I had somehow drifted out of physics, essentially because I was too busy elsewhere. But there were challenges for me in moving into numerical analysis, because I was no longer in the first flush of youth, yet at the time I made my unsuccessful submission as above I knew not a single person in the world in the field of numerical analysis. Admittedly there were Australian experts, some even in the area of integral equations (I am thinking of David Elliott, Bob Anderssen, Frank de Hoog, and Mike Osborne, all of them later good friends), but at that time I am sorry to say that I had never heard of any of them. This was in essence because of the remarkable cultural separation that exists between physics and mathematics: in the main they publish in different journals, attend different conferences, and almost never meet each other. For me, I recall that until 1975 I had never attended a conference with mathematics in the title and since my undergraduate days had never been in a mathematics department other than my own.

But all of that was to change quickly. In (I think it was) 1976 I was an invited speaker at the Australian Applied Mathematics Conference. Within a year or two of that, I was Editor of the Australian journal of the applied mathematicians (now the *ANZIAM Journal*) and in subsequent years have been heavily involved, to my pleasure, in all aspects of Australian and indeed international mathematics.

In particular, I find it pleasing to be able to report that this not-so-young new boy was accepted remarkably quickly into the professional community of numerical analysts. Many people helped in this—Philip Anselone and Kendall Atkinson certainly, also the great Ben Noble who happened to be at my lecture on "New

Methods for Integral Equations" at the Australian Applied Mathematics Conference, Zuhair Nashed who invited me to a meeting of the American Mathematical Society, Günther Hämmerlin, who invited me to more than one Oberwolfach conference, and many others, to all of whom I am forever grateful.

In the early 1980s, I first met Vidar Thomée, a renowned authority on the numerical solution of parabolic problems (e.g., the time-dependent problem of heat diffusion in some region). In the course of time, we published together some six papers, with him patiently teaching me the modern theory of partial differential equations and me contributing, I hope, some insight into integral equations, numerical integration, and superconvergence. Our first joint paper [241] was concerned with superconvergence for integral equations, but with the analysis informed by the analysis of finite element methods for partial differential equations. In addition to collaborating on papers, Thomée also introduced me to the whole new world of finite element methods, a world that included subsequent friends Lars Walhlbin and Al Schatz at Cornell, who took me further into the world of PDE and superconvergence [152].

In the early 1980s, I became interested in high-dimensional numerical integration, something that has become a major theme for me in recent years. It came about this way. While visiting old physics friends at Flinders University in Adelaide, the late Ian McCarthy introduced me to the amazing number-theoretic methods of Korobov and others, which they were trying out experimentally in their atomic physics codes. I had never heard of these things and was immediately captivated by them. But early on I had the idea that they could be extended from the classical constructions (nowadays called lattice rules of rank one) to more general constructions, so I applied, successfully, for an Australian Research Council grant to work on such a generalisation, which led eventually to the publication [224]. The classical constructions of point sets were to my eyes like crystal lattices, in that they are unchanged under special overall shifts (or translations), if you think of the point set as extended indefinitely. That paper, with my postdoc Philip Kachoyan, for the first time allowed the points to be any set which is invariant under translations. In developing a theory for such general lattice rules, I could take advantage of earlier experience in teaching the theory of group representations in quantum mechanics, and was also helped by looking back at my old books on solid-state physics, where the "dual lattice" plays an important role (e.g., in the scattering of X-rays from crystals). Incidentally, that is another paper which experienced great troubles in the refereeing process, but no doubt in the end the paper was all the better for it.

In 1984 I had an opportunity to present ideas on lattice methods for numerical integration at an international congress in Leuven. There I met James Lyness, who suggested that we could work together on this topic for 10 years and expect to publish one paper per year. James (regrettably now deceased) was a colourful personality, and a renowned expert on numerical integration, with whom as it happens I had shared an office at UNSW in my first year or two. While we never reached our 10 papers, we did some interesting work together, and in particular in [225] managed to classify all lattice rules according to "rank", a new concept at the time.

In 1987 I spent an important sabbatical at the University of Stuttgart with Wolfgang Wendland. Under his influence I became more interested in what are called boundary integral equations. (In brief, these might be described as ways of solving certain differential equation for a bounded region by an integral equation that lives only on the boundary.) For a number of years, this was a major preoccupation. One aspect was the development of "qualocation" methods (a quadrature-based extension of collocation methods), notably in a paper with Wendland, [248].

I have always been interested in the methods of approximation, which is the foundation subject for all work on the approximate solution of differential equations, integral equations, and numerical integration. In the early 1990s, having often used polynomial approximations for integral equations and numerical integration, I became interested in questions of polynomial approximation on spheres and other manifolds. I was intrigued by this conundrum: that whereas approximation of a periodic function on an interval (or what is equivalent, for a function on a circle), the approximation known as interpolation (which just means fitting a (trigonometric) polynomial through the function values at equally spaced points), has properties as good as those of the more famous orthogonal projection, yet for spheres of dimension more than one this is not the case: indeed interpolation on spheres remains to this day very problematic; see [286]. Yet a discrete approximation with the right properties (but admittedly using more points than interpolation) is available. This approximation, now called "hyperinterpolation", appeared in the paper [203]. I must say I was rather proud of this paper (I remember presenting it in an animated way to the significant mathematician Werner Reinboldt on the only occasion I met him. He urged me to find a name different from interpolation: hence the name "hyperinterpolation"). That paper had more trouble getting past the referees than almost any other paper I have written. Nevertheless it continues to attract some interest.

I began working with Robert Womersley in the late 1990s, often on polynomial approximation and point distribution problems on spheres. This has been an extremely fruitful partnership, in which we each bring different attributes: from me analysis and from him very high level skills in optimization and high performance computing. One early fruit was the paper [253], in which we proved that the hyperinterpolation approximation can be optimal in a space of considerable practical importance (that of continuous functions).

My interest in approximation has been invigorated from time to time by stumbling across fascinating ideas well known to others but not to me, such as "radial basis functions" and more recently "needlets". On the first of these, I had the opportunity to work with Holger Wendland, an authority on RBFs and the inventor of a special class of localised RBFs that carry his name. With him we were able to develop a successful theory for Wendland RBFs of progressively smaller scale: the idea was to use thinner and thinner RBFs (but correspondingly more and more of them) to get successive corrections to an initial approximation. Of the several papers we wrote on this topic, I especially like our recent "Zooming-in" paper [130]. (It is not as recent as it seems, having been lost for 4 years in the refereeing process.) In this project, as in many other projects over the years, my former student Q. Thong Le Gia was a valued participant. I said former "student", but actually Thong only did an undergraduate (honours) degree at UNSW before heading to the USA; nevertheless he did his undergraduate work well enough to result in a joint paper [122], in which hyperinterpolation was extended to spheres in an arbitrary number of dimensions. Needlets, invented by Joe Ward and Fran Narcowich (now friends), are something similar, but are localised (and "spiky") polynomials. A recent paper [284] developed a fully constructive theory of needlet approximation, one that needs only function values at discrete points on the sphere. Other participants in that project were Robert Womersley and my recent student and continuing collaborator Yu Guang Wang.

Another recent influence in the broad area of approximation theory has been Edward Saff, an expert on potential theory and on energies and point distributions on the sphere. By combining different areas of expertise, we were able (together with Robert Womersley and Johann Brauchart) to come up with a new concept (of so-called QMC designs) of point distributions on the sphere; see [29]. (The essence is that instead of characterising point distributions by geometrical properties, or e.g., by minimal energy, now sequences of point sets are characterised by asymptotic convergence properties.) Time will tell how useful this concept is.

I remained interested in high-dimensional integration problems and, around 1980, was invited by Oxford University Press to write book on the subject. At the time I had an excellent colleague (and former student) Stephen Joe. We delayed writing a book on lattice methods until we thought we had the subject wrapped up (though it turned out we were quite wrong—the topic was far from finished). The book eventually appeared as [1] and remains a useful reference to the classical theory.

In 1994 I had the good fortune of meeting Henryk Woźniakowski, a world leader in the field of information-based complexity. We soon started to work seriously on problems of high-dimensional integration. At that time I was familiar with the work in the 1950s and 1960s of the number theorists, on lattice and other methods for integration in many variables. They typically worked with an arbitrary number of variables, but (like the numerical analysts of the time) paid little or no attention to the way the accuracy reduces, or the cost increases, as the number of variables increases: the sole interest was instead in what happens as the number of function-evaluation points increases. Henryk, in contrast, always asked: what happens as the number of variables increases? It turned out that this was a very good question. In our first paper [259], we were able to show something surprising, that in one of the most popular theoretical settings of the number theorists, while there was no flaw in their predictions of the rate of convergence for a large enough number of function values, in the worst case there would be no improvement until the number of points was impossibly large. (Technically, the required number of points was approximately 2 raised to the power of the number of dimensions. Try it out for 100 dimensions!)

In the second paper with Woźniakowski, we took seriously an idea that for problems with large numbers of variables, those variables might not be equally important. We thought that if the variables are ordered in order of importance, we should be able to quantify that decreasing importance by assigning to each variable a parameter (a "weight"), with the weights becoming progressively smaller. We were able to carry through that programme, to the point of being able, for an important setting, to characterise completely the condition on the weights needed to get a result independent of dimension, that is, of the number of variables. (Technically, we were able to find a necessary and sufficient condition for the worst case error to be independent of the dimension.) That work, appearing in [260], gave us much pleasure and has been the foundation for a large amount of subsequent work by ourselves and others.

I have been privileged for the past two decades to have two outstanding young colleagues in high-dimensional computation, Frances Kuo (a student of Stephen Joe, and hence my doctoral granddaughter) and Josef Dick (who earlier came from Salzburg to be my PhD student). Together we wrote a major review [52] of certain methods for high-dimensional integration. I have had many good students, but Josef was exceptional, in that shortly before he was due to submit his thesis he abandoned the work already done and wrote a new thesis on novel joint work with Friedrich Pillichshammer. He became the teacher and I the student.

Our interest in high-dimensional problems led us a few years ago into a joint project with an Australian merchant bank. (Many problems in mathematical finance are high-dimensional because they involve many, even infinitely many, random variables.) I think the truth is that we made no contribution to the bank's bottom line, but the experience had a lasting influence on our subsequent research, because none of our high-dimensional theories can explain the apparent success of some of our methods for so-called option pricing. (The problem with options is that their value is considered to be zero if the final price drops below an agreed "strike price". For that reason the functions that need to be integrated have a kink, which places them outside almost all existing theory.) With Frances Kuo and Michael Griebel, we made some progress in finding theoretical answers to this conundrum in [69], and at the moment we are proposing a practical cure in joint work with Hernan Leövey and Andreas Griewank.

Many years earlier, my first PhD student after my physics days was Ivan Graham, who came from Northern Ireland. In 2007, as a Professor at the University of Bath, he directed my attention to the field of partial differential equations (PDEs) with random coefficients, as a burgeoning source of very challenging high-dimensional problems. Christoph Schwab from ETH Zurich was another who directed my interest in that direction. Such PDE application has become a major interest for all of us in subsequent years. I especially like the experimental paper [65] and the theoretical paper [116], both with Frances Kuo and with Ivan Graham and colleagues Dirk Nuyens and Robert Scheichl in the first paper and Christoph Schwab in the second. This work continues.

Is there a consistent theme? Perhaps there is, to the extent that many problems I have worked on are governed by the question of what can be done, and proved, when the underlying problems of physics and mathematics are intrinsically infinite-dimensional, yet our computations can use only a finite amount of information (e.g., of function values at points) and a limited amount of computer resources. But the truth is that I have always found that one interesting problem leads to another, and I have just done whatever I have found interesting and the things my colleagues have helped me to do.

In the limited space of this essay, I cannot do justice to my more than 100 lifetime collaborators and my many students and postdoctoral fellows. Among the influential collaborators not mentioned already are Mark Ainsworth, Xiaojun Chen, Ronald Cools, Mahadevan Ganesh, Michael Giles, Rolf Grigorieff, Rainer Kress, Kerstin Hesse, Fred Hickernell, Hrushikesh Mhaskar, Harald Niederreiter, Philip Rabinowitz, Alastair Spence, Sergei Pereverzyev, Siegfried Prössdorf, Ernst Stephan, Xiaoqun Wang, and Grzegorz Wasilkowski. I am grateful to all of them, and more, for the lessons they have taught me and for the wonderful ideas to which they have introduced me.

Ian H. Sloan

Publications of Professor Sloan

Books

1. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press, Oxford (1994)

Edited Books, Special Issues, and Conference Proceedings

2. Dahmen, W., Geronimo, J., Xin, L., Pritsker, I., Sloan, I., Lubinsky, D. (eds.): Special volume on constructive function theory. Electron. Trans. Numer. Anal. **25** (2006)
3. Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.): Monte Carlo and Quasi-Monte Carlo methods 2012. In: Springer Proceedings in Mathematics and Statistics, vol. 65 (2013)
4. Jeltsch, R., Li, T., Sloan, I.H. (eds.): Some Topics in Industrial and Applied Mathematics. Higher Education Press/World Scientic, Beijing (2007)
5. Sloan, I.H., Novak, E., Woźniakowski, H., Traub, J.F. (eds.): Essays on the Complexity of Continuous Problems. European Mathematical Society, Zurich (2009)

Journal Articles, Book Chapters, and Conference Proceedings Papers

6. Aarons, J.C., Sloan, I.H.: Krauss-Kowalski calculations of nucleon-deuteron polarization. Phys. Rev. C - Nucl. Phys. **5**, 582–585 (1972)
7. Aarons, J.C., Sloan, I.H.: Vector and tensor polarizations in nucleon-deuteron scattering. Nucl. Phys. Sect. A **182**, 369–384 (1972)
8. Adhikari, S.K., Sloan, I.H.: Method for three-body equations. Phys. Rev. C **12**, 1152–1157 (1975)
9. Adhikari, S.K., Sloan, I.H.: Separable expansion of the *t*-matrix in the $^3S_1$-$^3D_1$ channel. Nucl. Phys. Sect. A **251**, 297–304 (1975)
10. Adhikari, S.K., Sloan, I.H.: Separable expansion of the *t* matrix with analytic form factors. Phys. Rev. C **11**, 1133–1140 (1975)

11. Adhikari, S.K., Sloan, I.H.: Separable operator expansions for the *t*-matrix. Nucl. Phys. Sect. A **241**, 429–442 (1975)
12. Ainsworth, M., Kelly, D.W., Sloan, I.H., Wang, S.: Post-processing with computable error bounds for the finite element approximation of a nonlinear heat conduction problem. IMA J. Numer. Anal. **17**, 547–561 (1997)
13. Ainsworth, M., Grigorieff, R.D., Sloan, I.H.: Semi-discrete Galerkin approximation of the single layer equation by general splines. Numer. Math. **79**, 157–174 (1998)
14. Amini, S., Sloan, I.H.: Collocation methods for second kind integral equations with non-compact operators. J. Integral Equ. Appl. **2**, 1–30 (1989)
15. An, C., Chen, X., Sloan, I.H., Womersley, R.S.: Well conditioned spherical designs for integration and interpolation on the two-sphere. SIAM J. Numer. Anal. **48**, 2135–2157 (2010)
16. An, C., Chen, X., Sloan, I.H., Womersley, R.S.: Regularized least squares approximations on the sphere using spherical designs. SIAM J. Numer. Anal. **50**, 1513–1534 (2012)
17. An, C., Chen, X., Sloan, I.H., Womersley, R.S.: Erratum: Regularized least squares approximations on the sphere using spherical designs (SIAM Journal on Numerical Analysis 50, 1513–1534 (2012)). SIAM J. Numer. Anal. **52**, 2205–2206 (2014)
18. Anselone, P.M., Sloan, I.H.: Integral equations on the half line. J. Integral Equ. **9**, 3–23 (1985)
19. Anselone, P.M., Sloan, I.H.: Numerical solutions of integral equations on the half line I. The compact case. Numer. Math. **51**, 599–614 (1987)
20. Anselone, P.M., Sloan, I.H.: Numerical solutions of integral equations on the half line II. The Wiener-Hopf case. J. Integral Equ. Appl. **1**, 203–225 (1988)
21. Anselone, P.M., Sloan, I.H.: Spectral approximations for Wiener-Hopf operators. J. Integral Equ. Appl. **2**, 237–261 (1990)
22. Anselone, P.M., Sloan, I.H.: Spectral approximations for Wiener-Hopf operators II. J. Integral Equ. Appl. **4**, 465–489 (1992)
23. Atkinson, K.E., Sloan, I.H.: The numerical-solution of 1st-kind logarithmic-kernel integral-equations on smooth open arcs. Math. Comput. **56**, 119–139 (1991)
24. Atkinson, K., Graham, I., Sloan, I.: Piecewise continuous collocation for integral-equations. SIAM J. Numer. Anal. **20**, 172–186 (1983)
25. Brady, T.J., Sloan, I.H.: Padé approximants and nucleon-deuteron scattering. Phys. Lett. B **40**, 55–57 (1972)
26. Brady, T.J., Sloan, I.H.: Variational approach to breakup calculations in the Amado model. In: Slaus, I., Moszkowski, S.A., Haddock, R.P., van Oers, W.T.H. (eds.) Few Particle Problems in the Nuclear Interaction, pp. 364–367. North Holland/American Elsevier, Amsterdam (1972)
27. Brady, T.J., Sloan, I.H.: Variational calculations of 3-body amplitudes. Bull. Am. Phys. Soc. **18**, 18–18 (1973)
28. Brady, T.J., Sloan, I.H.: Variational method for off-shell three-body amplitudes. Phys. Rev. C **9**, 4–15 (1974)
29. Brauchart, J.S., Saff, E.B., Sloan, I.H., Womersley, R.S.: QMC designs: Optimal order Quasi Monte Carlo integration schemes on the sphere. Math. Comput. **83**, 2821–2851 (2014)
30. Brauchart, J.S., Dick, J., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces. J. Math. Anal. Appl. **431**, 782–811 (2015)
31. Brauchart, J.S., Reznikov, A.B., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Random point sets on the sphere—hole radii, covering, and separation. Exp. Math. **2016**, 1–20 (2016)
32. Brown, G., Chandler, G.A., Sloan, I.H., Wilson, D.C.: Properties of certain trigonometric series arising in numerical analysis. J. Math. Anal. Appl. **162**, 371–380 (1991)
33. Cahill, R.T., Sloan, I.H.: Neutron-deuteron breakup models. Phys. Lett. B **33**, 195–196 (1970)
34. Cahill, R.T., Sloan, I.H.: Neutron-deuteron breakup with Amado's model. In: McKee, J.S.C., Rolph, P.M. (eds.) The Three-Body Problem, pp. 265–274. North Holland, Amsterdam (1970)
35. Cahill, R.T., Sloan, I.H.: Neutron-deuteron scattering with soft-core. Phys. Lett. B **31**, 353–354 (1970)

36. Cahill, R.T., Sloan, I.H.: Theory of neutron-deuteron break-up at 14.4 mev. Nucl. Phys. Sect. A **165**, 161–179 (1971)
37. Cahill, R.T., Sloan, I.H.: The *n-d* initial-state interaction in *n-d* break-up. Nucl. Phys. Sect. A **194**, 589–598 (1972)
38. Cao, H.T., Kelly, D.W., Sloan, I.H.: Post-processing for pointwise local error bounds for derivatives in finite element solutions. In: Design, Simulation and Optimisation Reliability and Applicability of Computational Methods, pp. 25–36. Universität Stuttgart, Stuttgart (1997)
39. Cao, H.T., Kelly, D.W., Sloan, I.H.: Local error bounds for post-processed finite element calculations. Int. J. Numer. Methods Eng. **45**, 1085–1098 (1999)
40. Cao, H.T., Kelly, D.W., Sloan, I.H.: Pointwise error estimates for stress in two dimensional elasticity. In: ACAM 99 2nd Australasian Congress on Applied Mechanics. ADFA, Canberra (1999)
41. Cao, H., Pereverzyev, S.V., Sloan, I.H., Tkachenko, P.: Two-parameter regularization of ill-posed spherical pseudo-differential equations in the space of continuous functions. Appl. Math. Comput. **273**, 993–1005 (2016)
42. Chandler, C., Sloan, I.H.: Addendum: Spurious solutions to *N*-particle scattering equations. Nucl. Phys. Sect. A **361**, 521–522 (1981)
43. Chandler, G.A., Sloan, I.H.: Spline qualocation methods for boundary integral equations. Numer. Math. **58**, 537–567 (1990)
44. Chandler, G.A., Sloan, I.H.: Spline qualocation methods for boundary integral equations. Numer. Math. **62**, 295 (1992)
45. Chernih, A., Sloan, I.H., Womersley, R.S.: Wendland functions with increasing smoothness converge to a Gaussian. Adv. Comput. Math. **40**, 185–200 (2014)
46. Cools, R., Sloan, I.H.: Minimal cubature formulae of trigonometric degree. Math. Comput. **65**, 1583–1600 (1996)
47. de Hoog, F., Sloan, I.H.: The finite-section approximation for integral-equations on the half-line. J. Aust. Math. Soc. Ser. B Appl. Math. **28**, 415–434 (1987)
48. Dick, J., Sloan, I.H., Wang, X., Woźniakowski, H.: Liberating the weights. J. Complex. **20**, 593–623 (2004)
49. Dick, J., Kuo, F.Y., Pillichshammer, F., Sloan, I.H.: Construction algorithms for polynomial lattice rules for multivariate integration. Math. Comput. **74**, 1895–1921 (2005)
50. Dick, J., Sloan, I.H., Wang, X., Woźniakowski, H.: Good lattice rules in weighted Korobov spaces with general weights. Numer. Math. **103**, 63–97 (2006)
51. Dick, J., Kritzer, P., Kuo, F.Y., Sloan, I.H.: Lattice-Nyström method for Fredholm integral equations of the second kind with convolution type kernels. J. Complex. **23**, 752–772 (2007)
52. Dick, J., Kuo, F.Y., Sloan, I.H.: High dimensional numerical integration—the Quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
53. Disney, S., Sloan, I.H.: Error-bounds for the method of good lattice points. Math. Comput. **56**, 257–266 (1991)
54. Disney, S., Sloan, I.H.: Lattice integration rules of maximal rank formed by copying rank-1 rules. SIAM J. Numer. Anal. **29**, 566–577 (1992)
55. Doleschall, P., Aarons, J.C., Sloan, I.H.: Exact calculations of *n-d* polarization. Phys. Lett. B **40**, 605–606 (1972)
56. Elschner, J., Prössdorf, S., Sloan, I.H.: The qualocation method for Symm's integral equation on a polygon. Math. Nachr. **177**, 81–108 (1996)
57. Elschner, J., Jeon, Y., Sloan, I.H., Stephan, E.P.: The collocation method for mixed boundary value problems on domains with curved polygonal boundaries. Numer. Math. **76**, 355–381 (1997)
58. Ganesh, M., Sloan, I.H.: Optimal order spline methods for nonlinear differential and integro-differential equations. Appl. Numer. Math. **29**, 445–478 (1999)
59. Ganesh, M., Langdon, S., Sloan, I.H.: Efficient evaluation of highly oscillatory acoustic scattering surface integrals. J. Comput. Appl. Math. **204**, 363–374 (2007)

60. Ganesh, M., Le Gia, Q.T., Sloan, I.H.: A pseudospectral quadrature method for Navier-Stokes equations on rotating spheres. Math. Comput. **80**, 1397–1430 (2011)
61. Gilbert, A.D., Kuo, F.Y., Sloan, I.H.: Hiding the weights—CBC black box algorithms with a guaranteed error bound. Math. Comput. Simul. **143**, 202–214 (2018)
62. Graham, I.G., Sloan, I.H.: On the compactness of certain integral operators. J. Math. Anal. Appl. **68**, 580–594 (1979)
63. Graham, I.G., Sloan, I.H.: Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in R^3. Numer. Math. **92**, 289–323 (2002)
64. Graham, I.G., Joe, S., Sloan, I.H.: Iterated Galerkin versus iterated collocation for integral equations of the second kind. IMA J. Numer. Anal. **5**, 355–369 (1985)
65. Graham, I.G., Kuo, F.Y., Nuyens, D., Scheichl, R., Sloan, I.H.: Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. J. Comput. Phys. **230**, 3668–3694 (2011)
66. Graham, I.G., Kuo, F.Y., Nichols, J.A., Scheichl, R., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. Numer. Math. **131**, 329–368 (2015)
67. Griebel, M., Kuo, F.Y., Sloan, I.H.: The smoothing effect of the ANOVA decomposition. J. Complex. **26**, 523–551 (2010)
68. Griebel, M., Kuo, F.Y., Sloan, I.H.: The smoothing effect of integration in R^d and the ANOVA decomposition. Math. Comput. **82**, 383–400 (2013)
69. Griebel, M., Kuo, F.Y., Sloan, I.H.: The ANOVA decomposition of a non-smooth function of infinitely many variables can have every term smooth. Math. Comput. **86**, 1855–1876 (2017)
70. Griebel, M., Kuo, F.Y., Sloan, I.H.: Note on "The smoothing effect of integration in R^d and the ANOVA decomposition". Math. Comput. **86**, 1847–1854 (2017)
71. Grigorieff, R.D., Sloan, I.H.: High-order spline Petrov-Galerkin methods with quadrature. In: Zeitschrift Fur Angewandte Mathematik Und Mechanik, vol. 76, pp.15–18. Akademie Verlag GMBH, Hamburg (1996)
72. Grigorieff, R.D., Sloan, I.H.: Spline Petrov-Galerkin methods with quadrature. Numer. Funct. Anal. Optim. **17**, 755–784 (1996)
73. Grigorieff, R.D., Sloan, I.H.: Galerkin approximation with quadrature for the screen problem in R^3. J. Integral Equ. Appl. **9**, 293–319 (1997)
74. Grigorieff, R.D., Sloan, I.H.: Stability of discrete orthogonal projections for continuous splines. Bull. Aust. Math. Soc. **58**, 307–332 (1998)
75. Grigorieff, R.D., Sloan, I.H.: On qualocation and collocation methods for singular integral equations with piecewise continuous coefficients, using continuous splines on quasi-uniform meshes. Oper. Theory: Adv. Appl. **121**, 146–161 (2001)
76. Grigorieff, R.D., Sloan, I.H.: Discrete orthogonal projections on multiple knot periodic splines. J. Approx. Theory **137**, 201–225 (2005)
77. Grigorieff, R.D., Sloan, I.H.: Qualocation for boundary integral equations. J. Integral Equ. Appl. **18**, 117–140 (2006)
78. Grigorieff, R.D., Sloan, I.H., Brandts, J.: Superapproximation and commutator properties of discrete orthogonal projections for continuous splines. J. Approx. Theory **107**, 244–267 (2000)
79. Hesse, K., Sloan, I.H.: High order numerical integration on the sphere and extremal point sets. J. Comput. Technol. **9**, 4–12 (2004)
80. Hesse, K., Sloan, I.H.: Optimal lower bounds for cubature error on the sphere S^2. J. Complex. **21**, 790–803 (2005)
81. Hesse, K., Sloan, I.H.: Optimal order integration on the sphere. In: Li, T., Zhang, P. (eds.) Contemporary Applied Mathematics, pp. 59–70. World Scientific, Beijing (2005)
82. Hesse, K., Sloan, I.H.: Worst-case errors in a Sobolev space setting for cubature over the sphere S^2. Bull. Aust. Math. Soc. **71**, 81–105 (2005)
83. Hesse, K., Sloan, I.H.: Hyperinterpolation on the sphere. In: Govil, N.K., Mhasker, H.N., Mohapatra, R.N., Nashed, Z., Szabados, J. (eds.) Frontiers in Interpolation and Approximation, pp. 213–248. Chapman & Hall/CRC, Boca Raton (2006)

84. Hesse, K., Sloan, I.H.: Cubature over the sphere S^2 in Sobolev spaces of arbitrary order. J. Approx. Theory **141**, 118–133 (2006)
85. Hesse, K., Kuo, F.Y., Sloan, I.H.: A component-by-component approach to efficient numerical integration over products of spheres. J. Complex. **23**, 25–51 (2007)
86. Hesse, K., Mhaskar, H., Sloan, I.H.: Quadrature in Besov spaces on the Euclidean sphere. J. Complex. **23**, 528–552 (2007)
87. Hesse, K., Sloan, I.H., Womersley, R.S.: Numerical integration on the sphere. In: Freeden, W., Nashed, M., Sonar, T. (eds.) Handbook of Geomathematics, 1st edn., pp. 1187–1220. Springer, Berlin (2010)
88. Hesse, K., Sloan, I.H., Womersley, R.S.: Numerical integration on the sphere. In: Handbook of Geomathematics, 2nd edn, pp. 2671–2710. Springer, Berlin (2015)
89. Hesse, K., Sloan, I.H., Womersley, R.S.: Radial basis function approximation of noisy scattered data on the sphere. Numer. Math. **137**, 579–605 (2017)
90. Hickernell, F.J., Sloan, I.H., Wasilkowski, G.W.: On strong tractability of weighted multivariate integration. Math. Comput. **73**, 1903–1911 (2004)
91. Hickernell, F.J., Sloan, I.H., Wasilkowski, G.W.: On tractability of weighted integration over bounded and unbounded regions in R^s. Math. Comput. **73**, 1885–1901 (2004)
92. Hickernell, F.J., Sloan, I.H., Wasilkowski, G.W.: On tractability of weighted integration for certain Banach spaces of functions. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 51–71. Springer, Berlin (2004)
93. Hickernell, F.J., Sloan, I.H., Wasilkowski, G.W.: The strong tractability of multivariate integration using lattice rules. In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 259–293. Springer, Berlin (2004)
94. Hickernell, F.J., Sloan, I.H., Wasilkowski, G.W.: A piecewise constant algorithm for weighted L_1 approximation over bounded or unbounded regions in R^s. SIAM J. Numer. Anal. **43**, 1003–1020 (2005)
95. Jeon, Y.J., Sloan, I.H., Stephan, E., Elschner, J.: Discrete qualocation methods for logarithmic-kernel integral equations on a piecewise smooth boundary. Adv. Comput. Math. **7**, 547–571 (1997)
96. Joe, S., Sloan, I.H.: On Bateman's method for second kind integral equations. Numer. Math. **49**, 499–510 (1986)
97. Joe, S., Sloan, I.H.: Imbedded lattice rules for multidimensional integration. SIAM J. Numer. Anal. **29**, 1119–1135 (1992)
98. Joe, S., Sloan, I.H.: On computing the lattice rule criterion R. Math. Comput. **59**, 557–568 (1992)
99. Joe, S., Sloan, I.H.: Implementation of a lattice method for numerical multiple integration. ACM Trans. Math. Softw. **19**, 523–545 (1993)
100. Joe, S., Sloan, I.H.: Implementation of a lattice method for numerical multiple integration (vol 19, p. 523, 1993). ACM Trans. Math. Softw. **20**, 245–245 (1994)
101. Kress, R., Sloan, I.H.: On the numerical solution of a logarithmic integral equation of the first kind for the Helmholtz equation. Numer. Math. **66**, 199–214 (1993)
102. Kress, R., Sloan, I.H., Stenger, F.: A sinc quadrature method for the double-layer integral equation in planar domains with corners. J. Integral Equ. Appl. **10**, 291–317 (1998)
103. Kumar, S., Sloan, I.H.: A new collocation type method for Hammerstein integral equations. Math. Comput. **48**, 585–593 (1987)
104. Kuo, F.Y., Sloan, I.H.: Lifting the curse of dimensionality. Not. Am. Math. Soc. **52**, 1320–1328 (2005)
105. Kuo, F.Y., Sloan, I.H.: Quasi-Monte Carlo methods can be efficient for integration over products of spheres. J. Complex. **21**, 196–210 (2005)
106. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Lattice rules for multivariate approximation in the worst case setting. In: Niederreiter, H., Talay, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 289–330. Springer, Berlin (2006)
107. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Periodization strategy may fail in high dimensions. Numer. Algorithms **46**, 369–391 (2007)

108. Kuo, F.Y., Dunsmuir, W.T., Sloan, I.H., Wand, M.P., Womersley, R.S.: Quasi-Monte Carlo for highly structured generalised response models. Methodol. Comput. Appl. Probab. **10**, 239–275 (2008)
109. Kuo, F.Y., Giles, M.B., Sloan, I.H., Waterhouse, B.J.: Quasi-Monte Carlo for finance applications. ANZIAM J. **50**, C308–C323 (2008)
110. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Lattice rule algorithms for multivariate approximation in the average case setting. J. Complex. **24**, 283–323 (2008)
111. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. J. Complex. **26**, 135–160 (2010)
112. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: Liberating the dimension. J. Complex. **26**, 422–454 (2010)
113. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decompositions of multivariate functions. Math. Comput. **79**, 953–966 (2010)
114. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo methods for high-dimensional integration: the standard (weighted Hilbert space) setting and beyond. ANZIAM J. **53**, 1–37 (2011)
115. Kuo, F.Y., Schwab, C., Sloan, I.H.: Erratum: Quasi-Monte Carlo methods for high-dimensional integration: the standard (weighted Hilbert space) setting and beyond (ANZIAM Journal 53, 1–37 (2011)). ANZIAM J. **53**, 251 (2012)
116. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
117. Kuo, F.Y., Sloan, I.H., Schwab, C.: Erratum: Quasi-Monte Carlo methods for high-dimensional integration: the standard (weighted Hilbert space) setting and beyond (ANZIAM Journal 53, 1–37 (2011)). ANZIAM J. **54**, 216–219 (2013)
118. Kuo, F.Y., Schwab, C., Sloan, I.H.: Multi-level Quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. Found. Comput. Math. **15**, 411–449 (2015)
119. Kuo, F.Y., Nuyens, D., Plaskota, L., Sloan, I.H., Wasilkowski, G.W.: Infinite-dimensional integration and the multivariate decomposition method. J. Comput. Appl. Math. **326**, 217–234 (2017)
120. Kuo, F.Y., Scheichl, R., Schwab, C., Sloan, I.H., Ullmann, E.: Multilevel Quasi-Monte Carlo methods for lognormal diffusion problems. Math. Comput. **86**, 2827–2860 (2017)
121. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Multivariate integration for analytic functions with Gaussian kernels. Math. Comput. **86**, 829–853 (2017)
122. Le Gia, Q.T., Sloan, I.H.: The uniform norm of hyperinterpolation on the unit sphere in an arbitrary number of dimensions. Constr. Approx. **17**, 249–265 (2001)
123. Le Gia, Q.T., Sloan, I.H., Tran, T.: Overlapping additive Schwarz preconditioners for elliptic PDEs on the unit sphere. Math. Comput. **78**, 79–101 (2009)
124. Le Gia, Q.T., Tran, T., Sloan, I.H., Stephan, E.P.: Boundary integral equations on the sphere with radial basis functions: error analysis. Appl. Numer. Math. **59**, 2857–2871 (2009)
125. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Multiscale analysis in Sobolev spaces on the sphere. SIAM J. Numer. Anal. **48**, 2065–2090 (2010)
126. Le Gia, Q.T., Sloan, I.H., Wathen, A.J.: Stability and preconditioning for a hybrid approximation on the sphere. Numer. Math. **118**, 695–711 (2011)
127. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. Appl. Comput. Harmon. Anal. **32**, 401–412 (2012)
128. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Multiscale RBF collocation for solving PDEs on spheres. Numer. Math. **121**, 99–125 (2012)
129. Le Gia, Q.T., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Needlet approximation for isotropic random fields on the sphere. J. Approx. Theory **216**, 86–116 (2017)
130. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Zooming from global to local: a multiscale RBF approach. Adv. Comput. Math. **43**, 581–606 (2017)

131. Lin, Q., Sloan, I.H., Xie, R.: Extrapolation of the iterated-collocation method for integral-equations of the 2nd kind. SIAM J. Numer. Anal. **27**, 1535–1541 (1990)
132. Lubich, C.H., Sloan, I.H., Thomée, V.: Nonsmooth data error estimates for approximations of an evolution equation with a positive type memory term. Math. Comput. **65**, 1–17 (1996)
133. Lyness, J.N., Sloan, I.H.: Cubature rules of prescribed merit. SIAM J. Numer. Anal. **34**, 586–602 (1997)
134. Lyness, J.N., Sloan, I.H.: Some properties of rank-2 lattice rules. Math. Comput. **53**, 627–637 (1989)
135. Mauersberger, D., Sloan, I.H.: A simplified approach to the semi-discrete Galerkin method for the single-layer equation for a plate. In: Bonnet, M., Sandig, A., Wendland, W. (eds.) Mathematical Aspects of Boundary Element Methods, Ecole Polytech, Palaiseau. Chapman & Hall/CRC Research Notes in Mathematics Series, vol. 414, pp. 178–190. Chapman & Hall/CRC Press, Boca Raton (2000)
136. Mclean, W., Sloan, I.H.: A fully discrete and symmetric boundary element method. IMA J. Numer. Anal. **14**, 311–345 (1994)
137. Mclean, W., Sloan, I.H., Thomée, V.: Time discretization via Laplace transformation of an integro-differential equation of parabolic type. Numer. Math. **102**, 497–522 (2005)
138. Monegato, G., Sloan, I.H.: Numerical solution of the generalized airfoil equation for an airfoil with a flap. SIAM J. Numer. Anal. **34**, 2288–2305 (1997)
139. Niederreiter, H., Sloan, I.H.: Lattice rules for multiple integration and discrepancy. Math. Comput. **54**, 303–312 (1990)
140. Niederreiter, H., Sloan, I.H.: Quasi Monte Carlo methods with modified vertex weights. In: Brass, H., Hämmerlin, G. (eds.) Numerical Integration IV, Mathematical Research Institute, Oberwolfach. International Series of Numerical Mathematics, vol. 112, pp. 253–265. Birkhäuser Verlag, Basel (1993)
141. Niederreiter, H., Sloan, I.H.: Integration of nonperiodic functions of two variables by Fibonacci lattice rules. J. Comput. Appl. Math. **51**, 57–70 (1994)
142. Niederreiter, H., Sloan, I.H.: Variants of the Koksma-Hlawka inequality for vertex-modified quasi-Monte Carlo integration rules. Math. Comput. Model. **23**, 69–77 (1996)
143. Novak, E., Sloan, I.H., Woźniakowski, H.: Tractability of tensor product linear operators. J. Complex. **13**, 387–418 (1997)
144. Novak, E., Sloan, I.H., Woźniakowski, H.: Tractability of approximation for weighted Korobov spaces on classical and quantum computers. Found. Comput. Math. **4**, 121–156 (2004)
145. Pereverzyev, S.V., Sloan, I.H., Tkachenko, P.: Parameter choice strategies for least-squares approximation of noisy smooth functions on the sphere. SIAM J. Numer. Anal. **53**, 820–835 (2015)
146. Price, J.F., Sloan, I.H.: Pointwise convergence of multiple Fourier series: Sufficient conditions and an application to numerical integration. J. Math. Anal. Appl. **169**, 140–156 (1992)
147. Prössdorf, S., Sloan, I.H.: Quadrature method for singular integral equations on closed curves. Numer. Math. **61**, 543–559 (1992)
148. Prössdorf, S., Saranen, J., Sloan, I.H.: A discrete method for the logarithmic-kernel integral-equation on an open arc. J. Aust. Math. Soc. Ser. B: Appl. Math. **34**, 401–418 (1993)
149. Rabinowitz, P., Sloan, I.H.: Product integration in the presence of a singularity. SIAM J. Numer. Anal. **21**, 149–166 (1984)
150. Reztsov, A.V., Sloan, I.H.: On 2D packings of cubes in the torus. Proc. Am. Math. Soc. **125**, 17–26 (1997)
151. Saranen, J., Sloan, I.H.: Quadrature methods for logarithmic-kernel integral equations on closed curves. IMA J. Numer. Anal. **12**, 167–187 (1992)
152. Schatz, A.H., Sloan, I.H., Wahlbin, L.B.: Superconvergence in finite element methods and meshes that are locally symmetric with respect to a point. SIAM J. Numer. Anal. **33**, 505–521 (1996)
153. Schwab, C., Sloan, I.H.: Review of "Cubature Formulas and Modern Analysis: an Introduction" by S. L. Sobolev. Math. Comput. **64**, 1761–1763 (1995)

154. Sheen, D., Sloan, I.H., Thomée, V.: A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature. Math. Comput. **69**, 177–195 (1999)
155. Sheen, D., Sloan, I.H., Thomée, V.: A parallel method for time discretization of parabolic equations based on Laplace transformation and quadrature. IMA J. Numer. Anal. **23**, 269–299 (2003)
156. Sinescu, V., Kuo, F.Y., Sloan, I.H.: On the choice of weights in a function space for Quasi-Monte Carlo methods for a class of generalised response models in statistics. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, pp. 631–647. Springer, Berlin (2013)
157. Sloan, I.H.: Method of polarized orbitals for elastic scattering of slow electrons by ionized helium and atomic hydrogen. Proc. Roy. Soc. Lond. Ser. A-Math. Phys. Sci. **281**, 151–163 (1964)
158. Sloan, I.H.: The ionization of neutral helium by electron impact. Proc. Phys. Soc. **85**, 435–442 (1965)
159. Sloan, I.H.: Unitary modification of the impulse-pickup approximation. Phys. Lett. B **25**, 84–85 (1967)
160. Sloan, I.H.: Unitary modifications of the impulse approximation. Phys. Rev. **162**, 855–858 (1967)
161. Sloan, I.H.: Approximation method for three-body collisions. Phys. Rev. **165**, 1587–1594 (1968)
162. Sloan, I.H.: Method for the numerical solution of linear second-order differential equations. J. Comput. Phys. **3**, 40–45 (1968)
163. Sloan, I.H.: Note: errors in the Numerov and Runge-Kutta methods. J. Comput. Phys. **2**, 414–416 (1968)
164. Sloan, I.H.: The numerical evaluation of principal-value integrals. J. Comput. Phys. **3**, 332–333 (1968)
165. Sloan, I.H.: Multiple-scattering analysis on a soluble neutron-deuteron model. Phys. Rev. **185**, 1361–1370 (1969)
166. Sloan, I.H.: Tensor force in the separable potential model of neutron-deuteron collisions. Nucl. Phys. Sect. A **139**, 337–352 (1969)
167. Sloan, I.H.: Levinson's theorem and *S*-wave neutron-deuteron scattering. Phys. Lett. B **34**, 243–244 (1971)
168. Sloan, I.H.: Perturbation method for 3 body collisions. Bull. Am. Phys. Soc. **16**, 1349 (1971)
169. Sloan, I.H.: Phase parameters for nucleon-deuteron scattering. Nucl. Phys. Sect. A **168**, 211–224 (1971)
170. Sloan, I.H.: Equations for four-particle scattering. Phys. Rev. C **6**, 1945–1955 (1972)
171. Sloan, I.H.: Separable expansions and perturbation theory for three-body collisions. Nucl. Phys. Sect. A **182**, 549–557 (1972)
172. Sloan, I.H.: A three-nucleon scattering calculation using perturbation theory. Nucl. Phys. Sect. A **188**, 193–204 (1972)
173. Sloan, I.H.: Sturmian expansion of the Coulomb *t* matrix. Phys. Rev. A **7**, 1016–1023 (1973)
174. Sloan, I.H.: Error analysis for a class of degenerate-kernel methods. Numer. Math. **25**, 231–238 (1975)
175. Sloan, I.H.: Convergence of degenerate-kernel methods. J. Aust. Math. Soc. Ser. B Appl. Math. **19**, 422–431 (1976)
176. Sloan, I.H.: Improvement by iteration for compact operator equations. Math. Comput. **30**, 758–764 (1976)
177. Sloan, I.H.: Iterated Galerkin method for eigenvalue problems. SIAM J. Numer. Anal. **13**, 753–760 (1976)
178. Sloan, I.H.: Comment on "the collocation variational method for solving Fredholm integral equations…". J. Phys. A: Math. General **11**, 1195–1197 (1978)
179. Sloan, I.H.: On the numerical evaluation of singular integrals. BIT **18**, 91–102 (1978)

180. Sloan, I.H.: On choosing the points in product integration. J. Math. Phys. **21**, 1032–1039 (1979)
181. Sloan, I.H.: The numerical solution of Fredholm equations of the second kind by polynomial interpolation. J. Integral Equ. **2**, 265–279 (1980)
182. Sloan, I.H.: A review of numerical methods for Fredholm equations of the second kind. In: Anderssen, R.S., de Hoog, F., Lukas, M. (eds.) The Application and Numerical Solution of Integral Equations, pp. 51–74. Sijthoff and Noordhoff, Alphen aan de Rijn (1980)
183. Sloan, I.H.: Analysis of general quadrature methods for integral equations of the second kind. Numer. Math. **38**, 263–278 (1981)
184. Sloan, I.H.: Comment on "failure of the connected-kernel method". Phys. Rev. C **23**, 1289–1292 (1981)
185. Sloan, I.H.: Mathematical and computational methods. Nucl. Phys. Sect. A **353**, 365–374 (1981)
186. Sloan, I.H.: Quadrature methods for integral-equations of the 2nd kind over infinite intervals. Math. Comput. **36**, 511–523 (1981)
187. Sloan, I.H.: Superconvergence and the Galerkin method for integral equations of the second kind. In: Baker, C.T.H., Miller, G.F. (eds.) Treatment of Integral Equations by Numerical Methods, pp. 197–207. Academic, Cambridge (1982)
188. Sloan, I.H.: Nonpolynomial interpolation. J. Approx. Theory **39**, 97–117 (1983)
189. Sloan, I.H.: Fast convergence of the iterated Galerkin method for integral equations. In: Noye, J., Fletcher, C. (eds.) Computational Techniques and Applications, pp. 352–358. North Holland, Amsterdam (1984)
190. Sloan, I.H.: Four variants of the Galerkin method for integral equations of the second kind. IMA J. Numer. Anal. **4**, 9–17 (1984)
191. Sloan, I.H.: The iterated Galerkin method for integral equations of the second kind. In: Jefferies, B., McIntosh, A. (eds.) Miniconference on Operator Theory and Partial Differential Equations, pp. 153–161. Centre for Mathematical Analysis, Australian National University, Canberra (1984)
192. Sloan, I.H.: Lattice methods for multiple integration. J. Comput. Appl. Math. **12–13**, 131–143 (1985)
193. Sloan, I.H.: A quadrature-based approach to improving the collocation method. Numer. Math. **54**, 41–56 (1988)
194. Sloan, I.H.: Superconvergence in the collocation and qualocation methods. In: Agarwal, R. (ed.) Numerical Mathematics, pp. 429–441. Birkhäuser Verlag, Basel (1988)
195. Sloan, I.H.: Superconvergence. In: Golberg, M. (ed.) Numerical Solution of Integral Equations, pp. 35–70. Plenum Press, New York (1990)
196. Sloan, I.H.: Error bounds for the method of good lattice points. Math. Comput. **56**, 257–266 (1991)
197. Sloan, I.H.: Error analysis of boundary integral methods. Acta Numer. **1**, 287–339 (1992)
198. Sloan, I.H.: Numerical-integration in high dimensions – the lattice rule approach. In: Espelid, T., Genz, A. (eds.) Numerical Integration, Bergen. Nato Advanced Science Institutes Series, Series C, Mathematical and Physical Sciences, vol. 357, pp. 55–69. Kluwer Academic Publishers, Dordrecht (1992)
199. Sloan, I.H.: Unconventional methods for boundary integral-equations in the plane. In: D. Griffiths, G. Watson (eds.) Numerical Analysis 1991, Univ Dundee, Dundee. Pitman Research Notes in Mathematics Series, vol. 260, pp. 194–218. Longman Scientific & Technical, New York (1992)
200. Sloan, I.H.: Review of "Random Number Generation and Quasi-Monte Carlo Methods" by H. Niederreiter. SIAM Rev. **35**, 680–681 (1993)
201. Sloan, I.H.: Boundary element methods. In: Theory and Numerics of Ordinary and Partial Differential Equations, pp. 143–180. Clarendon Press, Oxford (1995)
202. Sloan, I.H.: Lattice rules of moderate order. In: Fang, K.T., Hickernell, F.J. (eds.) Proceedings Workshop on Quasi-Monte Carlo Methods and Their Applications, pp. 147–153. Statistics Research and Consultancy Unit, Hong Kong Baptist University (1995)

203. Sloan, I.H.: Polynomial interpolation and hyperinterpolation over general regions. J. Approx. Theory **83**, 238–254 (1995)
204. Sloan, I.H.: Interpolation and hyperinterpolation on the sphere. In: Haussmann, W., Jetter, K., Reimer, M. (eds.) Multivariate Approximation: Recent Trends and Results, pp. 255–268. Akademie Verlag, Berlin (1997)
205. Sloan, I.H.: Review of "Integral Equations: Theory and Numerical Treatment" by Wolfgang Hackbusch. SIAM Rev. **39**, 360 (1997)
206. Sloan, I.H.: Multiple integration is intractable but not hopeless. ANZIAM J. **42**, 3–8 (2000)
207. Sloan, I.H.: Qualocation. J. Comput. Appl. Math. **125**, 461–478 (2000)
208. Sloan, I.H.: QMC integration—beating intractability by weighting the coordinate directions. In: Fang, K.T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 103–123. Springer, Berlin (2002)
209. Sloan, I.H.: Review of "Approximating Integrals via Monte Carlo and Deterministic Methods" by M. Evans and T. Swartz. SIAM Rev. **44**, 742–743 (2002)
210. Sloan, I.H.: Finite-order integration weights can be dangerous. Comput. Methods Appl. Math. **7**, 239–254 (2007)
211. Sloan, I.H.: Review of "Counting Australia in: The People, Organizations, and Institutions of Australian Mathematics" by Graeme L. Coben. The Mathematical Intelligencer, vol. 30, pp. 63–65. Halstead Press, Sydney (2008)
212. Sloan, I.H.: How high is high-dimensional? In: Essays on the Complexity of Continuous Problems, pp. 73–88. European Mathematical Society, Zurich (2009)
213. Sloan, I.H.: Polynomial approximation on spheres—generalizing de la Vallée-Poussin. Comput. Methods Appl. Math. **11**, 540–552 (2011)
214. Sloan, I.H.: What's new in high-dimensional integration?—Designing Quasi Monte Carlo for applications. In: Guo, L., Ma, Z.M. (eds.) Proceedings of the 8th International Congress on Industrial and Applied Mathematics, pp. 365–386. Higher Education Press, Beijing (2015)
215. Sloan, I.H., Aarons, J.C.: Vector and tensor polarizations in nucleon-deuteron scattering (ii). Nucl. Phys. Sect. A **198**, 321–342 (1972)
216. Sloan, I.H., Adhikari, S.K.: Method for Lippmann-Schwinger equations. Nucl. Phys. Sect. A **235**, 352–360 (1974)
217. Sloan, I.H., Atkinson, K.: Semi-discrete Galerkin methods for the single layer equation on Lipschitz curves. J. Integral Equ. Appl. **9**, 279–292 (1997)
218. Sloan, I.H., Brady, T.J.: Variational approach to the on- and off-shell T matrix. Phys. Rev. C **6**, 701–709 (1972)
219. Sloan, I.H., Brady, T.J.: Variational calculations of off-shell t-matrix. Bull. Am. Phys. Soc. **17**, 608 (1972)
220. Sloan, I.H., Burn, B.J.: Collocation with polynomials for integral equations of the second kind. J. Integral Equ. **1**, 77–94 (1979)
221. Sloan, I.H., Burn, B.J.: An unconventional quadrature method for logarithmic-kernel integral equations on closed curves. J. Integral Equ. Appl. **4**, 117–151 (1992)
222. Sloan, I.H., Gray, J.D.: Separable expansions of the t-matrix. Phys. Lett. B **44**, 354–356 (1973)
223. Sloan, I.H., Kachoyan, P.J.: Lattices for multiple integration. In: Gustafson, S.A., Womersley, R.S. (eds.) Mathematical Programming and Numerical Analysis Workshop, pp. 147–165. Centre for Mathematical Analysis, Australian National University, Canberra (1984)
224. Sloan, I.H., Kachoyan, P.: Lattice methods for multiple integration: theory, error analysis and examples. SIAM J. Numer. Anal. **24**, 116–128 (1987)
225. Sloan, I.H., Lyness, J.N.: The representation of lattice quadrature-rules as multiple sums. Math. Comput. **52**, 81–94 (1989)
226. Sloan, I.H., Lyness, J.N.: Lattice rules – projection regularity and unique representations. Math. Comput. **54**, 649–660 (1990)
227. Sloan, I.H., Massey, H.S.W.: The exchange-polarization approximation for elastic scattering of slow electrons by atoms and ions: electron scattering by helium ions. In: McDowell, M.R.C. (ed.) Atomic Collision Processes, pp. 14–15. North Holland, Amsterdam (1964)

228. Sloan, I.H., Moore, E.J.: Integral equation approach to electron-hydrogen collisions. J. Phys. B: Atomic Mol. Phys. **1**, 414–422 (1968)
229. Sloan, I.H., Osborn, T.: Multiple integration over bounded and unbounded regions. J. Comput. Appl. Math. **17**, 181–196 (1987)
230. Sloan, I.H., Reztsov, A.V.: Component-by-component construction of good lattice rules. Math. Comput. **71**, 263–273 (2001)
231. Sloan, I.H., Smith, W.E.: Product-integration with the Clenshaw-Curtis and related points—convergence properties. Numer. Math. **30**, 415–428 (1978)
232. Sloan, I.H., Smith, W.E.: Product integration with the Clenshaw-Curtis points: implementation and error estimates. Numer. Math. **34**, 387–401 (1980)
233. Sloan, I.H., Smith, W.E.: Properties of interpolatory product integration rules. SIAM J. Numer. Anal. **19**, 427–442 (1982)
234. Sloan, I.H., Sommariva, A.: Approximation on the sphere using radial basis functions plus polynomials. Adv. Compos. Mater. **29**, 147–177 (2008)
235. Sloan, I.H., Spence, A.: Wiener-Hopf integral equations: finite-section approximation and projection methods. In: Hämmerlin, G., Hoffmann, K.H. (eds.) Constructive Methods for Practical Treatment of Integral Equations, pp. 256–272. Birkhäuser Verlag, Basel (1985)
236. Sloan, I.H., Spence, A.: Integral equations on the half-line: a modified finite-section approximation. Math. Comput. **47**, 589–595 (1986)
237. Sloan, I.H., Spence, A.: Projection methods for integral equations on the half-line. IMA J. Numer. Anal. **6**, 153–172 (1986)
238. Sloan, I.H., Spence, A.: The Galerkin method for integral equations of the first kind with logarithmic kernel: Applications. IMA J. Numer. Anal. **8**, 123–140 (1988)
239. Sloan, I.H., Spence, A.: The Galerkin method for integral equations of the first kind with logarithmic kernel: theory. IMA J. Numer. Anal. **8**, 105–122 (1988)
240. Sloan, I.H., Stephan, E.P.: Collocation with Chebyshev polynomials for Symm's integral-equation on an interval. J. Aust. Math. Soc. Ser. B Appl. Math. **34**, 199–211 (1992)
241. Sloan, I., Thomée, V.: Superconvergence of the Galerkin iterates for integral-equations of the 2nd kind. J. Integral Equ. **9**, 1–23 (1985)
242. Sloan, I.H., Thomée, V.: Time discretization of an integrodifferential equation of parabolic type. SIAM J. Numer. Anal. **23**, 1052–1061 (1986)
243. Sloan, I.H., Tran, T.: The tolerant qualocation method for variable-coefficient elliptic equations on curves. J. Integral Equ. Appl. **13**, 73–98 (2001)
244. Sloan, I.H., Walsh, L.: Lattice rules—classification and searches. In: Brass, H., Hämmerlin, G. (eds.) Numerical Integration III, pp. 251–260. Birkhäuser Verlag, Basel (1988)
245. Sloan, I.H., Walsh, L.: A computer-search of rank-2 lattice rules for multidimensional quadrature. Math. Comput. **54**, 281–302 (1990)
246. Sloan, I.H., Wang, X.: Low discrepancy sequences in high dimensions: How well are their projections distributed? J. Comput. Appl. Math. **213**, 366–386 (2008)
247. Sloan, I.H., Wendland, W.L.: A quadrature based approach to improving the collocation method for splines of even degree. Zeitschrift für Analysis und ihre Anwendungen **8**, 361–376 (1989)
248. Sloan, I.H., Wendland, W.: Qualocation methods for elliptic boundary integral equations. Numer. Math. **79**, 451–483 (1998)
249. Sloan, I.H., Wendland, W.: Commutator properties for periodic splines. J. Approx. Theory **97**, 254–281 (1999)
250. Sloan, I.H., Wendland, W.: Spline qualocation methods for variable-coefficient elliptic equations on curves. Numer. Math. **83**, 497–533 (1999)
251. Sloan, I.H., Wendland, H.: Inf-sup condition for spherical polynomials and radial basis functions on spheres. Math. Comput. **78**, 1319–1331 (2009)
252. Sloan, I.H., Womersley, R.S.: The uniform error of hyperinterpolation on the sphere. In: Jetter, K., Haussmann, W., Reimer, M. (eds.) Advances in Multivariate Approximation, pp. 289–306. Wiley, New York (1999)

253. Sloan, I.H., Womersley, R.S.: Constructive polynomial approximation on the sphere. J. Approx. Theory **103**, 91–118 (2000)
254. Sloan, I.H., Womersley, R.S.: The search for good polynomial interpolation points on the sphere. In: Griffiths, D., Watson, G. (eds.) 18th Dundee Biennial Conference on Numerical Analysis, pp. 211–229. University of Dundee, Dundee (2000)
255. Sloan, I.H., Womersley, R.S.: Good approximation on the sphere, with application to geodesy and the scattering of sound. J. Comput. Appl. Math. **149**, 227–237 (2002)
256. Sloan, I.H., Womersley, R.S.: Extremal systems of points and numerical integration on the sphere. Adv. Comput. Math. **21**, 107–125 (2004)
257. Sloan, I.H., Womersley, R.S.: A variational characterisation of spherical designs. J. Approx. Theory **159**, 308–318 (2009)
258. Sloan, I.H., Womersley, R.S.: Filtered hyperinterpolation: a constructive polynomial approximation on the sphere. Int. J. Geomath. **3**, 95–117 (2012)
259. Sloan, I.H., Woźniakowski, H.: An intractability result for multiple integration. Math. Comput. **66**, 1119–1124 (1997)
260. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complex. **14**, 1–33 (1998)
261. Sloan, I.H., Woźniakowski, H.: Multiple integrals in many dimensions. In: Guangzhou International Symposium, Zhongshan University, Guangzhou, pp. 507–516. Marcell Decker, New York (1999)
262. Sloan, I.H., Woźniakowski, H.: Tractability of multivariate integration for weighted Korobov classes. J. Complex. **17**, 697–721 (2001)
263. Sloan, I.H., Woźniakowski, H.: Tractability of integration in non-periodic and periodic weighted tenor product Hilbert spaces. J. Complex. **18**, 479–499 (2002)
264. Sloan, I.H., Woźniakowski, H.: When does Monte Carlo depend polynomially on the number of variables? In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 407–437. Springer, Berlin (2004)
265. Sloan, I.H., Burn, B.J., Datyner, N.: A new approach to the numerical solution of integral equations. J. Comput. Phys. **18**, 92–105 (1975)
266. Sloan, I.H., Noussair, E., Burn, B.J.: Projection methods for equations of the second kind. J. Math. Anal. Appl. **69**, 84–103 (1979)
267. Sloan, I.H., Tran, D., Fairweather, G.: A fourth-order cubic spline method for linear second-order two-point boundary-value problems. IMA J. Numer. Anal. **13**, 591–607 (1993)
268. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. SIAM J. Numer. Anal. **40**, 1650–1665 (2002)
269. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of Quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comput. **71**, 1609–1640 (2002)
270. Sloan, I.H., Wang, X., Woźniakowski, H.: Finite-order weights imply tractability of multivariate integration. J. Complex. **20**, 46–74 (2004)
271. Smith, W.E., Sloan, I.H.: Product-integration rules based on the zeros of Jacobi-polynomials. SIAM J. Numer. Anal. **17**, 1–13 (1980)
272. Smith, W.E., Sloan, I.H., Opie, A.H.: Product integration over infinite intervals I. Rules based on the zeros of Hermite polynomials. Math. Comput. **40**, 519–535 (1983)
273. Tran, T., Sloan, I.H.: Tolerant qualocation—a qualocation method for boundary integral equations with reduced regularity requirement. J. Integral Equ. Appl. **10**, 85–115 (1998)
274. Tran, T., Le Gia, Q.T., Sloan, I.H., Stephan, E.P.: Preconditioners for pseudodifferential equations on the sphere with radial basis functions. Numer. Math. **115**, 141–163 (2010)
275. Wang, X., Sloan, I.H.: Why are high-dimensional finance problems often of low effective dimension? SIAM J. Sci. Comput. **27**, 159–183 (2005)
276. Wang, X., Sloan, I.H.: Efficient weighted lattice rules with applications to finance. SIAM J. Sci. Comput. **28**, 728–750 (2006)
277. Wang, X., Sloan, I.H.: Brownian bridge and principal component analysis: towards removing the curse of dimensionality. IMA J. Numer. Anal. **27**, 631–654 (2007)

278. Wang, X., Sloan, I.H.: Quasi-Monte Carlo methods in financial engineering: an equivalence principle and dimension reduction. Oper. Res. **59**, 80–95 (2011)
279. Wang, H., Sloan, I.H.: On filtered polynomial approximation on the sphere. J. Fourier Anal. Appl. **23**, 863–876 (2017)
280. Wang, S., Sloan, I.H., Kelly, D.W.: Pointwise a posteriori upper bounds for derivatives of a Neumann problem. In: Computer Techniques and Applications (CTAC95), pp. 771–778. World Scientific, Singapore (1996)
281. Wang, S., Sloan, I.H., Kelly, D.W.: Computable error bounds for pointwise derivatives of a Neumann problem. IMA J. Numer. Anal. **18**, 251–271 (1998)
282. Wang, X., Sloan, I.H., Dick, J.: On Korobov lattice rules in weighted spaces. SIAM J. Numer. Anal. **42**, 1760–1779 (2004)
283. Wang, Y.G., Sloan, I.H., Womersley, R.S.: Riemann localisation on the sphere. J. Fourier Anal. Appl. **2016**, 1–43 (2016)
284. Wang, Y.G., Le Gia, Q.T., Sloan, I.H., Womersley, R.S.: Fully discrete needlet approximation on the sphere. Appl. Comput. Harmon. Anal. **43**, 292–316 (2017)
285. Waterhouse, B.J., Kuo, F.Y., Sloan, I.H.: Randomly shifted lattice rules on the unit cube for unbounded integrands in high dimensions. J. Complex. **22**, 71–101 (2006)
286. Womersley, R.S., Sloan, I.H.: How good can polynomial interpolation on the sphere be? Adv. Comput. Math. **14**, 195–226 (2001)
287. Yan, Y., Sloan, L.: On integral equations of the first kind with logarithmic kernels. J. Integral Equ. Appl. **1**, 549–579 (1988)
288. Yan, Y., Sloan, I.H.: Mesh grading for integral-equations of the 1st kind with logarithmic kernel. SIAM J. Numer. Anal. **26**, 574–587 (1989)

Contents

About the Editors

Josef Dick is Associate Professor in the School of Mathematics and Statistics at the University of New South Wales, Sydney, Australia. He is a former PhD student of Professor Ian Sloan and currently Head of the Applied Mathematics Department in the School.

Frances Y. Kuo is Associate Professor in the School of Mathematics and Statistics at the University of New South Wales, Sydney, Australia. She is an academic grandchild of Professor Ian Sloan and a collaborator for over 15 years, currently holding the highest number of joint publications.

Henryk Woźniakowski is Emeritus Professor of Columbia University in New York and of University of Warsaw in Poland. He has been collaborating with Professor Ian Sloan for the last 20 years.

On Quasi-Energy-Spectra, Pair Correlations of Sequences and Additive Combinatorics

Ida Aichinger, Christoph Aistleitner, and Gerhard Larcher

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The investigation of the pair correlation statistics of sequences was initially motivated by questions concerning quasi-energy-spectra of quantum systems. However, the subject has been developed far beyond its roots in mathematical physics, and many challenging number-theoretic questions on the distribution of the pair correlations of certain sequences are still open. We give a short introduction into the subject, recall some known results and open problems, and in particular explain the recently established connection between the distribution of pair correlations of sequences on the torus and certain concepts from additive combinatorics. Furthermore, we slightly improve a result recently given by Jean Bourgain in Aistleitner et al. (Isr. J. Math., to appear. Available at https://arxiv.org/abs/1606.03591).

1 Introduction

Some of Ian Sloan's first published papers dealt with topics from mathematical physics, in particular with theoretical nuclear physics. Later he moved his area of research to applied mathematics and numerical analysis, and in particular Ian's

I. Aichinger
CERN, European Organization for Nuclear Research, Meyrin, Switzerland
e-mail: ida.aichinger@cern.ch

C. Aistleitner
TU Graz, Institute for Analysis and Number Theory, Graz, Austria
e-mail: aistleitner@math.tugraz.at

G. Larcher (✉)
Johannes Kepler University Linz, Institute for Financial Mathematics and Applied Number Theory, Linz, Austria
e-mail: gerhard.larcher@jku.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_1

ground-breaking work on complexity theory, numerical integration and mathematical simulation is well-known and highly respected among the scientific community of mathematicians. The techniques developed and analyzed by Ian in these fields are often based on the use of deterministic point sets and sequences with certain “nice” distribution properties, a method which is nowadays widely known under the name of *quasi-Monte Carlo method* (QMC). In the present paper we will combine these two topics, mathematical physics and the distribution of point sets.

Ian’s first research paper appeared 1964 in the Proceedings of the Royal Society (London), entitled “The method of polarized orbitals for the elastic scattering of slow electrons by ionized helium and atomic hydrogen” [26]. In the same journal, but 13 years later, Berry and Tabor published a groundbreaking paper on “Level clustering in the regular spectrum” [4]. This paper deals with the investigation of conservative quantum systems that are chaotic in the classical limit. More precisely, the paper deals with statistical properties of the energy spectra of these quantum systems, and Berry and Tabor conjectured that for the distribution function of the spacings between neighboring levels of a generic integrable quantum system the exponential Poisson law holds. That means, roughly speaking, the following.

Let H be the Hamiltonian of a quantum system and let $\lambda_1 \leq \lambda_2 \leq \ldots$ be its discrete energy spectrum. We call the numbers λ_i the *levels* of this energy spectrum. If it is assumed that

$$\#\{i : \lambda_i \leq x\} \sim cx^{\gamma}$$

for $x \to \infty$ and some constants $c > 0$, $\gamma \geq 1$, then consider $X_i := c\lambda_i^{\gamma}$. The Berry–Tabor conjecture now states that if the Hamiltonian is classically integrable and “generic”, then the X_i have the same local statistical properties as independent random variables coming from a Poisson process. Here the word “generic” is a bit vague; it essentially means that one excludes the known obvious (and less obvious) counterexamples to the conjecture. For more material on energy spectra of quantum systems and the following two concrete examples see the original paper of Berry and Tabor [4] as well as [5, 8, 16] and Chapter 2 in [7]. For a survey on the Berry–Tabor conjecture see [15].

Two basic examples of quantum systems are the two-dimensional “harmonic oscillator” with Hamiltonian

$$H = p_x^2 + p_y^2 + w^2\left(x^2 + y^2\right)$$

and the “boxed oscillator”. This is a particle constrained by a box in x-direction and by a harmonic potential in y-direction; the Hamiltonian in this case is given by

$$H = -p_x^2 - p_y^2 + w^2 y^2.$$

The investigation of the distribution of the energy levels in these two examples leads to the investigation of the pair correlation statistics of certain sequences $(\theta_n)_{n\geq 1}$ in the unit interval. More specifically, one is led to study the pair correlations of the sequence $(\{n\alpha\})_{n\geq 1}$ in the case of the 2-dimensional harmonic oscillator, and the pair correlations of the sequence $\left(\left\{n^2\alpha\right\}\right)_{n\geq 1}$ in the case of the boxed oscillator; here, and in the sequel, we write $\{\cdot\}$ for the fractional part function. In particular, for these sequences one is led to study the quantity R_2, which is introduced below.

Let $(\theta_n)_{n\geq 1}$ be a sequence of real numbers in $[0, 1]$, and let $\|\cdot\|$ denote the distance to the nearest integer. For every interval $[-s, s]$ we set

$$R_2\big([-s,s],(\theta_n)_{n\geq 1},N\big) = \frac{1}{N}\#\left\{1 \leq j \neq k \leq N : \left\|\theta_j - \theta_k\right\| \leq \frac{s}{N}\right\}.$$

The subscript "2" of "R_2" refers to the fact that these are the *pair* correlations, that is, the correlations of order 2—in contrast to triple correlations or correlations of even higher order. Note that the average spacing between two consecutive elements of $\{\theta_1, \ldots, \theta_N\}$ (understood as a point set on the torus) is $1/N$, and thus for an "evenly distributed" point set one would expect to find roughly $2s$ other points within distance $[-s/N, s/N]$ around a given point θ_j, causing $R_2([-s, s], (\theta_n)_{n\geq 1}, N)$ to be approximately $2s$ for such a point set (after summing over all elements of the point set and then normalizing with division by N). Actually, for a sequence of independent, $[0, 1]$-uniformly distributed random variables $\theta_1, \theta_2, \ldots$ one can easily show that for every $s \geq 0$ we have

$$R_2([-s,s],(\theta_n)_{n\geq 1},N) \to 2s,$$

almost surely. If this asymptotic relation holds for the distribution of pair correlations of a certain sequence we say that the distribution of the pair correlations is asymptotically *Poissonian*. Informally speaking, a sequence whose distribution of the pair correlations is asymptotically Poissonian may be seen as a sequence showing "random" behavior, and the investigation of the asymptotic distribution of the pair correlations of a deterministic sequence may be seen as studying the pseudo-randomness properties of this sequence.

The systematic investigation of the asymptotic distribution of the pair correlation of sequences on the torus (motivated by the applications in quantum physics) was started by Rudnick and Sarnak in [20] for the case of sequences of the form $\left(\left\{n^d\alpha\right\}\right)_{n\geq 1}$ for integers $d \geq 1$. In the case $d = 1$, the distribution of the pair correlations is *not* asymptotically Poissonian (independent of the value of α); this was remarked for example in [20] with a hint to the well-known *Three Distance Theorem*, which goes back to Świerczkowski and Sós [27]. For $d \geq 2$ the distribution of the pair correlations is asymptotically Poissonian for almost all α, which has been proved by Rudnick and Sarnak [20]. The case $d = 2$ (which corresponds to the energy levels of the *boxed oscillator*) has received particular attention; see for example [13, 17, 22, 29]. A generalization from $(\{n^d\alpha\})_{n\geq 1}$ to the case of $(\{a(n)\alpha\})_{n\geq 1}$ with $a(x) \in \mathbb{Z}[x]$ is obtained in [6]; again the pair correlations are asymptotically Poissonian for almost all α, provided that the degree of $a(x)$ is at least 2. Another case which has been intensively investigated is that of $(\{a(n)\alpha\})_{n\geq 1}$ for $(a(n))_{n\geq 1}$ being a *lacunary* sequence; see for example [3, 10, 21].

In [1] a general result was proved which includes earlier results (polynomial sequences, lacunary sequences, and sequences satisfying certain Diophantine conditions) and gives a unifying explanation. This result links the distribution of the pair correlations of the sequence $(\{a(n)\alpha\})_{n\geq 1}$ to the additive energy of the truncations of the integer sequence $(a(n))_{n\geq 1}$, a well-known concept from additive combinatorics which has been intensively studied. Recall that the additive energy $E(A)$ of a set of real numbers A is defined as

$$E(A) := \sum_{a+b=c+d} 1, \tag{1}$$

where the sum is extended over all quadruples (a, b, c, d) $\in A^4$. Trivially one has the estimate $|A|^2 \leq E(A) \leq |A|^3$, assuming that the elements of A are distinct. The additive energy of sequences has been extensively studied in the combinatorics literature. We refer the reader to [28] for a discussion of its properties and applications. To simplify notations, in the sequel whenever a sequence $A := (a(n))_{n\geq 1}$ is fixed we will abbreviate $R_2\,(s, \alpha, N)$ for $R_2\left([-s, s]\,, (\{a(n)\alpha\})_{n\geq 1}\,, N\right)$. Furthermore we will let A_N denote the first N elements of A. The result states that if the truncations A_N of an integer sequence A satisfy $E\,(A_N) \ll N^{3-\varepsilon}$ for some $\varepsilon > 0$, then $(\{a(n)\alpha\})_{n\geq 1}$ has (asymptotically) Poissonian pair correlations for almost all α. More precisely, the following theorem is true.

Theorem 1 *Let $(a(n))_{n\geq 1}$ be a sequence of distinct integers, and suppose that there exists a fixed constant $\varepsilon > 0$ such that*

$$E\,(A_N) \ll N^{3-\varepsilon} \quad \text{as } N \to \infty. \tag{2}$$

Then for almost all α one has

$$R_2\,(s, \alpha, N) \to 2s \quad \text{as } N \to \infty \tag{3}$$

for all $s \geq 0$.

Note that the condition of Theorem 1 is close to optimality, since by the trivial upper bound we always have $E\,(A_N) \leq N^3$; thus an arbitrarily small power savings over the trivial upper bound assures the "quasi-random" behavior of the pair correlations of $(\{a(n)\alpha\})_{n\geq 1}$. On the other hand, in [1] Bourgain showed the following negative result.

Theorem 2 *If $E\,(A_N) = \Omega\left(N^3\right)$, then there exists a subset of $[0, 1]$ of positive measure such that for every α from this set the pair correlations of the sequence $(\{a(n)\alpha\})_{n\geq 1}$ are* not *asymptotically Poissonian.*

We conjecture that actually even the following much stronger assertion is true.

Conjecture 1 If $E\,(A_N) = \Omega\left(N^3\right)$ there is **no** α for which the pair correlations of the sequence $(\{a(n)\alpha\})_{n\geq 1}$ are Poissonian.

In this paper we will prove a first partial result which should support this conjecture. However, before stating and discussing our result (which will be done in Sect. 2 below), we want to continue our general discussion of pair correlation problems. As one can see from the previous paragraphs, the metric theory of pair correlation problems on the torus is relatively well-understood. In contrast, there are only very few corresponding results which hold for a *specific* value of α. The most interesting case is that of the sequence $(\{n^2\alpha\})_{n\geq 1}$, where it is assumed that there is a close relation between Diophantine properties of α and the pair correlations distribution. For example, it is conjectured that for a quadratic irrational α this sequence has a pair correlations distribution which is asymptotically Poissonian; however, a proof of this conjecture seems to be far out of reach. A first step towards a proof of the conjecture was made by Heath-Brown [13], whose method requires bounds on the number of solutions of certain quadratic congruences; this topic was taken up by Shparlinski [24, 25], who obtained some improvements, but new ideas seem to be necessary for further steps toward a solution of the conjecture.

It should also be noted that the investigation of pair correlation distributions is not restricted to sequences of the torus. For example, consider a positive definite quadratic form $P(x, y) = \alpha x^2 + \beta xy + \gamma y^2$, and its values at the integers $(x, y) = (m, n) \in \mathbb{Z}^2$. These values form a discrete subset of $\mathbb{R}$, and one can study the pair correlations of those numbers contained in a finite window $[0, N]$. See for example [23, 30]. Another famous occurrence of the pair correlation statistics of an unbounded sequence in $\mathbb{R}$ is in Montgomery's pair correlation conjecture for the normalized spacings between the imaginary parts of zeros of the Riemann zeta function. The statement of the full conjecture is a bit too long to be reproduced here; we just want to mention that it predicts a distribution of the pair correlations which is very different from "simple" random behavior. For more details see Montgomery's paper [18]. There is a famous story related to Montgomery's conjecture; he met the mathematical physicist Freeman Dyson at tea time at Princeton, where Freeman Dyson identified Montgomery's conjectured distribution as the typical distribution of the spacings between normalized eigenvalues of large random Hermitian matrices—an observation which has led to the famous (conjectural) connection between the theory of the Riemann zeta function and random matrix theory. The whole story and more details can be found in [19].

2 New Results

In the sequel we give a first partial result towards a solution of the conjecture made above. Before stating the result we introduce some notations, and explain the background from additive combinatorics. For $v \in \mathbb{Z}$ let $A_N(v)$ denote the cardinality of the set

$$\left\{(x, y) \in \{1, \dots, N\}^2, x \neq y : a(x) - a(y) = v\right\}.$$

Then

$$E\left(A_N\right) = \Omega\left(N^3\right) \tag{4}$$

is equivalent to

$$\sum_{v\in\mathbb{Z}} A_N^2(v) = \Omega\left(N^3\right), \tag{5}$$

which implies that there is a $\kappa > 0$ and positive integers $N_1 < N_2 < N_3 < \ldots$ such that

$$\sum_{v\in\mathbb{Z}} A_{N_i}^2(v) \geq \kappa N_i^3, \qquad i = 1, 2, \ldots. \tag{6}$$

It will turn out that sequences $(a(n))_{n\geq 1}$ satisfying (4) have a strong linear substructure. From (6) we can deduce by the Balog–Szemeredi–Gowers-Theorem (see [2] and [12]) that there exist constants $c, C > 0$ depending only on κ such that for all $i = 1, 2, 3, \ldots$ there is a subset $A_0^{(i)} \subset \left(a\left(n\right)\right)_{1\leq n\leq N_i}$ such that

$$\left|A_0^{(i)}\right| \geq cN_i \qquad \text{and} \qquad \left|A_0^{(i)} + A_0^{(i)}\right| \leq C\left|A_0^{(i)}\right| \leq CN_i.$$

The converse is also true: If for all i for a set $A_0^{(i)}$ with $A_0^{(i)} \subset \left(a\left(n\right)\right)_{1\leq n\leq N_i}$ with $\left|A_0^{(i)}\right| \geq cN_i$ we have $\left|A_0^{(i)} + A_0^{(i)}\right| \leq C\left|A_0^{(i)}\right|$, then

$$\sum_{v\in\mathbb{Z}} A_{N_i}^2(v) \geq \frac{1}{C}\left|A_0^{(i)}\right|^3 \geq \frac{c^3}{C}N_i^3$$

and consequently $\sum_{v\in\mathbb{Z}} A_N^2(v) = \Omega\left(N^3\right)$ (this an elementary fact, see for example Lemma 1 (iii) in [14]).

Consider now a subset $A_0^{(i)}$ of $(a(n))_{1\leq n\leq N_i}$ with

$$\left|A_0^{(i)}\right| \geq cN_i \qquad \text{and} \qquad \left|A_0^{(i)} + A_0^{(i)}\right| \leq C\left|A_0^{(i)}\right|.$$

By the theorem of Freiman (see [11]) there exist constants d and K depending only on c and C, i.e. depending only on κ in our setting, such that there exists a *d-dimensional arithmetic progression* P_i of size at most KN_i such that $A_0^{(i)} \subset P_i$. This means that P_i is a set of the form

$$P_i := \left\{ b_i + \sum_{j=1}^{d} r_j k_j^{(i)} \,\middle|\, 0 \leq r_j < s_j^{(i)} \right\}, \tag{7}$$

with $b_i, k_1^{(i)}, \ldots, k_d^{(i)}, s_1^{(i)}, \ldots, s_d^{(i)} \in \mathbb{Z}$ and such that $s_1^{(i)} s_2^{(i)} \ldots s_d^{(i)} \leq KN_i$.

In the other direction again it is easy to see that for any set $A_0^{(i)}$ of the form (7) we have

$$\left|A_0^{(i)} + A_0^{(i)}\right| \leq 2^d K N_i.$$

Based on these observations we make the following definition:

Definition 1 Let $(a\,(n))_{n\geq 1}$ be a strictly increasing sequence of positive integers. We call this sequence *quasi-arithmetic of degree* **d**, where d is a positive integer, if there exist constants $C, K > 0$ and a strictly increasing sequence $(N_i)_{i\geq 1}$ of positive integers such that for all $i \geq 1$ there is a subset $A^{(i)} \subset (a\,(n))_{1\leq n\leq N_i}$ with $\left|A^{(i)}\right| \geq CN_i$ such that $A^{(i)}$ is contained in a d-dimensional arithmetic progression $P^{(i)}$ of size at most KN_i.

The considerations above show that a sequence $(a\,(x))_{x\geq 1}$ is quasi-arithmetic of some degree d if and only if it satisfies (5).

So our conjecture is equivalent to

Conjecture 2 If $(a(n))_{n\geq 1}$ is a quasi-arithmetic sequence of integers then there is **no** α such that the pair correlations of $(\{a(n)\alpha\})_{x\geq 1}$ are asymptotically Poissonian.

In the remaining part of this paper we will prove a theorem which slightly improvements the Theorem 2 of Bourgain for the subclass of sequences $(a(n))_{n\geq 1}$ which are quasi-arithmetic of degree 1.

Theorem 3 *If the sequence of integers $(a(n))_{n\geq 1}$ is quasi-arithmetic of degree 1, then the set of α's for which the distribution of the pair correlations of $(\{a(n)\alpha\})_{n\geq 1}$ is not asymptotically Poissonian has full measure.*

Remark The class of quasi-arithmetic sequences $(a\,(n))_{n\geq 1}$ of degree 1 contains all strictly increasing sequences with positive upper density, i.e.

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{\substack{n=1 \\ m\in\{a(n)\,|\, n\geq 1\}}}^{N} 1 > 0.$$

In particular this class contains all strictly increasing sequences which are bounded above by a linear function.

We will first state two auxiliary results in Sect. 3, and then give the proof of Theorem 3 in Sect. 4.

3 Auxiliary Results

Lemma 1 *Let $(\lambda_n)_{n\geq 1}$ be a strictly increasing sequence of positive integers. Let μ_n be the number of fractions of the form $j\lambda_n^{-1}$ $(0 < j < \lambda_n)$ which are not of the form $k\lambda_q^{-1}$ with some $q < n$ and $k < \lambda_q$. Furthermore, let $(\psi_n)_{n\geq 1}$ be a* *non-increasing sequence of positive reals such that* $\sum_{n=1}^{\infty} \psi_n = \infty$ *and with the following property (*):*

There exists a sequence $(\tau_n)_{n\geq 1}$ *of positive reals tending monotonically to zero, but so slowly that* $\sum_{n=1}^{\infty} \psi_n \tau_n$ *still diverges, and such that there exist a constant* $c > 0$ *and infinitely many positive integers* N *with*

$$\sum_{n=1}^{N} \mu_n \lambda_n^{-1} \psi_n \tau_n > c \sum_{n=1}^{N} \psi_n \tau_n.$$

Then—if (*) holds—for almost all $\theta \in \mathbb{R}$ *there exist infinitely many positive integers* n*, and integers* m*, such that*

$$0 \leq \lambda_n \theta - m < \psi_n.$$

Proof This lemma is essentially the divergence part of Theorem IV in [9]. It is shown there that the assertion of our Lemma 1 is true under the slightly stronger condition that $(\psi_n)_{n\geq 1}$—as in our Lemma—is a non-increasing sequence of positive reals with $\sum_{n=1}^{\infty} \psi_n = \infty$, and that $(\lambda_n)_{n\geq 1}$ satisfies

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mu_n \lambda_n^{-1} > 0.$$

If we follow the proof of Theorem IV in [9] line by line we see that our slightly weaker condition (*) also is sufficient to obtain the desired result. In fact replacing Cassel's condition by our condition (*) is relevant only in the proof of Lemma 3 in [9], which is an auxiliary result for the proof of Theorem IV in [9]. □

Lemma 2 *For all* $\delta > 0$ *there is a positive constant* $c(\delta) > 0$*, such that for every infinite subset* A *of* $\mathbb{N}$ *with*

$$\underline{d}(A) := \liminf_{N\to\infty} \frac{1}{N} \# \{n \leq N \mid n \in A\} > \delta$$

we have

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{\substack{n\leq N \\ n\in A}} \frac{\varphi(n)}{n} \geq c(\delta).$$

Here φ *denotes the Euler totient function.*

Proof Let

$$B(t) := \lim_{N\to\infty} \frac{1}{N} \left| \left\{ n \leq N \,\middle|\, \frac{n}{\varphi(n)} \geq t \right\} \right|.$$

Then by the main theorem in [31] the limit $B(t)$ exists and satisfies

$$B(t) = \exp\left(e^{-te^{-\gamma}\left(1+O\left(t^{-2}\right)\right)}\right)$$

for t to infinity and with γ denoting Euler's constant. Here, and in the sequel, we write $\exp(x)$ for e^x.

So there is a constant $L > 0$ such that

$$B(t) \leq \exp\left(e^{-te^{-\gamma}\left(1-\frac{L}{t^2}\right)}\right)$$

for all $t \geq 1$. Hence

$$B(t) \leq \exp\left(e^{-\frac{1}{2}te^{-\gamma}}\right)$$

for all $t \geq \max\left(1, \sqrt{2L}\right)$. Now assume that $\delta > 0$ is so small that

$$t_0 := 2e^{\gamma} \log\left(-\log\frac{\delta}{4}\right) > \max\left(1, \sqrt{2L}\right).$$

Note that it suffices to prove the lemma for such δ. We have

$$B\left(t_0\right) = \lim_{N\to\infty} \frac{1}{N} \left| \left\{ n \leq N \,\middle|\, \frac{n}{\varphi(n)} \geq t_0 \right\} \right|$$

and

$$B\left(t_0\right) \leq \exp\left(e^{-\frac{1}{2}e^{-\gamma}t_0}\right) = \frac{\delta}{4}.$$

Hence there exists N_0 such that for all $N \geq N_0$

$$\frac{1}{N} \left| \left\{ n \leq N \,\middle|\, \frac{n}{\varphi(n)} \geq t_0 \right\} \right| \leq \frac{\delta}{3}.$$

Therefore, since $\underline{d}(A) > \delta$, for all sufficiently large N we have

$$\frac{1}{N} \left| \left\{ n \leq N, n \in A \,\middle|\, \frac{n}{\varphi(n)} \leq t_0 \right\} \right| \geq \frac{\delta}{3}$$

and consequently also

$$\frac{1}{N}\sum_{\substack{n\le N\\ n\in A}}\frac{\varphi(n)}{n}\ge\frac{\delta}{3}\frac{1}{t_0}=:c(\delta)>0.$$

□

4 Proof of Theorem 3

Let $(a(n))_{n\ge1}$ be quasi-arithmetic of degree one and let $C, K > 0$, $(N_i)_{i\ge1}$, $\left(A^{(i)}\right)_{i\ge1}$ and $\left(P^{(i)}\right)_{i\ge1}$ be as described in Definition 1. In the sequel we will define inductively a certain strictly increasing subsequence $(M_l)_{l\ge1}$ of $(N_i)_{i\ge1}$.

Set $M_1 := N_1$ and assume that $M_1, M_2, \dots, M_{l-1}$ already are defined. If $M_l = N_{i_l}$ (where i_l still has to be defined) to simplify notations we write $A_l := A^{(i_l)}$, $P_l := P^{(i_l)}$.

We set

$$P_l := \{a_l + r\kappa_l \mid 0\le r < KM_l\}$$

and

$$A_l := \left\{a_l + r_j^{(l)}\kappa_l \middle| j = 1,2,\dots,s_l\right\}$$

with certain fixed $r_j^{(l)}$ with $1\le r_1^{(l)} < r_2^{(l)} < \dots < r_{s_l}^{(l)} < KM_l$ and $s_l \ge CM_l$. Of course we have $s_l < KM_l$.

We consider

$$V_l := \left\{\left(r_i^{(l)} - r_j^{(l)}\right)\kappa_l \middle| 1\le j < i\le s_l\right\},$$

the set of positive differences of A_l. Here V_l is the set itself, whereas by $\widetilde{V_l}$ we will denote the same set of positive differences but counted with multiplicity (so strictly speaking $\widetilde{V_l}$ is a multi-set rather than a set). Hence $|V_l| < KM_l$, whereas

$$\left|\widetilde{V_l}\right| = \frac{s_l\,(s_l-1)}{2}\ge c_1M_l^2.$$

Here and in the sequel we write c_i for positive constants depending only on C and K. We note that a value $u\in V_l$ has multiplicity at most s_l.

Let x be the number of elements in V_l with multiplicity at least $c_2 M_l$ where $c_2 := \min\left(K, \frac{c_1}{2K}\right)$. Assume that $x < c_2 M_l$. Then

$$\begin{aligned} c_1 M_l^2 &\le \left|\widetilde{V}_l\right| \le x s_l + (|V_l| - x)\, c_2 M_l \\ &\le x K M_l + (K M_l - x)\, c_2 M_l \\ &= M_l\,(x\,(K - c_2) + K c_2 M_l) \\ &< M_l^2\,(c_2\,(K - c_2) + K c_2) \\ &< M_l^2 c_1, \end{aligned}$$

a contradiction.

So there are at least $c_2 M_l$ values $u \in V_l$ with multiplicity at least $c_2 M_l$. We take the $\frac{c_2}{2} M_l$ largest of these values and denote them by $T_1^{(l)} < T_2^{(l)} < \ldots < T_{w_l}^{(l)}$ with $w_l \ge \frac{c_2}{2} M_l$ and $T_j^{(l)} := R_j^{(l)} \kappa_l$. Note that

$$\frac{c_2}{2} M_l \le R_1^{(l)} < \ldots < R_{w_l}^{(l)} < K M_l. \tag{8}$$

Remember that we still have to choose $i_l > i_{l-1}$ and to define M_l as N_{i_l}. We choose now i_l so large that

$$M_l > \left(\sum_{p=1}^{l-1} \sum_{q=1}^{w_p} T_q^{(p)} \right)^2 . \tag{9}$$

So altogether we have constructed a strictly increasing sequence $\lambda_1 < \lambda_2 < \lambda_3 < \ldots$ of integers given by $T_1^{(1)} < \ldots < T_{w_1}^{(1)} < T_1^{(2)} < \ldots < T_{w_2}^{(2)} < T_1^{(3)} < \ldots$.

Furthermore we define a decreasing sequence $(\psi_n)_{n \ge 1}$ of positive reals in the following way. If λ_n is such that $T_1^{(l)} \le \lambda_n \le T_{w_l}^{(l)}$, then $\psi_n := \frac{1}{M_l}$.

Obviously we have

$$\lim_{n \to \infty} \psi_n = 0$$

and

$$\sum_{n=1}^{\infty} \psi_n \ge \sum_{l=1}^{\infty} w_l \frac{1}{M_l} \ge \sum_{l=1}^{\infty} \frac{c_2}{2} M_l \frac{1}{M_l} = \infty.$$

We will show below that (λ_n) and (ψ_n) satisfy the condition (*) of Lemma 1.

We choose $N := w_1 + \ldots + w_l$ and first estimate $\sum_{n\le N} \mu_n \lambda_n^{-1} \psi_n$ from below (for the definition of μ_n see Lemma 1). We have

$$\sum_{n\le N} \mu_n \lambda_n^{-1} \psi_n \ge \sum_{n=N-w_l+1}^{N} \mu_n \lambda_n^{-1} \psi_n .$$

In the following we estimate μ_n from below for n with $N - w_l + 1 \le n \le N$, i.e., $\lambda_n = T_i^{(l)} = R_i^{(l)} \kappa_l$ for some i with $1 \le i \le w_l$.

Consider first λ_q with $q \le w_1 + \ldots + w_{l-1}$. Then the number of j with $0 \le j < \lambda_n$ such that $j\lambda_n^{-1}$ is of the form $k\lambda_q^{-1}$ with $0 \le k < \lambda_q$ trivially is at most λ_q.

Now consider λ_q with $q > w_1 + \ldots + w_{l-1}$ and $\lambda_q < \lambda_n$, i.e.,

$$\lambda_q = T_h^{(l)} = R_h^{(l)} \kappa_l$$

for some h with $1 \le h < i$. Then the number of j with $0 \le j < \lambda_n$ such that $j\lambda_n^{-1}$ is **not** of the form $k\lambda_q^{-1}$ with $0 \le k < \lambda_q$, i.e., such that

$$\begin{aligned} \frac{j}{\lambda_n} = \frac{k}{\lambda_q} &\Leftrightarrow \frac{j}{R_i^{(l)} \kappa_l} = \frac{k}{R_h^{(l)} \kappa_l} \\ &\Leftrightarrow \frac{j}{R_i^{(l)}} = \frac{k}{R_h^{(l)}} \end{aligned}$$

does **not** hold, is at least $\varphi\left(R_i^{(l)}\right) \kappa_l$. Hence by (8) and by (9)

$$\begin{aligned} \mu_n &\ge \varphi\left(R_i^{(l)}\right) \kappa_l - \sum_{q=1}^{w_1+\ldots+w_{l-1}} \lambda_q \\ &\ge \varphi\left(R_i^{(l)}\right) \kappa_l - \sqrt{M_l} \ge \frac{1}{2} \varphi\left(R_i^{(l)}\right) \kappa_l \end{aligned}$$

for all l large enough, say $l \ge l_0$ (note that $R_i^{(l)} \ge \frac{c_2}{2} M_l$).

Therefore for $l \ge l_0$

$$\begin{aligned} \sum_{n\le N} \mu_n \lambda_n^{-1} \psi_n &\ge \sum_{n=N-w_l+1}^{N} \mu_n \lambda_n^{-1} \psi_n \\ &\ge \frac{1}{M_l} \sum_{i=1}^{w_l} \frac{1}{2} \varphi\left(R_i^{(l)}\right) \kappa_l \frac{1}{R_i^{(l)} \kappa_l} \\ &= \frac{1}{2M_l} \sum_{i=1}^{w_l} \frac{\varphi\left(R_i^{(l)}\right)}{R_i^{(l)}} . \end{aligned} \tag{10}$$

Later on we will use the same chain of inequalities starting from the second expression in (10).

We recall that $w_l \geq \frac{c_2}{2} M_l$, and $R_i^{(l)} \leq KM_l$ for all $i = 1, \ldots, w_l$. Hence $R_1^{(l)}, \ldots, R_{w_l}^{(l)}$ form a subset of $\{1, 2, \ldots, KM_l\}$ of density at least $c_3 := \frac{c_2}{2K}$. Hence by Lemma 2 we have for l large enough and with c from Lemma 2 that

$$\sum_{n \leq N} \mu_n \lambda_n^{-1} \psi_n \geq \frac{K}{2} c \left(\frac{c_2}{2K} \right) =: c_4 > 0. \tag{11}$$

This holds for all $N = w_1 + \ldots + w_l$ and all $l \geq l_0$.

Finally we have to choose the function $(\tau_n)_{n \geq 1}$ from condition (*) in Lemma 1 in a suitable way. If λ_n is such that $T_1^{(l)} \leq \lambda_n \leq T_{w_l}^{(l)}$, i.e., if $\psi_n = \frac{1}{M_l}$, then we set $\tau_n := \frac{1}{l}$. Then

$$\sum_{n=1}^{\infty} \psi_n \tau_n \geq \sum_{l=1}^{\infty} w_l \frac{1}{M_l} \frac{1}{l} \geq \sum_{l=1}^{\infty} \frac{c_2}{2} M_l \frac{1}{M_l} \frac{1}{l} = \infty.$$

Finally, on the one hand for all $N = w_1 + \ldots + w_l$ we have by (10) and (11) that

$$\sum_{n \leq N} \mu_n \lambda_n^{-1} \psi_n \tau_n \geq \sum_{l'=l_0}^{l} c_4 \frac{1}{l'} \geq c_5 \log l$$

for all $l \geq l_0$.

On the other hand we have

$$\sum_{n \leq N} \psi_n \tau_n \leq \sum_{l'=1}^{l} w_{i_l} \frac{1}{M_l} \frac{1}{l} = \sum_{l'=1}^{l} K \frac{1}{l} \leq c_6 \log l.$$

Consequently

$$\sum_{n \leq N} \mu_n \lambda_n^{-1} \psi_n \tau_n \geq c_5 \log l \geq \frac{c_5}{c_6} \sum_{n \leq N} \psi_n \tau_n$$

and the conditions of Lemma 1 are satisfied for $(\lambda_n)_{n \geq 1}$ and $(\psi_n)_{n \geq 1}$. We conclude from Lemma 1 that for almost all α there exist infinitely many n such that $\|\lambda_n \alpha\| \leq \psi_n$ holds. Let such an α be given, and let $n_1 < n_2 < n_3 < \ldots$ be such that $\|\lambda_{n_i} \alpha\| \leq \psi_{n_i}$ for all $i = 1, 2, 3, \ldots$. For any n_i let $l(n_i)$ be defined such that $w_1 + w_2 + \ldots + w_{l(n_i)-1} < n_i \leq w_1 + w_2 + \ldots + w_{l(n_i)}$, then $\psi_{n_i} = \frac{1}{M_{l(n_i)}}$, hence

$$0 \leq \|\lambda_{n_i} \alpha\| \, M_{l(n_i)} < 1$$

for all i.

Let ρ with $0 \leq \rho \leq 1$ be a limit point of $\left(\|\lambda_{n_i}\alpha\| M_{l(n_i)}\right)_{i=1,2,\ldots}$. We distinguish now between two cases.

First case: $\rho = 0$.

Then there exists a subsequence $m_1 < m_2 < m_3 < \ldots$ of $n_1 < n_2 < n_3 < \ldots$ such that

$$0 \leq \|\lambda_{m_i}\alpha\| < \frac{1}{M_{l(m_i)}} \frac{c_2}{4K^2}$$

for all i. λ_{m_i} is an element of $V_{l(m_i)}$ with multiplicity at least $c_2 M_{l(m_i)}$. Hence there exist at least $c_2 M_{l(m_i)}$ pairs (p, q) with

$$1 \leq p < q \leq s_{l(m_i)} < KM_{l(m_i)}$$

and

$$\|\{a(q)\alpha\} - \{a(p)\alpha\}\| < \frac{1}{M_{l(m_i)}} \frac{c_2}{4K^2}.$$

Let now $s = \frac{c_2}{4K}$ then for all $M = KM_{l(m_i)}$ we have

$$\frac{1}{M}\#\left\{1 \leq p \neq q \leq M : \|\{a(q)\alpha\} - \{a(p)\alpha\}\| \leq \frac{s}{M}\right\} \geq \frac{c_2}{K} = 4s,$$

and hence

$$R_2([-s, s], \alpha, M) \not\to 2s.$$

Second case: $\rho > 0$.

Let $\varepsilon := \min\left(\frac{\rho}{2}, \frac{c_2}{8K^2}\right) > 0$. Then there exists a subsequence $m_1 < m_2 < m_3 < \ldots$ of $n_1 < n_2 < n_3 < \ldots$ such that

$$0 \leq \left|M_{l(m_i)} \|\lambda_{m_i}\alpha\| - \rho\right| < \varepsilon$$

for all i. Hence there exist at least $c_2 M_{l(m_i)}$ pairs (p, q) with $1 \leq p < q \leq s_{l(m_i)} < KM_{l(m_i)}$ and

$$\|\{a(q)\alpha\} - \{a(p)\alpha\}\| \in \left[\frac{\rho - \varepsilon}{M_{l(m_i)}}, \frac{\rho + \varepsilon}{M_{l(m_i)}}\right].$$

Let $s_1 := K(\rho - \varepsilon)$ and $s_2 := K(\rho + \varepsilon)$, then $s_2 - s_1 = 2K\varepsilon \leq \frac{c_2}{4K}$. Let for $M := KM_{l(m_i)}$ and $j = 1, 2$:

$$\Lambda^{(j)} := \frac{1}{M}\#\left\{1 \leq p \neq q \leq M : \|\{a(q)\alpha\} - \{a(p)\alpha\}\| \leq \frac{s_j}{M}\right\}.$$

Then $\varLambda^{(2)} - \varLambda^{(1)} \geq \frac{1}{M} c_2 \frac{M}{K} = \frac{c_2}{K}$. Hence at least one of

$$\left|\varLambda^{(2)} - 2s_2\right| \geq \frac{c_2}{8K} \quad \text{or}$$

$$\left|\varLambda^{(1)} - 2s_1\right| \geq \frac{c_2}{8K} \quad \text{holds,}$$

since otherwise

$$\begin{aligned}\frac{c_2}{2K} &\leq \left|\varLambda^{(2)} - \varLambda^{(1)}\right| - 2\left(s_2 - s_1\right)\\ &\leq \left|\varLambda^{(2)} - 2s_2 - \varLambda^{(1)} + 2s_1\right| \leq \left|\varLambda^{(2)} - 2s_2\right| + \left|\varLambda^{(1)} - 2s_1\right|\\ &\leq \frac{c_2}{4K},\end{aligned}$$

which is a contradiction. Therefore either

$$R_2\left(\left[-s_1, s_1\right], \alpha, M\right) \not\to 2s_1 \quad \text{or}$$

$$R_2\left(\left[-s_2, s_2\right], \alpha, M\right) \not\to 2s_2,$$

which proves the theorem.

Acknowledgements The first author is supported by CERN, European Organization for Nuclear Research, Geneva, as part of the Austrian Doctoral Student Programme.

The second author is supported by the Austrian Science Fund (FWF), START-project Y-901.

The third author is supported by the Austrian Science Fund (FWF): Project F5507-N26, which is part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

References

1. Aistleitner, C., Larcher, G., Lewko, M.: Additive energy and the Hausdorff dimension of the exceptional set in metric pair correlation problems. With an appendix by Jean Bourgain. Isr. J. Math. (to appear). Available at https://arxiv.org/abs/1606.03591
2. Balog, A., Szemerédi, E.: A statistical theorem of set addition. Combinatorica **14**(3), 263–268 (1994).
3. Berkes, I., Philipp, W., Tichy, R.: Pair correlations and U-statistics for independent and weakly dependent random variables. Ill. J. Math. **45**(2), 559–580 (2001)
4. Berry, M., Tabor, M.: Level clustering in the regular spectrum. R. Soc. Lond. Proc. Ser. A **356**, 375–394 (1977)
5. Bleher, P.M.: The energy level spacing for two harmonic oscillators with golden mean ratio of frequencies. J. Statist. Phys. **61**(3–4), 869–876 (1990)
6. Boca, F.P., Zaharescu, A.: Pair correlation of values of rational functions (mod p). Duke Math. J. **105**(2), 267–307 (2000)
7. Bogomolny, E.: Quantum and arithmetic chaos. In: Proceedings of the Les Houches Winter School "Frontiers in Number Theory, Physics and Geometry" (2003)

8. Casati, G., Guarneri, I., Izraĭ lev, F.M.: Statistical properties of the quasi-energy spectrum of a simple integrable system. Phys. Lett. A **124**(4–5), 263–266 (1987)
9. Cassels, J.W.S.: Some metrical theorems in Diophantine approximation. I. Proc. Cambridge Philos. Soc. **46**, 209–218 (1950)
10. Chaubey, S., Lanius, M., Zaharescu, A.: Pair correlation of fractional parts derived from rational valued sequences. J. Number Theory **151**, 147–158 (2015)
11. Freĭ man, G.A.: Foundations of a Structural Theory of Set Addition. Translations of Mathematical Monographs, vol. 37. American Mathematical Society, Providence, RI (1973). Translated from the Russian
12. Gowers, W.T.: A new proof of Szemerédi's theorem for arithmetic progressions of length four. Geom. Funct. Anal. **8**(3), 529–551 (1998)
13. Heath-Brown, D.R.: Pair correlation for fractional parts of αn^2. Math. Proc. Cambridge Philos. Soc. **148**(3), 385–407 (2010)
14. Lev, V.: The (Gowers–)Balog–Szemerédi theorem: an exposition http://people.math.gatech.edu/~ecroot/8803/baloszem.pdf
15. Marklof, J.: The Berry-Tabor conjecture. In: European Congress of Mathematics, Vol. II, Barcelona, 2000. Progress in Mathematics, vol. 202, pp. 421–427. Birkhäuser, Basel (2001)
16. Marklof, J.: Energy level statistics, lattice point problems, and almost modular functions. In: Frontiers in Number Theory, Physics, and Geometry. I, pp. 163–181. Springer, Berlin (2006)
17. Marklof, J., Strömbergsson, A.: Equidistribution of Kronecker sequences along closed horocycles. Geom. Funct. Anal. **13**(6), 1239–1280 (2003)
18. Montgomery, H.L.: The pair correlation of zeros of the zeta function. In: Analytic Number Theory (Proceedings of Symposia in Pure Mathematics, Vol. XXIV, St. Louis University, St. Louis, MO, 1972), pp. 181–193. American Mathematical Society, Providence, RI (1973)
19. Paul, B.: Tea time at Princeton. Harv. Coll. Math. Rev. **4**, 41–53 (2012)
20. Rudnick, Z., Sarnak, P.: The pair correlation function of fractional parts of polynomials. Comm. Math. Phys. **194**(1), 61–70 (1998)
21. Rudnick, Z., Zaharescu, A.: A metric result on the pair correlation of fractional parts of sequences. Acta Arith. **89**(3), 283–293 (1999)
22. Rudnick, Z., Sarnak, P., Zaharescu, A.: The distribution of spacings between the fractional parts of $n^2\alpha$. Invent. Math. **145**(1), 37–57 (2001)
23. Sarnak, P.: Values at integers of binary quadratic forms. In: Harmonic Analysis and Number Theory (Montreal, PQ, 1996). CMS Conference Proceedings, vol. 21, pp. 181–203. American Mathematical Society, Providence, RI (1997)
24. Shparlinski, I.E.: On small solutions to quadratic congruences. J. Number Theory **131**(6), 1105–1111 (2011)
25. Shparlinski, I.E.: On the restricted divisor function in arithmetic progressions. Rev. Mat. Iberoam. **28**(1), 231–238 (2012)
26. Sloan, I.: The method of polarized orbitals for the elastic scattering of slow electrons by ionized helium and atomic hydrogen. Roy. Soc. Lond. Proc. Ser. A **281**, 151–163 (1964)
27. Sós, V.T.: On the distribution mod 1 of the sequence $n\alpha$. Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Math. **1**, 127–134 (1958)
28. Tao, T., Vu, V.: Additive Combinatorics. Cambridge Studies in Advanced Mathematics, vol. 105. Cambridge University Press, Cambridge (2006)
29. Truelsen, J.L.: Divisor problems and the pair correlation for the fractional parts of $n^2\alpha$. Int. Math. Res. Not. **2010**(16), 3144–3183 (2010)
30. Vanderkam, J.M.: Values at integers of homogeneous polynomials. Duke Math. J. **97**(2), 379–412 (1999)
31. Weingartner, A.: The distribution functions of $\sigma(n)/n$ and $n/\phi(n)$. Proc. Am. Math. Soc. **135**(9), 2677–2681 (2007)

Towards an Efficient Finite Element Method for the Integral Fractional Laplacian on Polygonal Domains

Mark Ainsworth and Christian Glusa

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We explore the connection between fractional order partial differential equations in two or more spatial dimensions with boundary integral operators to develop techniques that enable one to efficiently tackle the integral fractional Laplacian. In particular, we develop techniques for the treatment of the dense stiffness matrix including the computation of the entries, the efficient assembly and storage of a sparse approximation and the efficient solution of the resulting equations. The main idea consists of generalising proven techniques for the treatment of boundary integral equations to general fractional orders. Importantly, the approximation does not make any strong assumptions on the shape of the underlying domain and does not rely on any special structure of the matrix that could be exploited by fast transforms. We demonstrate the flexibility and performance of this approach in a couple of two-dimensional numerical examples.

1 Introduction

Large scale computational solution of partial differential equations has revolutionised the way in which scientific research is performed. Historically, it was

M. Ainsworth (✉)
Division of Applied Mathematics, Brown University, Providence, RI, USA

Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: Mark_Ainsworth@Brown.edu

C. Glusa
Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA

Division of Applied Mathematics, Brown University, Providence, RI, USA
e-mail: caglusa@sandia.gov

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_2

generally the case that the mathematical models, expressed in the form of partial differential equations involving operators such as the Laplacian, were impossible to solve analytically, and difficult to resolve numerically. This led to a concerted and sustained research effort into the development of efficient numerical methods for approximating the solution of partial differential equations. Indeed, many researchers who were originally interested in applications shifted research interests to the development and analysis of numerical methods. A case in point is Professor Ian H. Sloan who originally trained as physicist but went on to carry out fundamental research in a wide range of areas relating to computational mathematics. Indeed, one may struggle to find an area of computational mathematics in which Sloan has not made a contribution and the topic of the present article, fractional partial differential equations, may be one of the very few.

In recent years, there has been a burgeoning of interest in the use of non-local and fractional models. To some extent, this move reflects the fact that with present day computational resources coupled with state of the art numerical algorithms, attention is now shifting back to the fidelity of the underlying mathematical models as opposed to their approximation. Fractional equations have been used to describe phenomena in anomalous diffusion, material science, image processing, finance and electromagnetic fluids [30]. Fractional order equations arise naturally as the limit of discrete diffusion governed by stochastic processes [20].

Whilst the development of fractional derivatives dates back to essentially the same time as their integer counterparts, the computational methods available for their numerical resolution drastically lags behind the vast array of numerical techniques from which one can choose to treat integer order partial differential equations. The recent literature abounds with work on numerical methods for fractional partial differential equations in one spatial dimension and fractional order temporal derivatives. However, with most applications of interest being posed on domains in two or more spatial dimensions, the solution of fractional equations posed on complex domains is a problem of considerable practical interest.

The archetypal elliptic partial differential equation is the Poisson problem involving the standard Laplacian. By analogy, one can consider a fractional Poisson problem involving the fractional Laplacian. The first problem one encounters is that of how to define a fractional Laplacian, particularly in the case where the domain is compact, and a number of alternatives have been suggested. The *integral fractional Laplacian* is obtained by restriction of the Fourier definition to functions that have prescribed value outside of the domain of interest, whereas the spectral fractional Laplacian is based on the spectral decomposition of the regular Laplace operator. In general, the two operators are different [24], and only coincide when the domain of interest is the full space.

The approximation of the integral fractional Laplacian using finite elements was considered by D'Elia and Gunzburger [10]. The important work of Acosta and Borthagaray [1] gave regularity results for the analytic solution of the fractional Poisson problem and obtained convergence rates for the finite element approximation supported by numerical examples computed using techniques described in [2].

The numerical treatment of fractional partial differential equations is rather different from the integer order case owing to the fact that the fractional derivative is a non-local operator. This creates a number of issues including the fact that the resulting stiffness matrix is dense and, moreover, the entries in the matrix are given in terms of singular integrals. In turn, these features create issues in the numerical computation of the entries and the need to store the entries of a dense matrix, not to mention the fact that a solution of the resulting matrix equation has to be computed. The seasoned reader will readily appreciate that many of these issues are shared by boundary integral equations arising from classical integer order differential operators [26, 27, 31]. This similarity is not altogether surprising given that the boundary integral operators are pseudo-differential operators of fractional order.

A different, integer order operator based approach, was taken by Nochetto, Otárola and Salgado [21] for the case of the spectral Laplacian. Caffarelli and Silvestre [7] showed that the operator can be realised as a Dirichlet-to-Neumann operator of an extended problem in the half space in $d + 1$ dimensions.

In the present work, we explore the connection with boundary integral operators to develop techniques that enable one to efficiently tackle the integral fractional Laplacian. In particular, we develop techniques for the treatment of the stiffness matrix including the computation of the entries, the efficient storage of the resulting dense matrix and the efficient solution of the resulting equations. The main ideas consist of generalising proven techniques for the treatment of boundary integral equations to general fractional orders. Importantly, the approximation does not make any strong assumptions on the shape of the underlying domain and does not rely on any special structure of the matrix that could be exploited by fast transforms. We demonstrate the flexibility and performance of this approach in a couple of two-dimensional numerical examples.

2 The Integral Fractional Laplacian and Its Weak Formulation

The fractional Laplacian in $\mathbb{R}^d$ of order s, for $0 < s < 1$ and $d \in \mathbb{N}$, of a function u can be defined by the Fourier transform $\mathscr{F}$ as

$$(-\Delta)^s u = \mathscr{F}^{-1}\left[|\xi|^{2s}\mathscr{F}u\right].$$

Alternatively, this expression can be rewritten [29] in integral form as

$$(-\Delta)^s u(\mathbf{x}) = C(d,s)\,\mathrm{p.\,v.}\int_{\mathbb{R}^d} d\mathbf{y}\,\frac{u(\mathbf{x}) - u(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{d+2s}}$$

where

$$C(d,s)=\frac{2^{2s}s\Gamma\left(s+\frac{d}{2}\right)}{\pi^{d/2}\Gamma\left(1-s\right)}$$

is a normalisation constant and p. v. denotes the Cauchy principal value of the integral [19, Chapter 5]. In the case where $s = 1$ this operator coincides with the usual Laplacian. If $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, we define the *integral fractional Laplacian* $(-\Delta)^s$ to be the restriction of the full-space operator to functions with compact support in Ω. This generalises the homogeneous Dirichlet condition applied in the case $s = 1$ to the case $s \in (0, 1)$.

Define the usual fractional Sobolev space $H^s\left(\mathbb{R}^d\right)$ via the Fourier transform. If Ω is a sub-domain as above, then we define the Sobolev space $H^s\left(\Omega\right)$ to be

$$H^s\left(\Omega\right):=\left\{u\in L^2\left(\Omega\right)\mid \|u\|_{H^s(\Omega)}<\infty\right\},$$

equipped with the norm

$$\|u\|_{H^s(\Omega)}^2=\|u\|_{L^2(\Omega)}^2+\int_\Omega d\mathbf{x}\int_\Omega d\mathbf{y}\frac{(u(\mathbf{x})-u(\mathbf{y}))^2}{|\mathbf{x}-\mathbf{y}|^{d+2s}}.$$

The space

$$\widetilde{H}^s\left(\Omega\right):=\left\{u\in H^s\left(\mathbb{R}^d\right)\mid u=0\text{ in }\Omega^c\right\}$$

can be equipped with the energy norm

$$\|u\|_{\widetilde{H}^s(\Omega)}:=\sqrt{\frac{C(d,s)}{2}}\,|u|_{H^s\left(\mathbb{R}^d\right)},$$

where the non-standard factor $\sqrt{C(d,s)/2}$ is included for convenience. For $s > 1/2$, $\widetilde{H}^s\left(\Omega\right)$ coincides with the space $H_0^s\left(\Omega\right)$ which is the closure of $C_0^\infty\left(\Omega\right)$ with respect to the $H^s\left(\Omega\right)$-norm. For $s < 1/2$, $\widetilde{H}^s\left(\Omega\right)$ is identical to $H^s\left(\Omega\right)$. In the critical case $s = 1/2$, $\widetilde{H}^s\left(\Omega\right) \subset H_0^s\left(\Omega\right)$, and the inclusion is strict. (See for example [19, Chapter 3].)

The usual approach to dealing with elliptic PDEs consists of obtaining a weak form of the operator by multiplying the equation by a test function and applying integration by parts [13]. In contrast, for equations involving the fractional Laplacian $(-\Delta)^s u$, we again multiply by a test function $v \in \widetilde{H}^s\left(\Omega\right)$ and integrate over $\mathbb{R}^d$, and then, instead of integration by parts, we use the identity

$$\int_{\mathbb{R}^d} d\mathbf{x}\int_{\mathbb{R}^d} d\mathbf{y}\frac{\left(u\left(\mathbf{x}\right)-u\left(\mathbf{y}\right)v\left(\mathbf{x}\right)\right)}{|\mathbf{x}-\mathbf{y}|^{d+2s}}=-\int_{\mathbb{R}^d} d\mathbf{x}\int_{\mathbb{R}^d} d\mathbf{y}\frac{\left(u\left(\mathbf{x}\right)-u\left(\mathbf{y}\right)v\left(\mathbf{y}\right)\right)}{|\mathbf{x}-\mathbf{y}|^{d+2s}}.$$

Following this approach, since both u and v vanish outside of Ω, we arrive at the bilinear form

$$a(u,v) = b\,(u,v) + C(d,s)\int_{\Omega} d\mathbf{x}\int_{\Omega^c} d\mathbf{y}\frac{u\,(\mathbf{x})\,v\,(\mathbf{x})}{|\mathbf{x}-\mathbf{y}|^{d+2s}},$$

with

$$b(u,v) = \frac{C(d,s)}{2}\int_{\Omega} d\mathbf{x}\int_{\Omega} d\mathbf{y}\frac{(u\,(\mathbf{x})-u\,(\mathbf{y}))\,(v\,(\mathbf{x})-v\,(\mathbf{y}))}{|\mathbf{x}-\mathbf{y}|^{d+2s}},$$

corresponding to $(-\Delta)^s$ on $\widetilde{H}^s\,(\Omega)\times\widetilde{H}^s\,(\Omega)$. The bilinear form $a\,(\cdot,\cdot)$ is trivially seen to be $\widetilde{H}^s\,(\Omega)$-coercive and continuous and, as such, is amenable to treatment using the Lax-Milgram Lemma.

In this article we shall concern ourselves with the computational details needed to implement the finite element approximation of problems involving the fractional Laplacian. To this end, the presence of the unbounded domain Ω^c in the bilinear form $a\,(\cdot,\cdot)$ is somewhat undesirable. Fortunately, we can dispense with Ω^c using the following argument. The identity

$$\frac{1}{|\mathbf{x}-\mathbf{y}|^{d+2s}} = \frac{1}{2s}\nabla_{\mathbf{y}}\cdot\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^{d+2s}},$$

enables the second integral to be rewritten using the Gauss theorem as

$$\frac{C(d,s)}{2s}\int_{\Omega} d\mathbf{x}\int_{\partial\Omega} d\mathbf{y}\frac{u\,(\mathbf{x})\,v\,(\mathbf{x})\;\,\mathbf{n}_{\mathbf{y}}\cdot(\mathbf{x}-\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{d+2s}},$$

where $\mathbf{n}_y$ is the *inward* normal to $\partial\Omega$ at $\mathbf{y}$, so that the bilinear form can be expressed equivalently as

$$\begin{aligned} a(u,v) = {} & \frac{C(d,s)}{2}\int_{\Omega} d\mathbf{x}\int_{\Omega} d\mathbf{y}\frac{(u\,(\mathbf{x})-u\,(\mathbf{y}))\,(v\,(\mathbf{x})-v\,(\mathbf{y}))}{|\mathbf{x}-\mathbf{y}|^{d+2s}} \\ & + \frac{C(d,s)}{2s}\int_{\Omega} d\mathbf{x}\int_{\partial\Omega} d\mathbf{y}\frac{u\,(\mathbf{x})\,v\,(\mathbf{x})\;\,\mathbf{n}_{\mathbf{y}}\cdot(\mathbf{x}-\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{d+2s}}. \end{aligned}$$

As an aside, we note that the bilinear form $b\,(u,v)$ represents the so-called *regional fractional Laplacian* [5, 8]. The regional fractional Laplacian can be interpreted as a generalisation of the usual Laplacian with homogeneous Neumann boundary condition for $s=1$ to the case of fractional orders $s\in(0,1)$. It will transpire from our work that most of the presented techniques carry over to the regional fractional Laplacian by simply omitting the boundary integral terms.

3 Finite Element Approximation of the Fractional Poisson Equation

The *fractional Poisson problem*

$$\begin{aligned}(-\Delta)^s u &= f \quad \text{in}\Omega,\\ u &= 0 \quad \text{in}\Omega^c\end{aligned}$$

takes the variational form

$$\text{Find } u \in \widetilde{H}^s(\Omega): \quad a(u,v) = \langle f, v\rangle \quad \forall v \in \widetilde{H}^s(\Omega). \tag{1}$$

Henceforth, let Ω be a polygon, and let $\mathscr{P}_h$ be a family of shape-regular and globally quasi-uniform triangulations of Ω, and $\mathscr{P}_{h,\partial}$ the induced boundary meshes [13]. Let $\mathscr{N}_h$ be the set of vertices of $\mathscr{P}_h$ and h_K be the diameter of the element $K \in \mathscr{P}_h$, and h_e the diameter of $e \in \mathscr{P}_{h,\partial}$. Moreover, let $h := \max_{K\in\mathscr{P}_h} h_K$. Let ϕ_i be the usual piecewise linear basis function associated with a node $\mathbf{z}_i \in \mathscr{N}_h$, satisfying $\phi_i(\mathbf{z}_j) = \delta_{ij}$ for $\mathbf{z}_j \in \mathscr{N}_h$, and let $X_h := \operatorname{span}\{\phi_i \mid \mathbf{z}_i \in \mathscr{N}_h\}$. The finite element subspace $V_h \subset \widetilde{H}^s(\Omega)$ is given by $V_h = X_h$ when $s < 1/2$ and by

$$V_h = \{v_h \in X_h \mid v_h = 0 \text{ on } \partial\Omega\} = \operatorname{span}\{\phi_i \mid \mathbf{z}_i \notin \partial\Omega\}$$

when $s \geq 1/2$. The corresponding set of degrees of freedom $\mathscr{I}_h$ for V_h is given by $\mathscr{I}_h = \mathscr{N}_h$ when $s < 1/2$ and otherwise consists of nodes in the interior of Ω. In both cases we denote the cardinality of $\mathscr{I}_h$ by n. The set of degrees of freedom on an element $K \in \mathscr{P}_h$ is denoted by $\mathscr{I}_K$.

The stiffness matrix associated with the fractional Laplacian is defined to be $\boldsymbol{A}^s = \{a(\phi_i,\phi_j)\}_{i,j}$, where

$$\begin{aligned} a(\phi_i,\phi_j) = {}& \frac{C(d,s)}{2}\int_\Omega d\mathbf{x}\int_\Omega d\mathbf{y}\frac{(\phi_i(\mathbf{x})-\phi_i(\mathbf{y}))\left(\phi_j(\mathbf{x})-\phi_j(\mathbf{y})\right)}{|\mathbf{x}-\mathbf{y}|^{d+2s}}\\ &+\frac{C(d,s)}{2s}\int_\Omega d\mathbf{x}\int_{\partial\Omega} d\mathbf{y}\frac{\phi_i(\mathbf{x})\,\phi_j(\mathbf{x})\;\mathbf{n}_\mathbf{y}\cdot(\mathbf{x}-\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{d+2s}}.\end{aligned}$$

The existence of a unique solution to the fractional Poisson problem Eq. (1) and its finite element approximation follows from the Lax-Milgram Lemma.

The rate of convergence of the finite element approximation is given by the following theorem:

Theorem 1 ([1]) *If the family of triangulations $\mathscr{P}_h$ is shape regular and globally quasi-uniform, and $u \in H^\ell(\Omega)$, for $0 < s < \ell < 1$ or $1/2 < s < 1$ and $1 < \ell < 2$, then*

$$\|u-u_h\|_{\widetilde{H}^s(\Omega)} \leq C(s,d)\,h^{\ell-s}\,|u|_{H^\ell(\Omega)}. \tag{2}$$

In particular, by applying regularity estimates for u in terms of the data f, the solution satisfies

$$\|u-u_h\|_{\widetilde{H}^s(\Omega)} \leq \begin{cases} C(s)\, h^{1/2}\, |\log h|\, \|f\|_{C^{1/2-s}(\Omega)} & \text{if } 0<s<1/2, \\ Ch^{1/2}\, |\log h|\, \|f\|_{L^\infty(\Omega)} & \text{if } s=1/2, \\ \frac{C(s,\beta)}{2s-1} h^{1/2} \sqrt{|\log h|}\, \|f\|_{C^\beta(\Omega)} & \text{if } 1/2<s<1, \beta>0 \end{cases}$$

Moreover, using a standard Aubin-Nitsche argument [13, Lemma 2.31] gives estimates in $L^2(\Omega)$:

Theorem 2 ([6]) *If the family of triangulations $\mathscr{P}_h$ is shape regular and globally quasi-uniform, and, for $\epsilon>0$, $u\in H^{s+1/2-\epsilon}(\Omega)$, then*

$$\|u-u_h\|_{L^2} \leq \begin{cases} C(s,\epsilon) h^{1/2+s-\epsilon}\, |u|_{H^{s+1/2-\epsilon}(\Omega)} & \text{if } 0<s<1/2, \\ C(s,\epsilon) h^{1-2\epsilon}\, |u|_{H^{s+1/2-\epsilon}(\Omega)} & \text{if } 1/2\leq s<1. \end{cases}$$

When $s=1$ classical results [13, Theorems 3.16 and 3.18] show that if $u\in H^\ell(\Omega)$, $1<\ell\leq 2$,

$$\|u-u_h\|_{H_0^1(\Omega)} \leq Ch^{\ell-1}\, |u|_{H^\ell(\Omega)},$$
$$\|u-u_h\|_{L^2(\Omega)} \leq Ch^{\ell}\, |u|_{H^\ell(\Omega)},$$

so that (2) can be seen as a generalisation to the case $s\in(0,1)$. For $s=1$, $u\in H^2(\Omega)$ if the domain is of class C^2 or a convex polygon and if $f\in L^2(\Omega)$ [13, Theorems 3.10 and 3.12]. However, when $s\in(0,1)$, higher order regularity of the solution is not guaranteed under such conditions.

For example, consider the problem

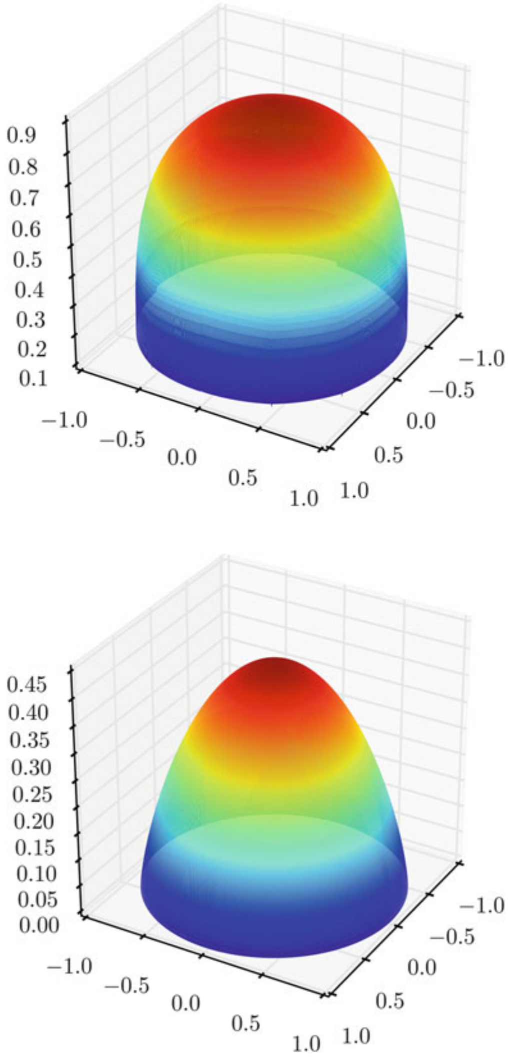
$$\begin{aligned} (-\Delta)^s u^s(\mathbf{x}) &= 1 \quad \text{in } \Omega=\left\{\mathbf{x}\in\mathbb{R}^2 \mid |\mathbf{x}|<1\right\}, \\ u^s(\mathbf{x}) &= 0 \quad \text{in } \Omega^c, \end{aligned}$$

with analytic solution [14]

$$u^s(\mathbf{x}) := \frac{2^{-2s}}{\Gamma(1+s)^2}\left(1-|\mathbf{x}|^2\right)^s.$$

Although the domain is C^∞ and the right-hand side is smooth, u^s is only in $H^{s+1/2-\epsilon}(\Omega)$ for any $\epsilon>0$. Sample solutions for $s\in\{0.25,0.75\}$ are shown in Fig. 1.

Fig. 1 Solutions u^s to the fractional Poisson equation with constant right-hand side for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*)



4 Computation of Entries of the Stiffness Matrix

The computation of entries of the stiffness matrix $\boldsymbol{A}^s$ in the case of the usual Laplacian ($s = 1$) is straightforward. However, for $s \in (0, 1)$, the bilinear form contains factors $|\mathbf{x} - \mathbf{y}|^{-d-2s}$ which means that simple closed forms for the entries are no longer available and suitable quadrature rules therefore must be identified. Moreover, the presence of a repeated integral over Ω (as opposed to an integral over just Ω in the case $s = 1$) means that the matrix needs to be assembled in a double loop over the elements of the mesh so that the computational cost is potentially much larger than in the integer $s = 1$ case. Additionally, every degree of freedom is coupled to all other degrees of freedom and the stiffness matrix is therefore dense.

4.1 Reduction to Smooth Integrals

In order to compute the entries of $\boldsymbol{A}^s = \left\{a\left(\phi_i, \phi_j\right)\right\}_{ij}$ we decompose the expression for the entries into contributions from elements $K, \tilde{K} \in \mathscr{P}_h$ and external edges $e \in \mathscr{P}_{h,\partial}$:

$$a(\phi_i, \phi_i) = \sum_K \sum_{\tilde{K}} a^{K\times\tilde{K}}(\phi_i, \phi_j) + \sum_K \sum_e a^{K\times e}(\phi_i, \phi_j),$$

where the contributions $a^{K\times\tilde{K}}$ and $a^{K\times e}$ are given by:

$$a^{K\times\tilde{K}}(\phi_i, \phi_j) = \frac{C(d,s)}{2} \int_K d\mathbf{x} \int_{\tilde{K}} d\mathbf{y} \frac{(\phi_i(\mathbf{x}) - \phi_i(\mathbf{y}))\left(\phi_j(\mathbf{x}) - \phi_j(\mathbf{y})\right)}{|\mathbf{x}-\mathbf{y}|^{d+2s}}, \tag{3}$$

$$a^{K\times e}(\phi_i, \phi_j) = \frac{C(d,s)}{2s} \int_K d\mathbf{x} \int_e d\mathbf{y} \frac{\phi_i\left(\mathbf{x}\right)\phi_j\left(\mathbf{x}\right)\ \mathbf{n}_e \cdot (\mathbf{x}-\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{d+2s}}. \tag{4}$$

Although the following approach holds for arbitrary spatial dimension d, we restrict ourselves to $d = 2$ dimensions. In evaluating the contributions $a^{K\times\tilde{K}}$ over element pairs $K \times \tilde{K}$, several cases need to be distinguished:

1. K and $\tilde{K}$ have empty intersection,
2. K and $\tilde{K}$ are identical,
3. K and $\tilde{K}$ share an edge,
4. K and $\tilde{K}$ share a vertex.

These cases are illustrated in Fig. 2. In case 1, where the elements do not touch, the Stroud conical quadrature rule [28] (or any other suitable Gauss rule on simplices)

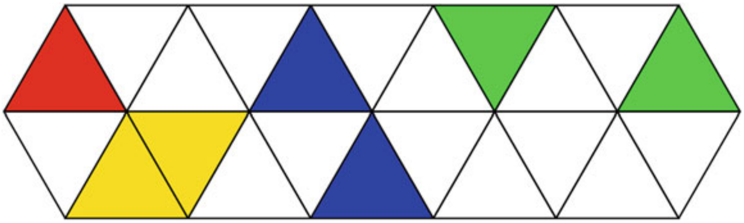


Fig. 2 Element pairs that are treated separately. We distinguish element pairs of identical elements (*red*), element pairs with common edge (*yellow*), with common vertex (*blue*) and separated elements (*green*)

of *sufficiently high order* can be used to approximate the integrals. More details as to what constitutes a sufficiently high order are given in Sect. 4.2.

Special care has to be taken in the remaining cases 2–4, in which the elements are touching, owing to the presence of a singularity in the integrand. Fortunately, the singularity is removable and can, as pointed out in [1], be treated using standard techniques from the boundary element literature [22]. More specifically, we write the integral as a sum of integrals over sub-simplices. Each sub-simplex is then mapped onto the hyper-cube $[0,1]^4$ using the Duffy transformation [11]. The advantage of pursuing this approach is that the singularity arising from the degenerate nature of the Duffy transformation offsets the singularity present in the integrals. For example, we obtain the following expressions

$$a^{K\times\tilde{K}}(\phi_i,\phi_j) = \frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|} \sum_{\ell=1}^{L_c}\int_{[0,1]^4} d\boldsymbol{\eta}\; \bar{J}^{(\ell,c)} \frac{\bar{\psi}_{k(i)}^{(\ell,c)}(\boldsymbol{\eta})\,\bar{\psi}_{k(j)}^{(\ell,c)}(\boldsymbol{\eta})}{\left|\sum_{k=0}^{6-c}\bar{\psi}_{k}^{(\ell,c)}(\boldsymbol{\eta})\,\mathbf{x}_k\right|^{2+2s}}, \tag{5}$$

and

$$a^{K\times e}(\phi_i,\phi_j) = \frac{C(2,s)}{2s}\frac{|K|}{\left|\hat{K}\right|}\frac{|e|}{|\hat{e}|} \sum_{\ell=1}^{L_c}\int_{[0,1]^3} d\boldsymbol{\eta}\; \bar{J}^{(\ell,c)} \frac{\phi_{k(i)}^{(\ell,c)}(\boldsymbol{\eta})\,\phi_{k(j)}^{(\ell,c)}(\boldsymbol{\eta})\;\sum_{k=0}^{5-c}\bar{\psi}_{k}^{(\ell,c)}(\boldsymbol{\eta})\,\mathbf{n}_e\cdot\mathbf{x}_k}{\left|\sum_{k=0}^{5-c}\bar{\psi}_{k}^{(\ell,c)}(\boldsymbol{\eta})\,\mathbf{x}_k\right|^{2+2s}} \tag{6}$$

in which the singularity $|\mathbf{x}-\mathbf{y}|^{-d-2s}$ is no longer present. The derivations of the terms involved can be found in [2, 22] and, for completeness, are summarised in the Appendix, along with the notations used in Eqs. (5) and (6). Removing the singularity means that the integrals in Eqs. (5) and (6) are amenable to approximation using standard Gaussian quadrature rules of *sufficiently high order* as discussed in Sect. 4.2. The same idea is applicable in any number of space dimensions.

4.2 Determining the Order of the Quadrature Rules

The foregoing considerations show that the evaluation of the entries of the stiffness matrix boils down to the evaluation of integrals with smooth integrands, i.e. expressions Eqs. (3) and (4) for case 1 and expressions Eqs. (5) and (6) for case 2–4. As mentioned earlier, it is necessary to use a *sufficiently high order* quadrature rule to approximate these integrals. We now turn to the question of how high is sufficient.

The arguments used to prove the ensuing estimates follow a pattern similar to the proofs of Theorems 5.3.29, 5.3.23 and 5.3.24 in [22]. The main difference from [22] is the presence of the boundary integral term. More details on the development of this type of quadrature rules in the context of boundary element methods can be found in the work of Erichsen and Sauter [12].

Theorem 3 *For $d = 2$, let $\mathscr{I}_K$ index the degrees of freedom on $K \in \mathscr{P}_h$, and define $\mathscr{I}_{K\times\tilde{K}} := \mathscr{I}_K \cup \mathscr{I}_{\tilde{K}}$. Let k_T (respectively $k_{T,\partial}$) be the quadrature order used for touching pairs $K \times \tilde{K}$ (respectively $K \times e$), and let $k_{NT}\left(K, \tilde{K}\right)$ (respectively $k_{NT,\partial}(K, e)$) be the quadrature order used for pairs that have empty intersection. Denote the resulting approximation to the bilinear form $a(\cdot,\cdot)$ by $a_Q(\cdot,\cdot)$. Then the consistency error due to quadrature is bounded by*

$$|a(u, v) - a_Q(u, v)| \le C\,(E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial})\,\|u\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} \qquad \forall u, v \in V_h,$$

where the errors are given by

$$\begin{aligned}
E_T &= h^{-2-2s}\rho_1^{-2k_T},\\
E_{NT} &= \max_{K,\tilde{K}\in\mathscr{P}_h,\overline{K}\cap\overline{\tilde{K}}=\emptyset} h^{-2} d_{K,\tilde{K}}^{-2s}\left(\rho_2 \frac{d_{K,\tilde{K}}}{h}\right)^{-2k_{NT}\left(K,\tilde{K}\right)},\\
E_{T,\partial} &= h^{-1-2s}\rho_3^{-2k_{T,\partial}},\\
E_{NT,\partial} &= \max_{K\in\mathscr{P}_h,e\in\mathscr{P}_{h,\partial},\overline{K}\cap\overline{e}=\emptyset} h^{-1} d_{K,e}^{-2s}\left(\rho_4 \frac{d_{K,e}}{h}\right)^{-2k_{NT,\partial}(K,e)},
\end{aligned}$$

$d_{K,\tilde{K}} := \inf_{\mathbf{x}\in K,\mathbf{y}\in\tilde{K}} |\mathbf{x}-\mathbf{y}|$, $d_{K,e} := \inf_{\mathbf{x}\in K,\mathbf{y}\in e} |\mathbf{x}-\mathbf{y}|$, and $\rho_j > 1$, $j = 1, 2, 3, 4$, are constants.

The proof of the Theorem is deferred to the Appendix.

The impact of the use of quadrature rules on the accuracy of the resulting finite element approximation can be quantified using Strang's first lemma [13, Lemma 2.27]:

$$\begin{aligned}
\|u - u_h\|_{\widetilde{H}^s(\Omega)} &\le C \inf_{v_h\in V_h}\left[\|u - v_h\|_{\widetilde{H}^s(\Omega)} + \sup_{w_h\in V_h}\frac{|a(v_h, w_h) - a_Q(v_h, w_h)|}{\|w_h\|_{\widetilde{H}^s(\Omega)}}\right]\\
&\le C \inf_{v_h\in V_h}\Bigg[\|u - v_h\|_{\widetilde{H}^s(\Omega)}\\
&\qquad + (E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial})\,\|v_h\|_{L^2(\Omega)} \sup_{w_h\in V_h}\frac{\|w_h\|_{L^2(\Omega)}}{\|w_h\|_{\widetilde{H}^s(\Omega)}}\Bigg]\\
&\le C \inf_{v_h\in V_h}\left[\|u - v_h\|_{\widetilde{H}^s(\Omega)} + (E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial})\,\|v_h\|_{L^2(\Omega)}\right],
\end{aligned}$$

where we used the Poincare inequality $\|w_h\|_{L^2(\Omega)} \leq C \|w_h\|_{\widetilde{H}^s(\Omega)}$ in the last step. We then use the Scott-Zhang interpolation operator Π_h [9, 23] and the estimate

$$\|u - \Pi_h u\|_{\widetilde{H}^s(\Omega)} \leq C h^{\ell - s} |u|_{H^\ell(\Omega)},$$

used in the proof of Theorem 1 to bound the first term on the right-hand side:

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C \left[h^{\ell-s} |u|_{H^\ell(\Omega)} + (E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial}) \|\Pi_h u\|_{L^2(\Omega)}\right].$$

We choose the quadrature rules in such a way that the remaining terms on the right-hand side are also of order $\mathcal{O}\left(h^{\ell-s}\right)$, i.e.

$$k_T \geq \frac{(\ell - s + 2 + 2s)}{2\log(\rho_1)} |\log h| - C, \tag{7}$$

$$k_{NT}\left(K, \tilde{K}\right) \geq \frac{((\ell - s)/2 + 1 + s) |\log h| - s \log \frac{d_{K,\tilde{K}}}{h} - C}{\log \frac{d_{K,\tilde{K}}}{h} + \log(\rho_2)}, \tag{8}$$

$$k_{T,\partial} \geq \frac{(\ell - s + 1 + 2s)}{2\log(\rho_3)} |\log h| - C, \tag{9}$$

$$k_{NT,\partial}(K, e) \geq \frac{((\ell - s)/2 + 1/2 + s) |\log h| - s \log \frac{d_{K,e}}{h} - C}{\log \frac{d_{K,e}}{h} + \log(\rho_4)}. \tag{10}$$

In particular, if the pair $K \times \tilde{K}$ (respectively $K \times e$) is well separated, so that $d_{K,\tilde{K}} \sim 1$ ($d_{K,e} \sim 1$), then

$$k_{NT}\left(K, \tilde{K}\right) \geq (\ell - s)/2 + 1,$$
$$k_{NT,\partial}(K, e) \geq (\ell - s)/2 + 1/2$$

is sufficient.

In practice, the quadrature order for non-touching element pairs can be chosen depending on $d_{K,\tilde{K}}$ using Eqs. (8) and (10), or an appropriate choice of cutoff distance D can be determined so that element pairs with $d_{K,\tilde{K}} < D$ are approximated using a quadrature rule with $\mathcal{O}(|\log h|)$ nodes, and pairs with $d_{K,\tilde{K}} \geq D$ are computed using a constant number of nodes.

It transpires from the expressions derived in the Appendix and the fact that $n \sim h^{-2}$ that the complexity to calculate the contributions by a single pair of elements K and $\tilde{K}$ scales like

- $\log n$ if the elements coincide,
- $(\log n)^2$ if the elements share only an edge,
- $(\log n)^3$ if the elements share only a vertex,
- $(\log n)^4$ if the elements have empty intersection, but are "near neighbours", and
- C if the elements are well separated.

Since $n \sim |\mathscr{P}_h|$, we cannot expect a straightforward assembly of the stiffness matrix to scale better than $\mathscr{O}\left(n^2\right)$. Similarly, its memory requirement is n^2, and a single matrix-vector product has complexity $\mathscr{O}\left(n^2\right)$, which severely limits the size of problems that can be considered.

5 Solving the Linear Systems

The fractional Poisson equation leads to the linear algebraic system

$$\boldsymbol{A}^s \mathbf{u} = \mathbf{b}, \tag{11}$$

whereas time-dependent problems (using implicit integration schemes) lead to systems of the form

$$\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \mathbf{u} = \mathbf{b}, \tag{12}$$

where Δt is the time-step size. In typical examples, the time-step will be chosen so that the orders of convergence in both spatial and temporal discretisation errors are balanced.

In both cases, the matrices are dense and the condition number of $\boldsymbol{A}^s$ grows as the mesh is refined ($h \to 0$). The cost of using a direct solver is prohibitively expensive, growing as $\mathscr{O}\left(n^3\right)$. An alternative is to use an iterative solver such as the conjugate gradient method but the rate of convergence will depend on the condition number. The following result quantifies how the condition number of $\boldsymbol{A}^s$ depends on the fractional order s and the mesh size h:

Theorem 4 ([4]) *For $s < d/2$, and a family of shape regular and globally quasi-uniform triangulations $\mathscr{P}_h$ with maximal element size h, the spectrum of the stiffness matrix satisfies*

$$ch^d \boldsymbol{I} \leq \boldsymbol{A}^s \leq Ch^{d-2s}\boldsymbol{I},$$

and hence the condition number of the stiffness matrix satisfies

$$\kappa\left(\boldsymbol{A}^s\right) = Ch^{-2s}.$$

The exponent of the growth of the condition number depends on the fractional order s. For small s, the matrix is better conditioned, similarly to the mass matrix in the case of integer order operators. As $s \to 1$, the growth of the condition number approaches $\mathscr{O}\left(h^{-2}\right)$, as for the usual Laplacian. Consequently, just as the conjugate gradient method fails to be efficient for the solution of equations arising from the discretisation of the Laplacian, CG becomes increasingly uncompetitive for the solution of equations arising from the fractional Laplacian.

In the integer order case, multigrid iterations have been used with great success for solving systems involving both the mass matrix and the stiffness matrix that arises from the discretisation of the regular Laplacian. It is therefore to be expected that the same will remain true for systems arising from the fractional Laplacian. In practice, a single multigrid iteration is much more expensive than a single iteration of conjugate gradient. The advantage of multigrid is, however, that the number of iterations is essentially independent of the number of unknowns n. Consequently, while the performance of CG degenerates as n increases, this will not be the case with multigrid making it attractive as a solver for the fractional Poisson problem.

Turning to the systems that arise from the discretisation of time-dependent problems, we first observe that an explicit scheme will lead to CFL conditions on the time-step size of the form $\Delta t \leq Ch^{2s}$. On the other hand, for implicit time-stepping, the following theorem shows that if the time-step $\Delta t = \mathcal{O}\left(h^{2s}\right)$, we can expect the conjugate gradient method to converge rapidly, at a rate which does not degenerate as n increases, in contrast with what is observed for steady problems:

Lemma 1 *For a shape regular and globally quasi-uniform family of triangulations $\mathscr{P}_h$ and time-step $\Delta t \leq 1$,*

$$\kappa\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \leq C\left(1 + \frac{\Delta t}{h^{2s}}\right).$$

Proof Since $ch^d \boldsymbol{I} \leq \boldsymbol{M} \leq Ch^d \boldsymbol{I}$, this also permits us to deduce that

$$c\left(h^d + \Delta t\, h^d\right)\boldsymbol{I} \leq \boldsymbol{M} + \Delta t \boldsymbol{A}^s \leq C\left(h^d + \Delta t\, h^{d-2s}\right)\boldsymbol{I}$$

and so

$$\kappa\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \leq C\left(1 + \frac{\Delta t}{h^{2s}}\right).$$

□

This shows that for a general time-step $\Delta t \geq h^{2s}$, the number of iterations the conjugate gradient method will require for systems of the form Eq. (12) will grow as $\sqrt{\Delta t / h^{2s}} \sim n^{s/d}\sqrt{\Delta t}$. Consequently, if Δt is large compared to h^{2s}, a multigrid solver outperforms conjugate gradient for the systems Eq. (12), but if Δt is on the same order as h^{2s}, conjugate gradient iterations will generally be more efficient than a multigrid method.

In this section we have concerned ourselves with the effect that the mesh and the fractional order have on the rate of convergence of iterative solvers. This, of course, ignores the cost of carrying out the iteration in which a matrix-vector multiply must be computed at each step. The complexity of both multigrid and conjugate gradient iterations depends on how efficiently the matrix-vector product $\boldsymbol{A}^s\mathbf{x}$ can be computed. By way of contrast, the mass matrix in Eq. (12) has $\mathcal{O}(n)$ entries, so its matrix-vector product scales linearly in the number of unknowns. Since all the basis functions ϕ_i interact with one another, the matrix $\boldsymbol{A}^s$ is dense and the associated matrix-vector product has complexity $\mathscr{O}\left(n^2\right)$. In the following section, we discuss a sparse approximation that will preserve the order of the approximation error of the fractional Laplacian, but display significantly better scaling in terms of both memory usage and operation counts for both assembly and matrix-vector product.

6 Sparse Approximation of the Matrix

The presence of a factor $|\mathbf{x}-\mathbf{y}|^{-d-2s}$ in the integrand in the expression for the entries of the stiffness matrix means that the contribution of pairs of elements that are well separated is significantly smaller than the contribution arising from pairs of elements that are close to one another. This suggests the use of the *panel clustering method* [17] from the boundary element literature, whereby such far field contributions are replaced by less expensive low-rank blocks rather than computing and storing all the individual entries from the original matrix. Conversely, the near-field contributions are more significant but involve only local couplings and hence the cost of storing the individual entries is a practical proposition. A full discussion of the panel clustering method is beyond the scope of the present work but can be found in [22, Chapter 7]. Here, we confine ourselves to stating only the necessary definitions and steps needed to describe our approach.

Definition 1 ([22]) A *cluster* is a union of one or more indices from the set of degrees of freedom $\mathscr{I}$. The nodes of a hierarchical cluster tree $\mathscr{T}$ are clusters. The set of all nodes is denoted by T and satisfies

1. $\mathscr{I}$ is a node of $\mathscr{T}$.
2. The set of leaves $\mathrm{Leaves}(\mathscr{T}) \subset T$ corresponds to the degrees of freedom $i \in \mathscr{I}$ and is given by

$$\mathrm{Leaves}(\mathscr{T}) := \{\{i\} : i \in \mathscr{I}\}.$$

3. For every $\sigma \in T\backslash\mathrm{Leaves}\,(\mathscr{T})$ there exists a minimal set $\Sigma\,(\sigma)$ of nodes in $T\backslash\{\sigma\}$ (i.e. of minimal cardinality) that satisfies

$$\sigma = \bigcup_{\tau\in\Sigma(\sigma)} \tau.$$

The set $\Sigma\,(\sigma)$ is called the sons of σ. The edges of the cluster tree $\mathscr{T}$ are the pairs of nodes $(\sigma,\tau) \in T \times T$ such that $\tau \in \Sigma\,(\sigma)$.

An example of a cluster tree for a one-dimensional problem is given in Fig. 3.

Definition 2 ([22]) The *cluster box* Q_σ of a cluster $\sigma \in T$ is the minimal hypercube which contains $\bigcup_{i\in\sigma} \operatorname{supp}\phi_i$. The *diameter* of a cluster is the diameter of its

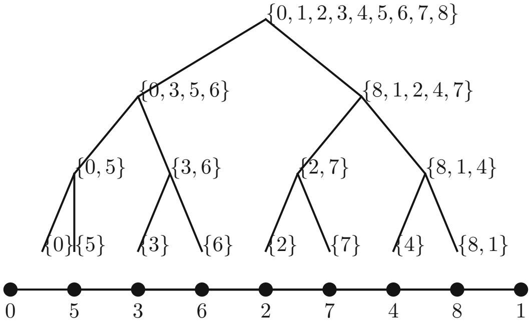


Fig. 3 Cluster tree for a one dimensional problem. For each cluster, the associated degrees of freedom are shown. The mesh with its nodal degrees of freedom is plotted at the bottom

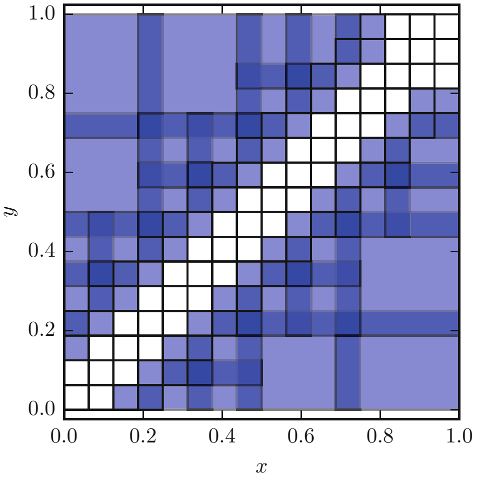


Fig. 4 Cluster pairs for a one dimensional problem. The cluster boxes of the admissible cluster pairs are coloured in light blue, and their overlap in darker blue. The diagonal cluster pairs are not admissible and are not approximated, but assembled in full

cluster box $\operatorname{diam}(\sigma) := \sup_{\mathbf{x},\mathbf{y}\in Q_\sigma} |\mathbf{x}-\mathbf{y}|$. The *distance* of two clusters σ and τ is $\operatorname{dist}(\sigma,\tau) := \inf_{\mathbf{x}\in Q_\sigma,\mathbf{y}\in Q_\tau} |\mathbf{x}-\mathbf{y}|$. The subspace V_σ of V_h is defined as $V_\sigma := \operatorname{span}\{\phi_i \mid i \in \sigma\}$.

For given $\eta > 0$, a pair of clusters (σ, τ) is called *admissible*, if

$$\eta \operatorname{dist}(\sigma,\tau) \geq \max\{\operatorname{diam}(\sigma), \operatorname{diam}(\tau)\}.$$

The admissible cluster pairs can be determined recursively. Cluster pairs that are not admissible and have no admissible sons are part of the near field and are assembled into a sparse matrix. The admissible cluster pairs for a one dimensional problem are shown in Fig. 4.

For admissible pairs of clusters σ and τ and any degrees of freedom $i \in \sigma$ and $j \in \tau$, the corresponding entry of the stiffness matrix is

$$(A^s)_{ij} = a\left(\phi_i, \phi_j\right) = -C(d,s) \int_\Omega \int_\Omega k(\mathbf{x},\mathbf{y})\, \phi_i(\mathbf{x})\, \phi_j(\mathbf{y})$$

with kernel $k(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|^{-(d+2s)}$. The kernel can be approximated on $Q_\sigma \times Q_\tau$ using Chebyshev interpolation of order m in every spatial dimension by

$$k_m(\mathbf{x},\mathbf{y}) = \sum_{\alpha,\beta=1}^{m^d} k\left(\boldsymbol{\xi}_\alpha^\sigma, \boldsymbol{\xi}_\beta^\tau\right) L_\alpha^\sigma(\mathbf{x}) L_\beta^\tau(\mathbf{y}).$$

Here, $\boldsymbol{\xi}_\alpha^\sigma$ are the tensor Chebyshev nodes on Q_σ, and L_α^σ are the associated Lagrange polynomials on the cluster box Q_σ with $L_\alpha^\sigma\left(\boldsymbol{\xi}_\beta^\sigma\right) = \delta_{\alpha\beta}$. This leads to the following approximation:

$$(\boldsymbol{A}^s)_{ij} \approx -C(d,s) \sum_{\alpha,\beta=1}^{m^2} k\left(\boldsymbol{\xi}_\alpha^\sigma, \boldsymbol{\xi}_\beta^\tau\right) \int_{\operatorname{supp}\phi_i} \phi_i(\mathbf{x}) L_\alpha^\sigma(\mathbf{x})\ d\mathbf{x} \int_{\operatorname{supp}\phi_j} \phi_j(\mathbf{y}) L_\beta^\tau(\mathbf{y})\ d\mathbf{y}$$

In fact, the expressions $\int_{\operatorname{supp}\phi_i} \phi_i(\mathbf{x}) L_\alpha^\sigma(\mathbf{x})\ d\mathbf{x}$ can be computed recursively starting from the finest level of the cluster tree, since for $\tau \in \Sigma(\sigma)$ and $\mathbf{x} \in Q_\tau$

$$L_\alpha^\sigma(\mathbf{x}) = \sum_\beta L_\alpha^\sigma\left(\boldsymbol{\xi}_\beta^\tau\right) L_\beta^\tau(\mathbf{x}).$$

This means that for all leaves $\sigma = \{i\}$, and all $1 \le \alpha \le m^d$, the *basis far-field coefficients*

$$\int_{\operatorname{supp}\phi_i} \phi_i(\mathbf{x}) L_\alpha^\sigma(\mathbf{x})\ d\mathbf{x}$$

need to be evaluated (e.g. by $m+1$-th order Gaussian quadrature). Moreover, the *shift coefficients*

$$L_\alpha^\sigma\left(\boldsymbol{\xi}_\beta^\tau\right)$$

for $\tau \in \Sigma(\sigma)$ must be evaluated, as well as the kernel approximations

$$k\left(\boldsymbol{\xi}_\alpha^\sigma, \boldsymbol{\xi}_\beta^\tau\right)$$

for every admissible pair of clusters (σ,τ). We refer the reader to [22] for further details.

The consistency error of this approximation is given by the following theorem:

Theorem 5 ([22], Theorems 7.3.12 and 7.3.18) *There exists* $\gamma \in (0,1)$ *such that*

$$|k(\mathbf{x},\mathbf{y}) - k_m(\mathbf{x},\mathbf{y})| \le \frac{C\gamma^m}{\operatorname{dist}(\sigma,\tau)^{d+2s}}.$$

The consistency error between the bilinear form $a(\cdot,\cdot)$ and the bilinear form $a_C(\cdot,\cdot)$ of the panel clustering method is

$$|a(u,v) - a_C(u,v)| \leq C\gamma^m (1+2\eta)^{d+2s} C_{d,s}(h) \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)},$$

where

$$C_{d,s}(h) = \begin{cases} h^{-2} & \text{if } d = 1 \text{ and } s < 1/2, \\ h^{-2}(1+|\log h|) & \text{if } d = 1 \text{ and } s = 1/2, \\ h^{-d-2s} & \text{otherwise.} \end{cases}$$

Again, by invoking Strang's Lemma, $\mathcal{O}\left(h^{\ell-s}\right)$ convergence is retained if the interpolation order m satisfies

$$m \geq \frac{(\ell - s + 2)|\log h|}{|\log \gamma|} \qquad \text{if } d = 1 \text{ and } s < 1/2$$

$$m \geq \frac{(\ell - s + 2)|\log h| + \log(1+|\log h|)}{|\log \gamma|} \qquad \text{if } d = 1 \text{ and } s = 1/2$$

$$m \geq \frac{(\ell - s + d + 2s)|\log h|}{|\log \gamma|} \qquad \text{otherwise.}$$

By following the arguments in [22], it can be shown that the number of near field entries, i.e. the entries that need to be assembled using the quadrature rules described in Sect. 4, scales linearly in n. The same conclusion holds for the number of far field cluster pairs. Since the four dimensional integral contributions $a^{K\times\tilde{K}}$ are evaluated using Gaussian quadrature rules with at most $k \sim \log n$ quadrature nodes per dimension, the assembly of the near field contributions scales with $n \log^{2d} n$. The far field kernel approximations and the shift coefficients have size $m^{2d} \sim \log^{2d} n$, and are also calculated in $\log^{2d} n$ complexity. This means that all the kernel approximations and shift coefficients are obtained in $n \log^{2d} n$ time. Finally, the nm^d basis far-field coefficients require the evaluation of integrals using $m+1$-th order Gaussian quadrature, leading to a complexity of $n \log^{2d} n$ as well. The overall complexity of the panel clustering method is therefore $\mathcal{O}\left(n \log^{2d} n\right)$, and the sparse approximation requires $\mathcal{O}\left(n \log^{2d} n\right)$ memory. In practice, this means that the assembly of the near-field matrix dominates the other steps but involves only local computations.

The computation of the matrix-vector product involving upward and downward recursion in the cluster tree and multiplication by the kernel approximations can also be shown to scale with $\mathcal{O}\left(n \log^{2d} n\right)$.

As an aside, we note that one could also opt to use a conventional dense approximation of the discretised fractional Laplacian such as the "hybrid" scheme described in [16] which reduces the far field computation to the computation of a "Nyström-type" approximation. While the complexity of this approach still scales as $\mathcal{O}\left(n^2\right)$, the constant is significantly smaller than if the dense matrix were to be used.

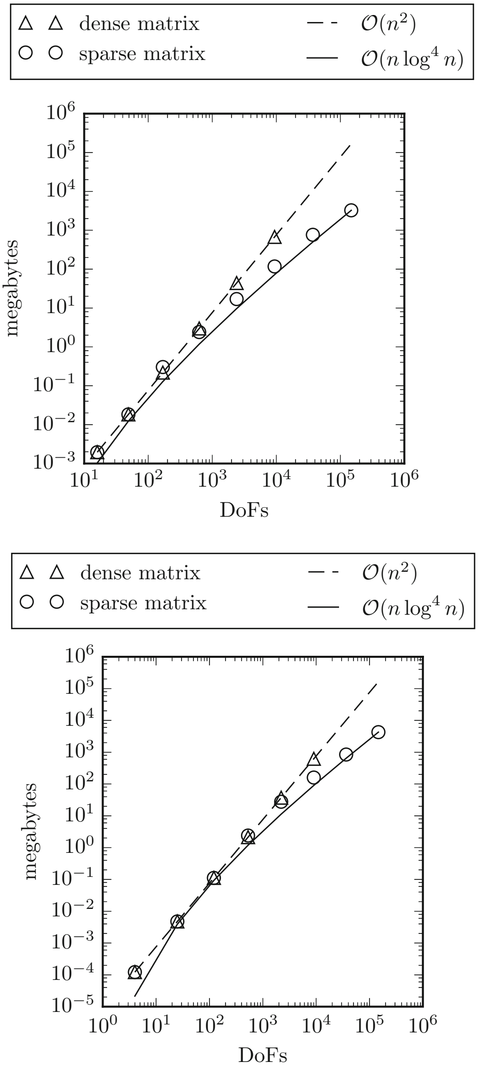
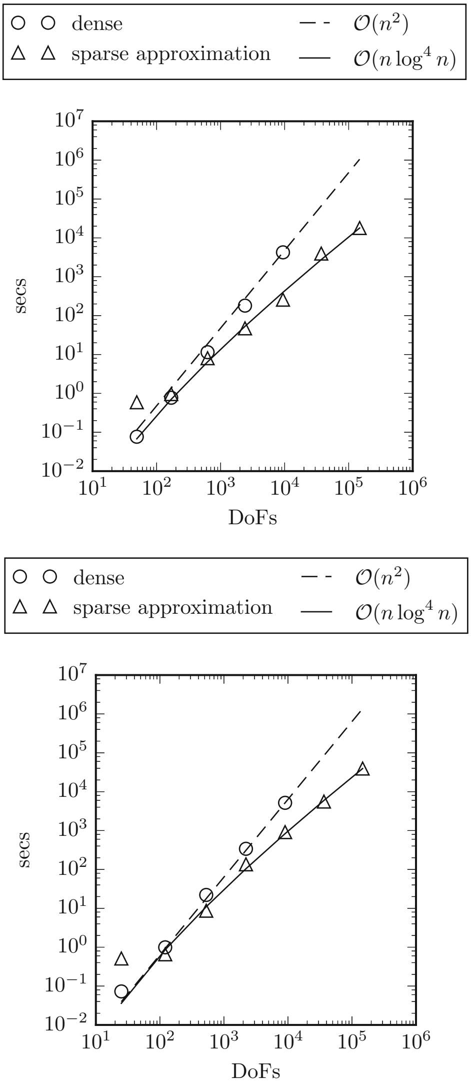


Fig. 5 Memory usage of the dense matrix and its sparse approximation. $s = 0.25$ (*top*), $s = 0.75$ (*bottom*). While the dense matrix uses n^2 floating-point numbers, the sparse approximation can be seen to require only $\mathcal{O}\left(n \log^4 n\right)$ memory. At roughly 2000 unknowns, the memory footprint of the sparse approximation separates from the $\mathcal{O}\left(n^2\right)$ curve

We illustrate the above results by assembling both the full matrix as well as its sparse approximation on the unit disk for fractional orders $s = 0.25$ and $s = 0.75$. The memory usage of the matrices are compared in Fig. 5. For low number of degrees of freedom, none of the cluster pairs are admissible, so the full matrix and its approximation have the same size. Starting with roughly 2000 degrees of freedom, the memory footprint of the sparse approximate starts to follow the $n \log^4 n$ curve and therefore outperforms the full assembly. The same behaviour can be observed for the assembly times, as seen in Fig. 6.

Fig. 6 Assembly time of the dense matrix and its sparse approximation. $s = 0.25$ (*top*), $s = 0.75$ (*bottom*). The time to assemble the full matrix grows quadratically in the number of unknowns, whereas the sparse approximation starts to follow the $n \log^4 n$ curve at about 2000 degrees of freedom



7 Applications

7.1 Fractional Poisson Equation

We consider the fractional Poisson problem

$$\begin{aligned} (-\Delta)^s u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{in } \Omega^c \end{aligned}$$

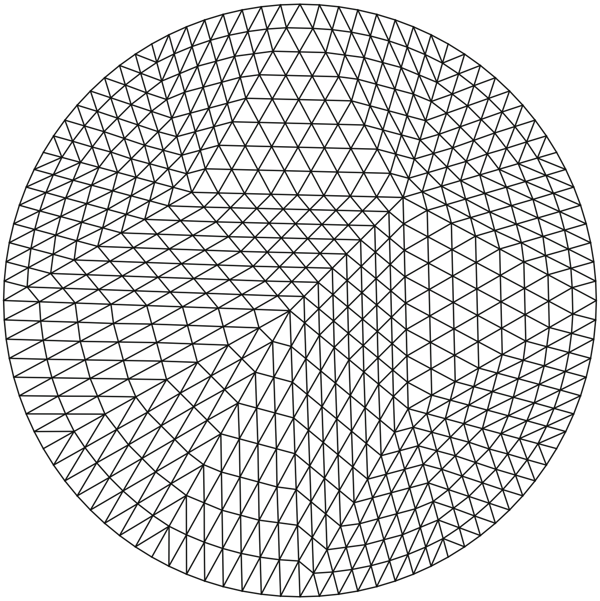


Fig. 7 A quasi-uniform triangulation of the disc domain, obtained through uniform refinement followed by projection of the resulting boundary nodes back onto the unit circle

on the unit disk $\Omega = \left\{\mathbf{x} \in \mathbb{R}^2 \mid |\mathbf{x}| \leq 1\right\}$. The discretised fractional Poisson problem then reads

$$\boldsymbol{A}^s\mathbf{u} = \mathbf{b}, \tag{13}$$

where $u_h = \sum_{i=1}^{n} u_i\phi_i \in V_h$ is the approximation to the solution u, and $b_i = \langle f, \phi_i\rangle$.

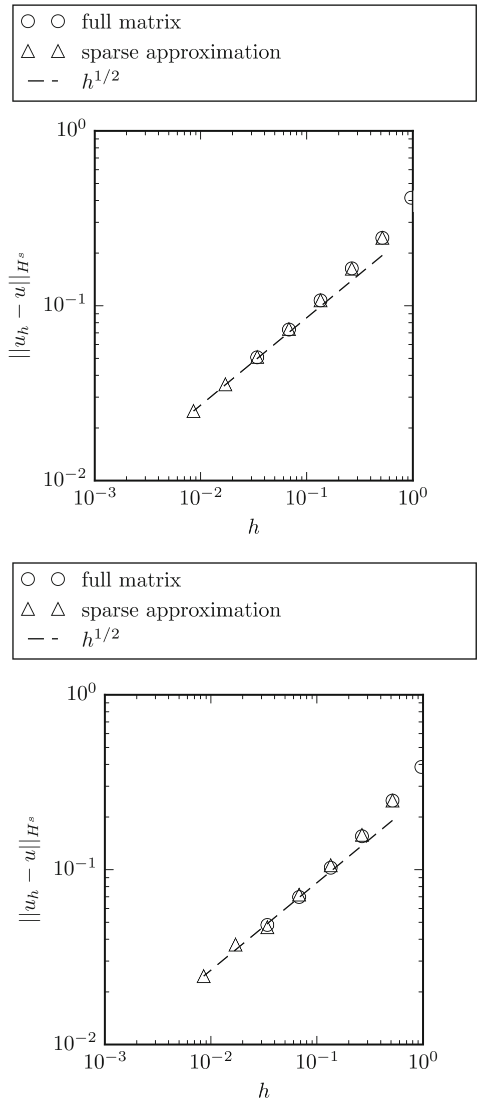
Triangulations of the disc are obtained through uniform refinement of a uniform initial mesh. After each refinement, the boundary nodes are projected onto the unit circle, resulting in triangulations of the type shown in Fig. 7.

We first consider the test case introduced in Sect. 3 where $f = 1$ with analytic solution [14] given by

$$u^s\left(\mathbf{x}\right) := \frac{2^{-2s}}{\Gamma\left(1+s\right)^2}\left(1-|\mathbf{x}|^2\right)^s.$$

Both the full matrix and its sparse approximation are assembled for $s \in \{0.25, 0.75\}$, and Eq. (13) is solved using LAPACK's `dgesv` routine and a multigrid solver in the dense case, and multigrid and conjugate gradient methods in the sparse case. Two steps of pre- and postsmoothing by Jacobi iteration are used on every level of the multigrid solver. Recall that solutions for $s = 0.25$ and $s = 0.75$ were shown in Fig. 1. In Figs. 8 and 9, the discretisation error is plotted in $\widetilde{H}^s\left(\Omega\right)$ and

Fig. 8 Error $\|u^s - u_h\|_{\widetilde{H}^s(\Omega)}$ for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*) in the case of solutions with singular behaviour close to the boundary. Both the full matrix and its sparse approximation are shown to achieve the predicted rate of $h^{1/2}$

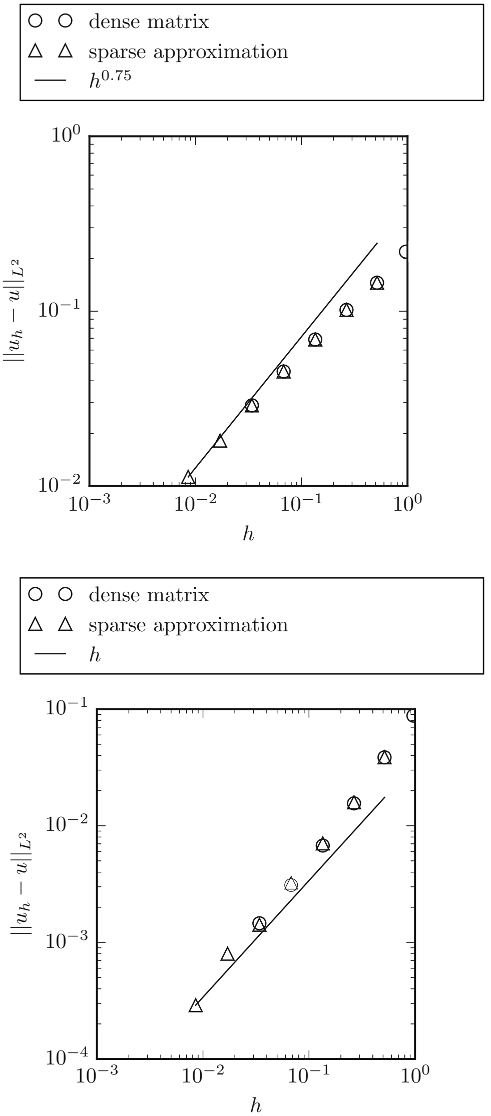


in L^2-norm. It can be seen that the rates predicted by Theorems 1 and 2 of $h^{1/2}$ and $h^{1/2+\min(1/2,s)}$ are indeed obtained, and that the error curves for the full matrix and its sparse approximation are essentially indistinguishable.

For a second example, the right-hand side f is chosen such that $u = 1 - |\mathbf{x}|^2 \in H^2(\Omega)$. The action of f on $v \in V_h$ is approximated by

$$(f, v) = a(I_{\underline{h}}u, v),$$

Fig. 9 Error $\|u^s - u_h\|_{L^2}$ for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*) in the case of solutions with singular behaviour close to the boundary. Both the full matrix and its sparse approximation are shown to achieve the predicted rate of $h^{1/2+\min\{s,1/2\}}$



where $I_{\underline{h}}$ is the interpolation operator onto a highly refined mesh with $\underline{h} < h$. The resulting consistency error in this case is

$$\sup_{v} \frac{\left|a(u,v) - a(I_{\underline{h}}u, v)\right|}{\|v\|_{\widetilde{H}^s(\Omega)}} \le C \left\|u - I_{\underline{h}}u\right\|_{\widetilde{H}^s(\Omega)} \le C\underline{h}^{2-s} \left|u\right|_{H^2} .$$

Therefore, if $\underline{h}$ is sufficiently smaller than h, the consistency error will be negligible compared to the discretisation error.

Fig. 10 Errors $\|u-u_h\|_{\widetilde{H}^s(\Omega)}$ and $\|u-u_h\|_{L^2}$ for $s=0.25$ (*top*) and $s=0.75$ (*bottom*) in the case of a smooth solution $u(\mathbf{x})=1-|\mathbf{x}|^2\in H^2(\Omega)$. Optimal orders are achieved both in $\widetilde{H}^s(\Omega)$- and L^2-norm

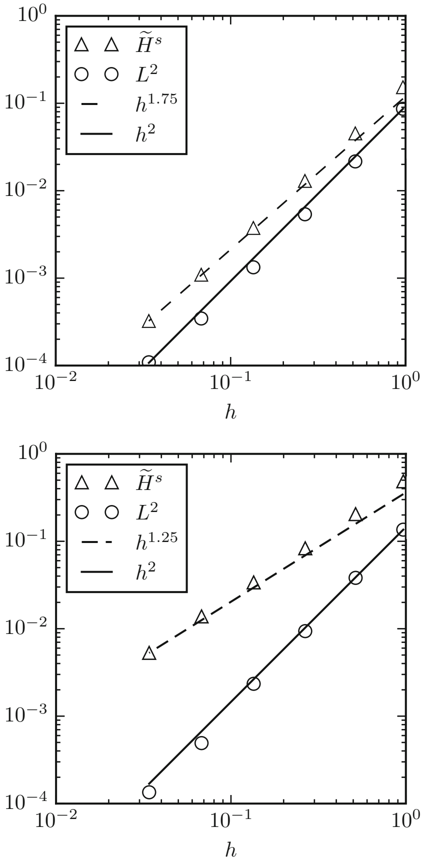


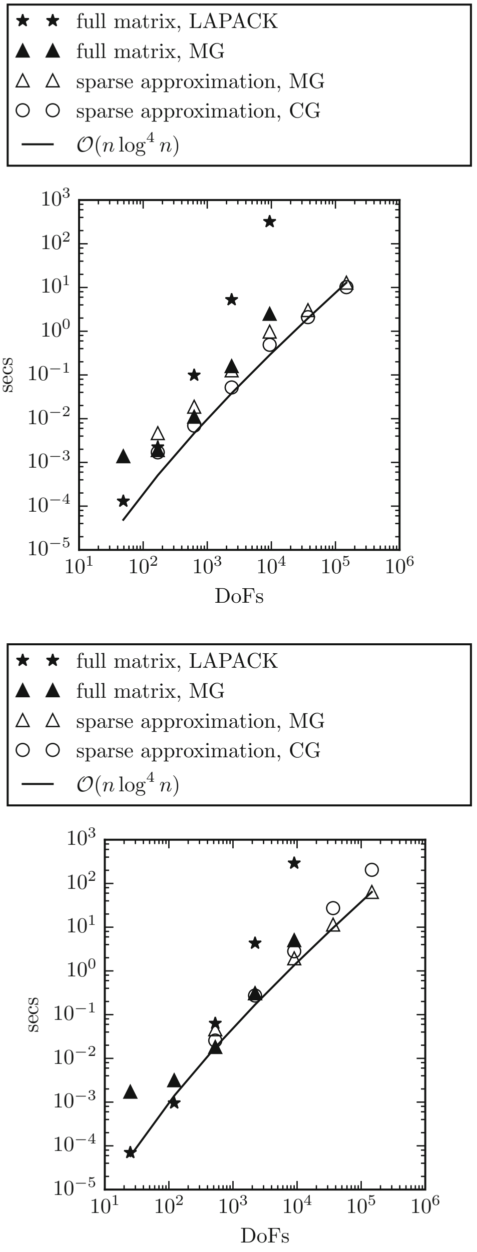
Table 1 Asymptotic complexities of different solvers for the discretised fractional Poisson problem $A^s\mathbf{u}=\mathbf{b}$

Method	Dense matrix	Sparse approximation
Dense solver	n^3	–
Conjugate gradient	$n^{2+s/d}$	$n^{1+s/d}(\log n)^{2d}$
Multigrid	n^2	$n(\log n)^{2d}$

The dependency of the error on the mesh size h can be seen in Fig. 10. The discretisation error decays as h^{2-s} in $\widetilde{H}^s(\Omega)$-norm, and as h^2 in L^2-norm, which are the optimal orders that we would expect based on estimate (2).

Summarising the results of Sects. 5 and 6, we expect different solvers for the fractional Laplacian to have complexities as given in Table 1. The timings for the different combinations of dense or sparse matrix with a solver are shown in Fig. 11. It can be observed that the sparse approximation asymptotically outperforms the dense solvers. Moreover, for the larger value of s, the multigrid solver starts to outperform the conjugate gradient method for increasingly smaller numbers of unknowns as one would expect based on earlier arguments.

Fig. 11 Solution time for the fractional Laplacian using different solvers and the full matrix and its sparse approximation for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*). The solvers using the full matrix are outperformed by the ones based on the sparse approximation. For larger fractional order s, the break-even between conjugate gradient and multigrid iteration occurs at a lower number of unknowns



7.2 Fractional Heat Equation

The fractional heat equation is given by

$$\begin{aligned} u_t + (-\Delta)^s u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{in } \Omega^c. \end{aligned}$$

We propose to approximate the problem using an implicit method in time. The simplest such scheme is the backward Euler method

$$(\boldsymbol{M} + \Delta t\, \boldsymbol{A}^s)\, \mathbf{u}^{k+1} = \boldsymbol{M}\mathbf{u}^k + \Delta t \mathbf{f}^{k+1},$$

where $u(\cdot, k\Delta t) \approx \sum_i u_i^k \phi_i$ and $f_i^k = (f(\cdot, k\Delta t), \phi_i)$.

More generally, let us assume that a scheme of order α is used in time. In order to obtain optimal convergence in L^2-norm, in view of Theorem 2, we shall choose $\Delta t^\alpha \sim h^{1/2+\min(1/2,s)}$, i.e.

$$\Delta t_{L^2} \sim h^{\min(2,1+2s)/(2\alpha)}.$$

On the other hand, if optimal $\widetilde{H}^s(\Omega)$-convergence is desired, we need $\Delta t_{\widetilde{H}^s(\Omega)} \sim h^{1/(2\alpha)}$, see Theorem 1. Consequently, if an order α scheme is used for time stepping, with optimal time step Δt_{L^2} or $\Delta t_{\widetilde{H}^s(\Omega)}$, we find by Lemma 1 that the condition numbers of the iteration matrix satisfy

$$\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\, \boldsymbol{A}^s\right) \leq C\left(1 + h^{\min(2,1+2s)/(2\alpha)-2s}\right),$$

$$\kappa\left(\boldsymbol{M} + \Delta t_{\widetilde{H}^s(\Omega)}\, \boldsymbol{A}^s\right) \leq C\left(1 + h^{1/(2\alpha)-2s}\right).$$

In particular, in the L^2 case, this shows that the condition number will not grow at all as the mesh size decreases if $s \in (0, 1/(4\alpha - 2)]$. For fractional orders s that are slightly larger than $1/(4\alpha - 2)$, the condition number only grows very slowly as the mesh size is decreased. The larger the fractional order, the faster the linear system becomes ill-conditioned. In the $\widetilde{H}^s(\Omega)$ case, the condition number of the linear system grows as the mesh size is decreased for $s > 1/(4\alpha)$.

We illustrate the consequences of the above result in the case of a second order accurate time stepping scheme ($\alpha = 2$), and for $s = 0.25$ and $s = 0.75$. In the case of $s = 0.25$, $\Delta t_{L^2} \sim h^{3/8}$ and $\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\, \boldsymbol{A}^s\right) \sim 1 + h^{-1/8}$. This suggests that the conjugate gradient method will deliver good results for a wide range of mesh sizes h, as the number of iterations will only grow as $\sqrt{\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\, \boldsymbol{A}^s\right)} \sim h^{-1/16}$. The convergence of the multigrid method does not depend on the condition number and is essentially independent of h. This is indeed what is observed in the top part of Fig. 12. In Fig. 13, the number of iterations is shown. It can be observed that for $s = 0.25$ both the multigrid and the conjugate gradient solver require an essentially constant number of iterations for varying values of Δt.

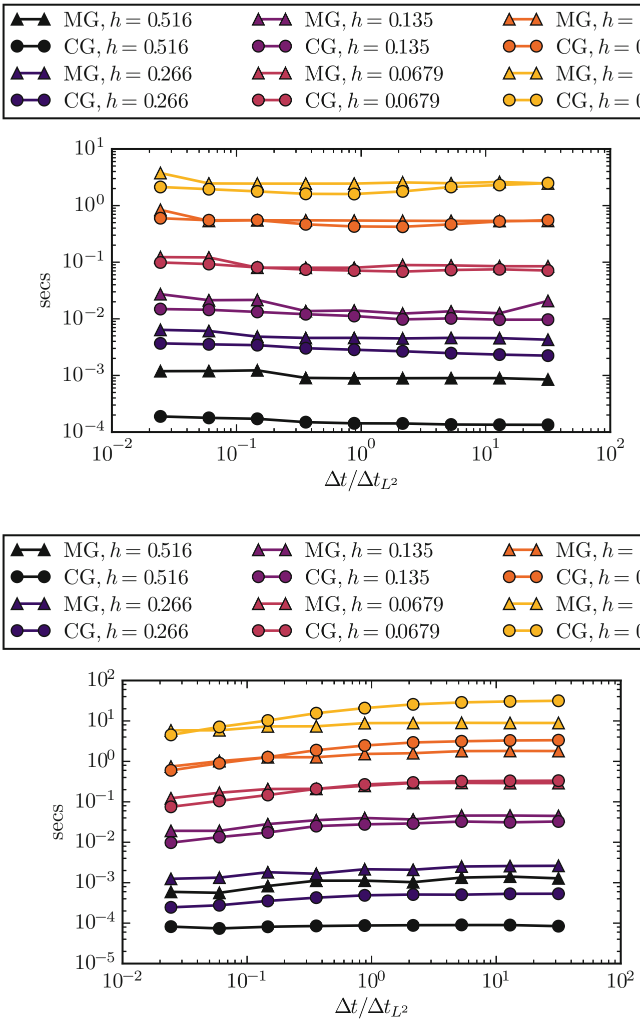


Fig. 12 Timings in seconds for CG and MG depending on Δt for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*). It can be observed that, for $s = 0.25$, the conjugate gradient method is essentially on par with the multigrid solver. For $s = 0.75$, the multigrid solver asymptotically outperforms the conjugate gradient method, since the condition number $\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\boldsymbol{A}^s\right)$ grows as h^{-1}

On the other hand, for $s = 0.75$, $\Delta t_{L^2} \sim h^{1/2}$ and $\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\,\boldsymbol{A}^s\right) \sim 1 + h^{-1}$. Therefore, the condition number increases a lot faster as h goes to zero, and we expect that multigrid asymptotically outperforms the CG solver. This is indeed what is observed in Figs. 12 and 13.

The complexities of the different solvers for different choices of time step size are summarised in Table 2.

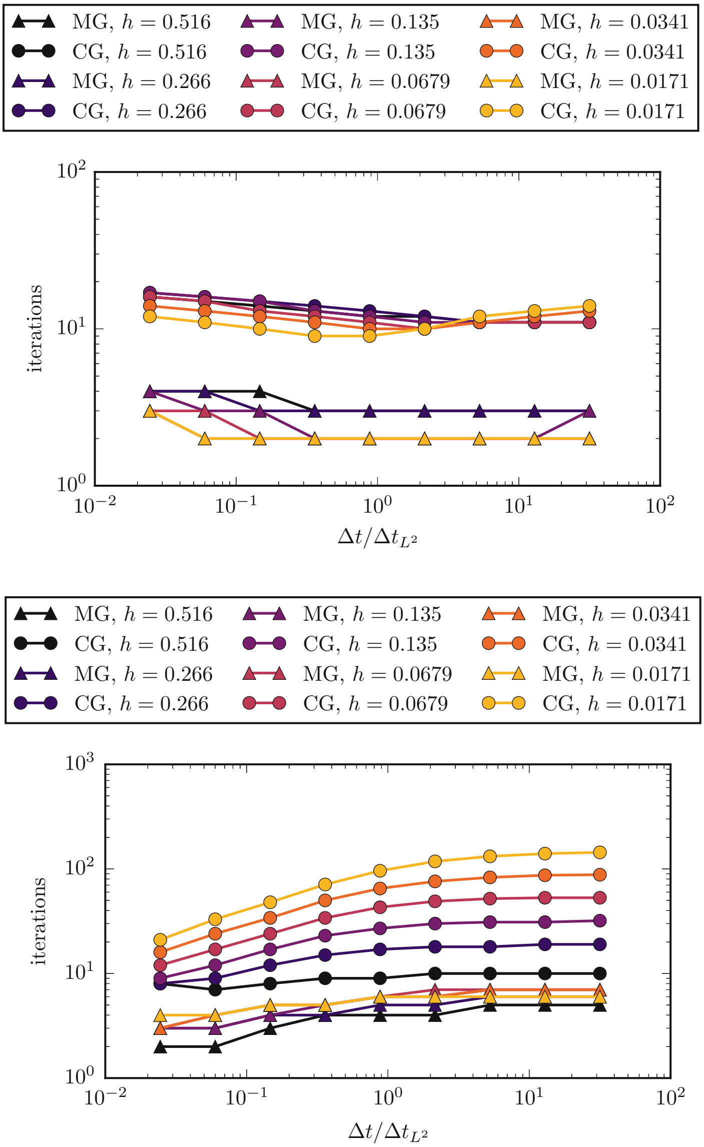


Fig. 13 Number of iterations for CG and MG depending on Δt for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*). For $s = 0.25$, the number of iterations is essentially independent of Δt. For $s = 0.75$, the number of iterations of the multigrid solver is independent of Δt, but the iterations count for conjugate gradient grows with $h^{-1/2}$

Table 2 Complexity of different solvers for $(\boldsymbol{M} + \Delta t \boldsymbol{A}^s)\,\mathbf{u} = \mathbf{b}$ for $\Delta t = \Delta t_{L^2}$ and $\Delta t = \Delta t_{\widetilde{H}^s(\Omega)}$ for an α-order time stepping scheme

Method	$\Delta t = \Delta t_{L^2}$	$\Delta t = \Delta t_{\widetilde{H}^s(\Omega)}$
Conjugate gradient	$n^{1+2s/d-\min(2,1+2s)/(2\alpha d)}\,(\log n)^{2d}$	$n^{1+2s/d-1/(2\alpha d)}\,(\log n)^{2d}$
Multigrid	$n\,(\log n)^{2d}$	$n\,(\log n)^{2d}$

Table 3 IMEX scheme by Koto

0	0			
1	0	1		
1/2	0	−1/2	1	
1	0	−1	1	1
	0	−1	1	1

0				
1	1			
1/2	0	0		
1	0	0	1	
	0	0	1	0

Implicit scheme on the left, explicit on the right

7.3 *Fractional Reaction-Diffusion Systems*

In [15], a space-fractional Brusselator model was analysed and compared to the classical integer-order case. The coupled system of equations is given by

$$\begin{aligned}\frac{\partial X}{\partial t} &= -D_X\,(-\Delta)^{\alpha}\,X + A - (B+1)X + X^2Y,\\ \frac{\partial Y}{\partial t} &= -D_Y\,(-\Delta)^{\beta}\,Y + BX - X^2Y.\end{aligned}$$

Here, D_X and D_Y are diffusion coefficients, A and B are reaction parameters, and α and β determine the type of diffusion. By rewriting the solutions as deviations from the stationary solution $X = A$, $Y = B/A$ and rescaling, one obtains

$$\frac{\partial u}{\partial t} = -\,(-\Delta)^{\alpha}\,u + (B-1)u + Q^2 v + \frac{B}{Q}u^2 + 2Quv + u^2 v, \tag{14}$$

$$\eta^2\frac{\partial v}{\partial t} = -\,(-\Delta)^{\beta}\,v - Bu - Q^2 v - \frac{B}{Q}u^2 - 2Quv - u^2 v, \tag{15}$$

with $\eta = \sqrt{D_Y/D_X^{\beta/\alpha}}$ and $Q = A\eta$.

In [15] the equations were augmented with periodic boundary conditions and approximated using a pseudospectral method for various different parameter combinations. Here, thanks to the foregoing developments, we have the flexibility to handle more general domains and, in particular, we consider the case where Ω corresponds to a Petri-dish, i.e. $\Omega = \left\{\mathbf{x} \in \mathbb{R}^2 \mid |\mathbf{x}| \leq 1\right\}$ is the unit disk. We solve the above set of equations using a second order accurate IMEX scheme proposed by Koto [18], whose Butcher tableaux are given by Table 3. The diffusive parts are treated implicitly and therefore require the solution of several systems all of which are of the type $\boldsymbol{M} + c\Delta t\boldsymbol{A}^s$ with appropriate values of c.

In order to verify the correct convergence behaviour, we add forcing functions f and g to the system, chosen such that the analytic solution is given by

$$u = \eta \sin(t) u^s(\mathbf{x}),$$
$$v = \eta^{-1} \cos(2t) u^s(\mathbf{x}),$$

for suitable initial conditions, where u^s is the solution of the fractional Poisson problem with constant right-hand side. We take $\alpha = \beta = 0.75$, and choose $\Delta t \sim h^{1/2}$, since we already saw that the rate of the spatial approximation in L^2-norm is of order h. We measure the error as

$$e^u_{L^2} = \max_{0 \le t_i \le 10} \left\| u(t_i, \cdot) - u^i_h \right\|_{L^2}, \qquad e^v_{L^2} = \max_{0 \le t_i \le 10} \left\| v(t_i, \cdot) - v^i_h \right\|_{L^2},$$
$$e^u_{\widetilde{H}^s(\Omega)} = \max_{0 \le t_i \le 10} \left\| u(t_i, \cdot) - u^i_h \right\|_{\widetilde{H}^s(\Omega)}, \quad e^v_{\widetilde{H}^s(\Omega)} = \max_{0 \le t_i \le 10} \left\| v(t_i, \cdot) - v^i_h \right\|_{\widetilde{H}^s(\Omega)}.$$

From the error plots in Fig. 14, it can be observed that $e_{L^2} \sim h$ and $e_V \sim h^{1/2}$, as expected.

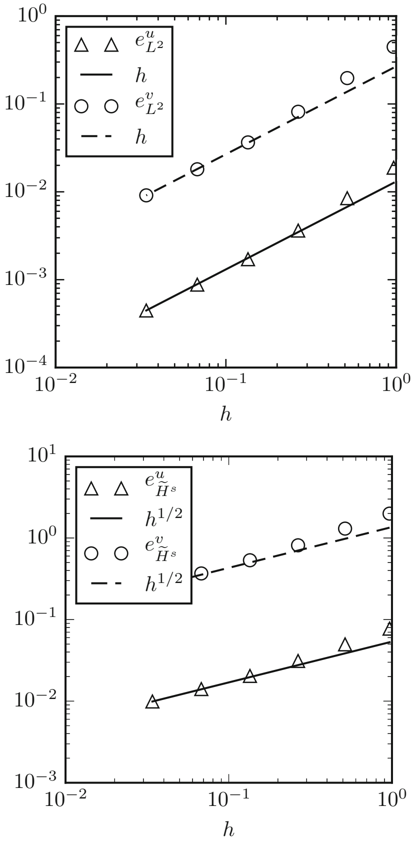


Fig. 14 Error in L^2-norm (*top*) and $\widetilde{H}^s(\Omega)$-norm (*bottom*) in the Brusselator model. Optimal orders of convergence are achieved (compare Theorems 1 and 2)

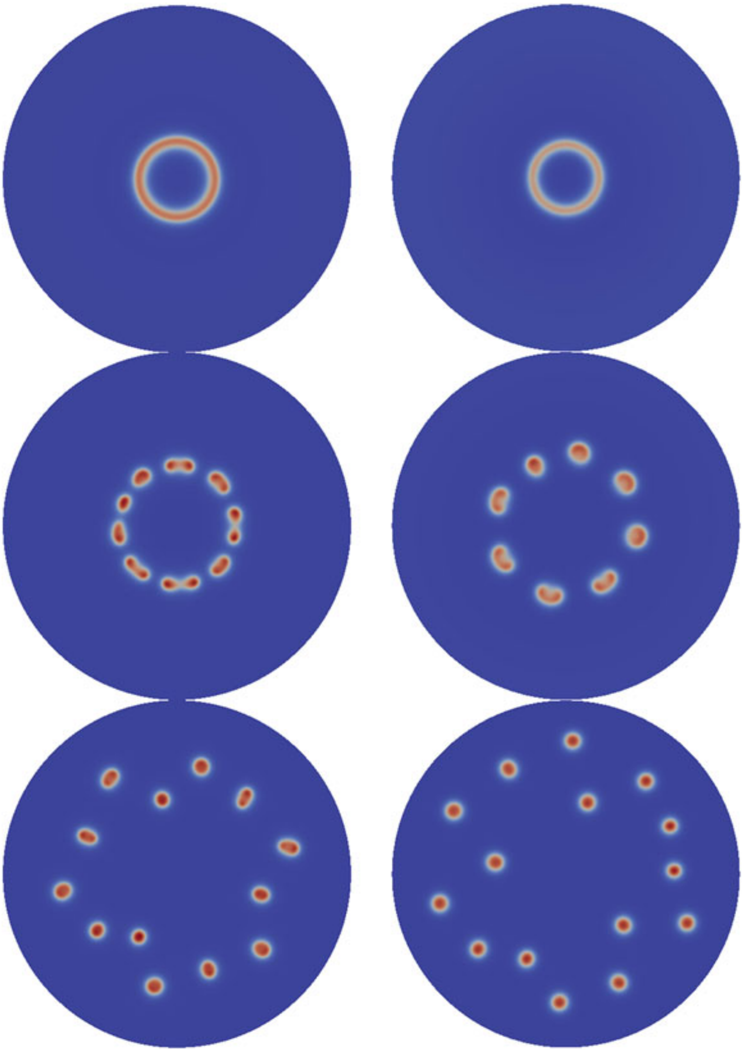


Fig. 15 Localised spot solutions of the Brusselator system with $\alpha = \beta = 0.625$ (*left*) and $\alpha = \beta = 0.75$ (*right*). u is shown in both cases, and time progresses from top to bottom. The initial perturbation was identical in both cases. The initial perturbation in the centre of the domain forms a ring, whose radius is bigger if the fractional orders of diffusion α, β are smaller. The ring breaks up into several spots, which start to replicate and spread out over the whole domain. $n \approx 50{,}000$ unknowns were used in the finite element approximation

Having verified the accuracy of the method, we turn to the solution of the system Eqs. (14) and (15) augmented with exterior Neumann conditions as described in Sect. 2. Golovin et al. [15] observed that for $\eta = 0.2$, $B = 1.22$ and $Q = 0.1$, a single localised perturbation would first form a ring and then break up into spots. The radius of the ring and the number of resulting spots increases as the fractional orders are decreased. In Fig. 15, simulation results for $\alpha = \beta = 0.625$

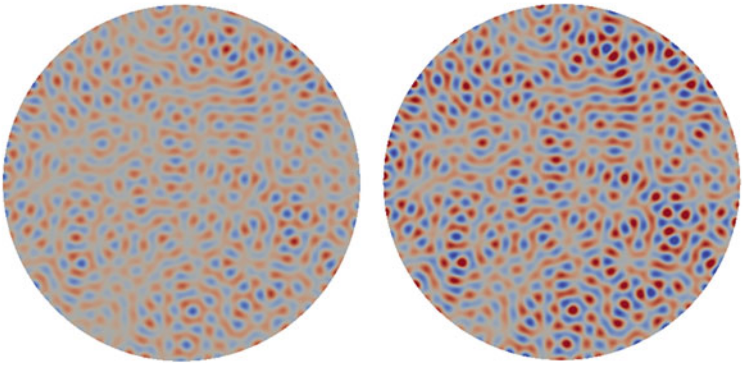


Fig. 16 Stripe solutions of the Brusselator system with $\alpha = \beta = 0.75$. u is shown on the left, and v on the right. The random initial condition leads to the formation of stripes throughout the domain. $n \approx 50{,}000$ unknowns were used in the finite element approximation

and $\alpha = \beta = 0.75$ are shown. We observe that in both cases, an initially circular perturbation develops into a ring. Lower diffusion coefficients do lead to a larger ring, which breaks up later and into more spots. In the last row, we can see that the resulting spots start to replicate and spread out over the whole domain.

Another choice of parameters leads to stripes in the solution. For $\alpha = \beta = 0.75$, $\eta = 0.2$, $B = 6.26$ and $Q = 2.5$, and a random initial condition, stripes without directionality form in the whole domain. This is in alignment with the theoretical considerations of Golovin et al. [15] (Fig. 16).

8 Conclusion

We have presented a reasonably complete and coherent approach for the efficient approximation of problems involving the fractional Laplacian, based on techniques from the boundary element literature. In particular, we discussed the efficient assembly and solution of the associated matrix, and demonstrated the feasibility of a sparse approximation using the panel clustering method. The potential of the approach was demonstrated in several numerical examples, and were used to reproduce some of the findings for a fractional Brusselator model. While we focused on the case of $d = 2$ dimensions, the generalisation to higher dimensions does not pose any fundamental difficulties. Moreover, the approach taken to obtain a sparse approximation to the dense system matrix for the fractional Laplacian does not rely strongly on the form of the interaction kernel $k(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|^{-(d+2s)}$, and generalisations to different kernels such as the one used in peridynamics [25] are therefore possible. In the present work we have confined ourselves to the discussion of quasi-uniform meshes. However, solutions of problems involving the fractional Laplacian exhibit line singularities in the neighbourhood of the boundary. The efficient resolution of such problems would require locally refined meshes which form the topic of forthcoming work [3].

Acknowledgements This work was supported by the MURI/ARO on "Fractional PDEs for Conservation Laws and Beyond: Theory, Numerics and Applications" (W911NF-15-1-0562).

Appendix 1: Derivation of Expressions for Singular Contributions

The contributions $a^{K\times\tilde{K}}$ and $a^{K\times e}$ as given in Eqs. (3) and (4) for touching elements K and $\tilde{K}$ contain removable singularities. In order to make these contributions amenable to numerical quadrature, the singularities need to be lifted. We outline the derivation for $d = 2$ dimensions.

The expression for $a^{K\times\tilde{K}}$ can be transformed into integrals over the reference element $\hat{K}$:

$$\begin{aligned}
&a^{K\times\tilde{K}}(\phi_i,\phi_j)\\
&=\frac{C(2,s)}{2}\int_K d\mathbf{x}\int_{\tilde{K}} d\mathbf{y}\frac{(\phi_i(\mathbf{x})-\phi_i(\mathbf{y}))\left(\phi_j(\mathbf{x})-\phi_j(\mathbf{y})\right)}{|\mathbf{x}-\mathbf{y}|^{2+2s}}\\
&=\frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\int_{\hat{K}} d\hat{\mathbf{x}}\int_{\hat{K}} d\hat{\mathbf{y}}\frac{(\phi_i(\mathbf{x}(\hat{\mathbf{x}}))-\phi_i(\mathbf{y}(\hat{\mathbf{y}})))\left(\phi_j(\mathbf{x}(\hat{\mathbf{x}}))-\phi_j(\mathbf{y}(\hat{\mathbf{y}}))\right)}{|\mathbf{x}(\hat{\mathbf{x}})-\mathbf{y}(\hat{\mathbf{y}})|^{2+2s}}.
\end{aligned}$$

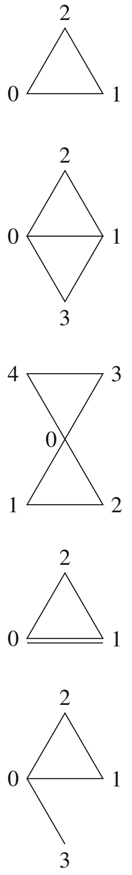
Similarly, by introducing the reference edge $\hat{e}$, we obtain

$$\begin{aligned}
a^{K\times e}(\phi_i,\phi_j) &= \frac{C(2,s)}{2s}\int_K d\mathbf{x}\int_e d\mathbf{y}\frac{\phi_i(\mathbf{x})\,\phi_j(\mathbf{x})\ \mathbf{n}_e\cdot(\mathbf{x}-\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{2+2s}}\\
&= \frac{C(2,s)}{2s}\frac{|K|}{\left|\hat{K}\right|}\frac{|e|}{|\hat{e}|}\int_{\hat{K}} d\hat{\mathbf{x}}\int_{\hat{e}} d\hat{\mathbf{y}}\frac{\phi_i(\mathbf{x}(\hat{\mathbf{x}}))\,\phi_j(\mathbf{x}(\hat{\mathbf{x}}))\ \mathbf{n}_e\cdot(\mathbf{x}(\hat{\mathbf{x}})-\mathbf{y}(\hat{\mathbf{y}}))}{|\mathbf{x}(\hat{\mathbf{x}})-\mathbf{y}(\hat{\mathbf{y}})|^{2+2s}}
\end{aligned}$$

for touching elements K and edges e. If K and $\tilde{K}$ or e have $c \geq 1$ common vertices, and if we designate by λ_k, $k = 0,\dots,6-c$ the barycentric coordinates of $K\cup\tilde{K}$ or $K\cup e$ respectively (cf. Fig. 17), we have

$$\lambda_{k(i)}(\hat{\mathbf{x}}) = \phi_i(\mathbf{x}(\hat{\mathbf{x}})),$$

Fig. 17 Numbering of local nodes for touching triangular elements K and $\tilde{K}$ or element K and edge e. (**a**) $K \cap \tilde{K} = K$. (**b**) $K \cap \tilde{K}$ =edge. (**c**) $K \cap \tilde{K}$ =vertex. (**d**) $K \cap e = e$. (**e**) $K \cap e$ =vertex



where $k(i)$ is the local index on $K \cup \tilde{K}$ or $K \cup e$ of the global degree of freedom i. Moreover, we have

$$\mathbf{x}(\hat{\mathbf{x}}) - \mathbf{y}(\hat{\mathbf{y}}) = \sum_{k=0}^{6-c} \lambda_k(\hat{\mathbf{x}})\,\mathbf{x}_k - \sum_{k=0}^{6-c} \lambda_k(\hat{\mathbf{y}})\,\mathbf{x}_k$$

$$= \sum_{k=0}^{6-c} \left[\lambda_k(\hat{\mathbf{x}}) - \lambda_k(\hat{\mathbf{y}})\right] \mathbf{x}_k.$$

Here, $\mathbf{x}_k$, $k = 0, \ldots, 6 - c$ are the vertices that span $K \cup \tilde{K}$ or $K \cup e$ respectively.

By setting

$$\psi_k(\hat{\mathbf{x}}, \hat{\mathbf{y}}) := \lambda_k(\hat{\mathbf{x}}) - \lambda_k(\hat{\mathbf{y}}),$$

we can therefore write

$$a^{K\times\tilde{K}}(\phi_i,\phi_j) = \frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\int_{\hat{K}} d\hat{\mathbf{x}}\int_{\hat{K}} d\hat{\mathbf{y}}\frac{\psi_{k(i)}\left(\hat{\mathbf{x}},\hat{\mathbf{y}}\right)\psi_{k(j)}\left(\hat{\mathbf{x}},\hat{\mathbf{y}}\right)}{\left|\sum_{k=0}^{6-c}\psi_k\left(\hat{\mathbf{x}},\hat{\mathbf{y}}\right)\mathbf{x}_k\right|^{2+2s}}.$$

By carefully splitting the integration domain $\hat{K}\times\hat{K}$ into L_c parts and applying a Duffy transformation to each part, the contributions can be rewritten into integrals over a unit hyper-cube, where the singularities are lifted.

$$\begin{aligned}a^{K\times\tilde{K}}(\phi_i,\phi_j) = {} & \frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\\ & \sum_{\ell=1}^{L_c}\int_{[0,1]^4} d\boldsymbol{\eta}\,\bar{J}^{(\ell,c)}\frac{\bar{\psi}_{k(i)}^{(\ell,c)}\left(\boldsymbol{\eta}\right)\bar{\psi}_{k(j)}^{(\ell,c)}\left(\boldsymbol{\eta}\right)}{\left|\sum_{k=0}^{2d-c}\bar{\psi}_k^{(\ell,c)}\left(\boldsymbol{\eta}\right)\mathbf{x}_k\right|^{2+2s}}.\end{aligned} \tag{16}$$

The details of this approach can be found in Chapter 5 of [22] for the interactions between K and $\tilde{K}$. We record the obtained expressions in this case.

- K and $\tilde{K}$ are identical, i.e. $c=3$

$$L_3 = 3, \qquad \bar{J}^{(1,3)} = \bar{J}^{(2,3)} = \bar{J}^{(3,3)} = \eta_0^{3-2s}\eta_1^{2-2s}\eta_2^{1-2s},$$

$$\bar{\psi}_k^{(1,3)} = \begin{cases}-\eta_3\\ \eta_3-1\\ 1\end{cases} \qquad \bar{\psi}_k^{(2,3)} = \begin{cases}-1\\ 1-\eta_3\\ \eta_3\end{cases} \qquad \bar{\psi}_k^{(3,3)} = \begin{cases}\eta_3\\ -1\\ 1-\eta_3\end{cases}$$

- K and $\tilde{K}$ share an edge, i.e. $c=2$

$$L_2 = 5, \qquad \bar{J}^{(1,2)} = \eta_0^{3-2s}\eta_1^{2-2s},$$
$$\bar{J}^{(2,2)} = \bar{J}^{(3,2)} = \bar{J}^{(4,2)} = \bar{J}^{(5,2)} = \eta_0^{3-2s}\eta_1^{2-2s}\eta_2$$

$$\bar{\psi}_k^{(1,2)} = \begin{cases}-\eta_2\\ 1-\eta_3\\ \eta_3\\ \eta_2-1\end{cases} \qquad \bar{\psi}_k^{(2,2)} = \begin{cases}-\eta_2\eta_3\\ \eta_2-1\\ 1\\ \eta_2\eta_3-\eta_2\end{cases} \qquad \bar{\psi}_k^{(3,2)} = \begin{cases}\eta_2\\ \eta_2\eta_3-1\\ 1-\eta_2\\ -\eta_2\eta_3\end{cases}$$

$$\bar{\psi}_k^{(4,2)} = \begin{cases}\eta_2\eta_3\\ 1-\eta_2\\ \eta_2-\eta_2\eta_3\\ -1\end{cases} \qquad \bar{\psi}_k^{(5,2)} = \begin{cases}\eta_2\eta_3\\ \eta_2-1\\ 1-\eta_2\eta_3\\ -\eta_2\end{cases}$$

- K and $\tilde{K}$ share a vertex, i.e. $c = 1$

$$L_1 = 2, \qquad \bar{J}^{(1,1)} = \bar{J}^{(2,1)} = \eta_0^{3-2s}\eta_2$$

$$\bar{\psi}_k^{(1,1)} = \begin{cases} \eta_2 - 1 \\ 1 - \eta_1 \\ \eta_1 \\ \eta_2\eta_3 - \eta_2 \\ -\eta_2\eta_3 \end{cases} \qquad \bar{\psi}_k^{(2,1)} = \begin{cases} 1 - \eta_2 \\ \eta_2 - \eta_2\eta_3 \\ \eta_2\eta_3 \\ \eta_1 - 1 \\ -\eta_1 \end{cases}$$

We notice that the contributions for identical elements only depend on η_3, so that in fact only one-dimensional integrals need to be computed. Similarly, the cases of common edges or common vertices only require two and three dimensional integration.

In a similar fashion, the integration domain of $a^{K\times e}$ can be split into several parts, so that the singularity can be lifted:

$$a^{K\times e}(\phi_i, \phi_j)$$
$$= \frac{C(2,s)}{2s} \frac{|K|}{|\hat{K}|} \frac{|e|}{|\hat{e}|} \int_{[0,1]^3} d\boldsymbol{\eta} \bar{J}^{(\ell,c)} \frac{\phi_{k(i)}^{(\ell,c)}(\boldsymbol{\eta})\, \phi_{k(j)}^{(\ell,c)}(\boldsymbol{\eta}) \;\; \sum_{k=0}^{5-c} \bar{\psi}_k^{(\ell,c)}(\boldsymbol{\eta})\, \mathbf{n}_e \cdot \mathbf{x}_k}{\left| \sum_{k=0}^{5-c} \bar{\psi}_k^{(\ell,c)}(\boldsymbol{\eta})\, \mathbf{x}_k \right|^{2+2s}}.$$

Here, $\phi_k^{(\ell,c)}$ are the expressions for the local shape functions under the Duffy transformations. The obtained expressions are

- e is an edge of K, i.e. $c = 2$

$$L_2 = 3, \qquad \bar{J}^{(1,2)} = \bar{J}^{(2,2)} = \bar{J}^{(3,2)} = \eta_0^{-2s}(1 - \eta_0),$$

$$\phi_k^{(1,2)} = \begin{cases} 1 - \eta_0 - \eta_2 + \eta_0\eta_2 \\ \eta_0 + \eta_2 - \eta_0\eta_1 - \eta_0\eta_1 \\ \eta_0\eta_1 \end{cases} \qquad \phi_k^{(2,2)} = \begin{cases} 1 - \eta_0 - \eta_2 + \eta_0\eta_2 \\ \eta_2 - \eta_0\eta_2 \\ \eta_0 \end{cases}$$

$$\phi_k^{(3,2)} = \begin{cases} 1 - \eta_2 + \eta_0\eta_2 - \eta_0\eta_1 \\ \eta_2 - \eta_0\eta_2 \\ \eta_0\eta_1 \end{cases}$$

$$\bar{\psi}_k^{(1,2)} = \begin{cases} -1 \\ 1 - \eta_1 \\ \eta_1 \end{cases} \qquad \bar{\psi}_k^{(2,2)} = \begin{cases} -\eta_1 \\ \eta_1 - 1 \\ 1 \end{cases} \qquad \bar{\psi}_k^{(3,2)} = \begin{cases} 1 - \eta_1 \\ -1 \\ \eta_1 \end{cases}$$

We notice that for $s \geq 1/2$, the integrand still contains a singularity. In this case, the finite element space V_h does not include the degrees of freedom on the boundary. For the interaction of the single degree of freedom that is not on the boundary ($k = 2$), we obtain

$$\bar{J}^{(1,2)} = \bar{J}^{(2,2)} = \bar{J}^{(3,2)} = \eta_0^{2-2s}\left(1 - \eta_0\right),$$

$$\phi_2^{(1,2)} = \eta_1 \qquad \phi_2^{(2,2)} = 1 \qquad \phi_2^{(3,2)} = \eta_1$$

and $\bar{\psi}_2^{\ell,c}$ as above.

- K and e share a vertex, i.e. $c = 1$

$$L_1 = 2, \qquad \bar{J}^{(1,1)} = \eta_0^{1-2s}, \bar{J}^{(2,1)} = \eta_0^{1-2s}\eta_1$$

$$\bar{\psi}_k^{(1,1)} = \begin{cases} \eta_2 - 1 \\ 1 - \eta_1 \\ \eta_1 \\ -\eta_2 \end{cases} \qquad \bar{\psi}_k^{(2,1)} = \begin{cases} 1 - \eta_1 \\ \eta_1 - \eta_1\eta_2 \\ \eta_1\eta_2 \\ -1 \end{cases}$$

Appendix 2: Proof of Consistency Error Due to Quadrature

Next, we give the proof for the consistency error of the quadrature approximation first stated in Sect. 4.2.

Theorem 3 *For $d = 2$, let $\mathscr{I}_K$ index the degrees of freedom on $K \in \mathscr{P}_h$, and define $\mathscr{I}_{K\times\tilde{K}} := \mathscr{I}_K \cup \mathscr{I}_{\tilde{K}}$. Let k_T (respectively $k_{T,\partial}$) be the quadrature order used for touching pairs $K \times \tilde{K}$ (respectively $K \times e$), and let $k_{NT}\left(K, \tilde{K}\right)$ (respectively $k_{NT,\partial}\left(K, e\right)$) be the quadrature order used for pairs that have empty intersection. Denote the resulting approximation to the bilinear form $a\left(\cdot,\cdot\right)$ by $a_Q\left(\cdot,\cdot\right)$. Then the consistency error due to quadrature is bounded by*

$$\left|a(u, v) - a_Q(u, v)\right| \leq C\left(E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial}\right)\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \qquad \forall u, v \in V_h,$$

where the errors are given by

$$E_T = h^{-2-2s}\rho_1^{-2k_T},$$

$$E_{NT} = \max_{K,\tilde{K}\in\mathscr{P}_h, \overline{K}\cap\overline{\tilde{K}}=\emptyset} h^{-2} d_{K,\tilde{K}}^{-2s}\left(\rho_2 \frac{d_{K,\tilde{K}}}{h}\right)^{-2k_{NT}\left(K,\tilde{K}\right)},$$

$$E_{T,\partial} = h^{-1-2s}\rho_3^{-2k_{T,\partial}},$$

$$E_{NT,\partial} = \max_{K\in\mathscr{P}_h, e\in\mathscr{P}_{h,\partial}, \overline{K}\cap\overline{e}=\emptyset} h^{-1}d_{K,e}^{-2s}\left(\rho_4\frac{d_{K,e}}{h}\right)^{-2k_{NT,\partial}(K,e)},$$

$d_{K,\tilde{K}} := \inf_{\mathbf{x}\in K,\mathbf{y}\in\tilde{K}}|\mathbf{x}-\mathbf{y}|$, $d_{K,e} := \inf_{\mathbf{x}\in K,\mathbf{y}\in e}|\mathbf{x}-\mathbf{y}|$, and $\rho_j > 1$, $j = 1, 2, 3, 4$, are constants.

Proof Let the quadrature rules for the pairs $K\times\tilde{K}$ and $K\times e$ be denoted by $a_Q^{K\times\tilde{K}}(\cdot,\cdot)$ and $a_Q^{K\times e}(\cdot,\cdot)$. Set

$$E_{K\times\tilde{K}}^{i,j} = a^{K\times\tilde{K}}\left(\phi_i,\phi_j\right) - a_Q^{K\times\tilde{K}}\left(\phi_i,\phi_j\right),$$
$$E_{K\times e}^{i,j} = a^{K\times e}\left(\phi_i,\phi_j\right) - a_Q^{K\times e}\left(\phi_i,\phi_j\right).$$

For $u, v \in V_h$, we set

$$E_{K\times\tilde{K}}(u,v) = \sum_{i\in\mathscr{I}_{K\times\tilde{K}}}\sum_{j\in\mathscr{I}_{K\times\tilde{K}}} u_i v_j E_{K\times\tilde{K}}^{i,j},$$
$$E_{K\times e}(u,v) = \sum_{i\in\mathscr{I}_K}\sum_{j\in\mathscr{I}_K} u_i v_j E_{K\times e}^{i,j}$$

so that

$$\begin{aligned}\left|E_{K\times\tilde{K}}(u,v)\right| &\leq \left(\max_{i,j}\left|E_{K\times\tilde{K}}^{i,j}\right|\right)\sum_{i\in\mathscr{I}_{K\times\tilde{K}}}|u_i|\sum_{j\in\mathscr{I}_{K\times\tilde{K}}}\left|v_j\right| \\ &\leq \left(\max_{i,j}\left|E_{K\times\tilde{K}}^{i,j}\right|\right)\left|\mathscr{I}_{K\times\tilde{K}}\right|\sqrt{\sum_{i\in\mathscr{I}_{K\times\tilde{K}}}|u_i|^2}\sqrt{\sum_{j\in\mathscr{I}_{K\times\tilde{K}}}\left|v_j\right|^2}, \\ \left|E_{K\times e}(u,v)\right| &\leq \left(\max_{i,j}\left|E_{K,e}^{i,j}\right|\right)\sum_{i\in\mathscr{I}_K}|u_i|\sum_{j\in\mathscr{I}_K}\left|v_j\right| \\ &\leq \left(\max_{i,j}\left|E_{K,e}^{i,j}\right|\right)\left|\mathscr{I}_K\right|\sqrt{\sum_{i\in\mathscr{I}_K}|u_i|^2}\sqrt{\sum_{j\in\mathscr{I}_K}\left|v_j\right|^2}\end{aligned}$$

Since

$$\sum_{i\in\mathscr{I}_{K\times\tilde{K}}}|u_i|^2 \leq C\left[h_K^{-d}\int_K u^2 + h_{\tilde{K}}^{-d}\int_{\tilde{K}} u^2\right],$$
$$\sum_{i\in\mathscr{I}_K}|u_i|^2 \leq Ch_K^{-d}\int_K u^2,$$

we find

$$
\begin{aligned}
|a(u,v)-a_Q(u,v)| &\le \sum_K \sum_{\tilde{K}} \left|E_{K\times\tilde{K}}(u,v)\right| + \sum_K \sum_e \left|E_{K\times e}(u,v)\right| \\
&\le C \sum_K \sum_{\tilde{K}} \left(\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right) h^{-d} \left[\|u\|^2_{L^2(K)} + \|u\|^2_{L^2(\tilde{K})}\right]^{1/2} \\
&\qquad \left[\|v\|^2_{L^2(K)} + \|v\|^2_{L^2(\tilde{K})}\right]^{1/2} \\
&\quad + C \sum_K \sum_e \left(\max_{i,j}\left|E^{i,j}_{K\times e}\right|\right) h^{-d} \|u\|_{L^2(K)} \|v\|_{L^2(K)} \\
&\le Ch^{-d}\left(\max_{K,\tilde{K}}\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right) \sum_K \sum_{\tilde{K}} \|u\|_{L^2(K\cup\tilde{K})} \|v\|_{L^2(K\cup\tilde{K})} \\
&\quad + Ch^{-d}\left(\max_{K,e}\max_{i,j}\left|E^{i,j}_{K\times e}\right|\right) \sum_K \sum_e \|u\|_{L^2(K)} \|v\|_{L^2(K)} .
\end{aligned}
$$

Because

$$
\begin{aligned}
\sum_K \sum_{\tilde{K}} \|u\|_{L^2(K\cup\tilde{K})} \|v\|_{L^2(K\cup\tilde{K})} &\le \sqrt{\sum_K \sum_{\tilde{K}} \|u\|^2_{L^2(K\cup\tilde{K})}}\sqrt{\sum_K \sum_{\tilde{K}} \|v\|^2_{L^2(K\cup\tilde{K})}} \\
&\le 2\left|\mathscr{P}_h\right| \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\le Ch^{-d} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_K \sum_e \|u\|_{L^2(K)} \|v\|_{L^2(K)} &\le \left|\mathscr{P}_{h,\partial}\right| \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\le Ch^{1-d} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} ,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
|a(u,v)-a_Q(u,v)| \le C\Bigg[& h^{-2d}\left(\max_{K,\tilde{K}}\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right) \\
& + h^{1-2d}\left(\max_{K,e}\max_{i,j}\left|E^{i,j}_{K\times e}\right|\right)\Bigg] \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} .
\end{aligned}
$$

For $d = 2$, using Theorem 6 stated below permits to conclude. □

Theorem 6 ([22], Theorems 5.3.23 and 5.3.24) *If K and $\tilde{K}$ (K and e) are touching elements, then*

$$\left|E^{i,j}_{K\times\tilde{K}}\right| \leq Ch^{2-2s}\rho_1^{-2k_T},$$

$$\left|E^{i,j}_{K\times e}\right| \leq Ch^{2-2s}\rho_3^{-2k_{T,\partial}},$$

where $\rho_1, \rho_3 > 1$ and k_T, $k_{T,\partial}$ are the quadrature orders in every dimension of Eqs. (5) *and* (6).

If K and $\tilde{K}$ (K and e) are not touching, then

$$\left|E^{i,j}_{K\times\tilde{K}}\right| \leq Ch^2 d_{K,\tilde{K}}^{-2s}\tilde{\rho}_2\left(K,\tilde{K}\right)^{-2k_{NT}},$$

$$\left|E^{i,j}_{K\times e}\right| \leq Ch^2 d_{K,e}^{-2s}\tilde{\rho}_4\left(K,e\right)^{-2k_{NT,\partial}},$$

where $d_{K,\tilde{K}} := dist(K,\tilde{K})$, $d_{K,e} := dist(K,e)$, $\tilde{\rho}_2(K,\tilde{K}) := \rho_2 \max\left\{\frac{d_{K,\tilde{K}}}{h}, 1\right\}$, and $\tilde{\rho}_4(K,\tilde{K}) := \rho_4 \max\left\{\frac{d_{K,e}}{h}, 1\right\}$, with $\rho_2, \rho_4 > 1$, and k_{NT}, $k_{NT,\partial}$ are the quadrature order in every dimension of Eqs. (3) *and* (4).

References

1. Acosta, G., Borthagaray, J.P.: A fractional Laplace equation: regularity of solutions and finite element approximations. ArXiv e-prints (2015)
2. Acosta, G., Bersetche, F.M., Borthagaray, J.P.: A short FE implementation for a 2d homogeneous Dirichlet problem of a Fractional Laplacian. ArXiv e-prints (2016)
3. Ainsworth, M., Glusa, C.: Aspects of an adaptive finite element method for the fractional Laplacian: a priori and a posteriori error estimates, efficient implementation and multigrid solver. Comput. Methods Appl. Mech. Eng. **327**, 4–35 (2017). Doi: 10.1016/j.cma.2017.08.019
4. Ainsworth, M., McLean, W., Tran, T.: The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling. SIAM J. Numer. Anal. **36**(6), 1901–1932 (1999)
5. Bogdan, K., Burdzy, K., Chen, Z.Q.: Censored stable processes. Probab. Theory Relat. Fields **127**(1), 89–152 (2003)
6. Borthagaray, J.P., Del Pezzo, L.M., Martinez, S.: Finite element approximation for the fractional eigenvalue problem. ArXiv e-prints (2016)
7. Caffarelli, L., Silvestre, L.: An extension problem related to the fractional Laplacian. Commun. Partial Differ. Equ. **32**(8), 1245–1260 (2007)
8. Chen, Z.Q., Kim, P.: Green function estimate for censored stable processes. Probab. Theory Relat. Fields **124**(4), 595–610 (2002)
9. Ciarlet, P.: Analysis of the Scott–Zhang interpolation in the fractional order Sobolev spaces. J. Numer. Math. **21**(3), 173–180 (2013)
10. D'Elia, M., Gunzburger, M.: The fractional Laplacian operator on bounded domains as a special case of the nonlocal diffusion operator. Comput. Math. Appl. **66**(7), 1245–1260 (2013)
11. Duffy, M.G.: Quadrature over a pyramid or cube of integrands with a singularity at a vertex. SIAM J. Numer. Anal. **19**(6), 1260–1262 (1982)

12. Erichsen, S., Sauter, S.A.: Efficient automatic quadrature in 3-d Galerkin BEM. Comput. Methods Appl. Mech. Eng. **157**(3–4), 215–224 (1998)
13. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York, NY (2004)
14. Getoor, R.K.: First passage times for symmetric stable processes in space. Trans. Am. Math. Soc. **101**(1), 75–90 (1961)
15. Golovin, A.A., Matkowsky, B.J., Volpert, V.A.: Turing pattern formation in the Brusselator model with superdiffusion. SIAM J. Appl. Math. **69**(1), 251–272 (2008)
16. Graham, I.G., Hackbusch, W., Sauter, S.A.: Hybrid Galerkin boundary elements: theory and implementation. Numer. Math. **86**(1), 139–172 (2000)
17. Hackbusch, W., Nowak, Z.P.: On the fast matrix multiplication in the boundary element method by panel clustering. Numer. Math. **54**(4), 463–491 (1989)
18. Koto, T.: IMEX Runge–Kutta schemes for reaction–diffusion equations. J. Comput. Appl. Math. **215**(1), 182–195 (2008)
19. McLean, W.C.H.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge, (2000)
20. Meerschaert, M.M., Sikorskii, A.: Stochastic Models for Fractional Calculus, de Gruyter Studies in Mathematics, vol. 43. Walter de Gruyter & Co., Berlin (2012)
21. Nochetto, R.H., Otárola, E., Salgado, A.J.: A PDE approach to fractional diffusion in general domains: a priori error analysis. Found. Comput. Math. **15**(3), 733–791 (2015)
22. Sauter, S.A., Schwab, C.: Boundary Element Methods, pp. 183–287. Springer, Berlin (2011)
23. Scott, L.R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**(190), 483–493 (1990)
24. Servadei, R., Valdinoci, E.: On the spectrum of two different fractional operators. Proc. Roy. Soc. Edinb.: Sect. A Math. **144**(04), 831–855 (2014)
25. Silling, S.A.: Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech. Phys. Solids **48**(1), 175–209 (2000)
26. Sloan, I.H.: Error analysis of boundary integral methods. Acta Numer. **1**, 287–339 (1992)
27. Sloan, I.H., Spence, A.: The Galerkin method for integral equations of the first kind with logarithmic kernel: theory. IMA J. Numer. Anal. **8**(1), 105–122 (1988)
28. Stroud, A.H.: Approximate Calculation of Multiple Integrals. Prentice-Hall, Englewood Cliffs, NJ (1971)
29. Valdinoci, E.: From the long jump random walk to the fractional Laplacian. SeMA J.: Boletín de la Sociedad Española de Matemática Aplicada **49**, 33–44 (2009)
30. West, B.J.: Fractional Calculus View of Complexity: Tomorrow's Science. CRC Press, New York (2016)
31. Yan, Y., Sloan, I.H., et al.: On integral equations of the first kind with logarithmic kernels. J. Integral Equ. Appl. **1**, 549–579 (1988)

Irregularities of Distributions and Extremal Sets in Combinatorial Complexity Theory

Christoph Aistleitner and Aicke Hinrichs

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In 2004 the second author of the present paper proved that a point set in $[0,1]^d$ which has star-discrepancy at most ε must necessarily consist of at least $c_{\text{abs}} d\varepsilon^{-1}$ points. Equivalently, every set of n points in $[0,1]^d$ must have star-discrepancy at least $c_{\text{abs}} dn^{-1}$. The original proof of this result uses methods from Vapnik–Chervonenkis theory and from metric entropy theory. In the present paper we give an elementary combinatorial proof for the same result, which is based on identifying a sub-box of $[0,1]^d$ which has approximately d elements of the point set on its boundary. Furthermore, we show that a point set for which no such box exists is rather irregular, and must necessarily have a large star-discrepancy.

1 Introduction and Statement of Results

Let $\mathscr{A}^*$ denote the class of all axis-parallel boxes in $[0,1]^d$ which have one vertex at the origin. The *star-discrepancy* of a point set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in [0,1]^d$ is defined as

$$D_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sup_{A \in \mathscr{A}^*} \left| \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_A(\mathbf{x}_k) - \text{vol}(A) \right|,$$

C. Aistleitner (✉)
TU Graz, Institute of Analysis and Number Theory, Graz, Austria
e-mail: aistleitner@math.tugraz.at

A. Hinrichs
Institute of Analysis, University Linz, Linz, Austria
e-mail: aicke.hinrichs@jku.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_3

where $\mathbf{1}_A$ denotes the indicator function of A.[1] The notion of the star-discrepancy is crucial for the *Quasi-Monte Carlo integration* method, in which the integral over $[0,1]^d$ of a d-variate function f is approximated by the average $\frac{1}{n}\sum_{k=1}^{n} f(\mathbf{x}_k)$. Famously, by the Koksma–Hlawka inequality the error in this numerical integration method can be estimated by the product of the variation of f in an appropriate sense and the star-discrepancy of the set of sampling points $\mathbf{x}_1,\ldots,\mathbf{x}_n$. More information on this topic can be found in the classical monographs [3, 4, 8].

The most famous open problem in discrepancy theory concerns the necessary degree of irregularity of a point distribution in the multidimensional unit cube. More precisely, the problem asks for the smallest possible order of the discrepancy of a set of n points in $[0,1]^d$, which was partially answered by the celebrated results of Roth [12] and Bilyk–Lacey–Vagharshakyan [2] (see [1] for a survey) on the one hand and by many constructions of so-called low-discrepancy point sets (see [3]) on the other hand. In the formulation of this problem it is understood that d is fixed and $n \to \infty$. Another important open problem, which recently has received some attention, asks for the order of the *inverse of the star-discrepancy*: given a number $\varepsilon > 0$, what is the minimal cardinality $n^*(\varepsilon, d)$ of a point set in $[0,1]^d$ achieving star-discrepancy at most ε? This problem can also been seen as an irregularities-of-distributions problem, but one where the role of the *simultaneous* dependence of the minimal size of the discrepancy on both d and n is emphasized.

Concerning the inverse of the discrepancy, it is known that

$$n^*(\varepsilon, d) \le c_{\mathrm{abs}} d \varepsilon^{-2} \tag{1}$$

from a fundamental paper of Heinrich–Novak–Wasilkowski–Woźniakowski [6], and that

$$n^*(\varepsilon, d) \ge c_{\mathrm{abs}} d \varepsilon^{-1} \qquad (\varepsilon < \varepsilon_0) \tag{2}$$

due to a result of the second author of the present paper [7].[2] Thus the inverse of the discrepancy depends *linearly* on the dimension d, while the dependence on ε constitutes an important open problem. Novak and Woźniakowski conjectured that the exponent 2 of ε^{-1} in (1) is optimal. In [11, p. 63] they write:

> How about the dependence on ε^{-1}? This is open and seems to be a difficult problem. [...] We think that as long as we consider upper bounds of the form $n^*(\varepsilon, d) \le c_{\mathrm{abs}} d^k \varepsilon^{-\alpha}$, the exponent $\alpha \ge 2$ and 2 cannot be improved.

See also [10, Open problem 7] and [5, Problem 3].

[1] It does not make any difference, but for convenience we will assume in this paper that the boxes in $\mathscr{A}^*$ are closed. We will also allow point sets to contain identical points, so strictly speaking our point sets are not sets, but multi-sets.

[2] Throughout this paper, c_{abs} denotes positive absolute constants, not always the same.

Note that (1) and (2) can be formulated in a different, alternative form. Equation (1) is equivalent to saying that for all d and n there exist $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0,1]^d$ such that

$$D_n^*(\mathbf{x}_1, \dots, \mathbf{x}_n) \le c_{\text{abs}} \frac{\sqrt{d}}{\sqrt{n}},$$

while (2) is equivalent to the statement that for all $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0,1]^d$ we have

$$D_n^*(\mathbf{x}_1, \dots, \mathbf{x}_n) \ge c_{\text{abs}} \frac{d}{n} \qquad (n \ge c_{\text{abs}} d). \tag{3}$$

The proof of (2) in [7] uses methods from combinatorial complexity theory (more precisely, Vapnik–Chervonenkis theory) together with methods from metric entropy theory. The purpose of the present paper is twofold. On the one hand, we want to give an elementary proof of (2), in the spirit of the "cheap proof" which will be sketched below. On the other hand, we will use Vapnik–Chervonenkis theory (VC theory) and metric entropy theory in order to show that point sets which prohibit an application of the "cheap proof" must necessarily have a rather simple combinatorial structure from the point of view of VC theory, and must consequently have particularly large discrepancy.

The idea of the "cheap proof" is very simple. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be points in $[0,1]^d$. Find a box $A \in \mathscr{A}^*$ such that d points of $\mathbf{x}_1, \dots, \mathbf{x}_n$ are situated on the "right upper" boundary of A (that is, on one of the faces which are *not* adjacent to the origin). Let A_1 be a box which is just a little bit smaller than A and let A_2 a box which is just a little bit larger than A. Then the volumes of A_1 and A_2 are essentially equal, while the difference in points is at least d. Thus the star-discrepancy of $\mathbf{x}_1, \dots, \mathbf{x}_n$ is at least $\frac{d}{2n}$.

The problem with the "cheap argument" clearly is that it is not always possible to find a box A which has d points on its boundary—see for example the point set in Fig. 1 below. However, our proof of Theorem 1 shows that a slight modification of the "cheap argument" can actually be successfully implemented. Furthermore, as Theorem 2 will show, a point set which does not allow the "cheap argument" must have a very strong internal structure, and in particular must have a small combinatorial complexity in the sense of VC theory.

Theorem 1 *Let* $\mathbf{x}_1, \dots, \mathbf{x}_n$ *be points in* $[0,1]^d$*. Then*

$$D_n^*(\mathbf{x}_1, \dots, \mathbf{x}_n) \ge \frac{d}{12n},$$

provided that $n \ge 250d$*.*

From Theorem 1 we can deduce that

$$n^*(\varepsilon, d) \ge \frac{d\varepsilon^{-1}}{12}, \qquad \left(\varepsilon < \frac{1}{3000}\right).$$

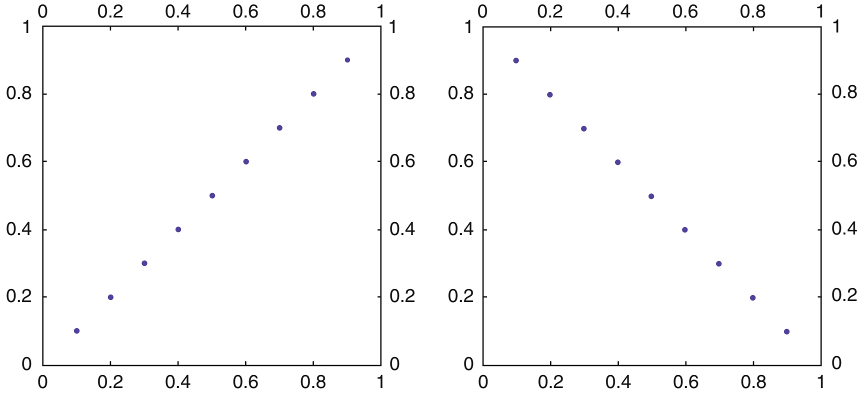


Fig. 1 Two "extremal" point sets. Note that point set (A) on the left is exactly the "opposite" of the point set (B) on the right

The constants appearing in Theorem 1 may be compared with those given in [7] for (2), where it was shown that

$$D_n^*(\mathbf{x}_1,\dots,\mathbf{x}_N) \geq \frac{d}{32e^2 n}, \qquad (n \geq d),$$

with $32e^2 \approx 236$. However, the reason for writing the present paper was not to improve the numerical constants in (2); rather, the purpose of this paper is to share some observations which we consider interesting.

The following theorem states, informally speaking, that a point configuration which does not allow one to apply the "cheap argument" must necessarily have a small combinatorial complexity, and consequently must have a large discrepancy. In other words, either the "cheap argument" is applicable straightforward or the point configuration must have even larger discrepancy especially because it prohibits the application of the "cheap argument". In the statement of the theorem, as the "right upper" boundary of an anchored axis-parallel box $A = [\mathbf{0}, \mathbf{a}]$ we mean the union of all those $(d-1)$-dimensional faces of A which are adjacent to the ("right upper") point $\mathbf{a}$.

Theorem 2 *Let $\mathbf{x}_1,\dots,\mathbf{x}_n$ be points in $[0,1]^d$, and assume that it is not possible to find a box in $\mathscr{A}^*$ such that the right upper boundary of this box contains at least $d/4$ of these points. Assume also that $n \geq d$. Then*

$$D_n^*(\mathbf{x}_1,\dots,\mathbf{x}_n) \geq \frac{d^{3/4}}{372 n^{3/4}}.$$

We finish the introduction with a discussion on the applicability of the "cheap proof" and on the combinatorial complexity of point sets. Here combinatorial complexity refers to the cardinality of the set

$$\{A \cap \{\mathbf{x}_1, \dots, \mathbf{x}_n\} : A \in \mathscr{A}^*\} \tag{4}$$

(this is a set of subsets of $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$). Since the class of anchored axis-parallel boxes $\mathscr{A}^*$ is a *Vapnik–Chervonenkis class* (VC class) of index d, the cardinality of the set (4) can be bounded by the Sauer–Shelah lemma, which asserts that this cardinality is at most

$$\sum_{i=0}^{d} \binom{n}{i}; \tag{5}$$

this is one of the main ingredients of the "entropy argument" in the proof of (2) in [7] (for definitions in the context of VC theory, see Sect. 4). This entropy argument gives better (that is, larger) lower bounds for the discrepancy the smaller the cardinality of (4) can be shown to be. As we will show in Sect. 4 below, a point set which prohibits the application of the "cheap argument" by having the property stated in the assumption of Theorem 2 must have a very small combinatorial complexity in the sense that the cardinality of (4) is much smaller than what could be deduced from the Sauer–Shelah lemma. However, in turn, if an improvement of the Sauer–Shelah lemma is not possible since the point set does not satisfy the assumptions of Theorem 2, then obviously the "cheap argument" is applicable to this point set. So two competing forces are at work here, both of which lead to a large discrepancy in one way or the other.

To illustrate the situation we present two extremal point sets in Fig. 1 (unfortunately the pictures are restricted to the less instructive two-dimensional case). The point set (A) on the left-hand side is extremal in the sense that it prohibits the application of the "cheap argument"—there is no anchored axis-parallel box which has two elements of the point set on its right-upper boundary. On the other hand, the point set (A) has very low complexity in the sense of VC theory: there are 9 points, and the cardinality of (4) is obviously 10 (note that the empty set also counts), which is smallest possible (unless we allow points to coincide). In contrast, for the point set (B) the cardinality of (4) can be calculated to be 46, which is $\binom{9}{0} + \binom{9}{1} + \binom{9}{2}$ and thus by (5) is largest possible. On the other hand, the "cheap argument" is obviously applicable to this point set, and actually there is a very large number of boxes which have two elements of (B) on their "right upper" boundary.

The outline of the remaining part of this paper is as follows. In Sect. 2 we use the "cheap argument" to prove Theorem 1 in the case $n \geq ed^2$, which is simpler than the general case and particularly instructive. In Sect. 3 we use the "cheap argument" to prove Theorem 1 in the general case. In Sect. 4 we introduce the necessary notions from VC theory and prove Theorem 2.

2 The “Cheap Proof” of Theorem 1 in the Case $n \geq ed^2$

In this section we will prove Theorem 1 under the additional assumption that $n \geq ed^2$, since in this case the proof is particularly simple. It also illustrates the idea of the proof in the general case, which, however, requires a more careful reasoning.

We may assume that $d \geq 2$. Let $\mathscr{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a point set in $[0,1]^d$ and let $\kappa = 1 - \frac{1}{d}$. Now consider the boxes $A, B \in \mathscr{A}^*$ given by

$$A = [0,1] \times [0,\kappa]^{d-1} \qquad \text{and} \qquad B = [0,\kappa]^d.$$

Observe that $B \subset A$ and that

$$\text{vol}\,(A \setminus B) = \frac{1}{d}\left(1 - \frac{1}{d}\right)^{d-1} > \frac{1}{ed} \geq \frac{d}{n}, \tag{6}$$

where we used the assumption that $n \geq ed^2$. If $A \cap \mathscr{P} = B \cap \mathscr{P}$ then we find

$$\begin{aligned} 2D_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) &\geq \left(\text{vol}\,(A) - \frac{\#(A \cap \mathscr{P})}{n}\right) + \left(\frac{\#(B \cap \mathscr{P})}{n} - \text{vol}\,(B)\right) \\ &= \text{vol}\,(A \setminus B) \geq \frac{d}{n}, \end{aligned}$$

which implies

$$D_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq \frac{d}{2n}. \tag{7}$$

On the other hand, if $A \cap \mathscr{P} \neq B \cap \mathscr{P}$, then there exists a point $\mathbf{y}_1 = (y_1^{(1)}, \ldots, y_1^{(d)}) \in \mathscr{P}$ in $A \setminus B$, i.e.

$$y_1^{(1)} > \kappa \qquad \text{and} \qquad y_1^{(k)} \leq \kappa \;\; \text{for } k \neq 1.$$

Arguing similarly for the other coordinates by modifying the set A such that it ranges all the way from 0 to 1 not in the first, but instead in the second, third, etc. coordinate, we either have already proved (7) or we find $y_1, \ldots, y_d \in \mathscr{P}$ such that

$$y_j^{(j)} > \kappa \qquad \text{and} \qquad y_j^{(k)} \leq \kappa \;\; \text{for } k \neq j.$$

Obviously, all these points are distinct and contained in the right upper boundary of the box

$$\left[0, y_1^{(1)}\right] \times \left[0, y_2^{(2)}\right] \times \cdots \times \left[0, y_d^{(d)}\right].$$

Thus we have found a box which contains at least d elements of $\mathscr{P}$ on its right upper boundary, and the "cheap argument" from above shows that (7) holds in this case. This proves Theorem 1 (with the value $1/2$ instead of $1/12$ for the constant) in the case $n \geq ed^2$.

3 The "Cheap Proof" of Theorem 1 in the General Case

The proof of Theorem 1 in the general case uses the same idea as the proof in the previous section; however, it requires a slightly more complicated combinatorial argument. The reason why the argument from the previous section fails is that the last inequality of (6) is no longer true, so that we can no longer guarantee that every box of the type $A \setminus B$ contains a point. Consequently we will use a slightly different construction, and distinguish between several cases.

As the reader will see our proof contains several numerical parameters, such as the number 25 in the definition of κ below. Of course we have chosen parameters which give a reasonable result. However, optimizing these parameter is fairly complicated, and we do not claim that we have found the optimal ones. Also, with this method there is a trade-off between the two constants appearing in the statement of the theorem, which are 1/12 and 250 in our formulation of Theorem 1. Decreasing one of them would possibly increase the other, and vice versa. In particular, we checked that the theorem also holds with constants 1/20 and 40.

We fix the point set $\mathscr{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and abbreviate

$$D_n^* = D_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

and $[d] = \{1, \ldots, d\}$. The trivial bound $D_n^* \geq \frac{1}{2n}$ from the one-dimensional case already proves the theorem in the case $d \leq 6$. So we may and do assume that $d \geq 7$.

We will need the reverse Bernoulli-type inequality

$$(1-x)^q \geq 1 - \frac{21}{20}qx \qquad \text{for } 0 \leq x \leq \frac{1}{10} \text{ and } \frac{1}{7} \leq q \leq \frac{1}{4}. \tag{8}$$

This inequality can be easily checked numerically and, what is more tedious, can be proved by elementary analysis. A proof can be found in the Appendix at the very end of the paper.

With

$$\kappa = \left(1 - \frac{25d}{n}\right)^{1/d}$$

we partition the point set $\mathscr{P}$ into subsets according to how many coordinates of the considered point are at least κ:

$$\mathscr{P}_0 = \left\{\mathbf{x} \in \mathscr{P} \ : \ x^{(j)} \le \kappa \text{ for all } j \in [d]\right\},$$
$$\mathscr{P}_1 = \left\{\mathbf{x} \in \mathscr{P} \ : \ x^{(j)} > \kappa \text{ for exactly one } j \in [d]\right\},$$
$$\mathscr{P}_2 = \left\{\mathbf{x} \in \mathscr{P} \ : \ x^{(j)} > \kappa \text{ for at least two } j \in [d]\right\}.$$

Furthermore, let

$$\mathscr{C} = \left\{j \in [d] \ : \ x^{(j)} > \kappa \text{ for some } \mathbf{x} \in \mathscr{P}_1\right\}$$

be the set of coordinates where at least one point in $\mathscr{P}_1$ has its largest coordinate.

We now distinguish between three cases.

Case 1 Assume that $\#\mathscr{C} \ge \frac{d}{6}$.

This is the simple case, when the "cheap proof" is directly applicable. Since every $\mathbf{x} \in \mathscr{P}_1$ has exactly one coordinate j with $x^{(j)} > \kappa$, there exists a box $A \in \mathscr{A}^*$ that contains $\#\mathscr{C} \ge \frac{d}{6}$ points of $\mathscr{P}_1 \subseteq \mathscr{P}$ on its right upper boundary. Hence the "cheap proof" shows that

$$D_n^* \ge \frac{d}{12n}$$

in this case.

Case 2 Assume that $\#\mathscr{C} < \frac{d}{6}$ and $\#\mathscr{P}_1 \ge \frac{107d}{24}$.

In this case there exist many points having exactly one large coordinate, but the "cheap proof" is not applicable since many of these points share the same few coordinate indices where they have their large coordinate. However, since too many points are located close to just a few right upper faces of the unit cube, there must also exist a large sub-box of $[0,1]^d$ (avoiding the proximity of these faces) which does not contain enough points.

More precisely, the box

$$A = [0,\kappa]^{\mathscr{C}} \times [0,1]^{[d]\setminus\mathscr{C}},$$

which extends from 0 to κ for those coordinate indices which are contained in $\mathscr{C}$ and from 0 to 1 for all other coordinates, has volume

$$\mathrm{vol}(A) = \kappa^{\#\mathscr{C}} \ge \kappa^{d/6} = \left(1 - \frac{25d}{n}\right)^{1/6} \ge 1 - \frac{35d}{8n},$$

where the last inequality follows from (8) and the assumption $n \geq 250d$. By definition of the sets $\mathscr{P}_i$, we have

$$\#(A \cap \mathscr{P}) \leq \#\mathscr{P}_0 + \#\mathscr{P}_2 = n - \#\mathscr{P}_1 \leq n - \frac{107d}{24}.$$

This implies

$$D_n^* \geq \operatorname{vol}(A) - \frac{\#(A \cap \mathscr{P})}{n} \geq \left(\frac{107}{24} - \frac{35}{8}\right)\frac{d}{n} = \frac{d}{12n}$$

also in this case.

Case 3 Assume that $\#\mathscr{P}_1 < \frac{107d}{24}$.

This is the most tricky case. Since the cardinality of $\mathscr{P}_1$ is small, the cardinality of $\mathscr{P}_2$ must be large. Thus we have a relatively large number of points which have multiple large coordinates, which means that these points cannot be assigned to different faces of the unit cube (as in Sect. 2 or as in Case 1) but that they are rather located in "corners" of the unit cube. Thus the "cheap proof" is not applicable. However, since many points are located in corners, this means that we can identify a large sub-box of $[0, 1]^d$, reaching all the way from 0 to 1 in many coordinates, which avoids these corners and contains an insufficient number of points of $\mathscr{P}$.

To give a detailed proof in this case, first we consider $A = [0, \kappa]^d$ which has volume $\kappa^d = 1 - \frac{25d}{n}$ and contains exactly those points of $\mathscr{P}$ that are in $\mathscr{P}_0$, that is

$$\#(A \cap \mathscr{P}) = \#\mathscr{P}_0 = n - \#\mathscr{P}_1 - \#\mathscr{P}_2.$$

Then it follows from

$$1 - \frac{\#\mathscr{P}_1 + \#\mathscr{P}_2}{n} - \operatorname{vol}(A) = \frac{\#(A \cap \mathscr{P})}{n} - \operatorname{vol}(A) \leq D_n^*$$

and from the assumption $\#\mathscr{P}_1 < \frac{107d}{24}$ that

$$M := \#\mathscr{P}_2 \geq \frac{493d}{24} - nD_n^*. \tag{9}$$

We now set up an inductive procedure to produce a large box which contains few points by successively removing points of $\mathscr{P}_2$. Let

$$S_0 = \mathscr{P}_2,\ R_0 = \emptyset,\ \mathscr{C}_0 = \emptyset,\ m_0 = \#R_0 = 0.$$

Now assume that $S_{k-1}, R_{k-1} \subset \mathscr{P}_2, \mathscr{C}_{k-1} = \{j_1, \ldots, j_{k-1}\} \subset [d]$ and $m_{k-1} = \#R_{k-1}$ are already defined. By definition of $\mathscr{P}_2$ and double counting we have

$$\begin{aligned}\sum_{j \in [d]\setminus\mathscr{C}_{k-1}} \#\left\{\mathbf{x} \in S_{k-1} \,:\, x^{(j)} > \kappa\right\} &= \sum_{\mathbf{x} \in S_{k-1}} \#\left\{j \in [d] \setminus \mathscr{C}_{k-1} \,:\, x^{(j)} > \kappa\right\} \\ &\geq 2\#S_{k-1}.\end{aligned}$$

Therefore, as long as $S_{k-1} \neq \emptyset$, we find $j_k \in [d] \setminus \mathscr{C}_{k-1}$ such that

$$R_k = \left\{\mathbf{x} \in S_{k-1} \, : \, x^{(j_k)} > \kappa\right\}$$

satisfies

$$m_k = \#R_k \geq \frac{2\#S_{k-1}}{\#[d] \setminus \mathscr{C}_{k-1}} \geq \frac{2\#S_{k-1}}{d}. \tag{10}$$

To complete the inductive construction, let

$$S_k = S_{k-1} \setminus R_k \quad \text{and} \quad \mathscr{C}_k = \mathscr{C}_{k-1} \cup \{j_k\}.$$

If $S_k = \emptyset$ for some $k < d$, we take $S_h = R_h = \emptyset$ for $h \geq k$ and choose j_h arbitrary among the remaining coordinates. Then the inductive process is defined for $k = 0, 1, \ldots, d$, and by (10) we have

$$m_k \geq \frac{2\#S_{k-1}}{d} = \frac{2}{d}\left[M - \sum_{h=1}^{k-1} m_h\right].$$

Fix k and let $q = \frac{k}{d}$. For the total number of points removed up to step k, we then have

$$\sum_{h=1}^{k} m_h \geq \frac{2k}{d}\left[M - \sum_{h=1}^{k} m_h\right] = 2q\left[M - \sum_{h=1}^{k} m_h\right],$$

which, by (9), implies

$$\sum_{h=1}^{k} m_h \geq \frac{2qM}{1+2q} \geq \frac{2q}{1+2q}\left(\frac{493d}{24} - nD_n^*\right).$$

We now consider the box

$$A = [0, \kappa]^{\mathscr{C}_k} \times [0, 1]^{[d] \setminus \mathscr{C}_k}.$$

which has volume

$$\operatorname{vol}(A) = \kappa^{\#\mathscr{C}_k} \geq \kappa^k = \left(1 - \frac{25d}{n}\right)^q \geq 1 - \frac{105q}{4} \cdot \frac{d}{n},$$

where the last inequality follows from (8) and the assumption $n \geq 250d$, provided that

$$\frac{1}{7} \leq q = \frac{k}{d} \leq \frac{1}{4}.$$

Since we assumed $d \geq 7$ and since d is an integer, we can satisfy this condition with the choice $k = \lceil \frac{d}{7} \rceil$.

By construction, none of the removed points in $R_1, \dots, R_k$ is contained in A, which implies

$$\#(A \cap \mathscr{P}) \leq n - \sum_{h=1}^{k} m_h \leq n - \frac{2q}{1+2q} \left(\frac{493d}{24} - nD_n^* \right).$$

Hence

$$D_n^* \geq \text{vol}(A) - \frac{\#(A \cap \mathscr{P})}{n} \geq \left(\frac{493q}{12(1+2q)} - \frac{105q}{4} \right) \frac{d}{n} - \frac{2q}{1+2q} D_n^*,$$

which in turn gives

$$D_n^* \geq \frac{1+2q}{1+4q} \left(\frac{493q}{12(1+2q)} - \frac{105q}{4} \right) \frac{d}{n} = \frac{q(89-315q)}{6(1+4q)} \frac{d}{n}.$$

It is easily verified that

$$\frac{q(89-315q)}{6(1+4q)} \geq \frac{1}{12}$$

for $\frac{1}{7} \leq q \leq \frac{1}{4}$, so that the theorem is also proved in this case.

4 Point Sets Which Prohibit the "Cheap Argument", and Combinatorial Complexity Theory

In Vapnik–Chervonenkis theory (VC theory) the notion of *shattering* plays a crucial role. Let $S = \{x_1, \dots, x_n\}$ be elements of some set X, and let $\mathscr{C}$ denote a collection of subsets of X. We say that $\mathscr{C}$ *shatters* S if

$$\# \{A \cap S : A \in \mathscr{C}\} = 2^n;$$

that is, if using the sets in $\mathscr{C}$ it is possible to pick out every possible subset from S. The *VC index* (or *VC dimension*) of $\mathscr{C}$ is the largest integer n for which there exists a set (of elements of X) of cardinality n which is shattered by $\mathscr{C}$. In our setting we have $X = [0, 1]^d$ and $\mathscr{C} = \mathscr{A}^*$, and the VC dimension of $\mathscr{A}^*$ is d.

Assume that $\mathscr{C}$ has VC dimension d, and that $\#S = n$. Then the Sauer–Shelah lemma asserts that

$$\# \{A \cap S : A \in \mathscr{C}\} \leq \sum_{i=0}^{d} \binom{n}{i} \tag{11}$$

(and this upper bound is in general optimal). We will sketch a proof of this lemma in the setting $X = [0,1]^d$ and $\mathscr{C} = \mathscr{A}^*$ below, since we will use this proof as a blueprint for the key inequality in our proof of Theorem 2.

Set

$$N(n,d) = \max_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in [0,1]^d} \#\{A \cap \{\mathbf{x}_1,\ldots,\mathbf{x}_n\} : A \in \mathscr{A}^*\}. \tag{12}$$

Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be any points in $[0,1]^d$. Assume, without loss of generality, that $\mathbf{x}_1$ has the largest first coordinate among all these points. Then for a given box A for the intersection $A \cap \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ there are two possibilities. Either $\mathbf{x}_1 \notin A$, which means that we "lose" one point. Or $\mathbf{x}_1 \in A$, which means that the first coordinate of the right upper corner of A is at least as large as the first coordinates of all other points as well, and we "lose" one dimension as well as the point $\mathbf{x}_1$ (which by construction is always contained in A in this case). Thus

$$N(n,d) \le N(n-1,d) + N(n-1,d-1), \qquad d,n \ge 2. \tag{13}$$

Together with the trivial initial values $N(1,d) = 2$ and $N(n,1) = n+1$ this leads to a recursion, whose solution gives (11). A detailed version of this proof can be found, for example, on page 46 of [9].

Now we are ready to prove Theorem 2. Let d be fixed. To avoid ambiguities, we write $\mathscr{A}^*(d)$ for the collection of axis-parallel boxes having one vertex at the origin, which are contained in $[0,1]^d$. For points $\mathbf{y}_1,\ldots,\mathbf{y}_m$ in $[0,1]^s$, we say that this collection of points has property $\mathbf{P}(r)$ if it is not possible to find a box in $\mathscr{A}^*(s)$ such that the right upper boundary of this box contains at least r of these points (where the term "right upper boundary" is defined as in the paragraph before the statement of Theorem 2). For the assumptions of Theorem 2 this means that we start with a set $\mathbf{x}_1,\ldots,\mathbf{x}_n$ in $[0,1]^d$ having property $\mathbf{P}(d/4)$.

Set

$$\hat{N}(m,s,r) = \max_{\substack{\mathbf{y}_1,\ldots,\mathbf{y}_m \in [0,1]^s, \\ \mathbf{y}_1,\ldots,\mathbf{y}_m \text{ has property } \mathbf{P}(r)}} \#\{A \cap \{\mathbf{y}_1,\ldots,\mathbf{y}_m\} : A \in \mathscr{A}^*(s)\}.$$

Assume that $r < s$, and let $\mathbf{y}_1,\ldots,\mathbf{y}_m \in [0,1]^s$ be points having property $\mathbf{P}(r)$. Upon a little reflection this implies that there must exist a point in $\mathbf{y}_1,\ldots,\mathbf{y}_m$ which has at least two maximal coordinates (that is, coordinate entries which are at least as large as the corresponding coordinate entries of all the other points). Without loss of generality, assume that this point is $\mathbf{y}_m$, and that its coordinates at positions $s-1$ and s are maximal in this sense.

$$\begin{aligned} &\#\{A \cap \{\mathbf{y}_1,\ldots,\mathbf{y}_m\} : A \in \mathscr{A}^*(s)\} \\ = &\#\{A \cap \{\mathbf{y}_1,\ldots,\mathbf{y}_m\} : A \in \mathscr{A}^*(s),\ \mathbf{y}_m \notin A\} \\ &\quad + \quad \#\{A \cap \{\mathbf{y}_1,\ldots,\mathbf{y}_m\} : A \in \mathscr{A}^*(s),\ \mathbf{y}_m \in A\} \end{aligned}$$

$$
\begin{aligned}
&= \#\{A \cap \{\mathbf{y}_1, \dots, \mathbf{y}_{m-1}\} :\ A \in \mathscr{A}^*(s)\} && (14)\\
&\qquad + \quad \#\{A \cap \{\mathbf{y}_1, \dots, \mathbf{y}_{m-1}\} :\ A \in \mathscr{A}^*(s),\ \mathbf{y}_m \in A\}. && (15)
\end{aligned}
$$

The term in line (14) is clearly dominated by $\hat{N}(m-1, s, r)$. To understand the term in line (15), let $\mathbf{y}^{(s-2)}$ denote the restriction (projection) of a point $\mathbf{y} \in [0,1]^s$ to its first $s-2$ coordinates, and define $A^{(s-2)}$ similarly as a projection of A. By construction $\mathbf{y}_m \in A$ implies that the coordinates at positions $s-1$ and s of all the points $\mathbf{y}_1, \dots, \mathbf{y}_{m-1}$ cannot exceed those of the right upper corner of A. Thus

$$
\begin{aligned}
&\#\{A \cap \{\mathbf{y}_1, \dots, \mathbf{y}_{m-1}\} :\ A \in \mathscr{A}^*(s),\ \mathbf{y}_m \in A\}\\
&= \#\left\{A^{(s-2)} \cap \{\mathbf{y}_1^{(s-2)}, \dots, \mathbf{y}_{m-1}^{(s-2)}\} :\ A \in \mathscr{A}^*(s),\ \mathbf{y}_m \in A\right\}\\
&\le \#\left\{A \cap \{\mathbf{y}_1^{(s-2)}, \dots, \mathbf{y}_{m-1}^{(s-2)}\} :\ A \in \mathscr{A}^*(s-2)\right\}. && (16)
\end{aligned}
$$

Furthermore, the point set $\left\{\mathbf{y}_1^{(s-2)}, \dots, \mathbf{y}_{m-1}^{(s-2)}\right\}$ has property $\mathbf{P}(r)$, which is inherited from the original point set $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$. Thus the term in line (16) is dominated by $\hat{N}(m-1, s-2, r)$, and in total we have

$$
\hat{N}(m, s, r) \le \hat{N}(m-1, s, r) + \hat{N}(m-1, s-2, r). \tag{17}
$$

This is an analogue of (13), except that now we "lose" two dimensions rather than only one, and that it is only valid as long as $r < s$.

Note that from the definition of $\hat{N}(m, s, r)$ we have $\hat{N}(m, s, r) \le N(m, s)$ for all m, s, r. Now we claim the following:

Claim We have $\hat{N}(n, d, r) \le N(n, r) \sum_{0 \le i \le d/2} \binom{n}{i}$.

The claim is obviously right whenever $r \ge d$, since then

$$
\hat{N}(n, d, r) \le N(n, d) \le N(n, r).
$$

On the other hand, whenever $r < d$, then by (17) we have

$$
\begin{aligned}
\hat{N}(n, d, r) &\le \hat{N}(n-1, d, r) + \hat{N}(n-1, d-2, r)\\
&\le N(n, r) \sum_{0 \le i \le d/2} \binom{n-1}{i} + N(n, r) \sum_{0 \le i \le d/2 - 1} \binom{n-1}{i}\\
&= N(n, r) \sum_{0 \le i \le d/2} \left(\binom{n-1}{i} + \binom{n-1}{i-1}\right) && (18)\\
&= N(n, r) \sum_{0 \le i \le d/2} \binom{n}{i}. && (19)
\end{aligned}
$$

where in line (18) we read $\binom{n-1}{-1} = 0$. Thus the claim is true by induction. Classically we have

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

for $n \geq d$ (see for example [9, Corollary 3.3]), so by (11) and (19) we have

$$\hat{N}(n,d,r) \leq \left(\frac{en}{r}\right)^r \left(\frac{en}{d/2}\right)^{d/2},$$

for $n \geq d$, which in particular yields

$$\hat{N}(n,d,d/4) \leq \left(\frac{4en}{d}\right)^{d/4} \left(\frac{2en}{d}\right)^{d/2} = 2^d \left(\frac{en}{d}\right)^{3d/4}. \tag{20}$$

The remaining part of the proof of Theorem 2 can be carried out similar to the proof of the main theorem in [7]. As shown in equation (8) of [7], for given $\varepsilon > 0$ there exists a collection $\mathscr{C}$ of at least $(8e\varepsilon)^{-d}$ anchored axis-parallel boxes in $[0,1]^d$ such that

$$\text{vol}(C_1 \Delta C_2) \geq \varepsilon \qquad \text{for all } C_1, C_2 \in \mathscr{C},$$

where Δ denotes the symmetric difference. Let $\mathbf{x}_1, \ldots, \mathbf{x}_d$ denote the points from the assumption of Theorem 2. Since $\mathscr{C}$ is a subset of $\mathscr{A}^*$, by (20) we have

$$\#\{C \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} : C \in \mathscr{C}\} \leq 2^d \left(\frac{en}{d}\right)^{3d/4}.$$

Thus by the pigeon hole principle there exist two sets C_1 and C_2 for which

$$C_1 \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = C_2 \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \qquad \text{and} \qquad \text{vol}(C_1 \Delta C_2) \geq \varepsilon, \tag{21}$$

provided that

$$2^d \left(\frac{en}{d}\right)^{3d/4} < \left(\frac{1}{8e\varepsilon}\right)^d. \tag{22}$$

It is easily seen that (21) implies that

$$D_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq \frac{\varepsilon}{4}$$

(see [7, Lemma 6]), and that because of $16e^{7/8} < 93$ inequality (22) is satisfied if we choose

$$\varepsilon = \frac{d^{3/4}}{93n^{3/4}}.$$

This proves Theorem 2.

Acknowledgements The first author is supported by the Austrian Science Fund (FWF), projects F5507-N26 and I1751-N26, and by the FWF START project Y-901-N35.

Appendix: Proof of Inequality (8)

Here we will prove that

$$(1-x)^q \geq 1 - \frac{21}{20}qx \qquad \text{for } 0 \leq x \leq \frac{1}{10} \text{ and } \frac{1}{7} \leq q \leq \frac{1}{4}.$$

We consider the function

$$f_q(x) = (1-x)^q + \frac{21}{20}qx$$

for fixed q with $\frac{1}{7} \leq q \leq \frac{1}{4}$. Since its second derivative

$$f_q''(x) = -q(1-q)(1-x)^{q-2}$$

is negative for $0 \leq x < 1$, the function $f_q(x)$ is strictly concave here. So the minimal value attained by f_q for $0 \leq x \leq \frac{1}{10}$ is either $f_q(0) = 1$ or $f_q(\frac{1}{10})$. It remains to verify that

$$g(q) = f_q\left(\frac{1}{10}\right) = \left(\frac{9}{10}\right)^q + \frac{21q}{200} \geq 1 \quad \text{for} \quad \frac{1}{7} \leq q \leq \frac{1}{4}.$$

Since the exponential function $\left(\frac{9}{10}\right)^q$ is strictly convex on $\mathbb{R}$, this is also true for $g(q)$. Since $g(0) = 1$, it is enough to verify that $g(\frac{1}{7}) \geq 1$ to ensure that $g(q) \geq 1$ also holds for $q \geq \frac{1}{7}$. Finally, $g(\frac{1}{7}) \geq 1$ is equivalent to

$$\frac{9}{10} \geq \left(1 - \frac{21}{1400}\right)^7 = \frac{115149904768984 13}{128000000000000000},$$

which is indeed true.

References

1. Bilyk, D.: Roth's orthogonal function method in discrepancy theory and some new connections. In: A Panorama of Discrepancy Theory. Lecture Notes in Mathematics, vol. 2107, pp. 71–158. Springer, Cham (2014)
2. Bilyk, D., Lacey, M.T., Vagharshakyan, A.: On the small ball inequality in all dimensions. J. Funct. Anal. **254**(9), 2470–2502 (2008)
3. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Cambridge University Press, Cambridge (2010).
4. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Lecture Notes in Mathematics, vol. 1651. Springer, Berlin (1997)
5. Heinrich, S.: Some open problems concerning the star-discrepancy. J. Complex. **19**(3), 416–419 (2003)
6. Heinrich, S., Novak, E., Wasilkowski, G.W., Woźniakowski, H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arith. **96**(3), 279–302 (2001)
7. Hinrichs, A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complex. **20**(4), 477–483 (2004)
8. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Wiley-Interscience [Wiley], New York, London, Sydney (1974)
9. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA (2012)
10. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Vol. 1: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society, Zürich (2008)
11. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Vol. 2: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society, Zürich (2010)
12. Roth, K.F.: On irregularities of distribution. Mathematika **1**, 73–79 (1954)

Importance Sampling and Stratification for Copula Models

Philipp Arbenz, Mathieu Cambou, Marius Hofert, Christiane Lemieux, and Yoshihiro Taniguchi

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract An importance sampling approach for sampling from copula models is introduced. The proposed algorithm improves Monte Carlo estimators when the functional of interest depends mainly on the behaviour of the underlying random vector when at least one of its components is large. Such problems often arise from dependence models in finance and insurance. The importance sampling framework we propose is particularly easy to implement for Archimedean copulas. We also show how the proposal distribution of our algorithm can be optimized by making a connection with stratified sampling. In a case study inspired by a typical insurance application, we obtain variance reduction factors sometimes larger than 1000 in comparison to standard Monte Carlo estimators when both importance sampling and quasi-Monte Carlo methods are used.

1 Introduction

Many applications in finance and insurance lead to the problem of calculating a functional of the form $\mu = \mathbb{E}(\Psi_0(\boldsymbol{X}))$, where $\boldsymbol{X} = (X_1, \dots, X_d) : \Omega \to \mathbb{R}^d$ is a random vector on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and $\Psi_0 : \mathbb{R}^d \to \mathbb{R}$ is a measurable function. If the components of $\boldsymbol{X}$ cannot be assumed to be independent, it is popular to model the distribution function H of $\boldsymbol{X}$ with a copula C, such that $H(x_1, \dots, x_d) =$

P. Arbenz
SCOR, Zurich, Switzerland
ETH Zurich, Zurich, Switzerland

M. Cambou
EdgeLab, Lausanne, Switzerland

M. Hofert · C. Lemieux (✉) · Y. Taniguchi
University of Waterloo, Waterloo, ON, Canada
e-mail: marius.hofert@uwaterloo.ca; clemieux@uwaterloo.ca; ytaniguc@uwaterloo.ca

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_4

$C(F_1(x_1), \dots, F_d(x_d))$, $\boldsymbol{x} \in \mathbb{R}^d$, where $F_j(x) = \mathbb{P}(X_j \le x)$, $j \in \{1, \dots, d\}$, are the univariate margins of H and $C : [0,1]^d \to [0,1]$ is a copula. A copula allows one to separate the dependence structure from the marginal distributions, which is useful for constructing multivariate stochastic models. We assume the reader to be familiar with copulas and refer to [16] or [17] for an introduction; see also Sect. 2 for important background information.

A drawback of using such flexible models is that an analytical form for the quantity of interest $\mathbb{E}(\Psi_0(\boldsymbol{X}))$ rarely exists, and thus numerical methods must be applied to evaluate it. Preferably, the employed techniques should be applicable to high-dimensional problems, which are common in finance. An advantage of Monte Carlo (MC) simulation is that the rate of convergence of its error is independent of the dimensionality of a given problem. Nevertheless, the convergence rate of plain MC is generally slow so that MC is often combined with some variance reduction technique (VRT) to improve the precision of estimators.

Importance sampling (IS) is a VRT often used for rare event simulations. IS attempts to reduce the variance of the MC estimator of $\mathbb{E}(\Psi_0(\boldsymbol{X}))$ by sampling $\boldsymbol{X}$ more frequently from the important region where $|\Psi_0(\boldsymbol{X})|$ is large. While there are many publications that design IS for Gaussian and t-copula models, [2, 8, 12, 19] for instance, not much attention has been given to IS for other types of copulas, including Archimedean copulas. To our knowledge, [3] is the only work which develops IS for Archimedean copulas.

The main contribution of this paper is the study of IS techniques that do not rely on a specific copula structure. We consider the case where the functional Ψ_0 of interest depends mainly on the behaviour of the random vector $\boldsymbol{X}$ when at least one of the components is large. Such problems often arise from dependence models in the realm of finance and insurance. We propose a new IS framework for this setup which can be implemented for all classes of copula models from which sampling is feasible. The main idea of our proposed IS approach is to oversample sets of the form $[0,1]^s \backslash [0, \lambda_k]^s$ for $0 \le \lambda_1 \le \dots \le \lambda_M \le 1$. Explicit algorithms are given in the case of Archimedean copulas. We also examine how to optimally choose the proposal distribution by making a connection with stratified sampling (SS), which is then used to propose yet another estimator based on our general IS setup.

While the plain MC method generates samples based on pseudo-random numbers, quasi-Monte Carlo (QMC) methods use a low-discrepancy sequence (LDS) to draw samples. An LDS has the property of covering the unit cube $[0,1)^d$ more uniformly than pseudo-random numbers generally do. This usually leads to approximations whose error converges to 0 faster than with MC. Furthermore, these methods can be randomized in a way that preserves their low discrepancy but allows for error estimation. Such randomized QMC methods can thus be seen as a VRT. QMC has been primarily used for multinormal models and has shown substantial improvements over plain MC. Recently, its effectiveness for sampling copula models was studied and demonstrated theoretically and empirically in [1]. Building up on that work, it is natural to try to combine QMC with our proposed IS approach.

The rest of this work is organized as follows. In Sect. 2, we motivate our proposed IS method and give the necessary background on Archimedean copulas and QMC methods for copulas. In Sect. 3 we introduce a general IS setup for copula models, and then show in Sect. 4 how to exploit the Marshall–Olkin stochastic representation of Archimedean copulas to design an efficient sampling algorithm for IS. We show that the proposed IS scheme is very similar to stratified sampling and then develop sampling methods for SS estimators. In Sect. 5, we derive variance expressions for the IS and SS estimators. By minimizing such variance expressions, we derive the optimal calibration for our proposal distribution, for both IS and SS estimators. In Sect. 6, we investigate the effectiveness of the proposed IS and SS schemes using numerical experiments. All proofs are deferred to the appendix.

2 Motivation and Background

In a copula model, we may write $\mu = \mathbb{E}(\Psi_0(\boldsymbol{X})) = \mathbb{E}(\Psi(\boldsymbol{U}))$, where $\boldsymbol{U} = (U_1, \dots, U_d) : \Omega \to \mathbb{R}^d$ is a random vector with distribution function C, $\Psi : [0, 1]^d \to \mathbb{R}$ is given by

$$\Psi(u_1, \dots, u_d) = \Psi_0(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)), \tag{1}$$

and $F_j^{-1}(p) = \inf\{x \in \mathbb{R} : F_j(x) \geq p\}$ for $j \in \{1, \dots, d\}$.

If C and $F_1, \dots, F_d$ are known, we can use MC simulation to estimate $\mathbb{E}(\Psi(\boldsymbol{U}))$. For a random sample $\{\boldsymbol{U}_i : i = 1, \dots, n\}$ of $\boldsymbol{U}$, the MC estimator of $\mathbb{E}(\Psi(\boldsymbol{U}))$ is

$$\hat{\mu}_{\mathrm{MC},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi(\boldsymbol{U}_i). \tag{2}$$

In this paper, we consider the case where Ψ is large only when at least one of its arguments is close to 1, or equivalently, if at least one of the components of $\boldsymbol{X}$ is large. This assumption is inspired by several applications in insurance:

- The fair premium of a stop loss cover with deductible D is $\mathbb{E}(\max\{\sum_{j=1}^{d} X_j - D, 0\})$. The corresponding functional is $\Psi(\boldsymbol{u}) = \max\{\sum_{j=1}^{d} F_j^{-1}(u_j) - D, 0\}$; see the left-hand side of Fig. 1 for a contour plot of Ψ for two Pareto margins.
- Risk measures for an aggregate sum $S = \sum_{j=1}^{d} X_j$, such as value-at-risk, $\mathrm{VaR}_\alpha(S)$, or expected shortfall, $\mathrm{ES}_\alpha(S)$, $\alpha \in (0, 1)$, cannot in general be written as an expectation of type $\mathbb{E}(\Psi_0(\boldsymbol{X}))$. However, they are functionals of the aggregate distribution function $F_S(x) = \mathbb{P}(S \leq x) = \mathbb{E}(\Psi(\boldsymbol{U}; x))$, where $\Psi(\boldsymbol{u}; x) = I_{\{F_1^{-1}(u_1) + \dots + F_d^{-1}(u_d) \leq x\}}$. We can therefore write

$$\begin{aligned} \mathrm{VaR}_\alpha(S) &= \inf\{x \in \mathbb{R} : \mathbb{E}(\Psi(\boldsymbol{U}; x)) \geq \alpha\}, \\ \mathrm{ES}_\alpha(S) &= \frac{1}{1-\alpha} \int_\alpha^1 \mathrm{VaR}_u(S)\, du, \end{aligned} \tag{3}$$

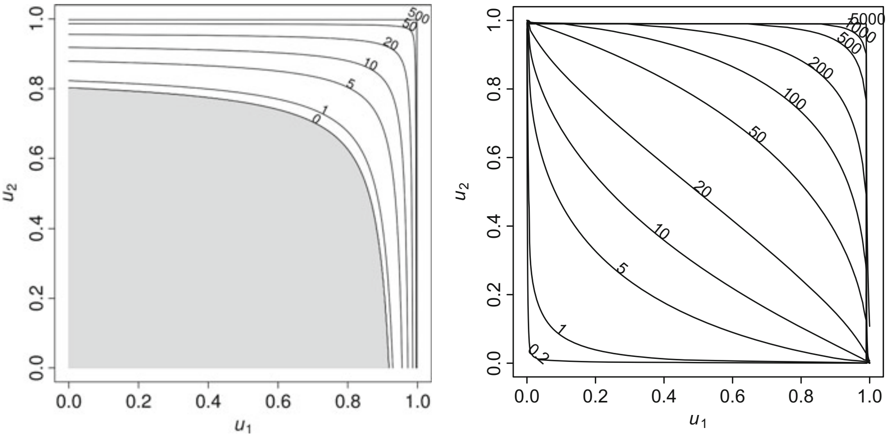


Fig. 1 *Left:* Contour lines for the excess function $\Psi(u_1, u_2) = \max\{F_1^{-1}(u_1) + F_2^{-1}(u_2) - 10, 0\}$, where the margins are Pareto distributed with $F_1(x) = 1 - (1 + x/4)^{-2}$ and $F_2(x) = 1 - (1 + x/8)^{-2}$. The grey area indicates where Ψ is zero. *Right:* Contour lines for the product function $\Psi(u_1, u_2) = F_1^{-1}(u_1)F_2^{-1}(u_2)$, where $X_1 \sim \mathrm{LN}(2, 1)$ and $X_2 \sim \mathrm{LN}(1, 1.5)$

which depend only on those x for which $\mathbb{E}(\Psi(\boldsymbol{U}; x)) \geq \alpha$ holds. This is determined by the tail behaviour of S, which is strongly influenced by the properties of the copula C when at least one component is close to 1. Note that capital allocation methods such as the Euler principle for expected shortfall behave similarly, see [21] and [16, p. 260].

Note that in this framework we follow the convention of [16, Remark 2.1] that $\boldsymbol{X}$ refers to a loss and $-\boldsymbol{X}$ to a profit, which is more common in an actuarial context. One could have equally well worked with the profit-and-loss random variable $-\boldsymbol{X}$ by changing the area of interest to where components of $\boldsymbol{X}$ are small.

2.1 Archimedean Copulas and Sampling Methods

Archimedean copulas form a popular class of copulas in actuarial science and risk management, as they can capture various types of tail dependence. An Archimedean copula admits the representation

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \tag{4}$$

where ψ is a univariate function called *generator* and is such that $\psi : [0, \infty) \to [0, 1]$ with $\psi(0) = 1$ and $\psi(\infty) = \lim_{t\to\infty} \psi(t) = 0$; also $\psi(t)$ is continuous and strictly decreasing on $[0, \inf\{t : \psi(t) = 0\}]$. We review two sampling techniques applicable to Archimedean copulas.

2.1.1 Conditional Distribution Method

The conditional distribution method (CDM) is a sampling technique that in principle works for any copula. For $j \in \{2, \dots, d\}$, let

$$C_{j|1\dots j-1}(u_j \mid u_1, \dots, u_{j-1}) = \mathbb{P}(U_j \le u_j \mid U_1 = u_1, \dots, U_{j-1} = u_{j-1})$$

be the conditional distribution of the jth component given the first $j-1$ components. As a function of u_j, $C_{j|1\dots j-1}(u_j \mid u_1, \dots, u_{j-1})$ is a univariate distribution function on $[0, 1]$ and thus can be sampled via inversion. Doing this iteratively in j based on the previously computed component samples leads to a sample from C according to the CDM. The efficiency of this sampling method depends on the computational cost required to evaluate the conditional quantile functions $C^{-1}_{j|1\dots j-1}(u_j \mid u_1, \dots, u_{j-1})$, which in many cases are not available in closed form. For Archimedean copulas, there exists a more efficient sampling method, which we now describe.

2.1.2 Marshall–Olkin Algorithm

It is well established that ψ induces an Archimedean copula for any dimension $d \ge 2$ if and only if ψ is the Laplace–Stieltjes transform of a distribution function of some positive random variable V, the so-called *frailty*. Based on V, one can derive the stochastic representation

$$\left(\psi\left(\frac{E_1}{V}\right), \dots, \psi\left(\frac{E_d}{V}\right)\right) \sim C, \tag{5}$$

where $E_1, \dots, E_d \overset{\text{ind.}}{\sim} \mathrm{Exp}(1)$ are independent of the positive frailty random variable V whose Laplace–Stieltjes transform is ψ. This sampling method is known as *Marshall–Olkin (MO) algorithm*; see [14].

For many popular Archimedean copulas, the frailty random variable V from the MO algorithm has a known distribution, for instance V is Gamma distributed for Clayton copulas; see Table 1 for information about five popular Archimedean

Table 1 Popular Archimedean generators and corresponding frailty distributions

Family	Parameter	$\psi(t)$	V
Ali-Mikhail-Haq	$\theta \in [0, 1)$	$(1-\theta)/(\exp(t)-\theta)$	$\mathrm{Geo}(1-\theta)$
Clayton	$\theta \in (0, \infty)$	$(1+t)^{-1/\theta}$	$\Gamma(1/\theta, 1)$
Frank	$\theta \in (0, \infty)$	$-\log(1-(1-e^{-\theta})\exp(-t))/\theta$	$\mathrm{Log}(1-e^{-\theta})$
Gumbel	$\theta \in [1, \infty)$	$\exp(-t^{1/\theta})$	$\mathrm{Stable}(1/\theta, 1, \cos^{\theta}(\pi/(2\theta)), I_{\{\theta=1\}}; 1)$
Joe	$\theta \in [1, \infty)$	$1-(1-\exp(-t))^{1/\theta}$	$\mathrm{Sibuya}(1/\theta)$

copulas and the corresponding frailty random variables V, and see [10, Table 1] for the details concerning Table 1. In Sect. 4 we develop an IS algorithm that exploits the MO representation of Archimedean copulas.

2.2 Quasi-Monte Carlo and Copula Models

The combination of QMC and copulas is studied in depth in [1]. To describe how it works, let $\eta : [0,1)^{d+k} \to [0,1)^d$ for $k \geq 0$ be some transformation function such that $\eta(\boldsymbol{U}') \sim C$ for $\boldsymbol{U}' \sim \mathrm{U}[0,1)^{d+k}$. The choice of η corresponds to the choice of sampling methods for C, such as CDM or MO. The plain MC estimator (2) thus becomes

$$\hat{\mu}_{\mathrm{MC},n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\eta(\boldsymbol{U}_i')), \quad \boldsymbol{U}_i' \overset{\text{ind.}}{\sim} \mathrm{U}[0,1)^{d+k}. \tag{6}$$

To use QMC, we replace the point set $\{\boldsymbol{U}_i',\ i = 1,\dots,n\}$ by a low-discrepancy point set. The choice of sampling algorithm η is not very important to control the MC error, but it is for QMC, as explained in [1]. The sampling algorithms we propose in this work are applicable to both MC and QMC, and numerical results for both methods are reported in Sect. 6. For QMC we use a Sobol' sequence [20] and apply to it a randomization based on a digital-shift (see [13, Section 6.2.2]) so that we can construct unbiased estimators and compute confidence intervals for the quantity of interest by using replication.

3 Importance Sampling for Copula Models

IS is a popular variance reduction technique for rare event simulations. Suppose we want to estimate $\mathbb{E}(\Psi(\boldsymbol{U}))$ where $\boldsymbol{U} \sim C$, for a d-dimensional copula C. In IS, we draw samples from some proposal distribution $\tilde{\boldsymbol{U}} \sim G$ and construct the estimator

$$\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\tilde{\boldsymbol{U}}_i) w(\tilde{\boldsymbol{U}}_i), \quad \tilde{\boldsymbol{U}}_i \overset{\text{ind.}}{\sim} G, \tag{7}$$

where $w(\boldsymbol{u}) = \frac{dC(\boldsymbol{u})}{dG(\boldsymbol{u})}$ is the Radon-Nikodym derivative of C with respect to G. The function w works as a weight function so that the estimator remains unbiased after changing the distribution. Intuitively, the variance of the IS estimator is smaller than the variance of the plain MC estimator if the proposal distribution is concentrated around the important region, which we characterized in Sect. 2 as the region where the maximal component of a sample point is close to 1.

In order to define the proposal distribution G, we suggest a mixing approach by taking a weighted average of a multivariate distribution function $C_\lambda : [0,1]^d \to [0,1]$ over different values of λ. Let F_Λ denote a discrete distribution function of a random variable $\Lambda : \Omega \mapsto [0,1)$, defined by $q_k := \mathbb{P}(\Lambda = \lambda_k)$, $k = 1, \dots, M$. We then define the distribution function G of $\tilde{\boldsymbol{U}}$ as a mixture of C_λ with respect to F_Λ:

$$G(\boldsymbol{u}) = \sum_{k=1}^{M} q_k C_{\lambda_k}(\boldsymbol{u}),$$

where C_λ is a distorted version of the copula C itself that concentrates samples in a region of the form $[0,1]^d \setminus [0,\lambda]^d$. Note that the C_λ we will construct (see (8)) is a copula only if $C(\lambda \mathbf{1}) = 0$, but C_λ does not need to be copula for our approach to work.

We will see that this mixture approach is natural in order to allow C to be absolutely continuous with respect to G. In particular, the absolute continuity is guaranteed for any copula C if the following assumption is satisfied.

Assumption 1 *The random variable* Λ *satisfies* $\mathbb{P}(\Lambda = 0) > 0$.

In order to obtain a well defined weight function w and an unbiased estimator $\hat{\mu}_{\mathrm{IS},n}$, Assumption 1 must be fulfilled. Note that this assumption does not require particular conditions on C. Although it seems restrictive, we will see that it is also needed to have a consistent estimator $\hat{\mu}_{\mathrm{IS},n}$. Moreover, ensuring $\mathbb{P}(\Lambda = 0) > 0$ can be seen as a form of defensive mixture sampling, where a fraction of samples are drawn from the original distribution [9]. Defensive sampling bounds the IS weights away from infinity (as will be seen in Lemma 2) so that the resulting estimator has a finite variance. To that end, we assume Assumption 1 to be satisfied in what follows.

The construction of the proposal distribution G as a C_λ-mixture directly yields a sampling method, as one can draw a realization of G by first drawing $\Lambda \sim F_\Lambda$ and then $\tilde{\boldsymbol{U}} \sim C_\Lambda$. Therefore, the following algorithm can be used to construct $\hat{\mu}_{\mathrm{IS},n}$:

Algorithm 1 General IS algorithm for copulas

```
Fix $n \in \mathbb{N}$.
Draw $\Lambda_i \sim F_\Lambda$, $i \in \{1, \dots, n\}$.
Draw $\tilde{\boldsymbol{U}}_i \sim C_{\Lambda_i}$, $i \in \{1, \dots, n\}$.
Calculate $w(\tilde{\boldsymbol{U}}_i) = dC(\tilde{\boldsymbol{U}}_i)/dG(\tilde{\boldsymbol{U}}_i)$, $i \in \{1, \dots, n\}$.
Return $\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\tilde{\boldsymbol{U}}_i) w(\tilde{\boldsymbol{U}}_i)$.
```

The following lemma establishes consistency and asymptotic normality of the estimator $\hat{\mu}_{\mathrm{IS},n}$.

Lemma 1 *Suppose that* $\mathrm{Var}(\Psi(\boldsymbol{U})) < \infty$ *and that* $w(\,\cdot\,) \le B$ *for a constant* $B < \infty$. *Then*

1. $\hat{\mu}_{\mathrm{IS},n}$ *converges* $\mathbb{P}$*-almost surely to* μ;
2. $\sigma^2 = \mathrm{Var}(\Psi(\tilde{\boldsymbol{U}}) w(\tilde{\boldsymbol{U}})) < \infty$ *and* $\sqrt{n}(\hat{\mu}_{\mathrm{IS},n} - \mu) \to \mathcal{N}(0, \sigma^2)$ *in distribution.*

We will later show that under some mild assumptions on F_Λ, the weight function will indeed be bounded on $[0, 1]$.

The form of C_λ we propose to work with is the distribution of $\boldsymbol{U}$ conditioned on the event that at least one of its components exceeds λ:

$$\begin{aligned} C_\lambda(\boldsymbol{u}) &= \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d \mid \max\{U_1, \ldots, U_d\} > \lambda) \\ &= \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d \mid \boldsymbol{U} \notin [0, \lambda]^d) \\ &= \frac{C(\boldsymbol{u}) - C\left(\min\{u_1, \lambda\}, \ldots, \min\{u_d, \lambda\}\right)}{1 - C(\lambda \mathbf{1})}, \end{aligned} \tag{8}$$

where $\lambda\mathbf{1} = \lambda(1, \ldots, 1) = (\lambda, \ldots, \lambda) \in [0, 1)^d$. By putting mass of Λ on $(0, 1)$, we can put more weight on the region of the copula where at least one component is large. For instance, if F_Λ is discrete and $\mathbb{P}(\Lambda = 0) = \mathbb{P}(\Lambda = 0.9) = 0.5$, then 50% of the samples of $\tilde{\boldsymbol{U}}$ are constrained to lie only in $[0, 1]^d \setminus [0, 0.9]^d$ while the other 50% of the samples will lie on $[0, 1]^d$. Note that the mass on $[0, 1]^d \setminus [0, 0.9]^d$ would then be higher than 50% since we can still sample from $[0, 1]^d \setminus [0, 0.9]^d$ when $\Lambda = 0$. On the other hand, the case $\mathbb{P}(\Lambda = 0) = 1$ yields $G = C$ since $C_\lambda = C$ for $\lambda = 0$.

We now describe how the weight function w based on the above choice for C_λ can be calculated.

Theorem 1 *The Radon–Nikodym derivative $w(\boldsymbol{u}) = dC(\boldsymbol{u})/dG(\boldsymbol{u})$ is given by*

$$w(\boldsymbol{u}) = \Big(\sum_{k=1}^{M} \frac{I_{\{\lambda_k \leq \max\{u_1, \ldots, u_d\}\}}}{1 - C(\lambda_k \mathbf{1})} q_k\Big)^{-1}.$$

In order to simplify the notation, let $\widetilde{w} : [0, 1] \to [0, \infty)$ be defined as

$$\widetilde{w}(u) = \Big(\sum_{k=1}^{M} \frac{I_{\{\lambda_k \leq u\}}}{1 - C(\lambda_k \mathbf{1})} q_k\Big)^{-1}. \tag{9}$$

Therefore we have that $w(\boldsymbol{u}) = \widetilde{w}(\max\{u_1, \ldots, u_d\})$. In order to evaluate $\widetilde{w}$, it is sufficient to calculate (or approximate) $C(\lambda_k \mathbf{1})$ for $k \in \{1, \ldots, M\}$. These values must be calculated only once and thus this approach is fast and can be easily implemented. In particular, the density of C does not have to be evaluated to calculate w (or $\tilde{w}$). This is in an advantage in comparison to most other IS algorithms, for which the existence of the density of C is required.

Lemma 2 *Under Assumption 1, $\widetilde{w}$ is bounded from above by $\mathbb{P}(\Lambda = 0)^{-1}$ on* $[0, 1]$.

As a consequence of Lemma 2, Assumption 1 is not only sufficient to obtain existence of the weights, but it also guarantees that they are bounded. In virtue of Lemma 1, this guarantees consistency and asymptotic normality of the IS estimator.

Note that our approach could be generalized to other forms of C_λ and F_Λ (e.g., not necessarily discrete). In such cases the evaluation of the weight function $\widetilde{w}$ might be more demanding and require the use of numerical integration schemes.

4 Importance Sampling Algorithm for Archimedean Copulas

While the IS method from the previous section can be applied to any copula, sampling from C_λ is in general difficult. A possible solution could be to use rejection sampling, but we do not pursue this approach here as we expect it would not work very well with QMC sampling. In this section, we instead focus on developing sampling algorithms for $\boldsymbol{U} \mid U_{(d)} := \max\{U_1, \ldots, U_d\} > \lambda$ when $\boldsymbol{U}$ follows an Archimedean copula with generator ψ. This corresponds to Step 3 in Algorithm 1. In light of (5), we have $(U_1, \ldots, U_d) \overset{\text{d}}{=} (\psi(\frac{E_1}{V}), \ldots, \psi(\frac{E_d}{V}))$ where $E_j \overset{\text{ind.}}{\sim} \text{Exp}(1)$ and V is the corresponding frailty random variable. The condition $U_{(d)} > \lambda$ can then be written as $E_{(1)} < \psi^{-1}(\lambda)V$, where $E_{(1)} \sim \text{Exp}(d)$ is the first order statistic of $\{E_1, \ldots, E_d\}$. In summary, sampling from $\boldsymbol{U} \mid U_{(d)} > \lambda$ is equivalent to sampling from $(E_1, \ldots, E_d, V) \mid E_{(1)} < \psi^{-1}(\lambda)V$. Algorithm 2 summarizes the sampling method for this conditional distribution where we let $\gamma = \psi^{-1}(\lambda)$. Proposition 1 asserts that samples from Algorithm 2 have the right distribution.

Proposition 1 *Let $E_1, \ldots, E_d$ be iid positive random variables and V be a positive random variable independent of the E_j's. Then a sample $(E_1, \ldots, E_d, V)$ constructed as in Steps 1–3 of Algorithm 2 has the distribution $(E_1, \ldots, E_d, V) \mid (E_{(1)} < \gamma V)$.*

While Proposition 1 holds for general (positive) E_j's and V, we now give detailed explanations of how to do the sampling for Steps 1 and 2 of Algorithm 2, i.e., when $E_j \overset{\text{ind.}}{\sim} \text{Exp}(1)$ and V is the frailty random variable. We assume that V is continuous for the derivations below. We need only minor modifications for the discrete case.

Step 1: Sample $(E_{(1)}, V) \mid (E_{(1)} < \gamma V)$

We want to sample from the joint distribution of $(E_{(1)}, V)$ conditioned on the event $(E_{(1)} < \gamma V)$. Let $f_{E_{(1)}}(x)$ and $f_V(v)$ be the density of $E_{(1)}$ and V, respectively. Further, let $f_{(E_{(1)},V)|(E_{(1)}<\gamma V)}(x, v)$ be the conditional joint density of $(E_{(1)}, V)$ given $E_{(1)} < \gamma V$. Then by independence of $E_{(1)}$ and V

$$f_{(E_{(1)},V)|(E_{(1)}<\gamma V)}(x, v) = \beta f_{E_{(1)}}(x) f_V(v) I(x < \gamma v), \tag{10}$$

Algorithm 2 Sampling step of the IS algorithm for Archimedean copulas

```
Require: $0 < \gamma = \psi^{-1}(\lambda) < \infty$.
1: Draw $(E_{(1)}, V) \mid (E_{(1)} < \gamma V)$.
2: Draw $(E_1, \ldots, E_d) \mid E_{(1)}$.
3: Let $U_j = \psi(E_j/V)$ for $j \in \{1, \ldots, d\}$.
4: return $(U_1, \ldots, U_d)$.
```

where $1/\beta = \mathbb{P}(E_{(1)} < \gamma V) = \mathbb{P}(U_{(d)} > \lambda) = (1 - C(\lambda\mathbf{1})) = (1 - \psi(d\psi^{-1}(\lambda)))$. We use conditional sampling to sample from this density, that is, we first sample V from the marginal conditional density $f_{V|(E_{(1)}<\gamma V)}$ of (10) then draw $E_{(1)}$ from (10) given V. Note that

$$f_{V|(E_{(1)}<\gamma V)}(v) = \beta f_V(v) \int_0^{\gamma v} f_{E_{(1)}}(x)\,dx = \beta f_V(v)(1 - \exp(-d\gamma v)). \tag{11}$$

Unfortunately, the density (11) does not belong to a known parametric family for most Archimedean copulas. Nonetheless, there exist efficient numerical algorithms that allow one to sample from a univariate distribution given its probability density function. For instance, the NINIGL Algorithm in [4] achieves this through numerical inversion techniques. Such algorithms could become costly if they had to be applied for several values of Λ. However in our numerical experiments, the threshold random variable Λ only takes a small number of distinct values, such as 10, which is much less than the number of simulations, which is of order 10,000. Furthermore for each value of $\Lambda = \lambda$, we sample from (11) thousands of times, which makes the overhead required to initialize the sampling algorithms negligible.

After sampling V from (11), we want to draw $E_{(1)}$ given V. Let the conditional density of $E_{(1)}$ be denoted by $f_{E_{(1)}|(E_{(1)}<\gamma V,V)}(x \mid V)$. Then

$$f_{E_{(1)}|(E_{(1)}<\gamma V)}(x \mid V) = \frac{d\exp(-dx)I(x < \gamma V)}{1 - \exp(-d\gamma V)}$$

and we can draw a sample from this density using the inversion technique. In particular, we generate $U \sim \mathrm{U}[0,1)$ and then let $E_{(1)} = -\frac{1}{d}\log(1 - U(1 - e^{-\gamma dV}))$.

Step 2: Sampling $(E_1, \ldots, E_d) \mid E_{(1)}$

Suppose we have drawn $E_{(1)} = x_{(1)}$ from Step 1. Let $f(x_1, \ldots, x_d) = \exp\left(-\sum_{i=1}^{d} x_i\right)$ be the joint density of $(E_1, \ldots, E_d)$. Note that each E_j is as likely to be the minimum. Consider the case where E_1 is the minimum. The conditional distribution is

$$\begin{aligned} f(x_1, \ldots, x_d \mid E_1 = E_{(1)},\ E_{(1)} = x_{(1)}) &= \frac{e^{-x_{(1)} - \sum_{j=2}^{d} x_j}}{(1/d)de^{-dx_{(1)}}} \\ &= e^{-\sum_{j=2}^{d}(x_j - x_{(1)})} \cdot I_{\{E_1 = x_{(1)}\}}. \end{aligned} \tag{12}$$

We can sample from (12) by letting $E_j = \mathrm{Exp}(1) + x_{(1)}$ independently for $j \in \{2, \ldots, d\}$.

Since any of the E_j's can be the minimum, we pick the index for the minimum component randomly from 1 to d and sample the rest of the components accordingly. This sampling method works for MC, but may not work very well for QMC. When randomly choosing the index for the minimum component, we potentially destroy the structure of the LDS. So, if we are working with an LDS, the CDM based on Proposition 2 below is preferred.

Proposition 2 *Suppose* $E_1, \dots, E_d$ *are iid* Exp(1). *Then*

$$\mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \dots, E_{k-1} = x_{k-1}, E_{(1)} = x)$$
$$= \begin{cases} 1 - \exp\{-(x_k - x)\}, & \textit{if } x_j = x \textit{ for some } j \in \{1, \dots, k-1\}, \\ \frac{1}{d-k+1} I_{\{x_k < x\}} + \frac{d-k}{d-k+1}(1 - \exp\{-(x_k - x)\}), & \textit{otherwise}. \end{cases} \tag{13}$$

To sample $E_1, \dots, E_d$, we let k take the successive values $k \in \{1, \dots, d\}$ in (13) and proceed by inversion.

4.1 Stratified Sampling Alternative to Importance Sampling

Recall from Algorithm 1 and the form of C_λ given in (8) that our IS scheme starts with sampling a threshold random variable $\Lambda = \lambda_k$ and then proceeds by sampling $\tilde{\boldsymbol{U}} \sim \boldsymbol{U} \,|\, (T = \max\{U_1, \dots, U_d\} > \lambda_k)$. Instead, we can construct a stratified sampling (SS) estimator based on the samples from $\boldsymbol{U} \,|\, (\lambda_{k+1} > T \ge \lambda_k)$. That is, we stratify the domain of $\boldsymbol{U}$ along with T. Suppose Λ takes M distinct values as $0 = \lambda_1 < \cdots < \lambda_M < 1$. Let $\lambda_{M+1} = 1$ for convenience. Then we can define M strata as

$$\Omega_k = \{\boldsymbol{u} \in [0, \lambda_{k+1})^d \,|\, \boldsymbol{u} \notin [0, \lambda_k)^d\}, \quad k = 1, \dots, M. \tag{14}$$

By construction, $\lambda_k \le T < \lambda_{k+1}$ if and only if $\boldsymbol{u} \in \Omega_k$. We can then construct the SS estimator

$$\hat{\mu}_{\mathrm{SS},n} = \sum_{k=1}^{M} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \Psi(\tilde{\boldsymbol{U}}_i^{(k)}), \tag{15}$$

where p_k is the stratum probability, n_k is the number of samples allocated to the stratum Ω_k, and $\tilde{\boldsymbol{U}}_i^{(k)} \overset{\text{ind.}}{\sim} \boldsymbol{U} \,|\, \Omega_k$. For Archimedean copulas, $p_k = \psi(d\psi^{-1}(\lambda_{k+1})) - \psi(d\psi^{-1}(\lambda_k))$. In Sect. 5, we show that the SS estimator has a smaller variance than the IS estimator. It is easy to show that sampling from Ω_k is equivalent to sampling from $(E_1, \dots, E_d, V) \,|\, \psi^{-1}(\lambda_{k+1})V < E_{(1)} \le \psi^{-1}(\lambda_k)V$. Let $\gamma_k = \psi^{-1}(\lambda_k)$ for all $k \in \{1, \dots, M+1\}$. Algorithm 3 summarizes the procedure to sample from each stratum.

In this algorithm, Step 2 is exactly the same as for the IS case (Algorithm 2). For Step 1, we use conditional sampling to draw samples from the joint conditional density of $(E_{(1)}, V) \,|\, (\gamma_{k+1}V < E_{(1)} \le \gamma_k V)$. By using an argument similar to the one used for Step 1 of Algorithm 2, we can show that the marginal conditional density of V is

$$f_{V|(E_{(1)}<\gamma V)}(v) = \beta f_V(v)(\exp(-d\gamma_{k+1}v) - \exp(-d\gamma_k v)), \tag{16}$$

Algorithm 3 Sampling $\boldsymbol{U}_{k,j}$ in SS algorithm for Archimedean copulas

```
Require: $0 < \gamma_{k+1} < \gamma_k < \infty$.
1: Draw $(E_{(1)}, V) \mid (\gamma_{k+1} V < E_{(1)} \leq \gamma_k V)$.
2: Draw $(E_1, \dots, E_d) \mid E_{(1)}$.
3: Let $U_j = \psi(E_j/V)$ for $j \in \{1, \dots, d\}$.
4: return $(U_1, \dots, U_d)$.
```

where $f_V(v)$ is the density of V and $\beta = 1/p_k = 1/\psi[d\psi^{-1}(\lambda_{k+1})) - \psi(d\psi^{-1}(\lambda_k)]$. Conditional on V drawn from (16), generate $U \sim \mathrm{U}[0,1)$ and then let $E_{(1)} = -\frac{1}{d}\log\left[e^{-\gamma_{k+1}dy} - U(e^{-\gamma_{k+1}dy} - e^{-\gamma_k dy})\right]$. Then $(E_{(1)}, V)$ follows the desired distribution.

Remark 1 We can follow Algorithm 3 to sample from the SS distribution under QMC, if the number of samples to be drawn is fixed. In some cases, however, we want to keep running simulations until some error criterion is met. Since SS requires to have a subset of points allocated to each stratum, combining it with QMC for n not fixed is challenging. This is because when the total sample size is increased by successive increments, it means possibly disjoint subsets of a QMC point set will be used in a given stratum, which is undesirable. Whether or not this allocation over successive increments can be done in a clever way that exploits the uniformity of low-discrepancy sequences is a question we leave for future research.

5 Variance Analysis and Calibration Method

In this section, we analyze the variance of the IS and SS estimators and then propose calibration methods for choosing the q_k's designed to minimize the variance of the respective estimators. We also show that the SS scheme is more flexible to calibrate and gives an estimate with a smaller variance than the IS estimator.

We define the strata $\Omega_1, \dots, \Omega_M$ as in (14) and $C_k = C(\lambda_k \mathbf{1})$ for $k \in \{1, \dots, M\}$. The following proposition gives the variance of the IS estimator.

Proposition 3 *Let $\hat{\mu}_{\mathrm{IS},n}$ be the IS estimator described in Algorithm 1 with C_λ given in (8). Then its variance is given by*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\left(\sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{q_l}{1 - C_l}\right)^{-1} \mu_k^{(2)} - \mu^2\right), \tag{17}$$

where $p_k = \mathbb{P}(\boldsymbol{U} \in \Omega_k)$, $q_k = \mathbb{P}(\Lambda = \lambda_k)$ and $\mu_k^{(2)} = \mathbb{E}(\Psi^2(\boldsymbol{U}) \mid \Omega_k)$.

For the optimal calibration, we want to choose the q_k's so that (17) is minimized. The following proposition gives an analytical expression for the optimal calibration. For convenience, we define $\mu_0^{(2)} = 0$.

Proposition 4 *The set of q_k's that minimize* (17) *under the condition $\mu_1^{(2)} \leq \cdots \leq \mu_M^{(2)}$ with $\mu_0^{(2)} = 0$ for convenience, is*

$$q_k^{\text{opt}} = \frac{(1-C_k)\left(\sqrt{\mu_k^{(2)}} - \sqrt{\mu_{k-1}^{(2)}}\right)}{\sum_{k=1}^{M}(1-C_k)\left(\sqrt{\mu_k^{(2)}} - \sqrt{\mu_{k-1}^{(2)}}\right)}, \quad k \in \{1, \ldots, M\}. \tag{18}$$

Remark 2 If the condition $\mu_1^{(2)} \leq \cdots < \mu_M^{(2)}$ is not met, some of the q_k^{opt}'s given by (18) will be negative, which makes the IS scheme infeasible. Note that $q_k^{\text{opt}} < 0$ means that ever having the event $[\Lambda = \lambda_k]$ makes the overall variance greater than when $q_k^{\text{opt}} = 0$. We propose to then remove λ_k from the support of Λ if $q_k^{\text{opt}} < 0$. Accordingly, the strata Ω_k's will change so the stratum second moments need to be recomputed for the optimal calibration.

Of course, we do not know the true values of the $\mu_k^{(2)}$'s in practice, so we have to replace them with estimates. As often done for Neyman allocation, we can first run a pilot study with a small number of simulations and estimate the $\mu_k^{(2)}$'s. The condition $\mu_1^{(2)} \leq \cdots < \mu_M^{(2)}$ means that the outer strata must have greater stratum second moments than the inner strata. We refer to this condition as *increasing second moment (ISM)* condition. Whether this ISM condition is met depends on the problem at hand. In this paper, we specifically work with $\Psi(\boldsymbol{U})$ which is large when at least one component of $\boldsymbol{U}$ is large. This assumption on Ψ and the ISM condition are not incompatible, although there is no guarantee that the former implies the latter. If the ISM is satisfied, then we can substitute (18) into (17) and obtain

$$\text{Var}(\hat{\mu}_{\text{IS},n}^{\text{opt}}) = \frac{1}{n}\left(\left(\sum_{k=1}^{M} p_k \sqrt{\mu_k^{(2)}}\right)^2 - \mu^2\right). \tag{19}$$

Using the Cauchy-Schwarz inequality, we can show that

$$\text{Var}_{\mathbb{Q}}(\hat{\mu}_{\text{IS},n}^{\text{opt}}) = \frac{1}{n}\left(\left(\sum_{k=1}^{M} p_k \sqrt{\mu_k^{(2)}}\right)^2 - \mu^2\right) \leq \frac{1}{n}\left(\sum_{k=1}^{M} p_k \mu_k^{(2)} - \mu^2\right) = \text{Var}(\hat{\mu}_{\text{MC},n}).$$

Equality holds only when $\mu_k^{(2)}$ is the same for all k. Except for this restrictive case, the IS estimator with the optimal choice of q_k's always has a smaller variance than the plain MC counterpart. If the ISM is not met, there is no analytical form for the optimal q_k's. We can still find the optimal values using widely available convex optimization solvers in this case. If we let $q_1 = 1$ and $q_k = 0$ for $k = 2, \ldots, M$, the proposal distribution is the same as the original distribution. That is, IS becomes plain MC. Hence, if we choose the q_k's appropriately, the IS estimator cannot do worse than plain MC. In this sense, the IS estimator is similar to an SS estimator.

Now that we have derived the variance expression and the optimal choice of q_k's, we move on to the stratified sampling estimator (15). Using simple algebra, one can show that $\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) = \sum_{k=1}^{M} p_k^2 \sigma_k^2 / n_k$, where $\sigma_k^2 = \mathrm{Var}(\Psi(\boldsymbol{U}) \mid \Omega_k)$, $k \in \{1, \dots, M\}$ are the stratum variances. The optimal n_k's are given by Neyman allocation

$$n_k = \frac{n p_k \sigma_k}{\sum_{k=1}^{M} p_k \sigma_k}. \tag{20}$$

Unlike the IS estimator, there is no restriction on this optimal allocation, i.e., we do not need σ_k to be increasing with k. In this sense, the SS estimator is more flexible.

Since the true strata variances are unknown, we have to replace them with estimates. Investigating the optimal calibration formula for IS (18) and SS (20), it appears that the estimation error of the strata moments (the $\mu_k^{(2)}$'s for IS and the σ_k^2's for SS) has greater impact on the estimated calibration for IS than for SS. Since q_k for IS depends on $\sqrt{\mu_k^{(2)}} - \sqrt{\mu_{k-1}^{(2)}}$, the estimation error comes from both estimating $\mu_{k-1}^{(2)}$ and $\mu_k^{(2)}$. On the other hand, for SS, n_k depends on σ_k, so the estimation error comes from estimating σ_k^2 alone. Consequently, the approximation is likely to deviate more from the actual optimal calibration for IS than for SS.

Going back to IS and as discussed in [9], instead of choosing $\Lambda = \lambda_k$ with probability q_k it is more efficient to stratify Λ. That is, take $n_k = nq_k$ observations with $\Lambda = \lambda_k$. Let $\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}}$ denote such a stratified IS estimator. Generally nq_k is not an integer and needs to be rounded. If each n_k is large enough, this rounding effect is negligible. The following proposition compares the variance of the three estimators.

Proposition 5 *Suppose we have an IS estimator with* $\mathbb{P}(\Lambda = \lambda_k) = q_k, 1 \le k \le M$. *If the* $\mu_k = \mathbb{E}(\Psi(\boldsymbol{U}) | \Omega_k)$ *are not all equal and* n *is large enough, then there exists some strata sample allocation* $(n_1, \dots, n_M)$ *for the SS estimator such that* $\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) \le \mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}}) \le \mathrm{Var}(\hat{\mu}_{\mathrm{IS},n})$.

This result trivially holds when we use the optimal q_k's (18) for stratified and unstratified IS and use the optimal allocation (20) for SS. Since the SS estimator is more flexible for calibration and it has a smaller variance than both IS estimators, the SS approach is preferred if the sampling efforts for (11) and (16) are not significantly different. Nonetheless, depending on the type of the underlying copula, sampling from the IS distribution could be much easier than sampling from the SS distribution.

6 Numerical Examples

In this section, we investigate the efficiency of the IS and SS estimators introduced in this paper. We consider the valuation of tail-related quantities of a portfolio consisting of stocks from companies in the financial industry listed on the S&P 100. The five stocks in the portfolio are AIG, Allstate Corp., American Express Inc., Bank of New York and Citigroup Inc. Their stock symbols are AIG, ALL, AXP, BK and C, respectively. We assume that the value of the portfolio is 100 and that all the portfolio weights are equal. The data are daily negative log-returns of these five companies from 2010-01-01 to 2016-04-01 (1571 data points). We fit GARCH(1,1)-models with t-innovations to each return series to filter out the volatility clustering effect using the R package “rugarch” [6]. The fitted standardized residuals do not exactly follow a t-distribution, so we fit a semi-parametric distribution to the residuals using the R package “spd” [7]. The fitted model uses a kernel density estimate for the centre of the distribution and fits a generalized Pareto distribution to the tails. The use of generalized Pareto distribution to model the GARCH filtered residuals to estimate tail-related risk measures in a univariate setting is studied by McNeil and Frey [15]. We let $S = \sum_{j=1}^{d} X_j$ denote the portfolio loss over a 1 day period with

$$X_j = 100\omega_j \left(1 - \sum_{j=1}^{d} \exp(a_j - b_j \tilde{F}_j^{-1}(U_j)) \right),$$

where d is the number of assets, ω_j’s are the portfolio weights, a_j’s are the means of the log-returns, b_j’s are the fitted conditional standard deviations from the GARCH(1,1) model, $\tilde{F}_j$’s are the fitted semi-parametric distributions from the R package “spd”, and $(U_1, \ldots, U_d)$ follows the fitted copula. We use the R package “distr” [18] to sample from (11) and (16).

Using the R package “copula” [11], we fit the Gumbel, Frank, Clayton and Joe copulas to the standardized residuals based on MLE. Among the four Archimedean copulas, the Gumbel copula with $\theta = 1.604$ gives the best fit in terms of log-likelihood, followed by a Frank copula with $\theta = 4.06$. Hence we proceed assuming that the model we consider is well approximated by a Gumbel or a Frank copula.

The three functionals we estimate are stop loss $\mathbb{E}(\max\{S - D, 0\})$ with $D = 3$ for Gumbel and $D = 2$ for Frank, $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.99}$ of S. To define C_Λ, we use $\lambda_k = 1 - \left(\frac{1}{2}\right)^{k-1}$ for $k = 1, \ldots, M$, with $M = 10$. When constructing an IS estimator, we stratify Λ regardless of whether we use MC or QMC. When we calibrate the q_k’s for IS according to (18) and SS according to (20), we use ES as our objective function.

Table 2 shows the estimates, variance reduction factors and computational times for the three functionals for the five different estimators based on Gumbel and Frank copulas, respectively. We used 30 randomizations to estimate the variance of each estimator (MC and QMC). The estimates shown are based on SS estimators with QMC. Variance reduction factors are defined to be the ratios of the variance of the plain MC estimators over the variance of the estimators with the respective VRTs. The last row of Table 2 shows the increase in computation time compared to plain MC. We see that both IS and SS reduce the variance by large amounts and this is amplified when combined with QMC. Note that SS estimators generally give smaller variances than the IS estimators, as suggested by Proposition 5. For IS and SS estimators with and without QMC, we see that the largest variance reduction factors are for ES. This makes sense as we calibrate the q_k’s to minimize the variance of the ES estimator.

We also repeat the same experiment but with a portfolio of 20 stocks from large companies in the financial industry traded on NYSE (the full list is available from the authors); the results are displayed under $d = 20$ in Table 2. Figures 2 and 3 show the log-variance of the three different MC-based estimators for different n.

Table 2 Estimates and variance reduction factors for the Gumbel and Frank copulas based on $n = 30{,}000$, $d = 5$

		Gumbel						Frank					
			MC		QMC				MC		QMC		
Objective function	d	Estimate	IS	SS	Plain	IS	SS	Estimate	IS	SS	Plain	IS	SS
$\mathbb{E}(\max\{S-D,0\})$	5	0.012	67	168	33	1730	8085	0.011	6.4	11	14	85	161
	20	0.010	49	40	51	1128	3488	0.0034	4.6	4.1	5.7	46	46
$\text{VaR}_{0.99}(S)$	5	3.2	10	26	8.4	39	98	2.4	9.7	9.0	2.6	32	26
	20	3.04	7.9	7.2	5.8	19	28	2.1	4.3	4.8	3.6	16	19
$\text{ES}_{0.99}(S)$	5	4.2	89	175	29	6019	16,989	2.8	17	21	7.1	250	373
	20	4.03	49	39	48	1296	4205	2.3	4.6	3.8	4.0	38	36
Run time	5		3.6	3.7	1.8	3.7	3.8		3.6	3.7	1.1	3.6	3.7
	20		2.0	1.9	1.2	1.9	2.0		1.7	1.8	1.1	1.9	2.0

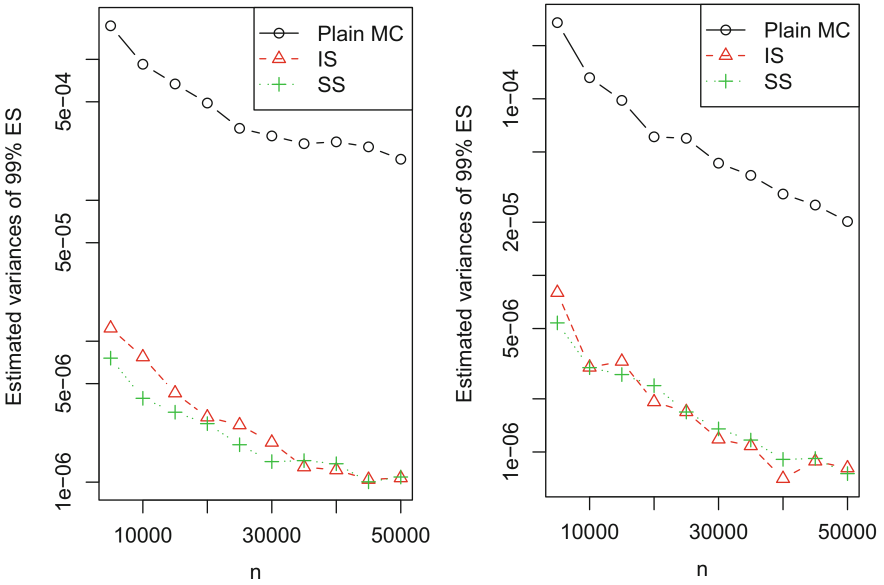


Fig. 2 Estimated variances of plain MC, IS and SIS estimators of $\mathrm{ES}_{0.99}$ for a Gumbel copula (left-hand side) and a Frank copula (right-hand side) for different n and for $d = 5$

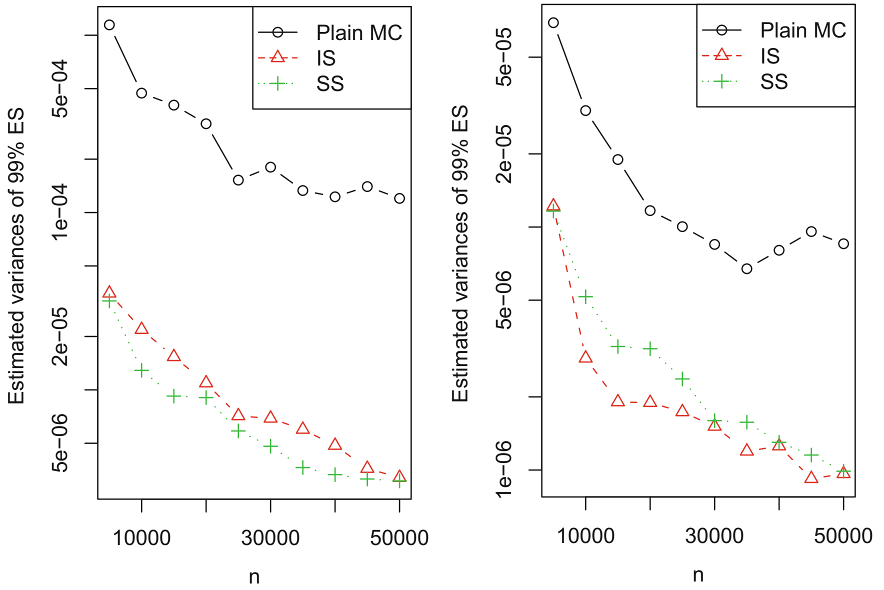


Fig. 3 Estimated variances of plain MC, IS and SIS estimators of 99% ES for a Gumbel copula (left-hand side) and a Frank copula (right-hand side) for different n and for $d = 5$

Acknowledgements We thank the anonymous referees for their helpful comments. The third and fourth author gratefully acknowledge the financial support of NSERC Canada through grant numbers #5010 and #238959, respectively.

Appendix

Proof of Lemma 1 Since $\mathbb{E}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}})) = \mathbb{E}(\Psi(\boldsymbol{U}))$, consistency follows directly from the Strong Law of Large Numbers. Note that $\mathbb{E}\left(\Psi(\tilde{\boldsymbol{U}})^2 w(\tilde{\boldsymbol{U}})^2\right) = \mathbb{E}\left(\Psi(\boldsymbol{U})^2 w(\boldsymbol{U})\right) \leq \mathbb{E}\left(\Psi(\boldsymbol{U})^2\right) B < \infty$, where the first equality is justified by a change of measure. We can immediately deduce asymptotic normality of $\hat{\mu}_{\mathrm{IS},n}$ by the Central Limit Theorem, see, for example, Section 2.4 in [5, p. 110]. □

Proof of Theorem 1 Due to Leibniz' integral rule, $dG(\boldsymbol{u}) = \int_0^1 dC_\lambda(\boldsymbol{u})dF_\Lambda(\lambda)$. From the definition of C_λ, we can derive the differential

$$dC_\lambda(\boldsymbol{u}) = \begin{cases} 0, & \boldsymbol{u} \in [0,\lambda]^d, \\ \frac{dC(\boldsymbol{u})}{1-C(\lambda\mathbf{1})}, & \text{otherwise.} \end{cases}$$

Using both identities, we obtain

$$dG(\boldsymbol{u}) = dC(\boldsymbol{u}) \int_0^1 \frac{I_{\{\lambda \leq \max\{u_1,\dots,u_d\}\}}}{1 - C(\lambda\mathbf{1})}\, dF_\Lambda(\lambda),$$

leading to the desired result. □

Proof of Lemma 2 Since $C(\lambda\mathbf{1}), \lambda \in [0,1]$, the diagonal section of the copula C and the distribution function F_Λ are both increasing functions. The weight function $\widetilde{w}$ is thus decreasing on $[0,1]$, and is bounded above by $\widetilde{w}(0) = \mathbb{P}(\Lambda = 0)^{-1} < \infty$. □

Proof of Proposition 1 We sample $(E_1,\dots,E_d,V) \mid (E_{(1)} < \gamma V)$ using conditional distribution sampling. That is, we first sample $(E_{(1)},V) \mid (E_{(1)} < \gamma V)$, which is the Step 1 of Algorithm 2. Given the $(E_{(1)},V)$ drawn, we then want to sample $(E_1,\dots,E_d) \mid (E_{(1)} < \gamma V, E_{(1)}, V)$ which is equivalent to sampling $(E_1,\dots,E_d) \mid E_{(1)}$ and this is the Step 2 of the algorithm. □

Proof of Proposition 2 First, consider the case where $x_j = x$ for some $j = 1,\dots,k-1$. Without loss of generality assume that $x_1 = x$, i.e., $E_1 = E_{(1)}$. So we want to find $\mathbb{P}(E_k \leq x_k \mid E_1 = x_1,\dots,E_{k-1} = x_{k-1}, E_{(1)} = E_1 = x)$. From (12), the conditional distribution of E_k is $x + \mathrm{Exp}(1)$. So the above probability equals

$$\mathbb{P}(E_k \leq x_k \mid E_1 = x_1,\dots,E_{k-1} = x_{k-1}, E_{(1)} = x) = 1 - e^{-(x_k - x)}. \tag{21}$$

Next, we consider the case $x_j \neq x$ for all $j = 1,\dots,k-1$. This means that $E_j = E_{(1)}$ for some $j = k,\dots,d$. Since all E_j are iid, there is a $\frac{1}{d-k+1}$ probability that $E_k = E_{(1)}$. In such a case $E_k = x$ with probability 1 as we are given $E_{(1)} = x$. Suppose $E_k \neq E_{(1)}$, which occurs with probability of $\frac{d-k}{d-k+1}$. Then we need to find the probability

$$\mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \ldots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_j \neq E_{(1)}, j = 1, \ldots k)$$
$$= \sum_{j=k+1}^{d} \frac{1}{d-k} \mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \ldots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_j = E_{(1)})$$
$$= \mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \ldots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_d = E_{(1)}) = 1 - e^{-(x_k - x)}.$$

The last equality again holds by (12) and the result follows. □

Proof of Proposition 3 Recall that the IS estimator (7) is

$$\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\tilde{\boldsymbol{U}}_i) w(\tilde{\boldsymbol{U}}_i) = \frac{1}{n}\sum_{i=1}^{n} \Psi(\tilde{\boldsymbol{U}}_i)\tilde{w}(t_i), \tag{22}$$

where $t_i = \max(\tilde{\boldsymbol{U}}_{i,1}, \ldots, \tilde{\boldsymbol{U}}_{i,d})$, and where the weight function (9) is

$$\widetilde{w}(u) = \Big(\sum_{k=1}^{M} \frac{I_{\{\lambda_k \le u\}}}{1 - C_k} q_k\Big)^{-1}. \tag{23}$$

Hence $\widetilde{w}$ is constant over each stratum Ω_k. Thus, for $\boldsymbol{u} \in \Omega_k$, we can define the stratum weight as

$$w_k = \Big(\sum_{l=1}^{k} \frac{q_l}{1 - C_l}\Big)^{-1}, \quad k \in \{1, \ldots, M\}. \tag{24}$$

The second moment of $w(\tilde{\boldsymbol{U}})\Psi(\tilde{\boldsymbol{U}})$ is

$$\mathbb{E}(w^2(\tilde{\boldsymbol{U}})\Psi^2(\tilde{\boldsymbol{U}})) = \mathbb{E}(w(\boldsymbol{U})\Psi^2(\boldsymbol{U})) = \sum_{k=1}^{M} p_k \mathbb{E}(w(\boldsymbol{U})\Psi^2(\boldsymbol{U}) \,|\, \boldsymbol{U} \in \Omega_k)$$
$$= \sum_{k=1}^{M} p_k w_k \mathbb{E}(\Psi^2(\boldsymbol{U}) \,|\, \boldsymbol{U} \in \Omega_k) = \sum_{k=1}^{M} p_k w_k \mu_k^{(2)} = \sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{1}{1 - C_l} q_l\right)^{-1} \mu_k^{(2)}.$$

The third equality holds because the weight function $\tilde{w}(t)$ is constant over each stratum. The last equality follows from (24). Then the variance of the IS estimator based on n samples is $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\left(\sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{1}{1-C_l} q_l \right)^{-1} \mu_k^{(2)} - \mu^2 \right)$. □

Proof of Proposition 4 Since the variance expression (17) is convex in q_k's, we can solve the minimization problem using Lagrange multipliers. First, we simplify (17) so that the minimization problem becomes easier. Let $\tilde{p}_k = \mathbb{P}(\tilde{\boldsymbol{U}} \in \Omega_k)$, the stratum probability under the proposal distribution. Observe that

$$\begin{aligned}\tilde{p}_k &= \sum_{l=1}^{M} q_l \cdot \mathbb{P}(\tilde{\boldsymbol{U}} \in \Omega_k \mid \Lambda = \lambda_l) = \sum_{l=1}^{k} q_l \cdot \mathbb{P}(\boldsymbol{U} \in \Omega_k \mid \max(U_1, \ldots, U_d) > \lambda_l) \\ &= \sum_{l=1}^{k} q_l \frac{p_k}{1 - C_l} = p_k \sum_{l=1}^{k} \frac{q_l}{1 - C_l}. \qquad (25)\end{aligned}$$

By (23) and (25), we can write $w_k = p_k/\tilde{p}_k$. The weight w_k is the ratio of probabilities of a sample falling onto stratum Ω_k under the original distribution and the proposal distribution. Plugging this expression into (17), we have

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\left(\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k} \mu_k^{(2)} - \mu^2\right). \qquad (26)$$

Using the Lagrange multiplier method, we can show that the optimal $\tilde{p}_k$ is

$$\tilde{p}_k^{\mathrm{opt}} = p_k \sqrt{\mu_k^{(2)}} \Big/ \sum_{k=1}^{M} p_k \sqrt{\mu_k^{(2)}}. \qquad (27)$$

Note that this optimal choice of $\tilde{p}_k$'s resembles the Neyman allocation, the optimal allocation under stratified sampling.

Using the relation $q_k = (1 - C_k)\left(\frac{\tilde{p}_k}{p_k} - \frac{\tilde{p}_{k-1}}{p_{k-1}}\right)$, (with $\tilde{p}_0/p_0 = 0$) the optimal q_k is

$$q_k^{\mathrm{opt}} \propto (1 - C_k)\left(\sqrt{\mu_k^{(2)}} - \sqrt{\mu_{k-1}^{(2)}}\right), (\text{with } \mu_0^{(2)} = 0). \qquad (28)$$

The assumption that $\mu_1^{(2)} \leq \cdots \leq \mu_M^{(2)}$ ensures that $q_k^{\mathrm{opt}} \geq 0$ for $k = 1, \ldots, M$. □

Proof of Proposition 5 We have $\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}} = \frac{1}{n}\sum_{k=1}^{M}\sum_{j=1}^{nq_k} \Psi(\tilde{\boldsymbol{U}}_{ki}) w(\tilde{\boldsymbol{U}}_{ki})$, $\tilde{\boldsymbol{U}}_{ki} \overset{iid}{\sim} \boldsymbol{U}|\Lambda = \lambda_k$. Thus $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}}) = \mathbb{E}\left[\mathrm{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \mid \Lambda)\right]/n + O(1/n^2)$ (term due to rounding nq_k). Since $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\mathrm{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}})$, we have $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{IS},n})$ as long as n is large enough for the $O(1/n^2)$ term due to rounding to be smaller than $\mathrm{Var}(\mathbb{E}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}})|\Lambda))/n > 0$. As shown before, $\tilde{p}_k = \mathbb{P}(\tilde{\boldsymbol{U}} \in \Omega_k) = p_k \sum_{l=1}^{k} \frac{q_l}{1-C_l}$. Consider an SS estimator with $n_k = n\tilde{p}_k$. Then $\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) = \frac{1}{n}\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k}\sigma_k^2$. Also $\mathrm{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \mid \Lambda = \lambda_k) = \mathrm{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \mid T > \lambda_k) \geq$

$\mathbb{E}[\mathrm{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \mid T > \lambda_k, T \in \Omega_j)] = \sum_{j=k}^{M} \frac{p_j}{1-C_k} w_j^2 \sigma_j^2$. Then, using (24) and $w_k = p_k/\tilde{p}_k$ we get

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}^{\mathrm{det}}) \geq \frac{1}{n}\sum_{k=1}^{M} q_k \sum_{j=k}^{M} \frac{p_j}{1-C_k} w_j^2 \sigma_j^2 = \frac{1}{n}\sum_{k=1}^{M} p_k w_k^2 \sigma_k^2 \sum_{j=1}^{k} \frac{q_j}{1-C_j}$$

$$= \frac{1}{n}\sum_{k=1}^{M} p_k w_k \sigma_k^2 = \frac{1}{n}\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k}\sigma_k^2 = \mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}).$$

□

References

1. Cambou, M., Hofert, M., Lemieux, C.: Quasi-random numbers for copula models. Stat. Comput. **27**(5), 1307–1329 (2017)
2. Chan, J., Kroese, D.: Efficient estimation of large portfolio loss probabilities in t-copula models. Eur. J. Oper. Res. **205**(2), 361–367 (2010)
3. Choe, G., Jang, H.: Efficient algorithms for basket default swap pricing with multivariate Archimedean copulas. Insurance: Math. Econ. **48**(2), 205–213 (2011)
4. Derflinger, G., Hörmann, W., Leydold, J.: Random variate generation by numerical inversion when only the density is known. ACM Trans. Model. Comput. Simul. **20**(4), 1–25 (2010)
5. Durrett, R.: Probability: Theory and Examples, 4th edn. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2010)
6. Ghalanos, A.: rugarch: Univariate GARCH models (2015). R package version 1.3-6
7. Ghalanos, A.: spd: Semi-Parametric Distribution (2015). R package version 2.0-1
8. Glasserman, P., Li, J.: Importance sampling for portfolio credit risk. Manag. Sci. **51**(11), 1643–1656 (2005)
9. Hesterberg, T.: Weighted average importance sampling and defensive mixture distributions. Technometrics **37**(2), 185–194 (1995)
10. Hofert, M., Mächler, M., et al.: Nested Archimedean copulas meet R: The nacopula package. J. Stat. Softw. **39**(9), 1–20 (2011)
11. Hofert, M., Kojadinovic, I., Maechler, M., Yan, J.: copula: Multivariate Dependence with Copulas (2016). R package version 0.999-15
12. Huang, P., Subramanian, D., Xu, J.: An importance sampling method for portfolio CVaR estimation with Gaussian copula models. In: Proceedings of the 2010 Winter Simulation Conference (WSC), pp. 2790–2800 (2010)
13. Lemieux, C.: Monte Carlo and Quasi-Monte-Carlo Sampling. Springer, New York (2009)
14. Marshall, A., Olkin, I.: Families of multivariate distributions. J. Am. Stat. Assoc. **83**(403), 834–841 (1988)
15. McNeil, A.J., Frey, R.: Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. J. Empir. Financ. **7**(3), 271–300 (2000)
16. McNeil, A., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques, Tools. Princeton University Press, Princeton (2005)
17. Nelsen, R.: An Introduction to Copulas, 2nd edn. Springer, New York (2006)
18. Ruckdeschel, P., Kohl, M., Stabla, T., Camphausen, F.: S4 classes for distributions. R News **6**(2), 2–6 (2006)
19. Sak, H., Hörmann, W., Leydold, J.: Efficient risk simulations for linear asset portfolios in the t-copula model. Eur. J. Oper. Res. **202**(3), 802–809 (2010)

20. Sobol, I.: On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput. Math. Math. Phys. **7**(4), 86–112 (1967)
21. Tasche, D.: Capital allocation to business units and sub-portfolios: the Euler principle. In: Resti, A. (ed.) Pillar II in the New Basel Accord: The Challenge of Economic Capital, pp. 423–453. Risk Books, London (2008)

A Spectral Method for the Biharmonic Equation

Kendall Atkinson, David Chien, and Olaf Hansen

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Let $\varOmega$ be an open, simply connected, and bounded region in $\mathbb{R}^d$, $d \geq 2$, with a smooth boundary $\partial\varOmega$ that is homeomorphic to $\mathbb{S}^{d-1}$. Consider solving $\varDelta^2 u + \gamma u = f$ over $\varOmega$ with zero Dirichlet boundary conditions. A Galerkin method based on a polynomial approximation space is proposed, yielding an approximation u_n. With sufficiently smooth problem parameters, the method is shown to be rapidly convergent. For $u \in C^\infty\left(\overline{\varOmega}\right)$ and assuming $\partial\varOmega$ is a C^∞ boundary, the convergence of $\|u - u_n\|_{H^2(\varOmega)}$ to zero is faster than any power of $1/n$. Numerical examples illustrate experimentally an exponential rate of convergence.

1 Introduction

Consider the biharmonic problem

$$\varDelta^2 u\left(s\right) + \gamma\left(s\right) u\left(s\right) = f\left(s\right), \qquad s \in \varOmega, \tag{1}$$

with the Dirichlet boundary conditions

$$u\left(s\right) = \frac{\partial u\left(s\right)}{\partial n_s} = 0, \qquad s \in \partial\varOmega. \tag{2}$$

K. Atkinson (✉)
The University of Iowa, Iowa City, IA, USA
e-mail: Kendall-Atkinson@uiowa.edu

D. Chien · O. Hansen
California State University San Marcos, San Marcos, CA, USA
e-mail: chien@csusm.edu; ohansen@csusm.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_5

The region $\Omega \subseteq \mathbb{R}^d$, $d \geq 2$, is to be bounded and simply-connected; and its boundary $\partial\Omega$ is to be smooth and homeomorphic to $\mathbb{S}^{d-1}$. Assume $f \in L^2(\Omega)$ and

$$\gamma_{\min} \equiv \min_{s\in\overline{\Omega}} \gamma(s) \geq 0. \tag{3}$$

This can be looked upon as a problem in the Sobolev space $H^4(\Omega)$. It can also be reformulated as a variational problem. For background on the use of this problem in mechanics, see [10], [16, Chap. 8].

Introduce the bilinear functional

$$\mathscr{A}(u,v) = \int_\Omega [\Delta u(s)\,\Delta v(s) + \gamma(s)\,u(s)\,v(s)]\;ds$$

and the linear functional

$$\ell_f(v) \equiv (f,v) = \int_\Omega f(s)\,v(s)\;ds, \qquad v \in L^2(\Omega).$$

Introduce the Hilbert space

$$H_0^2(\Omega) = \left\{ v \in H^2(\Omega) \mid v, \frac{\partial v}{\partial n} = 0, \text{ on } \partial\Omega \right\}.$$

For the norm, use

$$\|v\|_2 \equiv \|v\|_{H^2(\Omega)} = \sqrt{\sum_{|\mathbf{k}|\leq 2} \|D^{\mathbf{k}}v\|^2_{L^2(\Omega)}},$$

where $\mathbf{k} = (k_1, \ldots, k_d)$, $|\mathbf{k}| = k_1 + \cdots + k_d$, and

$$D^{\mathbf{k}}v(s) = \frac{\partial^{|\mathbf{k}|}v(s_1,\ldots,s_d)}{\partial s_1^{k_1}\cdots\partial s_d^{k_d}}.$$

The variational formulation of (1)–(2) is to find $u \in H_0^2(\Omega)$ for which

$$\mathscr{A}(u,v) = \ell_f(v), \qquad \forall v \in H_0^2(\Omega). \tag{4}$$

For a discussion of this reformulation, see Ciarlet [9, p. 28]. With the above assumptions and definitions, $\mathscr{A}$ is a strongly elliptic operator on $H_0^2(\Omega)$,

$$\mathscr{A}(v,v) \geq c_e\|v\|_2^2, \qquad v \in H_0^2(\Omega),$$

with $c_e > 0$. Also, $\mathscr{A}$ is a bounded bilinear operator,

$$|\mathscr{A}(v,w)| \le c_{\mathscr{A}} \|v\|_2 \|w\|_2, \qquad v,w \in H_0^2(\Omega),$$

for some finite $c_{\mathscr{A}} > 0$. Finally,

$$\|\ell_f\| \le \|f\|_{L^2(\Omega)}.$$

The Lax-Milgram Theorem (cf. [1, §8.3], [8, §2.7]) implies the existence of a unique solution u to (4) with

$$\|u\|_2 \le \frac{1}{c_e}\|\ell_f\|. \tag{5}$$

In Sect. 3 we present a Galerkin method for approximating (4), making use of multivariate orthonormal polynomial approximations. Numerical examples are given in Sect. 4.

2 Preliminaries

Assume the existence of an explicitly known continuously differentiable mapping

$$\Phi : \overline{\mathbb{B}}^d \underset{onto}{\xrightarrow{1-1}} \overline{\Omega} \tag{6}$$

and let $\Psi = \Phi^{-1} : \overline{\Omega} \underset{onto}{\xrightarrow{1-1}} \overline{\mathbb{B}}^d$ denote the inverse mapping. A very simple example of such a mapping is when Ω is the ellipse

$$\left(\frac{s_1}{a}\right)^2 + \left(\frac{s_2}{b}\right)^2 \le 1$$

with $a, b > 0$. Choose

$$\Phi(x) = (ax_1, bx_2), \qquad x \in \mathbb{B}^2.$$

It is necessary to know Φ explicitly, but not Ψ. The creation of such a mapping Φ is examined at length in [3].

Let

$$J(x) \equiv (D\Phi)(x) = \begin{bmatrix} \dfrac{\partial \Phi_1(x)}{\partial x_1} & \cdots & \dfrac{\partial \Phi_1(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \Phi_d(x)}{\partial x_1} & \cdots & \dfrac{\partial \Phi_d(x)}{\partial x_d} \end{bmatrix}, \qquad x \in \overline{\mathbb{B}}^d, \tag{7}$$

$$K(s) \equiv (D\Psi)(s) = \begin{bmatrix} \dfrac{\partial \Psi_1(s)}{\partial s_1} & \cdots & \dfrac{\partial \Psi_1(s)}{\partial s_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \Psi_d(s)}{\partial s_1} & \cdots & \dfrac{\partial \Psi_d(s)}{\partial s_d} \end{bmatrix}, \qquad s \in \overline{\Omega} \tag{8}$$

denote the Jacobian matrix of the transformations Φ and Ψ, respectively. Assume $J(x)$ is nonsingular on $\overline{\mathbb{B}}^d$,

$$\det J(x) \neq 0, \qquad x \in \overline{\mathbb{B}}^d;$$

and without loss of generality, assume

$$\det J(x) > 0, \qquad x \in \overline{\mathbb{B}}^d.$$

Differentiating the identity $\Phi(\Psi(s)) = s$ over Ω, or the identity $\Psi(\Phi(x)) = x$ over $\mathbb{B}^d$, leads to

$$J(x)K(s) = I, \qquad x = \Psi(s), \tag{9}$$

Thus the components of $K(s)$ can be obtained by using

$$K(s) = J(x)^{-1}, \qquad s = \Phi(x). \tag{10}$$

$$K(\Phi(x)) = J(x)^{-1}, \qquad x \in \overline{\mathbb{B}}^d. \tag{11}$$

Let v denote a general function defined over Ω. For the transformation $s = \Phi(x)$, introduce the notation $\widetilde{v}(x) = v(\Phi(x))$; or equivalently, $v(s) = \widetilde{v}(\Psi(s))$. Consider the derivatives with respect to s of $v(s)$. Let ∇_s denote the gradient with respect to the components of s; and do similarly for ∇_x. Then

$$\nabla_s v(s) = K(s)^{\mathrm{T}} \nabla_x \widetilde{v}(x), \qquad x = \Psi(s), \tag{12}$$

$$\nabla_x \widetilde{v}(x) = J(x)^{\mathrm{T}} \nabla_s v(s), \qquad s = \Phi(x),$$

with $\nabla_x \widetilde{v}(x)$ the gradient of $\widetilde{v}(x)$ written as a column vector, and analogously for $\nabla_s v(s)$. Further derivatives are considered later in Sect. 3.1.

2.1 Approximation Space

For $\Omega = \mathbb{B}^d$, introduce the approximation space

$$\widetilde{\mathscr{X}}_n = \left\{ \left(1 - |x|^2\right)^2 p\left(x\right) \mid p \in \Pi_n^d \right\},$$

where Π_n^d denotes the space of all polynomials in d variables and of degree $\leq n$. For general Ω, use the approximation space

$$\mathscr{X}_n = \left\{ \chi \circ \Phi^{-1} \mid \chi \in \widetilde{\mathscr{X}}_n \right\}.$$

Let $\mathscr{V}_k$ denote the space of all polynomials of degree k that are orthogonal in $L^2\left(\mathbb{B}^d\right)$ to all polynomials in Π_{k-1}^d using the standard inner product

$$(f, g) = \int_{\mathbb{B}^d} f\left(x\right) g\left(x\right)\, dx.$$

More precisely,

$$\mathscr{V}_k = \left\{ p \in \Pi_k^d \mid (p, q) = 0, \quad \forall q \in \Pi_{k-1}^d \right\}, \qquad k = 1, 2, \ldots,$$

and $\mathscr{V}_0$ is the set of all constant functions. Then

$$\Pi_n^d = \mathscr{V}_0 \oplus \mathscr{V}_1 \oplus \cdots \oplus \mathscr{V}_n,$$

is an orthogonal decomposition of Π_n^d within $L^2\left(\mathbb{B}^d\right)$. A basis for Π_n^d is defined by first defining a basis for each subspace $\mathscr{V}_k$, $k = 0, 1, \ldots, n$. Let $\left\{\varphi_{k,j}\right\}$ be an orthonormal basis of $\mathscr{V}_k$, let $M_k = \dim \mathscr{V}_k$, and define

$$\chi_{k,j}\left(x\right) = \left(1 - |x|^2\right)^2 \varphi_{k,j}\left(x\right), \qquad j = 1, \ldots, M_k.$$

Denote the corresponding basis for $\widetilde{\mathscr{X}}_n$ by $\left\{\chi_\ell\left(x\right) \mid 1 \leq \ell \leq N_n\right\}$,

$$N_n \equiv M_0 + \cdots + M_n.$$

Let $\left\{\psi_j \mid 1 \leq j \leq N_n\right\}$ be the corresponding basis for $\mathscr{X}_n$ using $\psi_\ell = \chi_\ell \circ \Phi^{-1}$. Note that for $d = 2$,

$$M_n = n + 1, \qquad N_n = \tfrac{1}{2}\left(n + 1\right)\left(n + 2\right).$$

Orthonormal bases $\left\{\varphi_{k,j}\right\}$ for $\mathscr{V}_k$, $k \geq 0$, are considered in [6, 11, 17].

3 The Numerical Method

The numerical method is a Galerkin method for approximating (4): find $u_n \in \mathscr{X}_n$ for which

$$\mathscr{A}\left(u_n, v\right) = \left(f, v\right), \qquad \forall v \in \mathscr{X}_n.$$

This is the standard variational framework used with finite element methods, with the approximating elements required to belong to $H_0^2(\Omega)$, a significant requirement.

Write

$$u_n\left(s\right) = \sum_{j=1}^{N_n} \alpha_j \psi_j\left(s\right).$$

Then the coefficients $\left\{\alpha_j\right\}$ must satisfy the linear system

$$\sum_{j=1}^{N_n} \alpha_j \mathscr{A}\left(\psi_j, \psi_i\right) = \ell_f\left(\psi_i\right), \qquad i = 1, \ldots, N_n. \tag{13}$$

The Lax-Milgram Theorem (cf. [1, §8.3], [8, §2.7], [9, p. 8]) implies the existence of u_n for all n, with

$$\|u_n\|_2 \le \frac{1}{c_e}\|\ell_f\|.$$

For the error in this Galerkin method, Cea's Lemma (cf. [1, p. 371], [8, p. 62], [9, p. 104]) implies the convergence of u_n to u, and moreover,

$$\|u - u_n\|_2 \le \frac{c_{\mathscr{A}}}{c_e} \inf_{v \in \mathscr{X}_n} \|u - v\|_2. \tag{14}$$

It remains to bound the best approximation error on the right side of this inequality. The error analysis is similar to that given in the earlier papers [2, 4, 5].

To bound the right side, make use of the following connection between norms in $H^k\left(\Omega\right)$ and $H^k\left(\mathbb{B}^d\right)$; the proof is omitted.

Lemma 1 *Assume $\Phi \in C^\infty\left(\Omega\right)$. Let $v \in H^k\left(\Omega\right)$ for some $k \ge 0$, $k \in \mathbb{N}_0$, and let $\widetilde{v}\left(x\right) = v\left(\Phi\left(x\right)\right)$. Then*

$$c_{1,k}\,\|\widetilde{v}\|_{H^k\left(\mathbb{B}^d\right)} \le \|v\|_{H^k(\Omega)} \le c_{2,k}\,\|\widetilde{v}\|_{H^k\left(\mathbb{B}^d\right)} \tag{15}$$

for some $c_{1,k}, c_{2,k} > 0$ independent of v.

In order to look at rates of convergence as a function of n, this lemma is used to convert the bound (14) to the equivalent bound

$$\|\widetilde{u}-\widetilde{u}_n\|_2 \le c \inf_{\widetilde{v\in\mathscr{X}_n}} \|\widetilde{u}-\widetilde{v}\|_2, \tag{16}$$

c a generic constant dependent on Φ, but not on u. Assume $u \in H_0^k(\Omega)$, and equivalently $\widetilde{u} \in H_0^k\left(\mathbb{B}^d\right)$, $k \ge 2$. Bounding the right side of (16) using [15, Thm 4.3] leads to the error bound

$$\|\widetilde{u}-\widetilde{u}_n\|_2 \le \frac{c}{n^{k-2}} \|\widetilde{u}\|_{H^k\left(\mathbb{B}^d\right)}. \tag{17}$$

Combined with Lemma 1 and (14),

$$\|u-u_n\|_2 \le \frac{c}{n^{k-2}} \|u\|_{H^k(\Omega)}, \tag{18}$$

again with c a generic constant. To obtain convergence for $k = 2$, it can be shown that

$$\inf_{\widetilde{v\in\mathscr{X}_n}} \|\widetilde{u}-\widetilde{v}\|_2 \to 0 \quad \text{as} \quad n \to \infty.$$

This follows because the polynomials $\cup_{n\ge 0}\widetilde{\mathscr{X}}_n$are dense in $H_0^2(\Omega)$ [note the comments following [15, Thm 4.3] and the denseness of the polynomials $\cup_{n\ge 0}\Pi_n^d$ in $H^k\left(\mathbb{B}^d\right)$].

3.1 Evaluating the Integrals

The integrals

$$\mathscr{A}\left(\psi_i, \psi_j\right) = \int_\Omega \left[\Delta\psi_i(s)\,\Delta\psi_j(s) + \gamma(s)\,\psi_i(s)\,\psi_j(s)\right] ds \tag{19}$$

must be computed. Begin by converting to an integral over $\mathbb{B}^d$ using the transformation $s = \Phi(x)$:

$$\mathscr{A}\left(\psi_i, \psi_j\right) = \int_{\mathbb{B}^d} \Big[\Delta_s\psi_i(s)\big|_{s=\Phi(x)}\, \Delta_s\psi_j(s)\big|_{s=\Phi(x)} \\ +\gamma(\Phi(x))\,\chi_i(x)\,\chi_j(x)\Big] \det J(x)\; dx.$$

The quantities $\Delta_s\psi_i(s)$, $i = 1, \ldots, N_n$, must be converted to functions involving derivatives with respect to x for $\chi_i(x)$.

For the transformation $x = \Psi(s)$, let $v(s) = \widetilde{v}(\Psi(s))$; or equivalently, $\widetilde{v}(x) = v(\Phi(x))$. Look at the derivatives with respect to s of $v(s)$. Then for $i = 1, \ldots, d$,

$$\begin{aligned}\frac{\partial v(s)}{\partial s_i} &= \sum_{j=1}^{d} \left.\frac{\partial \widetilde{v}(x)}{\partial x_j}\right|_{x=\Psi(s)} \times \frac{\partial \Psi_j(s)}{\partial s_i} \\ &= \left[\frac{\partial \Psi_1}{\partial s_i}, \ldots, \frac{\partial \Psi_d}{\partial s_i}\right] \nabla_x \widetilde{v}(x), \qquad x = \Psi(s).\end{aligned}$$

This is a proof of (12). Next,

$$\begin{aligned}\frac{\partial^2 v(s)}{\partial s_i^2} &= \frac{\partial}{\partial s_i}\left[\sum_{j=1}^{d} \left.\frac{\partial \widetilde{v}(x)}{\partial x_j}\right|_{x=\Psi(s)} \times \frac{\partial \Psi_j(s)}{\partial s_i}\right] \\ &= \sum_{j=1}^{d} \left.\frac{\partial \widetilde{v}(x)}{\partial x_j}\right|_{x=\Psi(s)} \times \frac{\partial^2 \Psi_j(s)}{\partial s_i^2} \\ &\quad + \sum_{j=1}^{d} \frac{\partial \Psi_j(s)}{\partial s_i} \sum_{k=1}^{d} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \frac{\partial \Psi_k(s)}{\partial s_i}.\end{aligned}$$

Summing over i,

$$\begin{aligned}\Delta_s v(s) &= \sum_{i,j=1}^{d} \left.\frac{\partial \widetilde{v}(x)}{\partial x_j}\right|_{x=\Psi(s)} \times \frac{\partial^2 \Psi_j(s)}{\partial s_i^2} \\ &\quad + \sum_{i,j,k=1}^{d} \frac{\partial \Psi_j(s)}{\partial s_i} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \frac{\partial \Psi_k(s)}{\partial s_i}.\end{aligned} \tag{20}$$

Look at the terms in (20). First,

$$\begin{aligned}\sum_{i,j=1}^{d} \left.\frac{\partial \widetilde{v}(x)}{\partial x_j}\right|_{x=\Psi(s)} \times \frac{\partial^2 \Psi_j(s)}{\partial s_i^2} &= \sum_{j=1}^{d} \frac{\partial \widetilde{v}(x)}{\partial x_j} \Delta_s \Psi_j(s) \\ &= \left[\Delta_s \Psi_1(s), \ldots, \Delta_s \Psi_d(s)\right] \nabla_x \widetilde{v}(x).\end{aligned} \tag{21}$$

Using the notation of (8),

$$\begin{aligned}\sum_{i,j,k=1}^{d} \frac{\partial \Psi_j(s)}{\partial s_i} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \frac{\partial \Psi_k(s)}{\partial s_i} &= \sum_{j,k=1}^{d} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \sum_{i=1}^{d} \frac{\partial \Psi_j(s)}{\partial s_i} \frac{\partial \Psi_k(s)}{\partial s_i} \\ &= \sum_{j,k=1}^{d} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \left[K(s)_{j,*}\right] \left[K(s)_{k,*}\right]^{\mathrm{T}} \qquad (22) \\ &= \sum_{j,k=1}^{d} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \left[K(s)\, K(s)^{\mathrm{T}}\right]_{j,k}.\end{aligned}$$

Returning to (20) and combining terms,

$$\begin{aligned}\Delta_s v(s) &= \left[\Delta_s \Psi_1(s), \dots, \Delta_s \Psi_d(s)\right] \nabla_x \widetilde{v}(x) \\ &\quad + \sum_{j,k=1}^{d} \frac{\partial^2 \widetilde{v}(x)}{\partial x_j \partial x_k} \left[K(s)_{j,*}\right] \left[K(s)_{k,*}\right]^{\mathrm{T}}. \qquad (23)\end{aligned}$$

Formula (22) can be evaluated from knowing $J(x)$; see (10) above. The formula (23) is to be evaluated with

$$\widetilde{v}(x) = \chi_\ell(x), \qquad 1 \le \ell \le N_n,$$

so as to create the elements $\mathscr{A}\left(\psi_i, \psi_j\right)$.

To evaluate (21), we need $\Delta_s \Psi_j(s)$, $1 \le j \le d$. The first derivatives of $\not{\Psi}$ can be obtained from $D_s \Psi(s) = \left[D_x \Phi(x)\right]^{-1}$ where $s = \Phi(x)$. How to obtain the functions $\Delta_s \Psi_j(s)$? Begin by differentiating

$$s = \Phi(\Psi(s)), \qquad s \in \Omega,$$

or

$$s_j = \Phi_j(\Psi_1(s), \dots, \Psi_d(s)), \qquad 1 \le j \le d.$$

The derivative with respect to s_i yields

$$\delta_{i,j} = \sum_{k=1}^{d} \left.\frac{\partial \Phi_j(x)}{\partial x_k}\right|_{x=\Psi(s)} \times \frac{\partial \Psi_k(s)}{\partial s_i}, \qquad 1 \le i,j \le d. \qquad (24)$$

Differentiate the components of (9), given in (24), with respect to s_ℓ: for $1 \leq i,j,\ell \leq d$,

$$0=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\frac{\partial^2\Psi_k(s)}{\partial s_i\partial s_\ell}+\sum_{k=1}^{d}\frac{\partial\Psi_k(s)}{\partial s_i}\sum_{m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\frac{\partial\Psi_m(s)}{\partial s_\ell}.$$

Let $\ell=i$,

$$\begin{aligned}0&=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\frac{\partial^2\Psi_k(s)}{\partial s_i^2}+\sum_{k=1}^{d}\frac{\partial\Psi_k(s)}{\partial s_i}\sum_{m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\frac{\partial\Psi_m(s)}{\partial s_i},\\&=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\frac{\partial^2\Psi_k(s)}{\partial s_i^2}+\sum_{k,m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\frac{\partial\Psi_k(s)}{\partial s_i}\frac{\partial\Psi_m(s)}{\partial s_i}.\end{aligned}$$

Sum over i: for $1\leq j\leq d$,

$$\begin{aligned}0&=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\Delta_s\Psi_k(s)+\sum_{k,m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\sum_{i=1}^{d}\frac{\partial\Psi_k(s)}{\partial s_i}\frac{\partial\Psi_m(s)}{\partial s_i}\\&=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\Delta_s\Psi_k(s)+\sum_{k,m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\left[K(s)_{k,*}\right]\left[K(s)_{m,*}\right]^{\mathrm{T}}\qquad(25)\\&=\sum_{k=1}^{d}\frac{\partial\Phi_j(x)}{\partial x_k}\Delta_s\Psi_k(s)+\sum_{k,m=1}^{d}\frac{\partial^2\Phi_j(x)}{\partial x_k\partial x_m}\left[K(s)K(s)^{\mathrm{T}}\right]_{k,m}.\end{aligned}$$

Introduce

$$\Delta_s\Psi(s)=\left[\Delta_s\Psi_1(s),\ldots,\Delta_s\Psi_d(s)\right]^{\mathrm{T}},$$

$$D^2\Phi_j(x)=\begin{bmatrix}\frac{\partial^2\Phi_j(x)}{\partial x_1\partial x_1}&\cdots&\frac{\partial^2\Phi_j(x)}{\partial x_1\partial x_d}\\\vdots&\ddots&\vdots\\\frac{\partial^2\Phi_j(x)}{\partial x_d\partial x_1}&\cdots&\frac{\partial^2\Phi_j(x)}{\partial x_d\partial x_d}\end{bmatrix},\qquad 1\leq j\leq d.$$

Introduce the dot product of two arrays of the same dimension:

$$A\odot B=\sum_{i,j}A_{i,j}B_{i,j}.$$

Then (25) can be written as

$$0 = \left[J(x)_{j,*}\right]\left[\Delta_s \Psi(s)\right] + D^2\Phi_j(x) \odot \left[K(s)\,K(s)^{\mathrm{T}}\right].$$

From (25),

$$\begin{aligned} 0 &= J(x)\,\Delta_s\Psi(s) + \begin{bmatrix} D^2\Phi_1(x) \odot \left[K(s)\,K(s)^{\mathrm{T}}\right] \\ \vdots \\ D^2\Phi_d(x) \odot \left[K(s)\,K(s)^{\mathrm{T}}\right] \end{bmatrix} \\ &\equiv J(x)\,\Delta_s\Psi(s) + \mathscr{D}^2\Phi(x) \odot \left[K(s)^{\mathrm{T}}\,K(s)\right]. \end{aligned}$$

which contains an implicit definition of $\mathscr{D}^2\Phi$ and an implicit notational extension of the operation $\odot$. Then

$$\Delta_s\Psi(s) = -J(x)^{-1}\left\{\mathscr{D}^2\Phi(x) \odot \left[K(s)\,K(s)^{\mathrm{T}}\right]\right\},$$

and recall (10) to compute $K(s)$.

This allows computing the coefficients $\mathscr{A}\left(\psi_i, \psi_j\right)$ of (13) by means of the change of variables $s = \Phi(x)$. Rewrite (23) as

$$\Delta_s v(s) = \left[\Delta_s\Psi(s)\right]^{\mathrm{T}} \nabla_x \widetilde{v}(x) + D^2\widetilde{v}(x) \odot \left[K(s)\,K(s)^{\mathrm{T}}\right]. \tag{26}$$

Returning to $\mathscr{A}\left(\psi_i, \psi_j\right)$, apply (26) with

$$\begin{aligned} \widetilde{v}(x) = \chi_{n,j}(x) &= \left(1 - |x|^2\right)^2 \varphi_{n,j}(x) \\ &= \left(1 - x_1^2 - \cdots - x_d^2\right)^2 \varphi_{n,j}(x) \end{aligned}$$

for $1 \le j \le M_k$, $0 \le k \le n$.

We need to find the first and second order derivatives of $\chi_{n,j}(x)$, and thus also of $\varphi_{n,j}(x)$.

$$\frac{\partial\chi_{n,j}(x)}{\partial x_k} = -4x_k\left(1 - x_1^2 - \cdots - x_d^2\right)\varphi_{n,j}(x) + \left(1 - |x|^2\right)^2 \frac{\partial\varphi_{n,j}(x)}{\partial x_k},$$

$$\frac{\partial^2\chi_{n,j}(x)}{\partial x_k^2} = \left\{-4\left(1 - x_1^2 - \cdots - x_d^2\right) + 8x_k^2\right\}\varphi_{n,j}(x)$$

$$\begin{aligned}&-8x_k\left(1-x_1^2-\cdots-x_d^2\right)\frac{\partial\varphi_{n,j}(x)}{\partial x_k}\\&+\left(1-|x|^2\right)^2\frac{\partial^2\varphi_{n,j}(x)}{\partial x_k^2}.\end{aligned}$$

For $\ell \neq k$,

$$\begin{aligned}\frac{\partial^2\chi_{n,j}(x)}{\partial x_k\partial x_\ell}&=8x_kx_\ell\varphi_{n,j}(x)\\&-4\left(1-x_1^2-\cdots-x_d^2\right)\left\{x_k\frac{\partial\varphi_{n,j}(x)}{\partial x_\ell}+x_\ell\frac{\partial\varphi_{n,j}(x)}{\partial x_k}\right\}\\&+\left(1-|x|^2\right)^2\frac{\partial^2\varphi_{n,j}(x)}{\partial x_k\partial x_\ell}.\end{aligned}$$

These can be combined with (26) to compute $\Delta_s\chi_j$ and thus to compute $\mathscr{A}\left(\psi_i,\psi_j\right)$ for $1 \leq i,j \leq N_n$.

The next step is to look at particular orthonormal polynomials $\left\{\varphi_{n,j}(x)\right\}$ and to compute

$$\begin{array}{cc}\varphi_{n,j}, & 1\leq j\leq M_n,\\ \dfrac{\partial\varphi_{n,j}(x)}{\partial x_k}, & \begin{array}{c}1\leq j\leq M_n\\ 1\leq k\leq d\end{array},\\ \dfrac{\partial^2\varphi_{n,j}(x)}{\partial x_k\partial x_\ell}, & \begin{array}{c}1\leq j\leq M_n\\ 1\leq k,\ell\leq d\end{array}.\end{array}$$

The best choice as regards speed of calculation is to use the polynomials discussed in [6], as they satisfy a triple recursion that allows for a rapid calculation. For the planar case, these are given by

$$\varphi_{n,k}(x)=\frac{1}{h_{k,n}}C_{n-k}^{k+1}(x_1)\left(1-x_1^2\right)^{\frac{k}{2}}C_k^{\frac{1}{2}}\left(\frac{x_2}{\sqrt{1-x_1^2}}\right),\qquad x\in\mathbb{B}^2,\qquad(27)$$

for $k=0,\ldots,n$, $n=0,1,\ldots$ The quantity $C_m^\lambda(t)$, $m \geq 0$, denotes the Gegenbauer polynomial of degree m and index λ.

For the three dimensional case, we use the polynomials

$$\begin{aligned}\varphi_{n,j,k}(x) &= \frac{1}{h_{j,k}} C_{n-j-k}^{j+k+3/2}(x_1)(1-x_1^2)^{j/2}\dots \\ &\times C_j^{k+1}\left(\frac{x_2}{\sqrt{1-x_1^2}}\right)(1-x_1^2-x_2^2)^{k/2} C_k^{1/2}\left(\frac{x_3}{\sqrt{1-x_1^2-x_2^2}}\right), \\ & x \in \mathbb{B}^3,\ 0 \le j+k \le n,\ n = 0,1,\dots \end{aligned} \tag{28}$$

which uses again the Gegenbauer polynomials. The numbers $h_{j,k}$ are normalization constants, see [11], and see [6] for the triple recursion.

4 Numerical Examples

4.1 Planar Examples

Our first examples are for Ω a planar region, and thus $\Phi : \mathbb{B}^2 \to \Omega$.

Example 1 Begin with the elliptical region Ω defined by the mapping $s = \Phi(x)$, $x \in \mathbb{B}^2$,

$$\begin{aligned} s_1 &= 2x_1 + x_2 \\ s_2 &= 3x_1 - 4x_2. \end{aligned} \tag{29}$$

Choose

$$f(s) = 10 \cos(s_1 - 0.1) \sin(s_2 + 0.1) \tag{30}$$

and $\gamma(s) \equiv 1$ over Ω. The solution is shown in Fig. 1. The true solution is unknown, so the error is estimated by using u_{n^*} as the 'true' solution with n^* much larger than n being used. In the present case, $n^* = 20$ was used. The maximum errors are shown in Fig. 2, and it appears to be an exponential decrease in the error. Figure 3 is a graph of $\log n$ vs. $\log(cond)$, with *cond* the condition number of (13). It indicates that the condition number is $\mathcal{O}(n^p)$ for some power p; experimentally and roughly, $p \approx 4.5$, and $p = 4$ seems most likely to be the theoretical power. That would be consistent with the condition number being $\mathcal{O}\left(N_n^2\right)$, as was observed earlier with the spectral method for the Neumann boundary value problem for second order equations.

Example 2 Consider the boundary mapping

$$\begin{aligned} \varphi(\theta) &= \rho(\theta)(\cos\theta, \sin\theta), \\ \rho(\theta) &= 3 + \cos\theta + 2\sin\theta, \qquad 0 \le \theta \le 2\pi. \end{aligned} \tag{31}$$

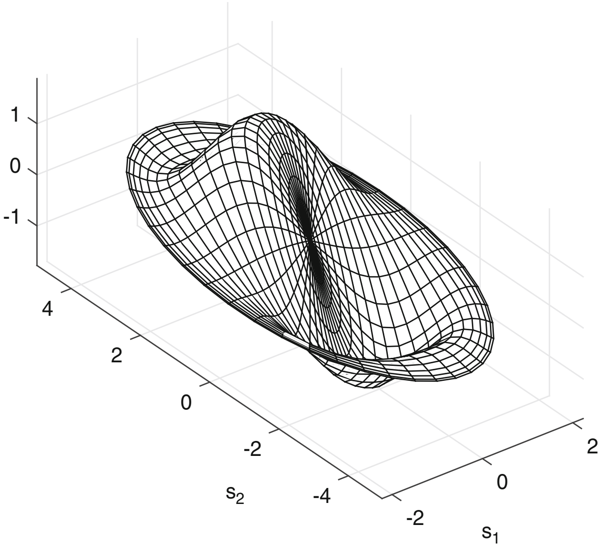


Fig. 1 Solution u with f given by (30), $\gamma(s) \equiv 1$, and the region (29)

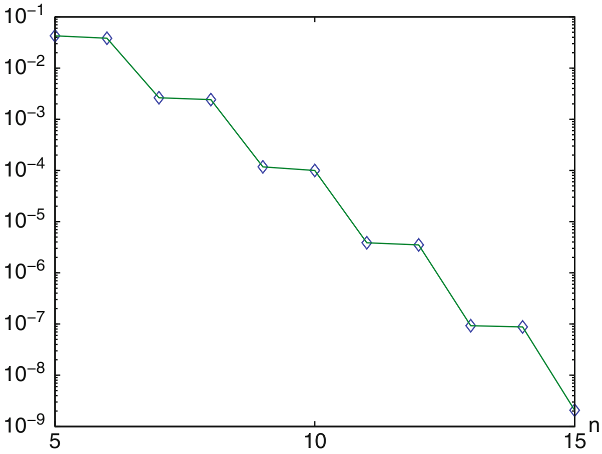


Fig. 2 Computed error in u_n with f given by (30) and $\gamma(s) \equiv 1$

This can be extended to a polynomial mapping of degree 2 in various ways, as discussed in [3], and one such mapping is illustrated in Fig. 4. This mapping Φ is obtained using (1) the interpolation/quadrature method of §3 in [3], followed by (2) computing the least squares polynomial approximation over $\mathbb{B}^2$ of degree 2 in each component.

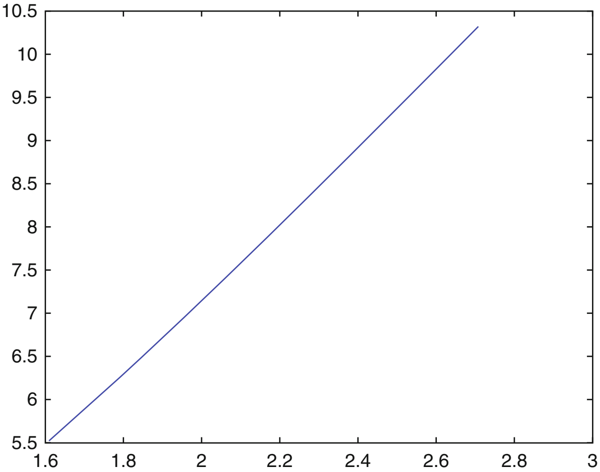


Fig. 3 $\log n$ vs. $\log(cond)$, with *cond* the condition number of (13), with the region (29)

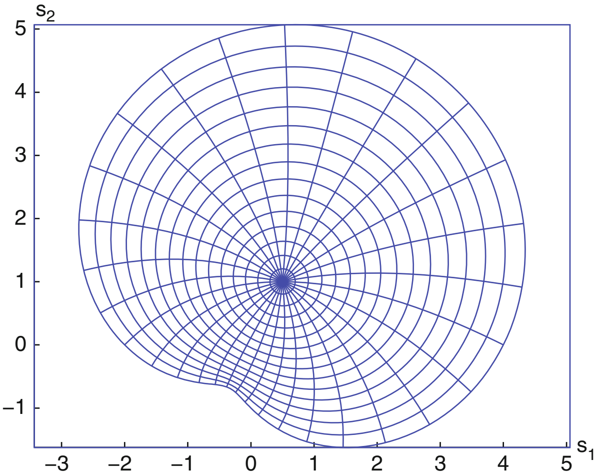


Fig. 4 Mappings for limacon boundary mapping (31)

Equation (1) is solved with the same choices for γ and f as in Example 1. Figure 5 illustrates the solution, using u_{20}. The errors are shown in Fig. 6. The condition numbers, shown in Fig. 7, appear to increase like $\mathscr{O}\left(N_n^2\right)$, as with Example 1.

Example 3 Consider the mapping

$$\Phi_1(x) = \left[x_1 - x_2 + ax_1^2,\, x_1 + x_2\right]^{\mathrm{T}}, \qquad x \in \overline{\mathbb{B}}^2, \tag{32}$$

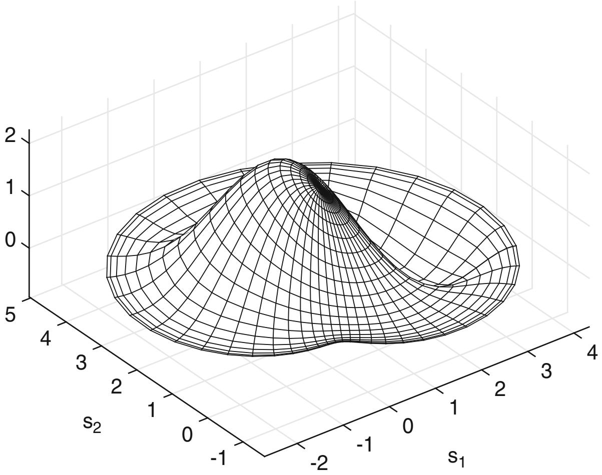


Fig. 5 Solution u with f given by (30), $\gamma(s) \equiv 1$, and $\partial\Omega$ given by (31)

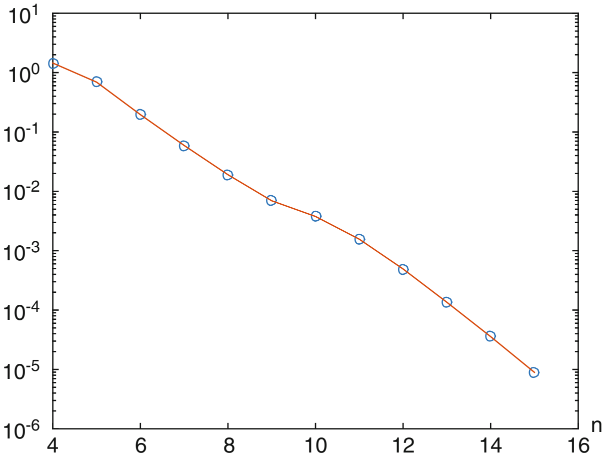


Fig. 6 Computed error in u_n with f given by (30), $\gamma(s) \equiv 1$, and the boundary mapping (31)

for a given $0 < a < 1$, with the image defining Ω. In addition, use the interpolation/quadrature method of §3 in [3] to create another mapping Φ_2 that agrees with Φ_1 on the boundary of $\mathbb{B}^2$. These mappings are illustrated in Fig. 8. Clearly Φ_2 is a 'better behaved' mapping as compared to Φ_1. We solve $\Delta^2 u + \gamma u = f$ as before, but now let

$$f(s) = 200\cos(st)\sin(t + 0.1). \tag{33}$$

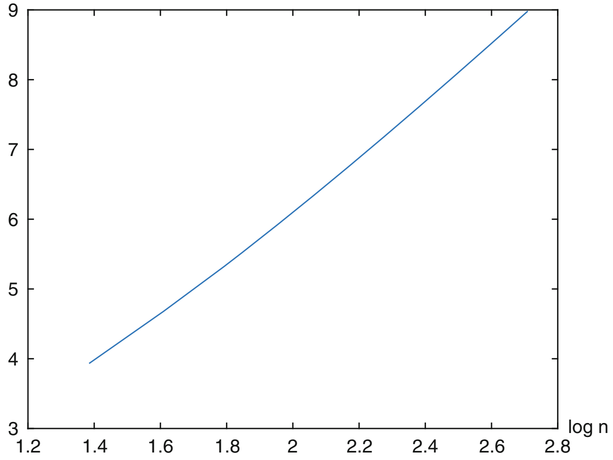


Fig. 7 $\log n$ vs. $\log(cond)$, with *cond* the condition number of (13), with the boundary mapping (31)

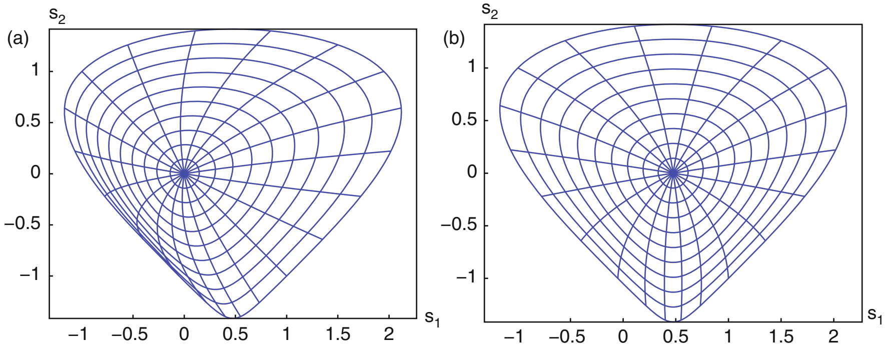


Fig. 8 The mapping Φ for boundary (32) with a $= 0.95$. (**a**) Φ_1. (**b**) Φ_2

The solution is shown in Fig. 9. The maximum errors are shown in Fig. 10, and there appears to be an exponential decrease in the error. The condition numbers are shown in Fig. 11, and again they appear to increase like $\mathscr{O}\left(N_n^2\right)$.

4.2 A Three Dimensional Example

Example 4 Here we consider the case of an ellipsoid

$$\Omega = \{(s_1, s_2, s_3) \mid s_1^2 + \left(\frac{s_2}{3}\right)^2 + \left(\frac{s_3}{2}\right)^2 \le 1\} \tag{34}$$

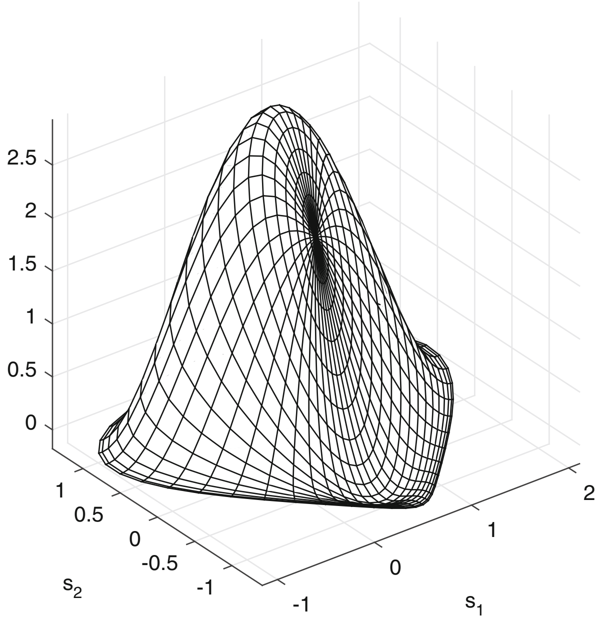


Fig. 9 Solution u with f given by (33), $\gamma(s) \equiv 1$, and $\partial\Omega$ given by (32)

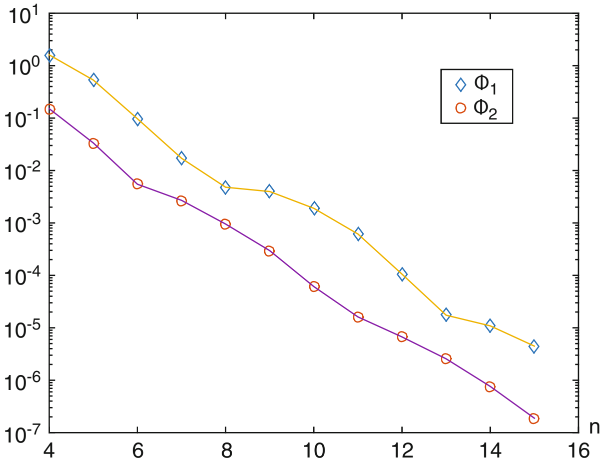


Fig. 10 Computed error in u_n with f given by (33), for the mappings Φ_1 and Φ_2 with the boundary specified by (32)

with the obvious mapping

$$\Phi(x_1, x_2, x_3) = [x_1, 3x_2, 2x_3], \ [x_1, x_2, x_3] \in \mathbb{B}^3. \tag{35}$$

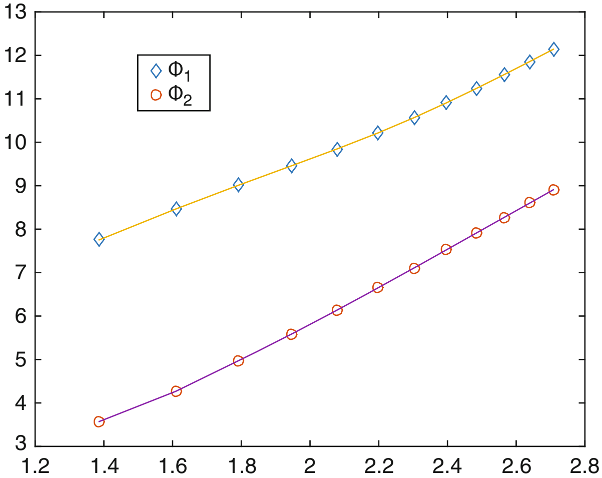


Fig. 11 $\log n$ vs. $\log(cond)$, with *cond* the condition number of (13), for the mappings Φ_1 and Φ_2 with the boundary specified by (32)

We solve Eq. (1) with $\gamma(s) \equiv 1$ and calculate the right hand side f_1 in such a way that the solution of (1) is given by

$$u(s_1, s_2, s_3) = \left(1 - s_1^2 - \left(\frac{s_2}{3}\right)^2 - \left(\frac{s_3}{2}\right)^2\right)^2 e^{3(s_1+s_2/3+s_3/2)}. \tag{36}$$

To study the influence of faster growing derivatives we use a second right hand side f_2 on the same domain Ω, such that the solution is given

$$v(s_1, s_2, s_3) = \left(1 - s_1^2 - \left(\frac{s_2}{3}\right)^2 - \left(\frac{s_3}{2}\right)^2\right)^2 e^{7(s_1+s_2/3+s_3/2)}. \tag{37}$$

We expect slower but still exponential convergence for the second example. This is confirmed in the numerical calculation, see Fig. 12, where the maximum errors are plotted versus n. The error graph for the solution u shows some saturation around $n = 22$, because we reach the precision limit of the Gauß–quadratures that we use for the evaluation of the integrals in Eq. (13). The graph of $\log(cond)$ versus $\log n$ in Fig. 13 shows again a polynomial behavior. From the numerical results we estimate a condition number of $\mathcal{O}\left(N_n^2\right)$, where we remember that $N_n = \mathcal{O}(n^3)$.

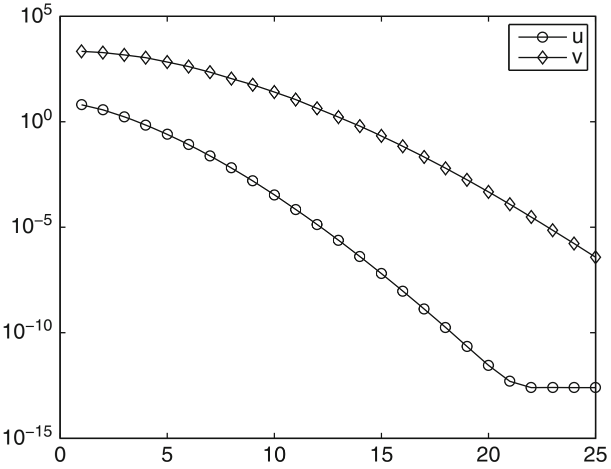


Fig. 12 Computed error in u_n and v_n, for the solutions u and v given in (36) and (37)

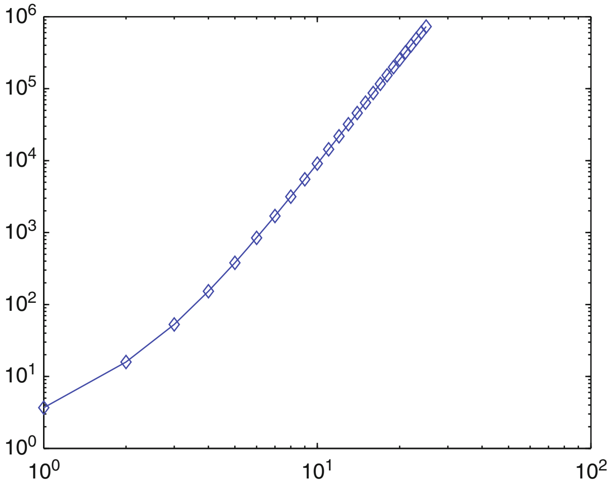


Fig. 13 $\log(cond)$ vs. $\log n$, with *cond* the condition number of (13), with $\gamma(s) \equiv 1$, and the mapping (35) for the domain (34)

5 Nonhomogeneous Boundary Conditions

Consider the Dirichlet biharmonic problem

$$\begin{gathered} \Delta^2 u(s) + \gamma(s)\,u(s) = f(s), \qquad s \in \Omega, \\ u(s) = g_1(s), \quad \frac{\partial u(s)}{\partial n_s} = g_2(s), \quad s \in \partial\Omega. \end{gathered} \tag{38}$$

This can be reduced to two simpler problems. Consider first the standard Dirichlet biharmonic problem

$$\begin{aligned} \Delta^2 w(s) &= 0, \qquad s \in \Omega, \\ w(s) = g_1(s), \quad \frac{\partial w(s)}{\partial n_s} &= g_2(s), \quad s \in \partial\Omega. \end{aligned} \tag{39}$$

Define $v = u - w$. Then v satisfies

$$\begin{aligned} \Delta^2 v(s) + \gamma(s) v(s) &= f(s) - \gamma(s) w(s) \equiv \widetilde{f}(s), \qquad s \in \Omega, \\ v(s) = \frac{\partial v(s)}{\partial n_s} &= 0, \quad s \in \partial\Omega. \end{aligned} \tag{40}$$

Begin by solving (39) numerically, obtaining an approximating solution $\widehat{w}(s) \approx w(s)$. Then solve (40) with $\widehat{w}(s)$ replacing $w(s)$ in the definition of $\widetilde{f}(s)$. The problem (40) can be solved using the methods given earlier in this paper. Solve for an approximating solution $v_n(s) \approx v(s)$, and then define

$$\widehat{u}(s) = v_n(s) + \widehat{w}(s), \qquad s \in \Omega,$$

as the approximating solution of (38).

To solve (39), a number of methods have been proposed, often using boundary integral equation reformulations. For a review of some of these, see [12, Chaps. 9,15], [13, 14].

Remark The eigenvalue problem for the biharmonic equation (1)–(2) is discussed and illustrated in the book [7, Chap. 9].

Traditional spectral methods use univariate approximations with a decomposition of the partial differential equation into univariate problems. Consider, for example, using a polar coordinates decomposition of the unit disk. But this leads to problems when treating the solution u at the center of the disk. The present spectral method makes use of the recently developed theory and tools for multivariate approximation over $\mathbb{B}^d$, avoiding artificial problems that can occur when using univariate approximations.

References

1. Atkinson, K., Han, W.: Theoretical Numerical Analysis: A Functional Analysis Framework, 3rd edn. Springer, New York (2009)
2. Atkinson, K., Hansen, O.: A spectral method for the eigenvalue problem for elliptic equations. Electron. Trans. Numer. Anal. **37**, 386–412 (2010)
3. Atkinson, K., Hansen, O.: Creating domain mappings. Electron. Trans. Numer. Anal. **39**, 202–230 (2012)

4. Atkinson, K., Chien, D., Hansen, O.: A spectral method for elliptic equations: the Dirichlet problem. Adv. Comput. Math. **33**, 169–189 (2010)
5. Atkinson, K., Hansen, O., Chien, D.: A spectral method for elliptic equations: the Neumann problem. Adv. Comput. Math. **34**, 295–317 (2011)
6. Atkinson, K., Chien, D., Hansen, O.: Evaluating polynomials over the unit disk and the unit ball. Numer. Algorithms **67**, 691–711 (2014)
7. Atkinson, K., Chien, D., Hansen, O.: Spectral methods using multivariate polynomials on the unit ball (submitted)
8. Brenner, S., Scott, L.: The Mathematical Theory of Finite Element Methods. Springer, New York (1994)
9. Ciarlet, P.: The Finite Element Method For Elliptic Problems. North-Holland, Amsterdam (1978)
10. Destuynder, P., Salaun, M.: Mathematical Analysis of Thin Plate Models. Springer, Berlin (1996)
11. Dunkl, C., Xu, Y.: Orthogonal Polynomials of Several Variables. Cambridge University Press, Cambridge (2001)
12. Jaswon, M., Symm, G.: Integral Equation Methods in Potential Theory and Elastostatics. Academic, New York (1977)
13. Jeon, Y.-M.: An indirect boundary integral equation for the biharmonic equation. SIAM J. Num. Anal. **31**, 461–476 (1992)
14. Jeon, Y.-M.: New boundary element formulas for the biharmonic equation. Adv. Comput. Math. **9**, 97–115 (1998)
15. Li, H., Xu, Y.: Spectral approximation on the unit ball. SIAM J. Num. Anal. **52**, 2647–2675 (2014)
16. Selvadurai, A.P.S.: Partial Differential Equations in Mechanics 2: The Biharmonic Equation The Poisson Equation. Springer, Berlin (2000)
17. Xu, Y.: Lecture notes on orthogonal polynomials of several variables. In: Advances in the Theory of Special Functions and Orthogonal Polynomials, pp. 135–188. Nova Science Publishers, New York (2004)

Quasi-Monte Carlo for an Integrand with a Singularity Along a Diagonal in the Square

Kinjal Basu and Art B. Owen

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Quasi-Monte Carlo methods are designed for integrands of bounded variation, and this excludes singular integrands. Several methods are known for integrands that become singular on the boundary of the unit cube $[0, 1]^d$ or at isolated possibly unknown points within $[0, 1]^d$. Here we consider functions on the square $[0, 1]^2$ that may become singular as the point approaches the diagonal line $x_1 = x_2$, and we study three quadrature methods. The first method splits the square into two triangles separated by a region around the line of singularity, and applies recently developed triangle QMC rules to the two triangular parts. For functions with a singularity 'no worse than $|x_1 - x_2|^{-A}$ is' for $0 < A < 1$ that method yields an error of $O((\log(n)/n)^{(1-A)/2})$. We also consider methods extending the integrand into a region containing the singularity and show that method will not improve upon using two triangles. Finally, we consider transforming the integrand to have a more QMC-friendly singularity along the boundary of the square. This then leads to error rates of $O(n^{-1+\epsilon+A})$ when combined with some corner-avoiding Halton points or with randomized QMC but it requires some stronger assumptions on the original singular integrand.

1 Introduction

Quasi-Monte Carlo (QMC) integration is designed for integrands of bounded variation in the sense of Hardy and Krause (BVHK). Such integrands must necessarily be bounded. Singular integrands cannot be BVHK; they cannot even be Riemann

K. Basu
LinkedIn Inc., Mountain View, CA, USA
e-mail: kbasu@linkedin.com

A. B. Owen (✉)
Stanford University, Stanford, CA, USA
e-mail: owen@stanford.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_6

integrable. It is known since [5] and [3] that for any integrand f on $[0,1]^d$ that is not Riemann integrable, there exists a sequence $\boldsymbol{x}_i \in [0,1]^d$ for which the star discrepancy $D_n^*(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) \to 0$ as $n \to \infty$ while $(1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$ fails to converge to $\int_{[0,1]^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

We are interested in problems where the singularity arises along a manifold in $[0,1]^d$. For motivation, see the engineering applications by Mishra and Gupta in [9] and several other papers. Apart from a few remarks, we focus solely on the problem where there is a singularity along the line $x_1 = x_2$ in $[0,1]^2$.

It is possible for QMC integration to succeed on unbounded integrands. Sobol' [14] noticed this when colleagues used his methods on such problems. He explained it in terms of QMC points that avoid a hyperbolic region around the lower boundary of the unit cube where the integrands became singular. Klinger [8] shows that Halton points and some digital nets avoid a cubical region around the origin. Halton points (after the zero'th) avoid hyperbolic regions around the boundary faces of the unit cube at a rate suitable to get error bounds for QMC [12]. Certain Kronecker sequences avoid hyperbolic regions around the boundary of the cube [7]. In all of these examples, avoiding the singularity should be understood as using points that approach it, but not too quickly, as the number n of function evaluations increases.

For plain Monte Carlo, the location of the singularity is not important. One only needs to consider the first two moments of the integrand. Because QMC exploits mild smoothness of the integrand, the nature of the singularity matters. Reference [13] considers randomized QMC (RQMC) methods for integrands with point singularities at unknown locations. In RQMC, the integrand is evaluated at points that, individually, are uniformly distributed on $[0,1]^d$ and this already implies a singularity avoidance property via the Borel-Cantelli lemma. If $\int f(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} < \infty$ then scrambled nets yield an unbiased estimate of $\mu = \int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ with RMSE $o(n^{-1/2})$ [10].

The analyses in [12] and [13] employ an extension $\tilde{f}$ of f from a set $K = K_n \subset [0,1]^d$ to $[0,1]^d$. The extension satisfies $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x})$ for $\boldsymbol{x} \in K$. Now the quadrature error is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) - \int_{[0,1]^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^n \tilde{f}(\boldsymbol{x}_i)\\ &\quad + \frac{1}{n}\sum_{i=1}^n \tilde{f}(\boldsymbol{x}_i) - \int_{[0,1]^d} \tilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\quad + \int_{[0,1]^d} \tilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{[0,1]^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

If all of the points satisfy $\boldsymbol{x}_i \in K$, then the first term drops out and we find that

$$\left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) - \int_{[0,1]^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \le \left|\frac{1}{n}\sum_{i=1}^n \tilde{f}(\boldsymbol{x}_i) - \int_{[0,1]^d} \tilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \int_{-K} |\tilde{f}(\boldsymbol{x}) - f(\boldsymbol{x})|\mathrm{d}\boldsymbol{x},$$

where $-K = [0,1]^d \setminus K$. The extension used in [12] and [13] is due to Sobol' [14]. It is particularly well suited to a Koksma-Hlawka bound for the first term above as $\tilde{f}$ has low variation.

In our case, we can isolate the singularity in the set $\{\boldsymbol{x} \mid |x_1 - x_2| < \epsilon\}$. A set $K \subset [0,1]^d$ is Sobol'-extensible to $[0,1]^d$ with anchor $\boldsymbol{c}$ if for every $\boldsymbol{x} \in K$ the rectangle $\prod_{j=1}^{d}[\min(x_j, c_j), \max(x_j, c_j)] \subset K$. In our case, the set $\{\boldsymbol{x} \mid |x_1 - x_2| \geq \epsilon\}$ in which f is bounded is not Sobol' extensible. The extension $\tilde{f}$ used in [12] and [13] cannot be defined for this problem.

Section 2 presents a strategy of avoiding a region near the singularity and integrating over two triangular regions using the method from [1]. The error is then a sum of two quadrature errors and one truncation error. We consider functions where the singularity is not more severe than that in $|x_1-x_2|^{-A}$ where $0 < A < 1$. Section 3 shows that the truncation error in this approach is $O(\epsilon^{-A})$ and the quadrature error is $O(\epsilon^{-A-1}\log(n)/n)$ using the points from [1] and a Koksma-Hlawka bound from [4]. The result is that we can attain a much better quadrature error bound of $O((\log(n)/n)^{(1-A)/2})$. Section 4 shows that an approach based on finding an extension $\tilde{f}$ of f would not yield a better rate for this problem. Section 5 transforms the problem so that each triangular region becomes the image of a unit square, with the singularity now on the boundary of the square. The singularity may be too severe for QMC. However, with an additional assumption on the nature of the singularity it is possible to attain a quadrature error of $O(n^{-1+\epsilon+A})$. Section 6 summarizes the findings and relates them to QMC-friendliness as discussed by several authors, including Ian Sloan in his work with Xiaoqun Wang.

2 Background

In the context of a Festschrift for Ian Sloan, we presume that the reader is familiar with quasi-Monte Carlo, discrepancy and variation. Modern approaches to QMC and discrepancy are covered in [6]. See [11] for an outline of variation for QMC including variation in the senses of Vitali and of Hardy and Krause.

We will use a notion of functions that are singular but not too badly singular.

Definition 1 The function f defined on $[0,1]^2$ has a diagonal singularity no worse than $|x_1 - x_2|^{-A}$ for $0 < A < 1$, if

$$\begin{aligned} |f(\boldsymbol{x})| &\leq B|x_1 - x_2|^{-A} \\ \left|\frac{\partial f(\boldsymbol{x})}{\partial x_j}\right| &\leq B|x_1 - x_2|^{-A-1}, \quad j \in \{1,2\}, \quad \text{and} \\ \left|\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_k}\right| &\leq B|x_1 - x_2|^{-A-2}, \quad j,k \in \{1,2\} \end{aligned} \tag{1}$$

all hold for some $B < \infty$.

We take $A > 0$ in order to allow a singularity and $A < 1$ because f must be integrable. Smaller values of A describe easier cases to handle. The value of A to use for a given integrand may be evident from its analytical form. If $A < 1/2$ then f^2 is integrable. Definition 1 is modeled on some previous notions:

Definition 2 The function $f(\boldsymbol{x})$ defined on $[0,1]^d$ has a lower edge singularity no worse than $\prod_{j=1}^d x_j^{-A_j}$, for constants $0 < A_j < 1$, if

$$|\partial^u f(\boldsymbol{x})| \le B \prod_{j=1}^d x_j^{-A_j - 1_{j\in u}},$$

holds for some $B < \infty$ and all $u \subseteq \{1, 2, \dots, d\}$.

Definition 3 The function $f(\boldsymbol{x})$ defined on $[0,1]^d$ has a point singularity no worse than $\|\boldsymbol{x} - \boldsymbol{z}\|^{-A}$, for $\boldsymbol{z} \in [0,1]^d$, if

$$|\partial^u f(\boldsymbol{x})| \le B \|\boldsymbol{x} - \boldsymbol{z}\|^{-A-|u|}$$

holds for some $B < \infty$ and all $u \subseteq \{1, 2, \dots, d\}$.

Definition 2 is one of several conditions in [12] for singularities that arise as $\boldsymbol{x}$ approaches the boundary of the unit cube. Definition 3 is used in [13] for isolated point singularities. Definition 1 is more stringent than Definitions 2 and 3 are, because it imposes a constraint on partial derivatives taken twice with respect to x_1 or x_2.

To estimate $\mu = \int_{[0,1]^2} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ we will sample points $\boldsymbol{x}_i \in [0,1]^2$. The points we use will avoid a region near the singularity by sampling only within

$$S_\epsilon = \left\{\boldsymbol{x} \in [0,1]^2 \;\middle|\; |x_1 - x_2| \ge \epsilon\right\}$$

where $0 < \epsilon < 1$. The set S_ϵ is the union of two disjoint triangles:

$$T_\epsilon^u = \left\{\boldsymbol{x} \in [0,1]^2 \;\middle|\; x_2 \ge x_1 + \epsilon\right\}, \quad \text{and}$$
$$T_\epsilon^d = \left\{\boldsymbol{x} \in [0,1]^2 \;\middle|\; x_2 \le x_1 - \epsilon\right\}.$$

We let $-S_\epsilon$ denote the set $[0,1]^2 \backslash S_\epsilon$. As remarked in the introduction, the set $T_u \cup T_d$ is not Sobol' extensible to $[0,1]^2$.

We will choose points $\boldsymbol{x}_{i,u} \in T_\epsilon^u$ for $i = 1, \dots, n$ and estimate $\mu_{\epsilon,u} = \int_{T_\epsilon^u} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ by

$$\hat{\mu}_{\epsilon,u} = \frac{\mathrm{vol}(T_\epsilon^u)}{n} \sum_{i=1}^n f(\boldsymbol{x}_{i,u}).$$

Using a similar estimate for T^d_ϵ we arrive at our estimate of μ,

$$\hat{\mu}_\epsilon = \hat{\mu}_{\epsilon,u} + \hat{\mu}_{\epsilon,d}.$$

Our error then consists of two quadrature errors and a truncation error and it satisfies the bound

$$|\hat{\mu}_\epsilon - \mu| \le \left|\hat{\mu}_{\epsilon,u} - \int_{T^u_\epsilon} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \left|\hat{\mu}_{\epsilon,d} - \int_{T^d_\epsilon} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \left|\int_{-S_\epsilon} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|. \quad (2)$$

3 Error Bounds

We show in Proposition 1 below that the truncation error bound $|\int_{-S_\epsilon} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}|$ is $O(\epsilon^{1-A})$ as $\epsilon \to 0$. We will use the construction from [1] and the Koksma-Hlawka inequality from [4] to provide an upper bound for the integration error over T^u_ϵ. That bound grows as $\epsilon \to 0$ and so to trade them off we will tune the way ϵ depends on n.

Proposition 1 *Under the regularity conditions* (1),

$$\left|\int_{-S_\epsilon} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \le \frac{2B\epsilon^{1-A}}{1-A}.$$

Proof We take the absolute value inside the integral and obtain

$$\int_{-S_\epsilon} |f(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} \le \int_{-S_\epsilon} B|x_1 - x_2|^{-A}\mathrm{d}\boldsymbol{x} \le B\int_0^1 2\int_0^\epsilon x_2^{-A}\mathrm{d}x_2\mathrm{d}x_1$$

from which the conclusion follows. □

Next we turn to the quadrature error over T^u_ϵ. Of course, T^d_ϵ is similar. The Koksma-Hlawka bound in [4] has

$$|\hat{\mu}_{\epsilon,u} - \mu_{\epsilon,u}| \le D^*_{T^u_\epsilon}(\boldsymbol{x}_{1,u},\dots,\boldsymbol{x}_{n,u})V_{T^u_\epsilon}(f)$$

where $D^*_{T_{u,\epsilon}}$ and $V_{T^u_\epsilon}$ are measures of discrepancy and variation suited to the triangle. Basu and Owen [1] provide a construction in which $D^*_{T^u_\epsilon} = O(\log(n)/n)$, the best possible rate.

Brandolini et al. [4, p. 46] provide a bound for $V_{T^u_\epsilon}$, the variation on the simplex as specialized to the triangle. To translate their bound into our setting, we introduce the notation $f_{rs} = \partial^{r+s}f/\partial^r x_1\partial^s x_2$. Specializing their bound to the domain T^u_ϵ we find that the variation is

$$
\begin{aligned}
O\Big(&|f(0,1)| + |f(0,\epsilon)| + |f(1-\epsilon,1)| \\
&+ \int_\epsilon^1 |f(0,x_2)|\mathrm{d}x_2 + \int_0^{1-\epsilon} |f(x_1,1)|\mathrm{d}x_1 + \int_0^{1-\epsilon} |f(x_1,x_1+\epsilon)|\mathrm{d}x_1 \\
&+ \int_\epsilon^1 |f_{01}(0,x_2)|\mathrm{d}x_2 + \int_0^{1-\epsilon} |f_{10}(x_1,1)|\mathrm{d}x_1 \\
&+ \int_0^{1-\epsilon} |f_{10}(x_1,x_1+\epsilon)|\mathrm{d}x_1 + \int_0^{1-\epsilon} |f_{01}(x_1,x_1+\epsilon)|\mathrm{d}x_1 \\
&+ \int_{T_\epsilon^u} |f(\boldsymbol{x})| + |f_{01}(\boldsymbol{x})| + |f_{10}(\boldsymbol{x})| + |f_{20}(\boldsymbol{x})| + |f_{02}(\boldsymbol{x})| + |f_{11}(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\Big)
\end{aligned}
\tag{3}
$$

as $\epsilon \to 0$. The implied constant in (3) includes their unknown constant C_2, the reciprocals of edge lengths of T_ϵ^u, the reciprocal of the area of T_ϵ^u, some small integers and some factors involving $\sqrt{2}(1-\epsilon)$, the length of the hypotenuse of T_ϵ^u.

Proposition 2 *Let f satisfy the regularity condition* (1)*. Then the trapezoidal variation of f over T_ϵ^u satisfies*

$$V_{T_\epsilon^u}(f) = O(\epsilon^{-A-1})$$

as $\epsilon \to 0$.

Proof Under condition (1),

$$|f(0,1)| + \int_\epsilon^1 |f(0,x_2)|\mathrm{d}x_2 + \int_0^{1-\epsilon} |f(x_1,1)|\mathrm{d}x_1 + \int_{T_\epsilon^u} |f(\boldsymbol{x})| = O(1).$$

Next

$$|f(0,\epsilon)| + |f(1-\epsilon,1)| + \int_0^{1-\epsilon} |f(x_1,x_1+\epsilon)|\mathrm{d}x_1 = O(\epsilon^{-A})$$

and

$$\int_\epsilon^1 |f_{01}(0,x_2)|\mathrm{d}x_2 + \int_0^{1-\epsilon} |f_{10}(x_1,1)|\mathrm{d}x_1 = O(\epsilon^{-A})$$

as well. Continuing through the terms, we find that

$$\int_0^{1-\epsilon} |f_{10}(x_1,x_1+\epsilon)|\mathrm{d}x_1 + \int_0^{1-\epsilon} |f_{01}(x_1,x_1+\epsilon)|\mathrm{d}x_1 = O(\epsilon^{-A-1}).$$

The remaining terms are integrals of absolute partial derivatives of f over T_ϵ^u. They are dominated by integrals of second derivatives and those terms obey the bound

$$\int_0^{1-\epsilon}\int_{x_1+\epsilon}^1 B_2|x_1-x_2|^{-A-2}\mathrm{d}x_2\mathrm{d}x_1 = O(\epsilon^{-A-1}).$$

□

Theorem 1 *Under the regularity conditions* (1), *we may choose* $\epsilon \propto \sqrt{\log(n)/n}$ *and get*

$$|\hat{\mu}-\mu| = O\Bigl(\Bigl(\frac{\log(n)}{n}\Bigr)^{(1-A)/2}\Bigr). \tag{4}$$

Proof From Propositions 1 and 2 we get

$$|\hat{\mu}-\mu| = O\Bigl(\epsilon^{1-A}+\frac{\log(n)}{n}\epsilon^{-1-A}\Bigr).$$

Taking ϵ to be a positive multiple of $\sqrt{\log(n)/n}$ yields the result. □

The choice of $\epsilon \propto \sqrt{\log(n)/n}$ optimizes the upper bound in (4).

4 Extension Based Approaches

Another approach to this problem is to construct a function $\tilde{f}$ where $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x})$ for $\boldsymbol{x} \in S_\epsilon$ and apply QMC to $\tilde{f}$. The function $\tilde{f}$ can smoothly bridge the gap between T_ϵ^u and T_ϵ^d. With such a function, the quadrature error satisfies

$$\Bigl|\frac{1}{n}\sum_{i=1}^n \tilde{f}(\boldsymbol{x}_i) - \int_{[0,1]^2} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Bigr| \le D_n^*(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)V_{\mathrm{HK}}(\tilde{f}) + \int_{-S_\epsilon}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} \tag{5}$$

where $V_{\mathrm{H}K}$ is total variation in the sense of Hardy and Krause.

Our regularity condition (1) allows for f to take the value ϵ^{-A} along the line $x_2 = x_1-\epsilon$ and to take the value $-\epsilon^{-A}$ along $x_2 = x_1+\epsilon$. By placing squares of side 2ϵ along the main diagonal we then find that the Vitali variation of an extension $\tilde{f}$ is at least $\lfloor(2\epsilon)^{-1}\rfloor 2\epsilon^{-A} \sim \epsilon^{-1-A}$. Therefore the Hardy-Krause variation of $\tilde{f}$ grows at least this quickly for some of the functions f that satisfy (1). More generally, for singular functions along a linear manifold M within $[0, 1]^d$, and no worse than $\mathrm{dist}(\boldsymbol{x}, M)^{-A}$, an extension over the region within ϵ of M could have a variation lower bound growing as fast as $\epsilon^{-(d-1)-A}$.

This result is much less favorable than the one for isolated point singularities [13]. For integrands on $[0,1]^d$ no worse than $\|\boldsymbol{x}-\boldsymbol{z}\|^{-A}$, where $\boldsymbol{z} \in [0,1]^d$, Sobol's low variation extension yields a function $\tilde{f}$ that agrees with f for $\|\boldsymbol{x}-\boldsymbol{z}\| \geq \epsilon > 0$ having $V_{\mathrm{H}K}(\tilde{f}) = O(\epsilon^{-A})$. Here we see that no extension can have such low variation for this type of singularity.

Owen [12] considers functions with singularities along the lower boundary of $[0,1]^d$ that are no worse than $\prod_{j=1}^{d} x_j^{-A_j}$. Sobol's extension from the region where $\prod_j x_j \geq \epsilon$ has variation $O(\epsilon^{-\max A_j})$ when the A_j are distinct (otherwise logarithmic factors enter). So that problem with singularities along the boundary also has a more accurate extension than can be obtained for singularities along the diagonal.

No extension $\tilde{f}$ from S_ϵ to $[0,1]^2$ can yield a bound (5) with a better rate than $O((\log n/n)^{(1-A)/2})$. To show this we first clarify one of the rules we impose on extensions. When we extend f from $\boldsymbol{x} \in S$ to values of $\boldsymbol{x} \notin S$ we do not allow the construction of $\tilde{f}$ to depend on $f(\boldsymbol{x})$ for $\boldsymbol{x} \notin S$. That is, we cannot peek outside the set we are extending from. Some such rule must be necessary or we could trivially get 0 error from an extension based on an oracle that uses the value of μ to define $\tilde{f}$. With our rule, any two functions f_1 and f_2 with $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ on S_ϵ have the same extension $\tilde{f}$. From the triangle inequality,

$$\max_{j=1,2}\left(\int_{-S_\epsilon} |\tilde{f}(\boldsymbol{x}) - f_j(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}\right) \geq \frac{1}{2}\int_{-S_\epsilon} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}.$$

Now let

$$f_1(\boldsymbol{x}) = \begin{cases} -|x_1 - x_2|^{-A}, & x_2 - x_1 > 0 \\ |x_1 - x_2|^{-A}, & x_2 - x_1 < 0, \end{cases}$$

and

$$f_2(\boldsymbol{x}) = \begin{cases} |x_1 - x_2|^{-A}, & x_2 - x_1 > 0 \\ \phi(x_2 - x_1), & 0 > x_2 - x_1 \geq -\epsilon \\ |x_1 - x_2|^{-A}, & -\epsilon > x_2 - x_1, \end{cases}$$

for a quadratic polynomial ϕ with $\phi(-\epsilon) = \epsilon^{-A}$, $\phi'(-\epsilon) = -A\epsilon^{-A-1}$, and $\phi''(-\epsilon) = A(A+1)\epsilon^{-A-2}$. Both f_1 and f_2 satisfy (1) and $\int_{-S_\epsilon} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}$ is larger than a constant times ϵ^{1-A}. That is the same rate as the truncation error from Proposition 1 and the quadrature error from this approach also attains the same rate as the error in Proposition 2. As a result, we conclude that even if we could construct the best extension $\tilde{f}$, it would not lead to a bound with a better rate than the one in Theorem 1.

5 Transformation

Here we consider applying a change of variable to move the singularity from the diagonal to an edge of the unit square. We focus on integrating $f(\boldsymbol{x})$ over $T^u = \{(x_1, x_2) \in [0, 1]^2 \mid 0 \le x_1 \le x_2 \le 1\}$ for f with a singularity no worse than $|x_1 - x_2|^{-A}$. The same strategy and same convergence rate hold on $T^d = \{(x_1, x_2) \in [0, 1]^2 \mid 0 \le x_2 \le x_1 \le 1\}$. Using a standard change of variable we have

$$\int_{T^u} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \frac{1}{2}\int_0^1 \int_0^1 f((1-u_1)\sqrt{u_2}, \sqrt{u_2})\mathrm{d}\boldsymbol{u},$$

which we then write as

$$\frac{1}{2}\int_{[0,1]^2} g(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \quad \text{for } g(\boldsymbol{u}) = f((1-u_1)\sqrt{u_2}, \sqrt{u_2}).$$

That is $g(\boldsymbol{u}) = f(\tau(\boldsymbol{u}))$ for a transformation $\tau : [0, 1]^2 \to T_u \subset [0, 1]^2$ given by $\tau_1(\boldsymbol{u}) = (1 - u_1)\sqrt{u_2}$ and $\tau_2(\boldsymbol{u}) = \sqrt{u_2}$.

The archetypal function with diagonal singularity satisfying Definition 1 is $f(\boldsymbol{x}) = |x_1 - x_2|^{-A}$. The corresponding function g for this f is

$$g(\boldsymbol{u}) = |\tau_1(\boldsymbol{u}) - \tau_2(\boldsymbol{u})|^{-A} = u_1^{-A} u_2^{-A/2}.$$

We see that the change of variable has produced an integrand with a singularity no worse than $u_1^{-A} u_2^{-A/2}$ according to Definition 2. Taking $\boldsymbol{u}_i$ to be the Halton points leads to a quadrature error at rate $O(n^{-1+\epsilon+A})$ for any $\epsilon > 0$, because Halton points (after the zeroth one) avoid the origin at a suitable rate [12, Corollary 5.6]. For this integrand g, randomized quasi-Monte Carlo points for will attain the mean error rate $\mathbb{E}(|\hat{\mu} - \mu|) = O(n^{-1+\epsilon+A})$ as shown in Theorem 5.7 of [12].

We initially thought that the conversion from a diagonal singularity to a lower edge singularity no worse than $u_1^{-A} u_2^{-A/2}$ would follow for other functions satisfying Definition 1. Unfortunately, that is not necessarily the case.

Let f be defined on $[0, 1]^2$ with a diagonal singularity no worse than $|x_1 - x_2|^{-A}$ for $0 < A < 1$. First,

$$|g(\boldsymbol{u})| = |f((1-u_1)\sqrt{u_2}, \sqrt{u_2})| \le B|u_1 u_2^{1/2}|^{-A}$$

which fits Definition 2. Similarly,

$$g_{10}(\boldsymbol{u}) = f_{10}(\tau_1(\boldsymbol{u}), \tau_2(\boldsymbol{u}))\frac{\partial \tau_1(\boldsymbol{u})}{\partial u_1} = O(|\tau_1 - \tau_2|^{-A-1}) \times u_2^{1/2} = O(u_1^{-A-1} u_2^{-A/2})$$

which also fits Definition 2. However,

$$\begin{aligned} g_{01}(\boldsymbol{u}) &= f_{10}(\tau(\boldsymbol{u}))\frac{\partial\tau_1(\boldsymbol{u})}{\partial u_2} + f_{01}(\tau(\boldsymbol{u}))\frac{\partial\tau_2(\boldsymbol{u})}{\partial u_2} \\ &= \bigl(f_{10}(\tau(\boldsymbol{u})) + f_{01}(\tau(\boldsymbol{u}))\bigr)\frac{1}{2}u_2^{-1/2} - f_{10}(\tau(\boldsymbol{u}))\frac{1}{2}u_1u_2^{-1/2}. \end{aligned} \tag{6}$$

Now f_{10} and f_{01} appearing in (6) are both $O(u_1^{-A-1}u_2^{-A/2-1/2})$. Therefore the two terms there are $O(u_1^{-A-1}u_2^{-A/2-1})$ and $O(u_1^{-A}u_2^{-A/2-1})$ respectively. The first term is too large by a factor of u_1^{-1} to suit Definition 2. We would need $(f_{01} + f_{10})(\tau(\boldsymbol{u}))$ to be only $O(u_1^{-A}u_2^{-A/2-1/2})$. Definition 1 is also not strong enough for g_{11} to be $O(u_1^{-A-1}u_2^{-A/2-1})$ as it would need to be under Definition 2. That definition yields only $O(u_1^{-A-2}u_2^{-A/2-1})$ without stronger assumptions. Theorem 2 below gives a sufficient condition where f is a modulated version of $|x_1 - x_2|^{-A}$.

Theorem 2 *Let $f(\boldsymbol{x}) = |x_1 - x_2|^{-A}h(\boldsymbol{x})$ for $\boldsymbol{x} \in [0,1]^2$ and $0 < A < 1$ where h and its first two derivatives are bounded. Then $g(\boldsymbol{u}) = f((1-u_1)\sqrt{u_2}, \sqrt{u_2})$ satisfies Definition 2 with $A_1 = A$ and $A_2 = A/2$.*

Proof We begin with

$$g(\boldsymbol{u}) = u_1^{-A}u_2^{-A/2}h((1-u_1)u_2^{1/2}, u_2^{1/2}) = O(u_1^{-1}u_2^{-A/2})$$

by boundedness of h. Next because u_1 is not in the second argument to h,

$$\begin{aligned} g_{10}(\boldsymbol{u}) &= -Au_1^{-A-1}u_2^{-A/2}h(\tau(\boldsymbol{u})) + u_1^{-A}u_2^{-A/2}h_{10}(\tau(\boldsymbol{u}))\partial\tau_1(\boldsymbol{u})/\partial u_1 \\ &= -Au_1^{-A-1}u_2^{-A/2}h(\tau(\boldsymbol{u})) - u_1^{-A}u_2^{-A/2+1/2}h_{10}(\tau(\boldsymbol{u})) \\ &= O(u_1^{-A-1}u_2^{-A/2}) \end{aligned}$$

as required. Similarly,

$$\begin{aligned} g_{01}(\boldsymbol{u}) &= -(A/2)u_1^{-A}u_2^{-A/2-1}h(\tau(\boldsymbol{u})) \\ &\quad + u_1^{-A}u_2^{-A/2}\bigl(h_{10}(\tau(\boldsymbol{u}))(1-u_1) + h_{01}(\tau(\boldsymbol{u}))\bigr)(1/2)u_2^{-1/2} \\ &= O(u_1^{-A}u_2^{-A/2-1}) \end{aligned}$$

as required. Finally,

$$\begin{aligned} g_{11}(\boldsymbol{u}) = {} & (A^2/2)u_1^{-A-1}u_2^{-A/2-1}h(\tau(\boldsymbol{u})) \\ & - (A/2)u_1^{-A}u_2^{-A/2-1}h_{10}(\tau(\boldsymbol{u}))(-u_2^{1/2}) \\ & - (A/2)u_1^{-A-1}u_2^{-A/2-1/2}\bigl(h_{10}(\tau(\boldsymbol{u}))(1-u_1) + h_{01}(\tau(\boldsymbol{u}))\bigr) \end{aligned}$$

$$
\begin{aligned}
&+ (u_1^{-A}u_2^{-A/2-1/2}/2)\big(-h_{10}(\tau(\boldsymbol{u})) + (1-u_1)h_{20}(\tau(\boldsymbol{u}))(-u_2^{1/2}) \\
&\qquad\qquad + h_{11}(\tau(\boldsymbol{u}))(-u_2^{1/2})\big) \\
&= O(u_1^{-A-1}u_2^{-A/2-1})
\end{aligned}
$$

as required. □

6 Discussion

We find that for an integrand with a singularity 'no worse than $|x_1 - x_2|^{-A}$' along the line $x_1 = x_2$ we can get a QMC estimate with error $O((\log(n)/n)^{(1-A)/2})$ by splitting the square into two triangles and ignoring a region in between them. The same method applies to singularities along the other diagonal of $[0,1]^2$. Moreover, the result extends to singularities along other lines intersecting the square. One can partition the square into rectangles, of which one has the singularity along the diagonal while the others have no singularity, and then integrate f over each of those rectangles.

That result does not directly extend to singularities along a linear manifold in $[0,1]^d$ for $d \geq 3$. The reason is that the QMC result for integration in the triangle from [1] has not been extended to the simplex. In a personal communication, Dimitry Bilyk told us that such an extension would imply a counterexample to the Littlewood conjecture, which is widely believed to be true. Basu and Owen [2] present some algorithms for RQMC over simplices, but they come without a Koksma-Hlawka bound that would be required for limiting arguments using sequences of simplices.

The rate $O((\log(n)/n)^{(1-A)/2})$ is a bit disappointing. We do much better by transforming the problem to place the singularity along the boundary of a square region, for then we can attain $O(n^{-1+\epsilon+A})$, under a stronger assumption that f is our prototypical singular function $|x_1 - x_2|^{-A}$ possibly modulated by a function h with bounded second derivatives on $[0,1]^2$. As a result we find that there is something to be gained by engineering QMC-friendly singularities in much the same way that benefits of QMC-friendly discontinuities have been found valuable by Wang and Sloan [15].

Acknowledgements This work was supported by the US National Science Foundation under grants DMS-1407397 and DMS-1521145. We thank two anonymous reviewers for helpful comments and we thank Frances Kuo for her work handling our paper.

References

1. Basu, K., Owen, A.B.: Low discrepancy constructions in the triangle. SIAM J. Num. An. **53**(2), 743–761 (2015)
2. Basu, K., Owen, A.B.: Scrambled geometric net integration over general product spaces. Found. Comput. Math. **17**(2), 1–30 (2015)
3. Binder, C.: Über einen Satz von de Bruijn und Post. Öst. Akad. der Wiss. Math.-Natur. Klasse. Sitz. Abteilung II **179**, 233–251 (1970)
4. Brandolini, L., Colzani, L., Gigante, G., Travaglini, G.: A Koksma–Hlawka inequality for simplices. In: Trends in Harmonic Analysis, pp. 33–46. Springer, Berlin (2013)
5. de Bruijn, N.G., Post, K.A.: A remark on uniformly distributed sequences and Riemann integrability. Indag. Math. **30**, 149–150 (1968)
6. Dick, J., Pillichshammer, F.: Digital Sequences, Discrepancy and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
7. Klinger, B.: Discrepancy of point sequences and numerical integration. Ph.D. thesis, Technische Universität Graz (1997)
8. Klinger, B.: Numerical integration of singular integrands using low-discrepancy sequences. Computing **59**, 223–236 (1997)
9. Mishra, M., Gupta, N.: Application of quasi Monte Carlo integration technique in EM scattering from finite cylinders. Prog. Electromagn. Res. Lett. **9**, 109–118 (2009)
10. Owen, A.B.: Monte Carlo variance of scrambled equidistribution quadrature. SIAM J. Numer. Anal. **34**(5), 1884–1910 (1997)
11. Owen, A.B.: Multidimensional variation for quasi-Monte Carlo. In: Fan, J., Li, G. (eds.) International Conference on Statistics in Honour of Professor Kai-Tai Fang's 65th Birthday (2005)
12. Owen, A.B.: Halton sequences avoid the origin. SIAM Rev. **48**, 487–583 (2006)
13. Owen, A.B.: Quasi-Monte Carlo for integrands with point singularities at unknown locations. In: Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 403–417. Springer, Berlin (2006)
14. Sobol', I.M.: Calculation of improper integrals using uniformly distributed sequences. Sov. Math. Dokl. **14**(3), 734–738 (1973)
15. Wang, X., Sloan, I.H.: Quasi-Monte Carlo methods in financial engineering: an equivalence principle and dimension reduction. Oper. Res. **59**(1), 80–95 (2011)

There Is No Strongly Regular Graph with Parameters (460, 153, 32 , 60)

Andriy Bondarenko, Anton Mellit, Andriy Prymak, Danylo Radchenko, and Maryna Viazovska

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We prove that there is no strongly regular graph (SRG) with parameters $(460, 153, 32, 60)$. The proof is based on a recent lower bound on the number of 4-cliques in a SRG and some applications of Euclidean representation of SRGs.

1 Introduction

A finite, undirected, simple graph $G = (V, E)$ with vertices V and edges E is called *strongly regular* with parameters (v, k, λ, μ) if G is k-regular on v vertices, and, in addition, any two adjacent vertices of G have exactly λ common neighbors, while any two non-adjacent vertices of G have exactly μ common neighbors.

The parameters (v, k, λ, μ) of a *SRG* must satisfy certain known conditions (see [5]), but in general it is an open question to determine parameters (v, k, λ, μ) for which strongly regular graphs (SRGs) exist, and, in case when they do exist, to classify such graphs. A list of known results for $v \leq 1300$ is maintained at [3].

A. Bondarenko (✉)
Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

A. Mellit
IST Austria, Klosterneuburg, Austria

A. Prymak
Department of Mathematics, University of Manitoba, Winnipeg, MB, Canada

D. Radchenko
International Centre for Theoretical Physics, Trieste, Italy

M. Viazovska
Humboldt University of Berlin, Berlin Mathematical School, Berlin, Germany

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_7

As an application of a recently established lower bound on the number of 4-cliques in a SRG (see [1] and [2]) and of Euclidean representation of SRGs, we obtain the following non-existence result in a very special case.

Theorem 1 *There is no strongly regular graph with parameters* $(460, 153, 32, 60)$.

Some general background on SRGs can be found in [4, Chapter 9] and [6], while [4, Chapter 8] and [5] contain details on Euclidean representation of SRGs. The argument with the Gram matrix used here has been extensively applied in [1].

2 Proof of Theorem 1

Assume that a SRG $G = (V, E)$ with parameters $(v, k, \lambda, \mu) = (460, 153, 32, 60)$ exists. Then adjacency matrix A of G satisfies the equations $AJ = 153J$ and $A^2 + 28A - 93I = 60J$, where I is the identity matrix and J is the matrix having all entries equal to 1. Consequently, A has the following spectrum: $153^1\, 3^{414}\, (-31)^{45}$. The Euclidean representation of G defines a mapping $V \ni u \mapsto x_u \in \mathbb{R}^{45}$ such that all x_u are unit vectors and the dot products between these vectors depend only on the adjacency between the corresponding vertices of G. More precisely, for two different vertices $u, w \in V$, we have

$$\langle x_u, x_w \rangle = \begin{cases} p, & \text{if } u \text{ is adjacent to } w, \\ q, & \text{if } u \text{ is not adjacent to } w, \end{cases} \qquad \text{where} \quad \begin{array}{l} p = \frac{-31}{153}, \\ q = \frac{5}{51}, \end{array}$$

and $\langle x, y \rangle$ is the Euclidean dot product in $\mathbb{R}^{45}$.

The first step is to show that G has at least 228,111 complete subgraphs of size 4. This follows from [1] (see also the bounds on the number of 4-cliques in [2]). Let us give a brief outline of the proof. For each edge $e = \{u, w\} \in E$ we consider the unit vector $y_e = \frac{x_u + x_w}{\|x_u + x_w\|}$. A simple calculation shows that the *distribution* of dot products between $\{x_u\}_{u \in V}$ and $\{y_e\}_{e \in E}$ depends only on the SRG parameters and the number of 4-cliques. Since the Gegenbauer polynomials $C_t^{(d-2)/2}(x)$ are positive definite on S^{d-1}, by applying them to the Gram matrix of $\{x_u\}_{u \in V} \cup \{y_e\}_{e \in E}$ we get a positive definite matrix (here $d = 45$, $t = 4$). To get an inequality for the number of 4-cliques we simply compute the value of the corresponding quadratic form on the vector that takes value 1 on x_u's and a on y_e's and optimize the parameter a.

Next, for any two adjacent vertices $u, w \in V$ (i.e., $\{u, w\} \in E$), let $V_{u,w}$ be the set of vertices $t \in V$ adjacent to both u and w. Note that any pair of adjacent vertices in $V_{u,w}$ forms a 4-clique together with u and w. Now choose u, w so that the number of edges in the subgraph of G induced by $V_{u,w}$ is largest possible (among all possible edges $\{u, w\} \in E$). Let $\mathscr{V} := V_{u,w}$, $\widetilde{G}$ be the subgraph of G induced by $\mathscr{V}$, and let m be the number of edges in $\widetilde{G}$. Since G has 35,190 edges, we get the following inequality on m from the lower bound on the number of 4-cliques: $m \geq \frac{6 \cdot 228111}{35190}$, so $m \geq 39$. For an upper bound on m, we will make use of the above Euclidean representation. Define $X_1 := \sum_{t\in\mathscr{V}} x_t$ and $X_2 := x_u + x_w$. The Gram matrix $M := (\langle X_i, X_j\rangle)_{i,j=1}^2$ is positive semi-definite, therefore, $\det M \geq 0$. Explicitly, in terms of m and graph parameters we have

$$M = \begin{pmatrix} \lambda + 2mp + (\lambda^2 - \lambda - 2m)q & 2\lambda p \\ 2\lambda p & 2 + 2p \end{pmatrix} = \frac{1}{153^2}\begin{pmatrix} 19776 - 92m & -1984 \\ -1984 & 244 \end{pmatrix}.$$

The inequality $\det M \geq 0$ leads to $m \leq \frac{2416}{61}$, therefore $m \leq 39$.

Thus, $m = 39$, and the graph $\widetilde{G}$ on 32 vertices has 39 edges. Let $\mathscr{W}$ be a set of 14 vertices of $\widetilde{G}$ which have the largest degrees (in $\widetilde{G}$). We claim that the sum α of the degrees of vertices of $\mathscr{W}$ in $\widetilde{G}$ is at least 42. Indeed, if each such degree is at least 3, then we are clearly done. Otherwise, the sum of the degrees of vertices not in $\mathscr{W}$ is at most $(32 - 14)2 = 36$, which means that $\alpha \geq 2 \cdot 39 - 36 = 42$. Denote by β the number of edges in the subgraph induced by $\mathscr{W}$. Then we have $\alpha - 2\beta$ edges between $\mathscr{W}$ and $\mathscr{V} \setminus \mathscr{W}$, and $39 + \beta - \alpha$ edges in $\mathscr{V} \setminus \mathscr{W}$. We take $Y_1 := \sum_{t\in\mathscr{V}\setminus\mathscr{W}} x_t$, $Y_2 := \sum_{t\in\mathscr{W}} x_t$, and $Y_3 := x_u + x_w$ and apply previous considerations. For the Gram matrix $\widetilde{M} := (\langle Y_i, Y_j\rangle)_{i,j=1}^3$ we clearly have $\det\widetilde{M} \geq 0$. On the other hand, we compute

$$\begin{aligned}
\langle Y_1, Y_1\rangle &= 18 + 2(39 + \beta - \alpha)p + (18 \cdot 17 - 2(39 + \beta - \alpha))q, \\
\langle Y_2, Y_2\rangle &= 14 + 2\beta p + (14 \cdot 13 - 2\beta)q, \\
\langle Y_1, Y_2\rangle &= (\alpha - 2\beta)p + (18 \cdot 14 - (\alpha - 2\beta))q, \\
\langle Y_1, Y_3\rangle &= 18 \cdot 2p, \quad \langle Y_2, Y_3\rangle = 14 \cdot 2p, \quad \langle Y_3, Y_3\rangle = 2 + 2p,
\end{aligned}$$

and therefore

$$\begin{aligned}
\det\widetilde{M} &= \left(-\frac{516304}{3581577}\alpha^2 + \frac{35785792}{3581577}\alpha\right) - \left(\frac{1252672}{3581577}\beta + \frac{198599296}{1193859}\right) \\
&=: \phi(\alpha) - \psi(\beta).
\end{aligned}$$

The quadratic function ϕ is decreasing for $\alpha \geq \frac{35785792}{2\cdot 516304} = \frac{2114}{61}$, in particular for $\alpha \geq 42$. The linear function ψ is clearly increasing. Since $39 + \beta - \alpha \geq 0$, we have $\beta \geq 3$. Now, since

$$0 \leq \det\widetilde{M} = \phi(\alpha) - \psi(\beta) \leq \phi(42) - \psi(3) = \frac{-270848}{132651} < 0,$$

we get a contradiction and hence Theorem 1 is proved.

3 Conclusion

Let us remark that the exact same reasoning from the proof above can be applied to some other strongly regular graphs. For instance, with some trivial changes we obtain non-existence of strongly regular graphs with parameters $(5929, 1482, 275, 402)$ and $(6205, 858, 47, 130)$, for which the number of 4-cliques is bounded from below by 4805 and 113 respectively. The key property that these three graphs have in common is that they have a very small (but strictly positive) value of the Krein parameter q_{22}^2 (see [4, Chapter 11] for the definition). The proof also goes through for some strongly regular graphs that satisfy $q_{22}^2 = 0$, or equivalently

$$(s + 1)(k + s + rs) = (k + s)(r + 1)^2,$$

where $r > 0$ and $s < 0$ are eigenvalues of the adjacency matrix. In this case the above reasoning shows that all λ-subgraphs must be regular. The smallest set of parameters that can be ruled out in this way is $(2950, 891, 204, 297)$. Alternatively, the non-existence in this case can be shown by noting that the first subconstituent must be strongly regular, but there exist no strongly regular graphs on 891 vertices of degree 204 (since there are no feasible parameters with $v = 891$ and $k = 204$).

References

1. Bondarenko, A., Prymak, A., Radchenko, D.: Non-existence of (76, 30, 8, 14) strongly regular graph and some structural tools. Linear Algebra Appl. **527**, 53–72 (2017)
2. Bondarenko, A., Prymak, A., Radchenko, D.: Supplementary files for the proof of non-existence of SRG(76,30,8,14). Preprint available at http://prymak.net/SRG-76-30-8-14/
3. Brouwer, A. E.: Parameters of strongly regular graphs, Electronically published tables. http://www.win.tue.nl/~aeb/graphs/srg/srgtab.html
4. Brouwer, A.E., Haemers, W.H.: Spectra of Graphs. Universitext. Springer, Berlin (2012)
5. Brouwer, A.E, van Lint, J.H.: Strongly regular graphs and partial geometries. In: Proceedings of the Conference Enumeration and Design, Waterloo, ON, 1982, pp. 85–122, Academic, Toronto, ON (1984)
6. Cameron, P. J.: Strongly Regular Graphs. Topics in Algebraic Graph Theory, Cambridge University Press, Cambridge (2004)

Low-Discrepancy Sequences for Piecewise Smooth Functions on the Torus

Luca Brandolini, Leonardo Colzani, Giacomo Gigante, and Giancarlo Travaglini

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We produce low-discrepancy infinite sequences which can be used to approximate the integral of a smooth periodic function restricted to a smooth convex domain with positive curvature in $\mathbb{R}^d$. The proof depends on simultaneous Diophantine approximation and on appropriate estimates of the decay of the Fourier transform of characteristic functions.

1 Introduction

In [6], the following version of the classical Koksma-Hlawka inequality has been shown. Its main feature is the possibility to numerically approximate the integral of piecewise smooth functions, which have discontinuities along rather general hypersurfaces in $\mathbb{R}^d$.

Theorem 1 *Let $h(t) = f(t)\,\chi_\Omega(t)$, where f is a smooth $\mathbb{Z}^d$-periodic function on $\mathbb{R}^d$ and χ_Ω is the characteristic function of a bounded Borel set Ω in $\mathbb{R}^d$. Let $1 \leq p, q \leq +\infty$, $1/p + 1/q = 1$. Let us call the quantity*

$$V_q(f) := \sum_{\alpha\in\{0,1\}^d} 2^{d-|\alpha|} \left\| \left(\frac{\partial}{\partial t}\right)^{\alpha} f \right\|_{L^q(\mathbb{T}^d)},$$

L. Brandolini · G. Gigante (✉)
Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione, Università degli Studi di Bergamo, Dalmine, Italy
e-mail: luca.brandolini@unibg.it; giacomo.gigante@unibg.it

L. Colzani · G. Travaglini
Dipartimento di Matematica e Applicazioni, Università di Milano-Bicocca, Milano, Italy
e-mail: leonardo.colzani@unimib.it; giancarlo.travaglini@unimib.it

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_8

q-variation *of the function* f. *Let* $\{t(j)\}_{j=1}^{N} \subset \mathbb{R}^d$, *for any* $s \in (0,1)^d$ *and for any* $x \in \mathbb{R}^d$ *let*

$$I(s,x) = \cup_{m\in\mathbb{Z}^d}\left([0,s_1]\times\ldots\times[0,s_d]+x+m\right)$$

be the periodization of the axis parallel box anchored at x *and with edges given by* s, *and let us call the quantity*

$$D_p\left(\Omega,\{t(j)\}_{j=1}^{N}\right) := \left\| \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{I(s,x)\cap\Omega}\left(t\left(j\right)+m\right) - \left|I\left(s,x\right)\cap\Omega\right| \right\|_{L^p\left((0,1)^d\times\mathbb{T}^d,\, ds\, dx\right)}$$

the p-discrepancy *of the point set* $\{t\left(j\right)+m\}_{j=1,m\in\mathbb{Z}^d}^{N}$ *with respect to the set* Ω. *Then*

$$\left|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d} h\left(t\left(j\right)+m\right) - \int_{\mathbb{R}^d} h\left(t\right)\,dt\right| \le V_q(f)\,D_p\left(\Omega,\{t(j)\}_{j=1}^{N}\right) .$$

Here we are interested in the case when Ω is a convex set. Of course, when $d=1$, a convex set Ω is just an interval, and therefore the error in numerical integration can be efficiently estimated by means of the classical Koksma inequality, see e.g. [13]. For this reason from now on we will assume $d\ge 2$.

Throughout the paper we will make use of the following *localized discrepancy*

$$D\left(\Omega,\{t(j)\}_{j=1}^{N},s,x\right) := \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{I(s,x)\cap\Omega}\left(t\left(j\right)+m\right) - \left|I\left(s,x\right)\cap\Omega\right|.$$

In order to introduce our results, we need some definitions.

Fix $a,b>0$. Define $\mathscr{O}(a,b)$ as the class of all smooth compact convex sets Ω that can be written as $\left\{x\in\mathbb{R}^d : \Phi(x)\le 0\right\}$ for some function $\Phi(x)$ with continuous derivatives up to the order $[(d+3)/2]$ (the integer part of $(d+3)/2$), and with the following properties:

1. The Gaussian curvature of $\partial\Omega$ is greater than a at every point.
2. $|\nabla\Phi(x)|\ge 1$ when $\Phi(x)=0$, and $|D^\alpha\Phi(x)|\le b$ for every $x\in\mathbb{R}^d$ and every multiindex α with $|\alpha|\le[(d+3)/2]$.

Definition 1 Let Ω be a non-empty compact subset of $\mathbb{R}^d$. The signed distance function δ_Ω is defined by

$$\delta_\Omega(x) = \begin{cases} \mathrm{dist}(x,\,\partial\Omega) & \text{if } x\in\Omega, \\ -\mathrm{dist}(x,\,\partial\Omega) & \text{if } x\notin\Omega. \end{cases}$$

For any real number u, define

$$\Omega^u = \left\{x \in \mathbb{R}^d : \delta_\Omega(x) \geq u\right\}.$$

It is easy to see that the signed distance function is Lipschitz continuous with constant 1, and it can be shown that when $\partial\Omega$ is smooth, then δ_Ω is smooth in a suitable neighborhood of $\partial\Omega$ (see [10, Section 14.6]). In particular, a suitable modification of this signed distance function away from the boundary can be taken as a natural choice for the defining function Φ of the set Ω.

The meaning of the above definitions comes from the following classical result.

Lemma 1 *For every $a, b > 0$ there exists $c > 0$ such that for every Ω in $\mathscr{O}(a, b)$ and every $\xi \in \mathbb{R}^d$ the Fourier transform of χ_Ω satisfies the estimate*

$$|\widehat{\chi}_\Omega(\xi)| \leq c\,(1 + |\xi|)^{-(d+1)/2}.$$

Proof The decay of the Fourier transform for a fixed single set can be found in [11, 12, 18]. A careful reading of the proof shows that the above constant c can be chosen independently of the set Ω as long as the Gaussian curvature is bounded from below, and a finite number of derivatives of the defining function are bounded.

□

Our main result is the following.

Theorem 2 *Assume that $\alpha_1, \ldots, \alpha_d$ are real algebraic numbers and assume that $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over $\mathbb{Q}$. Set $\alpha = (\alpha_1, \ldots, \alpha_d)$. Then for any $\eta, a, b > 0$ there is a constant c such that*

$$\sup_{\Omega \in \mathscr{O}(a,b)} \sup_{y \in \mathbb{R}^d} \sup_{s \in (0,1)^d} \sup_{x \in \mathbb{R}^d} \left| D\left(\Omega, \{j\alpha + y\}_{j=1}^N, s, x\right) \right| \leq c\, N^{-\frac{2}{d+1}+\eta}. \tag{1}$$

The above constant c depends on η, a, b, and on α.

In [7] a sharper version of this result has been proved in the two-dimensional case. In particular, a decay of order $N^{-2/3}\log N$ can be obtained under the hypothesis that the numbers $\{1, \alpha_1, \alpha_2\}$ form a basis of a number field of degree 3 over $\mathbb{Q}$. The difference with the above theorem can be explained as follows. One key ingredient in the proof of this type of results is an estimate of the decay of the Fourier transform of the characteristic function of the intersection of an axis parallel box I with a set Ω with smooth boundary and nonvanishing curvature. While in the two-dimensional case it is possible to give the sharp estimate

$$|\widehat{\chi}_{\Omega\cap I}(\xi)| \leq c\frac{1}{(1 + |\xi|)^{3/2}} + c\frac{1}{(1 + |\xi_1|)}\frac{1}{(1 + |\xi_2|)},$$

(see [7, Lemma 11]), in the general d-dimensional case we are only able to prove the following weaker, though simpler, estimate contained in Lemma 2,

$$|\widehat{\chi}_{\Omega\cap I}(\xi)| \leq c\prod_{i=1}^{d}\frac{\log\left(2+|\xi_i|\right)}{\left(1+|\xi_i|\right)^{\frac{d+1}{2d}}}.$$

There is a second difference between the proof given here and the proof in [7]. In order to properly apply the Fourier analytic methods, it is necessary to approximate from above and from below characteristic functions with smooth functions. In [7] this was obtained by means of entire functions of finite exponential type, and the discrepancy was estimated using a generalization of the Erdős-Turan inequality proved in [8]. This part has been simplified here, where the approximation is obtained by taking the convolution of the inner (and the outer) parallel body with a bump function with small support.

Another observation about the above result concerns the exponent $-2/\left(d+1\right)$. It follows from a result of J. Beck [3] and H. Montgomery [14] (see also Theorem 4.1 in [5]) that for any collection $\{t\left(j\right)\}_{j=1}^{N}$ of N points in the unit box there is a set $\Omega \in \mathscr{O}(a,b)$ contained in the unit box such that the localized discrepancy

$$\left|D\left(\Omega,\{t\left(j\right)\}_{j=1}^{N},\left(1,\ldots,1\right),\left(0,\ldots,0\right)\right)\right| \geq c\,N^{-\frac{d+1}{2d}}.$$

Therefore the exponent $-\left(d+1\right)/\left(2d\right)$ is the best one can hope to have in a result like Theorem 2. It should also be emphasized that if one replaces the collection $\mathscr{O}(a,b)$ with a smaller collection, things may change drastically. In particular, it has been proved in [1], see also [2], that if Ω is a Borel measurable set with positive Lebesgue measure contained in the unit box, then there is a collection $\{t\left(j\right)\}_{j=1}^{N}$ of N points in the unit box such that the ∞-discrepancy satisfies the bound

$$D_{\infty}\left(\Omega,\{t\left(j\right)\}_{j=1}^{N}\right) \leq cN^{-1}\left(\log N\right)^{3d+1}.$$

Unfortunately, this result is not constructive.

In an attempt to approach the optimal exponent $-\left(d+1\right)/2d$, we will show the following result, where the supremum in the variable y is replaced with an L^p norm in this variable. This allows us to replace the exponent $-2/(d+1)$ in (1) with the better exponent $-(d+1)/2d$.

Theorem 3 *Assume that $\alpha_1,\ldots,\alpha_d$ are real algebraic numbers and assume that $1,\alpha_1,\ldots,\alpha_d$ are linearly independent over $\mathbb{Q}$. Set $\alpha = (\alpha_1,\ldots,\alpha_d)$. Then for any $1 \leq p \leq +\infty$ and $\eta,a,b > 0$ there is a constant c such that the localized* *discrepancy satisfies*

$$\sup_{\Omega\in\mathscr{O}(a,b)}\ \sup_{s\in(0,1)^d}\ \sup_{x\in\mathbb{R}^d}\left\|D\left(\Omega,\{j\alpha+y\}_{j=1}^N,x,s\right)\right\|_{L^p\left(\mathbb{T}^d,\,dy\right)}$$
$$\leq\begin{cases} c\,N^{-\frac{d+1}{2d}+\eta} & \textit{if}\ 1\leq p\leq 2d/\left(d-1\right),\\ c\,N^{-\frac{2}{d+1}-\frac{d-1}{(d+1)p}+\eta} & \textit{if}\ 2d/\left(d-1\right)\leq p\leq+\infty.\end{cases}$$

The above constant c depends on p,η,a,b, and on α.

Theorem 3 has an immediate application to numerical integration. Indeed, translating by y the sequence $\{j\alpha\}$ and integrating, by Theorem 1 one obtains

$$\left\|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}h\left(j\alpha+y+m\right)-\int_{\mathbb{R}^d}h\left(t\right)\,dt\right\|_{L^p\left(\mathbb{T}^d,dy\right)}$$
$$\leq V_q(f)\ \left\|D_p\left(\Omega,\{j\alpha+y\}_{j=1}^N\right)\right\|_{L^p\left(\mathbb{T}^d,dy\right)}.$$

By Theorem 3, for any Ω in $\mathscr{O}(a,b)$

$$\left\|D_p\left(\Omega,\{j\alpha+y\}_{j=1}^N\right)\right\|_{L^p\left(\mathbb{T}^d,dy\right)}$$
$$=\left(\int_{\mathbb{T}^d}\int_{(0,1)^d}\int_{\mathbb{T}^d}\left|D\left(\Omega,\{j\alpha+y\}_{j=1}^N,s,x\right)\right|^p\,dy\,ds\,dx\right)^{1/p}$$
$$\leq\sup_{s\in(0,1)^d}\ \sup_{x\in\mathbb{R}^d}\left(\int_{\mathbb{T}^d}\left|D\left(\Omega,\{j\alpha+y\}_{j=1}^N,s,x\right)\right|^p\,dy\right)^{1/p}$$
$$\leq\begin{cases} cN^{-\frac{d+1}{2d}+\eta} & 1\leq p\leq\frac{2d}{d-1},\\ cN^{-\frac{2}{d+1}-\frac{d-1}{(d+1)p}+\eta} & \frac{2d}{d-1}\leq p\leq+\infty.\end{cases}$$

Therefore, Theorems 2 and 3 imply the following result.

Theorem 4 *Assume that $\alpha_1,\ldots,\alpha_d$ are real algebraic numbers and assume that $1,\alpha_1,\ldots,\alpha_d$ are linearly independent over $\mathbb{Q}$. Set $\alpha=(\alpha_1,\ldots,\alpha_d)$. Let $1\leq p,q\leq+\infty$, $1/p+1/q=1$. Then for any $\eta,a,b>0$ there is a constant c such that if f is a smooth $\mathbb{Z}^d$-periodic function on $\mathbb{R}^d$ then*

$$\sup_{\Omega\in\mathscr{O}(a,b)}\left\|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}f\left(j\alpha+y+m\right)\chi_\Omega\left(j\alpha+y+m\right)-\int_\Omega f\left(t\right)\,dt\right\|_{L^p\left(\mathbb{T}^d,dy\right)}$$
$$\leq\begin{cases} cV_q(f)N^{-\frac{d+1}{2d}+\eta} & \textit{if}\ 1\leq p\leq\frac{2d}{d-1},\\ cV_q(f)N^{-\frac{2}{d+1}-\frac{d-1}{(d+1)p}+\eta} & \textit{if}\ \frac{2d}{d-1}\leq p\leq+\infty.\end{cases}$$

Observe that the above result gives essentially the desired exponent whenever $1 \leq p \leq 2d/\,(d-1)$. Also observe that the exponent $(d+1)/2d$ is larger than $1/2$ for all d, while the exponent $2/(d+1) + (d-1)/((d+1)p)$ is larger than $1/2$ for all $p \leq +\infty$ when $d = 2$, for $p < +\infty$ when $d = 3$, and for $p < 2\,(d-1)\,/\,(d-3)$ when $d > 3$. Thus, in these ranges, the above estimates beat the decay $N^{-1/2}$ for the mean square error given by the classical Monte Carlo method with independent uniformly distributed sampling points.

As usual, in what follows we shall denote by c a constant that may vary from line to line.

2 Proofs and Auxiliary Results

One of the main ingredients is the following lemma, to be compared with Lemma 1.

Lemma 2 *For any $a, b > 0$ there exists $c > 0$ such that for any Ω in $\mathscr{O}(a,b)$ and for any hyper-rectangle I with edges of length at most 2 and parallel to the axes, and for any $n = (n_1, \ldots, n_d) \in \mathbb{R}^d$,*

$$|\widehat{\chi}_{\Omega\cap I}(n)| \leq c \prod_{i=1}^{d} \frac{\log(2+|n_i|)}{(1+|n_i|)^{\frac{d+1}{2d}}}. \tag{2}$$

Proof Since $\chi_{\Omega\cap I} = \chi_\Omega \chi_I$ and $\widehat{\chi_\Omega \chi_I} = \widehat{\chi}_\Omega * \widehat{\chi}_I$, then

$$|\widehat{\chi}_{\Omega\cap I}(n)| \leq (|\widehat{\chi}_\Omega| * |\widehat{\chi}_I|)\,(n)\,.$$

By Lemma 1

$$|\widehat{\chi}_\Omega\,(n)| \leq c \frac{1}{(1+|n|)^{\frac{d+1}{2}}},$$

while an explicit computation of the Fourier transform of χ_I as the product of d one-dimensional Fourier transforms of characteristic functions of intervals gives

$$|\widehat{\chi}_I\,(n)| \leq c \prod_{i=1}^{d} \frac{1}{1+|n_i|},$$

where c is a constant depending only on the dimension, since the edges of I have length at most 2. Since

$$1+|x| \geq c\,(1+|x_1|+\ldots+|x_d|) \geq c\,(1+|x_1|)^{1/d}\ldots(1+|x_d|)^{1/d}\,,$$

then

$$(|\widehat{\chi}_\Omega| * |\widehat{\chi}_I|)(n) \le c \int_{\mathbb{R}^d} \left(\prod_{i=1}^{d} \frac{1}{1+|n_i - x_i|} \right) \frac{1}{(1+|x|)^{\frac{d+1}{2}}} dx$$
$$\le c \prod_{i=1}^{d} \int_{\mathbb{R}} \frac{1}{1+|n_i - x_i|} \frac{1}{(1+|x_i|)^{\frac{d+1}{2d}}} dx_i$$

Thus, if $|n_i| \le 1$, then

$$\int_{\mathbb{R}} \frac{1}{1+|n_i - x_i|} \frac{1}{(1+|x_i|)^{\frac{d+1}{2d}}} dx_i \le c \int_{\mathbb{R}} \frac{1}{1+|x_i|} \frac{1}{(1+|x_i|)^{\frac{d+1}{2d}}} dx_i \le c,$$

while if $n_i \ge 1$, then

$$\int_{\mathbb{R}} \frac{1}{1+|n_i - x_i|} \frac{1}{(1+|x_i|)^{\frac{d+1}{2d}}} dx_i \le c \int_{0}^{n_i/2} \frac{1}{n_i} \frac{1}{(1+x_i)^{\frac{d+1}{2d}}} dx_i$$
$$+ c \int_{n_i/2}^{2n_i} \frac{1}{1+|n_i - x_i|} \frac{1}{n_i^{\frac{d+1}{2d}}} dx_i + c \int_{2n_i}^{+\infty} \frac{1}{x_i^{\frac{3d+1}{2d}}} dx_i \le c n_i^{-\frac{d+1}{2d}} \log(2+n_i).$$

□

Definition 2 Let B be the closed unit ball centered at the origin. If K is a convex body in $\mathbb{R}^d$, then the outer parallel body of K at distance r is defined as the Minkowski sum of K and rB,

$$K + rB = \{x + y : x \in K,\ |y| \le r\},$$

while the inner parallel body of K at distance r is defined as the Minkowski difference of K and rB,

$$K \div rB = \{x : x + rB \subset K\}.$$

Lemma 3 *Let K be a convex body in $\mathbb{R}^d$ and let K^u be as in Definition 1.*

(i) For any real number u, the set K^u is the outer or the inner parallel body of K at distance $|u|$, according to whether u is negative or positive, that is

$$K^u = K + |u|B,\ \textit{if}\ u \le 0,$$
$$K^u = K \div uB,\ \textit{if}\ u > 0.$$

(ii) For any real number u, the set K^u is convex (possibly empty).

(iii) If M is another convex body, then

$$(M \cap K)^u = M^u \cap K^u \text{ if } u \geq 0,$$
$$(M \cap K)^u \subset M^u \cap K^u \text{ if } u < 0.$$

Proof Points (i) and (iii) follow easily from the definitions, while the proof of (ii) can be found in [17, Chapter 3]. □

Lemma 4 *For every $a, b > 0$ there exists $\varepsilon > 0$ such that if Ω is in $\mathscr{O}(a, b)$ and if $|u| < \varepsilon$ then Ω^u is in $\mathscr{O}(a/2, 2b)$. In particular, Lemma 1 gives uniform estimates on the decay of the Fourier transform of Ω^u.*

Proof This is essentially a reformulation of Lemmas 14.16 and 14.17 in [10] for the case of convex bodies. In particular, in those lemmas it is proved that if Ω is a convex body in $\mathbb{R}^d$ with $\mathscr{C}^k$ boundary, if $\kappa_{\max}$ is the maximum of all the principal curvatures of $\partial\Omega$, and if one defines

$$\Gamma = \Gamma(\Omega, \kappa_{\max}) = \{x : -(2\kappa_{\max})^{-1} < \delta_\Omega(x) < (2\kappa_{\max})^{-1}\},$$

then $\delta_\Omega \in \mathscr{C}^k(\Gamma)$ and $|\nabla \delta_\Omega| = 1$ in Γ. Furthermore, the level set

$$\Omega_u = \left\{x \in \mathbb{R}^d : \delta_\Omega(x) = u\right\}$$

is $\mathscr{C}^k$ whenever $|u| < (2\kappa_{\max})^{-1}$ and its principal curvatures at a point x are given by

$$\kappa_i(x) = \frac{\kappa_i(y)}{1 - u\kappa_i(y)}, i = 1, \ldots, d-1,$$

where y is the unique point of $\partial\Omega$ such that $\text{dist}(x, y) = |u|$ and $\kappa_i(y)$ are the principal curvatures of $\partial\Omega$ at y. □

We are now ready to proceed with the proof of Theorem 2.

Proof of Theorem 2 Let us fix the set Ω, the translation parameters x and y in $\mathbb{R}^d$, and the shape parameter $s \in (0, 1)^d$. Call $m_1, \ldots, m_Q$ the lattice points for which the sets

$$([0, s_1] \times \ldots \times [0, s_d] + x - y + m_i) \cap (\Omega - y)$$

are nonempty, and for $i = 1, \ldots, Q$, let

$$K_i = ([0, s_1] \times \ldots \times [0, s_d] + x - y + m_i) \cap (\Omega - y),$$

then of course

$$\cup_{m\in\mathbb{Z}^d}\left(([0,s_1]\times\ldots\times[0,s_d]+x-y+m)\cap(\Omega-y)\right)=\cup_{i=1}^{Q}K_i.$$

The number Q is bounded by the maximum number of unit cubes with integer vertices that intersect any given translate of Ω in $\mathbb{R}^d$. This number is of course bounded by $(\operatorname{diam}(\Omega)+2)^d$ and $\operatorname{diam}(\Omega)$ is uniformly bounded in the class $\mathscr{O}(a,b)$. We recall that we need a uniform estimate with respect to Ω, x, y and s. The discrepancy

$$\left|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{I(s,x)\cap\Omega}\left(j\alpha+y+m\right)-\left|I\left(s,x\right)\cap\Omega\right|\right|$$
$$=\left|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{I(s,x-y)\cap(\Omega-y)}\left(j\alpha+m\right)-\left|I\left(s,x-y\right)\cap\left(\Omega-y\right)\right|\right|$$

is bounded by the sum of the discrepancies of the sets K_i,

$$\sum_{i=1}^{Q}\left|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{K_i}\left(j\alpha+m\right)-\left|K_i\right|\right|.$$

We shall therefore study the discrepancy of a single piece $K_i=K$. Assume that $K=I\cap(\Omega-y)$ where I is a box with edges parallel to the coordinate axes and length at most 1. For the sake of simplicity, we will call Ω the set $\Omega-y$. Let $\varepsilon>0$ be small enough so that $\Omega^{\pm\varepsilon}\in\mathscr{O}(a/2,2b)$, and let us call $I\left(-\varepsilon\right)$ the box that contains I and has facets at distance ε from the corresponding facets of I, and $K\left(-\varepsilon\right)=I\left(-\varepsilon\right)\cap\Omega^{-\varepsilon}$. Observe that $I\left(-\varepsilon\right)$ contains the outer parallel body of I at distance ε and therefore $K\left(-\varepsilon\right)$ contains the outer parallel body $K^{-\varepsilon}$ of K at distance ε. On the other hand, observe that $K^{\varepsilon}=I^{\varepsilon}\cap\Omega^{\varepsilon}$. Notice that K^{ε} may be empty and that I^{ε} is a box. In all cases, $K\left(-\varepsilon\right)$ and K^{ε} are the intersection of a box with edges parallel to the axes and length at most 2 and a smooth convex body in $\mathscr{O}(a/2,2b)$. Let φ be a smooth function with integral 1 and supported on the unit ball, and let

$$\varphi_{\varepsilon}\left(x\right)=\varepsilon^{-d}\varphi\left(x/\varepsilon\right).$$

By the above observations, $K^{\varepsilon}+\varepsilon B\subset K$ and $K\subset K\left(-\varepsilon\right)\div\varepsilon B$, so that for any $x\in\mathbb{R}^d$

$$\chi_{K^{\varepsilon}}*\varphi_{\varepsilon}\left(x\right)\leq\chi_{K}\left(x\right)\leq\chi_{K(-\varepsilon)}*\varphi_{\varepsilon}\left(x\right).$$

Thus, by the Poisson summation formula applied to the smooth compactly supported function $\chi_{K(-\varepsilon)} * \varphi_\varepsilon$,

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_K\left(j\alpha+m\right)-|K| \\
&\leq \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{K(-\varepsilon)} * \varphi_\varepsilon\left(j\alpha+m\right)-|K| \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\left(\chi_{K(-\varepsilon)} * \varphi_\varepsilon\right)^{\wedge}(m)\, e^{2\pi ijm\cdot\alpha}-|K| \\
&= |K(-\varepsilon)|-|K|+\frac{1}{N}\sum_{j=1}^{N}\sum_{m\neq 0}\widehat{\chi}_{K(-\varepsilon)}(m)\widehat{\varphi}_\varepsilon(m)\, e^{2\pi ijm\cdot\alpha} \\
&\leq |K(-\varepsilon)|-|K|+\sum_{m\neq 0}\left|\widehat{\chi}_{K(-\varepsilon)}(m)\right|\left|\widehat{\varphi}(\varepsilon m)\right|\left|\frac{1}{N}\sum_{j=1}^{N}e^{2\pi ijm\cdot\alpha}\right|.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_K\left(j\alpha+m\right)-|K| \\
&\geq \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{K^\varepsilon} * \varphi_\varepsilon\left(j\alpha+m\right)-|K| \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\left(\chi_{K^\varepsilon} * \varphi_\varepsilon\right)^{\wedge}(m)\, e^{2\pi ijm\cdot\alpha}-|K| \\
&= |K^\varepsilon|-|K|+\frac{1}{N}\sum_{j=1}^{N}\sum_{m\neq 0}\widehat{\chi}_{K^\varepsilon}(m)\widehat{\varphi}_\varepsilon(m)\, e^{2\pi ijm\cdot\alpha} \\
&\geq |K^\varepsilon|-|K|-\sum_{m\neq 0}\left|\widehat{\chi}_{K^\varepsilon}(m)\right|\left|\widehat{\varphi}(\varepsilon m)\right|\left|\frac{1}{N}\sum_{j=1}^{N}e^{2\pi ijm\cdot\alpha}\right|.
\end{aligned}$$

Observe first that $0 \leq |K|-|K^\varepsilon| \leq c\varepsilon$. Indeed, since $K \setminus K^\varepsilon \subset (I \setminus I^\varepsilon) \cup (\Omega \setminus \Omega^\varepsilon)$, it follows from the coarea formula and from the Archimedean postulate (if a convex body A is contained in a convex body B then the surface area of A is smaller than or equal to the surface area of B, see [4, Property 5, p. 52]) that

$$\begin{aligned}|K|-|K^{\varepsilon}| &= |K\setminus K^{\varepsilon}| \le |I\setminus I^{\varepsilon}|+|\Omega\setminus\Omega^{\varepsilon}|\\ &\le 2d\varepsilon+\int_0^{\varepsilon}\left(\int_{\{x:\delta_{\Omega}(x)=t\}} d\sigma(x)\right)dt\\ &\le 2d\varepsilon+\varepsilon\left|\{x:\delta_{\Omega}(x)=0\}\right|_{d-1}\le c\varepsilon.\end{aligned}$$

Similarly, since $K(-\varepsilon)\setminus K\subset (I(-\varepsilon)\setminus I)\cup(\Omega^{-\varepsilon}\setminus\Omega)$, then

$$\begin{aligned}|K(-\varepsilon)|-|K| &= |K(-\varepsilon)\setminus K| \le |I(-\varepsilon)\setminus I|+|\Omega^{-\varepsilon}\setminus\Omega|\\ &\le 2d(1+2\varepsilon)^{d-1}\varepsilon+\int_{-\varepsilon}^{0}\left(\int_{\{x:\delta_{\Omega}(x)=t\}} d\sigma(x)\right)dt\\ &\le 2d(1+2\varepsilon)^{d-1}\varepsilon+\varepsilon\left|\{x:\delta_{\Omega}(x)=-\varepsilon\}\right|_{d-1}\le c\varepsilon.\end{aligned}$$

The estimate of the exponential sums is standard,

$$\left|\frac{1}{N}\sum_{j=1}^{N}e^{2\pi ijm\cdot\alpha}\right| = \left|\frac{1}{N}\frac{\sin(\pi Nm\cdot\alpha)}{\sin(\pi m\cdot\alpha)}\right| \le \frac{1}{N\|m\cdot\alpha\|},$$

where $\|u\|$ is the distance of u from the closest integer. Finally, by Lemma 4, for both sets $K(-\varepsilon)$ and K^{ε} the estimate (2) for the Fourier transform in Lemma 2 holds uniformly in ε. It follows that the goal becomes to estimate

$$\begin{aligned}&c\varepsilon+\sum_{m\neq 0}\left(c\prod_{i=1}^{d}\frac{\log(2+|m_i|)}{(1+|m_i|)^{\frac{d+1}{2d}}}\left|\widehat{\varphi}(\varepsilon m)\right|\right)\frac{1}{N\|m\cdot\alpha\|}\\ &\le c\varepsilon+\frac{c}{N}\sum_{m\neq 0}\left(\prod_{i=1}^{d}\frac{\log(2+|m_i|)}{(1+|m_i|)^{\frac{d+1}{2d}}}\prod_{i=1}^{d}\frac{1}{1+\varepsilon|m_i|}\right)\frac{1}{\|m\cdot\alpha\|}.\end{aligned}$$

Let us first rearrange the above series by partitioning $\mathbb{Z}^d$ into sets where $m_i\neq 0$ if and only if $i\in S$, as S ranges over all possible nonempty subsets of $\{1,\ldots,d\}$. Thus

$$\begin{aligned}&\sum_{m\neq 0}\left(\prod_{i=1}^{d}\left(\frac{\log(2+|m_i|)}{(1+|m_i|)^{\frac{d+1}{2d}}}\frac{1}{1+\varepsilon|m_i|}\right)\right)\frac{1}{\|m\cdot\alpha\|}\\ &=\sum_{S\subset\{1,\ldots,d\},\,S\neq\emptyset}\ \sum_{m_i\neq 0\text{ iff }i\in S}\left(\prod_{i=1}^{d}\left(\frac{\log(2+|m_i|)}{(1+|m_i|)^{\frac{d+1}{2d}}}\frac{1}{1+\varepsilon|m_i|}\right)\right)\frac{1}{\|m\cdot\alpha\|}.\end{aligned}$$

Now, for any nonempty subset S of the set $\{1, \ldots, d\}$ with cardinality j, we need to estimate

$$\sum_{m_i \neq 0 \text{ iff } i \in S} \left(\prod_{i \in S} \left(\frac{\log(2 + |m_i|)}{(1 + |m_i|)^{\frac{d+1}{2d}}} \frac{1}{1 + \varepsilon |m_i|} \right) \right) \frac{1}{\|m \cdot \alpha\|}.$$

A dyadic decomposition along all directions $i_1, \ldots, i_j$ in S gives

$$\sum_{m_i \neq 0 \text{ iff } i \in S} \left(\prod_{i \in S} \left(\frac{\log(2 + |m_i|)}{(1 + |m_i|)^{\frac{d+1}{2d}}} \frac{1}{1 + \varepsilon |m_i|} \right) \right) \frac{1}{\|m \cdot \alpha\|}$$

$$\leq c \sum_{k_1=0}^{+\infty} \cdots \sum_{k_j=0}^{+\infty} \left(\frac{k_1}{2^{k_1 \frac{d+1}{2d}}} \frac{1}{1 + \varepsilon 2^{k_1}} \cdots \frac{k_j}{2^{k_j \frac{d+1}{2d}}} \frac{1}{1 + \varepsilon 2^{k_j}} \right) \sum_{m_{i_1} = 2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j} = 2^{k_j}}^{2^{k_j+1}-1} \frac{1}{\|m \cdot \alpha\|}.$$

Let us study the sum

$$\sum_{m_{i_1} = 2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j} = 2^{k_j}}^{2^{k_j+1}-1} \frac{1}{\|m \cdot \alpha\|}.$$

By the celebrated result of W.M. Schmidt [15], see also [16, Theorem 7C], since $1, \alpha_1, \ldots, \alpha_d$ are algebraic and linearly independent over $\mathbb{Q}$, for any $\eta > 0$ there is a constant $\gamma > 0$ such that for any $m \neq 0$,

$$\|m \cdot \alpha\| > \frac{\gamma}{(1 + |m_1|)^{1+\eta} \ldots (1 + |m_d|)^{1+\eta}}.$$

Then, arguing as in [9], in any interval of the form

$$\left[\frac{(h-1)\gamma}{(1 + 2^{k_1+1})^{1+\eta} \ldots \left(1 + \left|2^{k_j+1}\right|\right)^{1+\eta}}, \frac{h\gamma}{(1 + 2^{k_1+1})^{1+\eta} \ldots \left(1 + \left|2^{k_j+1}\right|\right)^{1+\eta}} \right),$$

where h is a positive integer, there are at most two numbers of the form $\|m \cdot \alpha\|$, with $m_{i_1} < 2^{k_1+1}, \ldots, m_{i_j} < 2^{k_j+1}$, and all other indices equal to zero. Indeed, assume by contradiction that there are three such numbers. Then for two of them, say $\|n \cdot \alpha\|$ and $\|m \cdot \alpha\|$, the fractional parts of $n \cdot \alpha$ and $m \cdot \alpha$ belong either to $(0, 1/2]$ or to $(1/2, 1)$. Assume without loss of generality that they belong to $(0, 1/2]$.

Then

$$\begin{aligned}\frac{\gamma}{\left(1+2^{k_1+1}\right)^{1+\eta}\dots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}} &> \left|\|n\cdot\alpha\|-\|m\cdot\alpha\|\right| \\ &= \left|(n\cdot\alpha-p)-(m\cdot\alpha-q)\right| \\ &\ge \|(n-m)\cdot\alpha\| \\ &> \frac{\gamma}{\left(1+2^{k_1+1}\right)^{1+\eta}\dots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}}.\end{aligned}$$

By the same type of argument, in the first interval $\left[0, \frac{\gamma}{\left(1+2^{k_1+1}\right)^{1+\eta}\dots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}}\right)$ there are no points of the form $\|m\cdot\alpha\|$ with $m_{i_1} < 2^{k_1+1}, \dots, m_{i_j} < 2^{k_j+1}$, and all other indices equal to zero. It follows that

$$\begin{aligned}\sum_{m_1=2^{k_1}}^{2^{k_1+1}-1}\cdots\sum_{m_j=2^{k_j}}^{2^{k_j+1}-1}\frac{1}{\|m\cdot\alpha\|} &\le c\sum_{h=1}^{2^{k_1+\dots+k_j}}\frac{2^{\left(k_1+\dots+k_j\right)(1+\eta)}}{h\gamma} \\ &\le c\,2^{\left(k_1+\dots+k_j\right)(1+\eta)}\left(k_1+\dots+k_j\right) \le c\,2^{\left(k_1+\dots+k_j\right)(1+2\eta)}.\end{aligned}$$

Thus

$$\begin{aligned}&\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}\left(\frac{k_1}{2^{k_1\frac{d+1}{2d}}}\frac{1}{1+\varepsilon 2^{k_1}}\cdots\frac{k_j}{2^{k_j\frac{d+1}{2d}}}\frac{1}{1+\varepsilon 2^{k_j}}\right)\sum_{m_1=2^{k_1}}^{2^{k_1+1}-1}\cdots\sum_{m_j=2^{k_j}}^{2^{k_j+1}-1}\frac{1}{\|m\cdot\alpha\|} \\ &\le c\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}\left(\frac{1}{2^{k_1\frac{d+1}{2d}}}\frac{1}{1+\varepsilon 2^{k_1}}\cdots\frac{1}{2^{k_j\frac{d+1}{2d}}}\frac{1}{1+\varepsilon 2^{k_j}}\right)2^{\left(k_1+\dots+k_j\right)(1+2\eta)} \\ &\le c\left(\sum_{k=0}^{+\infty}\frac{1}{2^{k\frac{d+1}{2d}}}\frac{1}{1+\varepsilon 2^{k}}2^{k(1+2\eta)}\right)^{j} \\ &\le c\left(\sum_{k=0}^{-\log_2\varepsilon}2^{k\left(\frac{d-1}{2d}+2\eta\right)}+\sum_{k=-\log_2\varepsilon}^{+\infty}\frac{2^{k\left(-\frac{d+1}{2d}+2\eta\right)}}{\varepsilon}\right)^{d} \\ &\le c\left(\varepsilon^{-\frac{d-1}{2d}-2\eta}+\varepsilon^{\frac{d+1}{2d}-2\eta-1}\right)^{d} \le c\varepsilon^{-\frac{d-1}{2}-2\eta d}.\end{aligned}$$

Choosing ε so that $\varepsilon = \varepsilon^{-\frac{d-1}{2}-2\eta d}N^{-1}$ gives the desired estimate $cN^{-\frac{2}{d+1}+\eta'}$. □

Proof of Theorem 3 Everything proceeds as in the proof of the Theorem 2, up until the point where we need to estimate

$$\left\|\sum_{i=1}^{Q}\left|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{K_i}\left(j\alpha+y+m\right)-\left|K_i\right|\right|\right\|_{L^p\left(\mathbb{T}^d,dy\right)}$$
$$\leq\sum_{i=1}^{Q}\left\|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_{K_i}\left(j\alpha+y+m\right)-\left|K_i\right|\right\|_{L^p\left(\mathbb{T}^d,dy\right)}.$$

Once again, take a single piece K_i and call it K, for simplicity. First consider $2\leq p<2d/(d-1)$. In this case, the smoothing argument with the convolution with the function φ_ε is superfluous. By the Hausdorff-Young inequality with $1/p+1/q=1$,

$$\left\|\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}^d}\chi_K\left(j\alpha+y+m\right)-|K|\right\|_{L^p\left(\mathbb{T}^d,dy\right)}$$
$$\leq\frac{1}{N}\left(\sum_{m\neq0}\left(\sum_{j=1}^{N}e^{2\pi ij\alpha\cdot m}\right)^q\left|\widehat{\chi}_K\left(m\right)\right|^q\right)^{1/q}$$
$$\leq\frac{c}{N}\left(\sum_{m\neq0}\min\left(N^q,\frac{1}{\|m\cdot\alpha\|^q}\right)\left(\prod_{i=1}^{d}\frac{\log\left(2+|m_i|\right)}{\left(1+|m_i|\right)^{\frac{d+1}{2d}}}\right)^q\right)^{1/q}$$
$$=\frac{c}{N}\sum_{S\subset\{1,\ldots,d\},\,S\neq\emptyset}\left(\sum_{m_i\neq0\text{ iff }i\in S}\left(\prod_{i=1}^{d}\frac{\log^q\left(2+|m_i|\right)}{\left(1+|m_i|\right)^{\frac{d+1}{d}\frac{q}{2}}}\right)\min\left(N^q,\frac{1}{\|m\cdot\alpha\|^q}\right)\right)^{1/q}$$

Now, for any nonempty subset S of the set $\{1,\ldots,d\}$ with cardinality j, we need to estimate

$$\sum_{m_i\neq0\text{ iff }i\in S}\left(\prod_{i\in S}\left(\frac{\log^q\left(2+|m_i|\right)}{\left(1+|m_i|\right)^{\frac{d+1}{d}\frac{q}{2}}}\right)\right)\min\left(N^q,\frac{1}{\|m\cdot\alpha\|^q}\right)$$

A dyadic decomposition along all the relevant directions in S gives

$$\sum_{m_i\neq0\text{ iff }i\in S}\left(\prod_{i\in S}\left(\frac{\log^q\left(2+|m_i|\right)}{\left(1+|m_i|\right)^{\frac{d+1}{d}\frac{q}{2}}}\right)\right)\min\left(N^q,\frac{1}{\|m\cdot\alpha\|^q}\right)$$
$$\leq c\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}\left(\frac{k_1^q}{2^{k_1\frac{d+1}{d}\frac{q}{2}}}\cdots\frac{k_j^q}{2^{k_j\frac{d+1}{d}\frac{q}{2}}}\right)\sum_{m_{i_1}=2^{k_1}}^{2^{k_1+1}-1}\cdots\sum_{m_{i_j}=2^{k_j}}^{2^{k_j+1}-1}\min\left(N^q,\frac{1}{\|m\cdot\alpha\|^q}\right).$$

Let us study the sum

$$\sum_{m_{i_1}=2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j}=2^{k_j}}^{2^{k_j+1}-1} \min\left(N^q, \frac{1}{\|m\cdot\alpha\|^q}\right).$$

The proof proceeds as in Theorem 2, *mutatis mutandis*. Since $1, \alpha_1, \ldots, \alpha_d$ are algebraic linearly independent over $\mathbb{Q}$, for any $\eta > 0$ there is a constant $\gamma > 0$ such that for any $m \neq 0$,

$$\|m\cdot\alpha\| > \frac{\gamma}{(1+|m_1|)^{1+\eta}\ldots(1+|m_d|)^{1+\eta}}.$$

In any interval of the form

$$\left[\frac{(h-1)\,\gamma}{(1+2^{k_1+1})^{1+\eta}\ldots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}}, \frac{h\gamma}{(1+2^{k_1+1})^{1+\eta}\ldots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}}\right),$$

where h is a positive integer, there are at most two numbers of the form $\|m\cdot\alpha\|$, with $m_{i_1} < 2^{k_1+1}, \ldots, m_{i_j} < 2^{k_j+1}$, and all other indices equal to zero. Also, in the first interval

$$\left[0, \frac{\gamma}{(1+2^{k_1+1})^{1+\eta}\ldots\left(1+\left|2^{k_j+1}\right|\right)^{1+\eta}}\right),$$

there are no points of the form $\|m\cdot\alpha\|$ with $m_{i_1} < 2^{k_1+1}, \ldots, m_{i_j} < 2^{k_j+1}$, and all other indices equal to zero. It follows that

$$\sum_{m_{i_1}=2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j}=2^{k_j}}^{2^{k_j+1}-1} \min\left(N^q, \frac{1}{\|m\cdot\alpha\|^q}\right) \leq c \sum_{h=1}^{2^{k_1+\ldots+k_j}} \min\left(N^q, \frac{2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}}{h^q}\right).$$

If $2^{\left(k_1+\ldots+k_j\right)(1+\eta)} \leq N$, then the sum reduces to

$$\sum_{h=1}^{2^{k_1+\ldots+k_j}} \min\left(N^q, \frac{2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}}{h^q}\right) = \sum_{h=1}^{2^{k_1+\ldots+k_j}} \frac{2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}}{h^q}$$

$$\leq c2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}.$$

If, on the other hand, $2^{\left(k_1+\ldots+k_j\right)(1+\eta)} \geq N$, then the sum reduces to

$$\begin{aligned}
&\sum_{h=1}^{2^{k_1+\ldots+k_j}} \min\left(N^q, \frac{2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}}{h^q}\right) \\
&\leq \sum_{1\leq h\leq 2^{\left(k_1+\ldots+k_j\right)(1+\eta)}/N} N^q + \sum_{2^{\left(k_1+\ldots+k_j\right)(1+\eta)}/N\leq h\leq 2^{k_1+\ldots+k_j}} \frac{2^{q\left(k_1+\ldots+k_j\right)(1+\eta)}}{h^q} \\
&\leq cN^{q-1}2^{\left(k_1+\ldots+k_j\right)(1+\eta)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{m_{i_1}=2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j}=2^{k_j}}^{2^{k_j+1}-1} \min\left(N^q, \frac{1}{\|m\cdot\alpha\|^q}\right) \\
&\leq c2^{\left(k_1+\ldots+k_j\right)(1+\eta)} \min\left(N^{q-1}, 2^{(q-1)\left(k_1+\ldots+k_j\right)(1+\eta)}\right).
\end{aligned}$$

Finally, if η is positive but small, then

$$\begin{aligned}
&\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}\left(\frac{k_1^q}{2^{k_1\frac{d+1}{d}\frac{q}{2}}}\cdots\frac{k_j^q}{2^{k_j\frac{d+1}{d}\frac{q}{2}}}\right)\sum_{m_{i_1}=2^{k_1}}^{2^{k_1+1}-1} \cdots \sum_{m_{i_j}=2^{k_j}}^{2^{k_j+1}-1} \min\left(N^q, \frac{1}{\|m\cdot\alpha\|^q}\right) \\
&\leq c\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}\left(\frac{k_1^q}{2^{k_1\frac{d+1}{d}\frac{q}{2}}}\cdots\frac{k_j^q}{2^{k_j\frac{d+1}{d}\frac{q}{2}}}\right)2^{\left(k_1+\ldots+k_j\right)(1+\eta)} \\
&\quad\times\min\left(N^{q-1}, 2^{(q-1)\left(k_1+\ldots+k_j\right)(1+\eta)}\right) \\
&\leq c\sum_{k_1=0}^{+\infty}\cdots\sum_{k_j=0}^{+\infty}2^{\left(k_1+\ldots+k_j\right)\left(\eta-\frac{d+1}{d}\frac{q}{2}\right)}2^{\left(k_1+\ldots+k_j\right)(1+\eta)} \\
&\quad\times\min\left(N^{q-1}, 2^{(q-1)\left(k_1+\ldots+k_j\right)(1+\eta)}\right) \\
&\leq c\sum_{s=0}^{+\infty}\left(\sum_{k_1+\ldots+k_j=s}1\right)2^{s\left(2\eta+1-\frac{d+1}{d}\frac{q}{2}\right)}\min\left(N^{q-1}, 2^{s(1+\eta)(q-1)}\right) \\
&\leq c\sum_{s=0}^{+\infty}2^{s\left(3\eta+1-\frac{d+1}{d}\frac{q}{2}\right)}\min\left(N^{q-1}, 2^{s(1+\eta)(q-1)}\right)
\end{aligned}$$

$$\begin{aligned}
&\leq c \sum_{0\leq s\leq \frac{1}{1+\eta}\log_2(N)} 2^{s\left(q-\frac{d+1}{d}\frac{q}{2}+(q+2)\eta\right)} + cN^{q-1} \sum_{\frac{1}{1+\eta}\log_2(N)\leq s\leq +\infty} 2^{s\left(3\eta+1-\frac{d+1}{d}\frac{q}{2}\right)} \\
&\leq c \left(2^{\frac{1}{1+\eta}\log_2 N}\right)^{q\frac{d-1}{2d}+(q+2)\eta} + cN^{q-1}\left(2^{\frac{1}{1+\eta}\log_2 N}\right)^{1-\frac{d+1}{d}\frac{q}{2}+3\eta} \\
&\leq cN^{\frac{1}{1+\eta}\left(q\frac{d-1}{2d}+(q+2)\eta\right)} \leq cN^{q\left(\frac{1}{2}-\frac{1}{2d}+\eta'\right)}.
\end{aligned}$$

This takes care of the case $p < 2d/(d-1)$. The case $p = +\infty$ is contained in Theorem 2. The intermediate cases $2d/(d-1) \leq p < +\infty$ follow by the interpolation

$$\int_Y |D(y)|^p \, dy \leq \sup_{y\in Y} |D(y)|^{p-s} \int_Y |D(y)|^s \, dy,$$

where $s < p < +\infty$, (Y, dy) is a measure space, and D is a measurable function on Y. □

References

1. Aistleitner, C., Dick, J.: Low-discrepancy point sets for non-uniform measures. Acta Arith. **163**, 345–369 (2014)
2. Aistleitner, C., Dick, J.: Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. Acta Arith. **167**, 143–171 (2015)
3. Beck, J.: Irregularities of distribution I. Acta Math. **159**, 1–49 (1987)
4. Bonnesen, T., Fenchel, W.: Theorie der konvexen Körper. Springer, Berlin (1974)
5. Brandolini, L., Colzani, L., Travaglini, G.: Average decay of Fourier transforms and integer points in polyhedra. Ark. Mat. **35**, 235–275 (1997)
6. Brandolini, L., Colzani, L., Gigante, G., Travaglini, G.: On the Koksma-Hlawka inequality. J. Complex. **29**, 158–172 (2013)
7. Brandolini, L., Colzani, L., Gigante, G., Travaglini, G.: Low-discrepancy sequences for piecewise smooth functions on the two-dimensional torus. J. Complex. **33**, 1–13 (2016)
8. Colzani, L., Gigante, G., Travaglini, G.: Trigonometric approximation and a general form of the Erdős Turán inequality. Trans. Am. Math. Soc. **363**, 1101–1123 (2011)
9. Davenport, H.: Notes on irregularities of distribution. Mathematika **3**, 131–135 (1956)
10. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, Reprint of the 1998 Edition. Springer, Berlin (2001)
11. Herz, C.S.: Fourier transforms related to convex sets. Ann. Math. **75**, 81–92 (1962)
12. Hlawka, E.: Integrale auf konvexen Körpern. I–II. Monatsh. Math. **54**, 1–36, 81–99 (1950)
13. Kuipers, L., Niederreiter H.: Uniform Distribution of Sequences. Dover, New York (2006)
14. Montgomery, H.: Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis, CBMS Regional Conference Series in Mathematics, vol. 84. American Mathematical Society, Providence (1994)
15. Schmidt, W.M.: Simultaneous approximation to algebraic numbers by rationals. Acta Math. **125** , 189–201 (1970)
16. Schmidt, W.M.: Approximation to algebraic numbers. Enseign. Math. **17**(2), 187–253 (1971)

17. Schneider, R.: Convex Bodies: The Brunn-Minkowski Theory, Second Expanded Edition (Encyclopedia of Mathematics and Its Applications). Cambridge University Press, Cambridge (2014)
18. Stein, E.M.: Harmonic Analysis: Real Variable Methods, Orthogonality and Oscillatory Integrals. Princeton University Press, Princeton (1993)

Explicit Families of Functions on the Sphere with Exactly Known Sobolev Space Smoothness

Johann S. Brauchart

Dedicated to Ian H. Sloan mentor and friend on the occasion of his 80th birthday.

Abstract We analyze explicit trial functions defined on the unit sphere $\mathbb{S}^d$ in the Euclidean space $\mathbb{R}^{d+1}$, $d \geq 1$, that are integrable in the $\mathbb{L}_p$-sense, $p \in [1, \infty)$. These functions depend on two free parameters: one determines the support and one, a critical exponent, controls the behavior near the boundary of the support. Three noteworthy features are: (1) they are simple to implement and capture typical behavior of functions in applications, (2) their integrals with respect to the uniform measure on the sphere are given by explicit formulas and, thus, their numerical values can be computed to arbitrary precision, and (3) their smoothness can be defined a priori, that is to say, they belong to Sobolev spaces $\mathbb{H}^s(\mathbb{S}^d)$ up to a specified index $\bar{s}$ determined by the parameters of the function. Considered are zonal functions $g(\boldsymbol{x}) = h(\boldsymbol{x} \cdot \boldsymbol{p})$, where $\boldsymbol{p}$ is some fixed pole on $\mathbb{S}^d$. The function $h(t)$ is of the type $[\max\{t, T\}]^\alpha$ or a variation of a truncated power function $x \mapsto (x)_+^\alpha$ (which assumes 0 if $x \leq 0$ and is the power x^α if $x > 0$) that reduces to $[\max\{t - T, 0\}]^\alpha$, $[\max\{t^2 - T^2, 0\}]^\alpha$, and $[\max\{T^2 - t^2, 0\}]^\alpha$ if $\alpha > 0$. These types of trial functions have as support the whole sphere, a spherical cap centered at $\boldsymbol{p}$, a bi-cap centered at the antipodes $\boldsymbol{p}$, $-\boldsymbol{p}$, or an equatorial belt. We give inclusion theorems that identify the critical smoothness $\bar{s} = \bar{s}(T, \alpha)$ and explicit formulas for the integral over the sphere. We obtain explicit formulas for the coefficients in the Laplace-Fourier expansion of these trial functions and provide the leading order term in the asymptotics for large index of the coefficients.

J. S. Brauchart (✉)
Institute of Analysis and Number Theory, Graz University of Technology, Kopernikusgasse 24/II, 8010 Graz, Austria
e-mail: j.brauchart@tugraz.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_9

1 Introduction and Statement of Results

Trial functions are used in simulation experiments, e.g., to test numerical integration, interpolation, and approximation methods. The purpose of our paper is to prove previously in the literature not available properties of functions that can be used as trial functions on the non-standard domain the sphere.

Let $\mathbb{S}^d$ be the unit sphere in the Euclidean space $\mathbb{R}^{d+1}$, $d \geq 1$, provided with the uniform normalized surface area measure σ_d (i.e., $\int_{\mathbb{S}^d} \mathrm{d}\sigma_d = 1$). We analyze four families of zonal trial functions defined on $\mathbb{S}^d$, $d \geq 1$; i.e., functions of the form $g(\boldsymbol{x}) = h(\boldsymbol{x} \cdot \boldsymbol{p}), \boldsymbol{p} \in \mathbb{S}^d$ fixed. These functions depend on one parameter controlling the support of the function and a critical exponent determining the behavior near the boundary of the support; cf. Fig. 1. Three noteworthy features are:

1. simple to implement and capture of typical behavior of functions in applications (kinks and threshold values that are features of, say, option pricing functions);
2. the integral with respect to σ_d on the sphere is given by an explicit formula and, thus, the numerical values can be computed to arbitrary precision; and
3. the smoothness can be defined a priori, that is to say, the Sobolev space classes $\mathbb{H}^s(\mathbb{S}^d)$ are specified by the parameters of the trial function.

We provide inclusion theorems that identify the least upper bound (critical index $\bar{s}$) of the smoothness of the Sobolev spaces over $\mathbb{S}^d$ to which the trial function of a family belong. The proofs employ asymptotic analysis of the coefficients of the Laplace-Fourier expansion of the trial function and, thus, rely on the Hilbert space structure of the underlying function space of square-integrable functions over $\mathbb{S}^d$. For this purpose, we derive explicit expressions of the coefficients $\widehat{h}_\ell$ and give their leading order term in the asymptotics for large index ℓ. In particular, we obtain explicit formulas of the integrals with respect to σ_d. For technical details, including representations of the coefficients in terms of special functions, see Sect. 3.

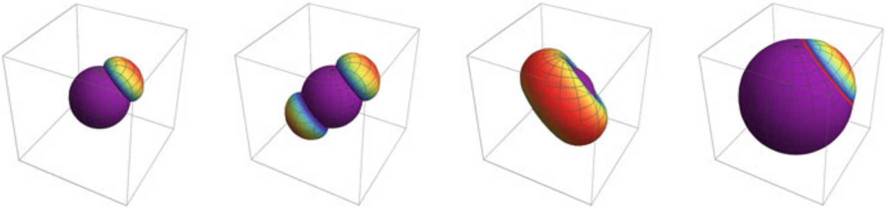


Fig. 1 Qualitative behavior of the trial functions g_1, g_2, g_3, and g_4 given in (1)–(4) for same value of $\alpha > 0$ and $0 < T < 1$. The function values are represented as suitable rescaled distances normal to the surface of the unit sphere

1.1 Trial Functions and Numerical Integration on Spheres

The testing of numerical integration schemes makes use of suitable trial functions with prescribed features. Numerical analysis of function approximation schemes (see also 'optimal recovery', cf. [43]) also makes use of trial functions. We cite recent work on hyperinterpolation [25], filtered hyperinterpolation [37], and filtered polynomial approximation [42]. We refer the reader to [32], the work of Genz (see, e.g., [17][1]), and online test suites of functions and data sets like [39]. See also the Virtual Library of Simulation Experiments: Test Functions and Datasets https://www.sfu.ca/~ssurjano/index.html.

Here, we shall focus on numerical integration on spheres. Regarding general references, we refer to [26]; see also [18, 28].

A frequently chosen trial function for the 2-sphere is *Franke's test function* [34],

$$\begin{aligned} f\big(x,y,z\big) := {} & 0.75\exp(-(9x-2)^2/4-(9y-2)^2/4-(9z-2)^2/4) \\ & + 0.75\exp(-(9x+1)^2/49-(9y+1)/10-(9z+1)/10) \\ & + 0.5\exp(-(9x-7)^2/4-(9y-3)^2/4-(9z-5)^2/4) \\ & - 0.2\exp(-(9x-4)^2-(9y-7)^2-(9z-5)^2), \qquad (x,y,z)\in\mathbb{S}^2, \end{aligned}$$

which has two Gaussian peaks of different heights, and a smaller dip and which is in $C^\infty(\mathbb{S}^2)$. Its integral is computable to arbitrary precision,

$$\int_{\mathbb{S}^2} f(\boldsymbol{x})\,\mathrm{d}\sigma_2(\boldsymbol{x}) = 0.5328652500843890\ldots.$$

The functions considered in this paper will be useful as trial functions on $\mathbb{S}^d$ for all dimensions $d \geq 1$ with a precise range of Sobolev space smoothness whose integrals are exactly known with values that can be given to arbitrary precision.

Sobolev spaces over $\mathbb{S}^d$ emerge naturally when dealing with numerical integration in the worst-case error setting in the continuous regime $s > d/2$. A *Quasi-Monte Carlo (QMC) method* is an equal weight numerical integration formula with *deterministic* node set: the integral $I(f)$ of a given continuous real function f on $\mathbb{S}^d$ is approximated by a QMC method $Q[X_N](f)$ for a node set $X_N = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\} \subset \mathbb{S}^d$,

$$I(f) := \int_{\mathbb{S}^d} f(\boldsymbol{x})\mathrm{d}\sigma_d(\boldsymbol{x}) \approx \frac{1}{N}\sum_{k=1}^{N} f(\boldsymbol{x}_k) =: Q[X_N](f).$$

[1] For the related MATLAB program TESTPACK, see http://people.sc.fsu.edu/~jburkardt/m_src/testpack/testpack.html.

A node set X_N is deterministically chosen in a sensible way so as to guarantee "small" *worst-case error* of numerical integration

$$\mathrm{wce}(Q[X_N];\mathbb{H}^s(\mathbb{S}^d)) := \sup\left\{\left|Q[X_N](f) - I(f)\right| : f \in \mathbb{H}^s(\mathbb{S}^d), \|f\|_{\mathbb{H}^s} \leq 1\right\}.$$

Reproducing kernel Hilbert space techniques (see [27] for the case of the unit cube) provide a convenient way to explicitly compute this error for a given node set whenever the reproducing kernel has a suitable closed form. Indeed (see [4, 7, 8, 12, 35] for generalizations), the worst-case error for the Sobolev space $\mathbb{H}^s(\mathbb{S}^d)$ for $s = (d+1)/2$ endowed with the reproducing kernel $1 - C_d|\boldsymbol{x} - \boldsymbol{y}|$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^d$, C_d given in (5), satisfies the following invariance principle named after Stolarsky [38],

$$\frac{1}{N^2}\sum_{j,k=1}^{N}\left|\boldsymbol{x}_j - \boldsymbol{x}_k\right| + \frac{1}{C_d}\left[\mathrm{wce}(Q[X_N];\mathbb{H}^s(\mathbb{S}^d))\right]^2 = \int_{\mathbb{S}^d}\int_{\mathbb{S}^d}|\boldsymbol{x} - \boldsymbol{y}|\,\mathrm{d}\sigma_d(\boldsymbol{x})\,\mathrm{d}\sigma_d(\boldsymbol{y});$$

i.e., points that maximize their sum of mutual Euclidean distances are excellent nodes for a QMC method that minimizes the worst-case error in the above setting. The distance maximization problem for the points is highly non-trivial which limits the usability of this approach to produce good node sets for QMC methods. It is known (see [10]) that a sequence (X_N^*) of maximal sum-of-distance N-point sets define QMC methods satisfying

$$\left|Q[X_N^*](f) - I(f)\right| \leq c_{s',d}\,\frac{\|f\|_{\mathbb{H}^s}}{N^{s/d}} \qquad \text{for all } f \in \mathbb{H}^s(\mathbb{S}^d) \text{ and all } \frac{d}{2} < s \leq \frac{d+1}{2}.$$

The order of N cannot be improved. It is an open problem if the range of s can be enlarged; i.e., the *strength* of (X_N^*), which is the supremum of the maximal range for s, is unknown. Determining the strength of a given sequence of N-point sets on $\mathbb{S}^d$ is a highly unresolved question. In contrast, consider *spherical t-designs*, introduced in the seminal paper [15], that integrate spherical polynomials of degree $\leq t$ exactly. A sequence $(Z_{N_t}^*)$ of spherical t-designs with exactly the optimal order of points, $N_t \asymp t^d$, has the remarkable property that

$$\left|Q[Z_{N_t^*}](f) - I(f)\right| \leq c_{s,d}\,\frac{\|f\|_{\mathbb{H}^s}}{N_t^{s/d}} \qquad \text{for all } f \in \mathbb{H}^s(\mathbb{S}^d) \text{ and } \textbf{all } s > \frac{d}{2}.$$

The order of N_t cannot be improved (see [9, 21–24]) and the strength of $(Z_{N_t}^*)$ is infinite. The existence of such sequences $(Z_{N_t}^*)$ follows from [5]. Spherical designs can be obtained by minimizing certain "energy functionals" (see [1, 2, 13, 14, 19, 36]). A fundamental unresolved question in the theory of numerical integration on the sphere concerns the explicit construction of good node sets with provable small or even optimal worst-case error bounds (for conjectures, see [6]). We remark that (X_N^*) is a QMC design sequence for $\mathbb{H}^s(\mathbb{S}^d)$ for $d/2 < s \leq (d+1)/2$ and $(Z_{N_t}^*)$ is a

QMC design sequence for $\mathbb{H}^s(\mathbb{S}^d)$ for all $s > d/2$; see [10, 11]. In general, a *QMC design sequence* (X_N) *for* $\mathbb{H}^s(\mathbb{S}^d)$, $s > d/2$, has the property

$$|Q[X_N](f) - I(f)| \leq \frac{c'_{s',d}}{N^{s'/d}} \, \|f\|_{\mathbb{H}^{s'}} \quad \text{for all } f \in \mathbb{H}^{s'}(\mathbb{S}^d) \text{ and all } \frac{d}{2} < s' \leq s.$$

The asymptotic behavior of the error of integration of one of the following trial functions should be understood in the context of above estimates.

1.2 Trial Functions

The first three families of trial functions are defined via variations of a truncated power function $x \mapsto (x)_+^\alpha$ that is 0 if $x \leq 0$ and is the power x^α if $x > 0$. Throughout the paper we shall assume that $0^\alpha = 0$ if $\alpha > 0$. All trial functions considered here (see definitions (1)–(4)) are non-negative and obey a power law of the following form: Let $g = g(\alpha, T; \cdot)$. Then $|g(\alpha, T; \cdot)|^p = g(p\alpha, T; \cdot)$ for $p \in \mathbb{R}$. Hence, the $\mathbb{L}_p$-norm of g is closely related to its rescaled integral; i.e.,

$$\|g\|_{\mathbb{L}_p(\mathbb{S}^d)}^p = \int_{\mathbb{S}^d} |g(\alpha, T; \boldsymbol{x})|^p \, \mathrm{d}\sigma_d(\boldsymbol{x}) = \int_{\mathbb{S}^d} g(p\,\alpha, T; \boldsymbol{x}) \, \mathrm{d}\sigma_d(\boldsymbol{x}).$$

Explicit expressions for these integrals (with α changed to $p\,\alpha$ and ℓ set to 0) are given in the corollaries below.

Proposition 1 *Let* $1 \leq p < \infty$. *Then the trial functions given in* (1)–(4) *are in* $\mathbb{L}_p(\mathbb{S}^d)$ *if the parameters* T *and* α *obey the relations in the table*

Eq.	$-1 < T < 0$	$T = 0$	$0 < T < 1$
(1)	$\alpha > -1/p$	$\alpha > -1/p$	$\alpha > -1/p$
(2)	–	$\alpha > -1/(2p)$	$\alpha > -1/p$
(3)	–	–	$\alpha > -1/p$
(4)	–	–	$\alpha \in \mathbb{R}$

The following inclusion theorems (one for each trial function type) concern Sobolev spaces $\mathbb{H}^s(\mathbb{S}^d)$ over $\mathbb{S}^d$ and provide sharp bounds on the smoothness parameter s of the space in terms of the critical exponent α. We remark that $s > 0$ is natural in the sense that $\mathbb{H}^0(\mathbb{S}^d)$ coincides with $\mathbb{L}_2(\mathbb{S}^d)$. However, the condition $s > d/2$ required for continuous embedding restricts the range of α from below.

Let $\boldsymbol{p} \in \mathbb{S}^d$, $T \geq 0$, and $\alpha \in \mathbb{R}$. The functions of the first family are defined as:

$$g_1(\boldsymbol{x}) := f_1(\boldsymbol{x} \cdot \boldsymbol{p}), \quad \boldsymbol{x} \in \mathbb{S}^d, \qquad f_1(t) := \big(t - T\big)_+^\alpha, \quad -1 \leq t \leq 1. \tag{1}$$

For $\alpha > 0$, we have $f_1(t) = [\max\{t - T, 0\}]^\alpha$. The function g_1 is supported on the spherical cap $\{\boldsymbol{x} \in \mathbb{S}^d : \boldsymbol{x} \cdot \boldsymbol{p} \geq T\}$ for $T \geq 0$; cf. first display in Fig. 1.

Theorem 1 *Let $-1 < T < 1$ and $\alpha > -1/2$. Then $g_1 \in \mathbb{H}^s(\mathbb{S}^d)$ iff $0 < s < \alpha + 1/2$.*

The Laplace-Fourier coefficients and their asymptotics are given in Sect. 3.2.

Corollary 1 *Let $\alpha > -1$. If $T = 0$, then*

$$\int_{\mathbb{S}^d} g_1(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{1}{2} \frac{\Gamma((d+1)/2)\Gamma((\alpha+1)/2)}{\sqrt{\pi}\,\Gamma((d+1+\alpha)/2)}.$$

If $0 < |T| < 1$, then (in terms of the Ferrers function P_ν^μ defined in (22))

$$\int_{\mathbb{S}^d} g_1(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = 2^{d/2-1} \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\sqrt{\pi}} \left(1 - T^2\right)^{(\alpha+d/2)/2} \mathrm{P}_{d/2-1}^{-(\alpha+d/2)}(T).$$

Let $\boldsymbol{p} \in \mathbb{S}^d$, $T \geq 0$, and $\alpha \in \mathbb{R}$. The functions of the second family are defined as:

$$g_2(\boldsymbol{x}) := f_2(\boldsymbol{x} \cdot \boldsymbol{p}), \quad \boldsymbol{x} \in \mathbb{S}^d, \qquad f_2(t) := \left(t^2 - T^2\right)_+^\alpha, \quad -1 \leq t \leq 1. \tag{2}$$

For $\alpha > 0$, we have $f_2(t) = \left[\max\{t^2 - T^2, 0\}\right]^\alpha$. The function g_2 is supported on the bi-cap $\{\boldsymbol{x} \in \mathbb{S}^d : |\boldsymbol{x} \cdot \boldsymbol{p}| \geq T\}$; cf. second display in Fig. 1.

Theorem 2 *Suppose $T = 0$.*

1. *Let $\alpha > -1/4$ with $\alpha \neq 0, 1, 2, \dots$. Then $g_2 \in \mathbb{H}^s(\mathbb{S}^d)$ iff $0 < s < 2\alpha + 1/2$.*
2. *Let $\alpha = 0, 1, 2, \dots$. Then $g_2 \in \mathbb{H}^s(\mathbb{S}^d)$ for all $s \geq 0$.*

Suppose $0 < T < 1$ and $\alpha > -1/2$. Then $g_2 \in \mathbb{H}^s(\mathbb{S}^d)$ iff $0 < s < \alpha + 1/2$.

The Laplace-Fourier coefficients and their asymptotics are given in Sect. 3.3.

Corollary 2 *If $T = 0$ and $\alpha > -1/2$, then*

$$\int_{\mathbb{S}^d} g_2(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1/2)}{\sqrt{\pi}\,\Gamma(\alpha+(d+1)/2)}.$$

If $0 < T < 1$ and $\alpha > -1$, then (in terms of the Jacobi function $P_\nu^{(\alpha,\beta)}$ given in (31))

$$\int_{\mathbb{S}^d} g_2(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\Gamma(\alpha+(d+1)/2)} \left(1 - T^2\right)^{\alpha+d/2} P_{-1/2}^{(\alpha+d/2,-\alpha-1/2)}(2T^2-1).$$

Remark 1 For $\alpha > -1$ and $0 < T < 1$ the integrals can also be written as

$$\int_{\mathbb{S}^d} g_2(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\sqrt{\pi}\,\Gamma(\alpha+1+d/2)} \left(1 - T^2\right)^{\alpha+d/2}$$
$$\times\ {}_2\mathrm{F}_1\left(\begin{matrix}1/2, d/2\\ \alpha+1+d/2\end{matrix}; 1 - T^2\right)$$

and for $d = 2$ take the form

$$\int_{\mathbb{S}^2} g_2(\boldsymbol{x}) \mathrm{d}\sigma_2(\boldsymbol{x}) = \frac{1}{\alpha+1} (1-T)^{\alpha+1} (1+T)^{\alpha} {}_2\mathrm{F}_1\left(\begin{matrix} -\alpha, 1 \\ \alpha+2 \end{matrix}; \frac{1-T}{1+T}\right).$$

The hypergeometric function reduces to a polynomial of degree α if $\alpha = 0, 1, 2, \ldots$. If $\alpha = -1/2, 1/2, 3/2, \ldots$ (i.e., $\alpha = \nu - 3/2$ for $\nu = 1, 2, 3, \ldots$), then

$$\int_{\mathbb{S}^2} g_2(\boldsymbol{x}) \mathrm{d}\sigma_2(\boldsymbol{x}) = \frac{(-1)^{\nu-1}}{2\nu-1} \frac{(1/2)_\nu}{(\nu-1)!} T^{2\nu-2} \Big(2 \operatorname{atanh}(\sqrt{1-T^2}) + \left(1-T^2\right)^{-1/2} \sum_{k=1}^{\nu-1} \frac{(k-1)!}{(1/2)_k} \left(1 - \frac{1}{T^2}\right)^k \Big).$$

Let $\boldsymbol{p} \in \mathbb{S}^d$, $T > 0$, and $\alpha \in \mathbb{R}$. The functions of the third family are defined as:

$$g_3(\boldsymbol{x}) := f_3(\boldsymbol{x} \cdot \boldsymbol{p}), \quad \boldsymbol{x} \in \mathbb{S}^d, \qquad f_3(t) := \left(T^2 - t^2\right)_+^{\alpha}, \qquad -1 \le t \le 1. \tag{3}$$

For $\alpha > 0$, the function f_3 reduces to $f_3(t) = \left[\max\{T^2 - t^2, 0\}\right]^\alpha$. The function g_3 is supported on the equatorial belt $\{\boldsymbol{x} \in \mathbb{S}^d : |\boldsymbol{x} \cdot \boldsymbol{p}| \le T\}$; cf. third display in Fig. 1.

Theorem 3 *Let* $0 < T < 1$ *and* $\alpha > -1/2$. *Then* $g_3 \in \mathbb{H}^s(\mathbb{S}^d)$ *iff* $0 < s < \alpha + 1/2$.

The Laplace-Fourier coefficients and their asymptotics are given in Sect. 3.4.

Corollary 3 *Let* $0 < T < 1$ *and* $\alpha > -1$. *Then*

$$\int_{\mathbb{S}^d} g_3(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\Gamma((d+1)/2+\alpha)} T^{2\alpha+1} P_{d/2-1}^{(\alpha+1/2, -\alpha-d/2)}(1-2T^2),$$

where the Jacobi function $P_\nu^{(\alpha,\beta)}$ *is defined in* (31).

Let $\boldsymbol{p} \in \mathbb{S}^d$, $T > 0$, and $\alpha \in \mathbb{R}$. The functions of the fourth family are defined as:

$$g_4(\boldsymbol{x}) := f_4(\boldsymbol{x} \cdot \boldsymbol{p}), \quad \boldsymbol{x} \in \mathbb{S}^d, \qquad f_4(t) := \left[\max\{t, T\}\right]^\alpha, \quad -1 \le t \le 1. \tag{4}$$

The support of g_4 is the whole sphere $\mathbb{S}^d$. The function g_4 attains the constant value T^α on the spherical cap $\{\boldsymbol{x} \in \mathbb{S}^d : \boldsymbol{x} \cdot \boldsymbol{p} \le T\}$; cf. fourth display in Fig. 1.

Theorem 4 *Let* $0 < T < 1$ *and* $\alpha \ge 0$. *Then* $g_4 \in \mathbb{H}^s(\mathbb{S}^d)$*,* $s > 0$*, iff* $0 < s < 3/2$.

The Laplace-Fourier coefficients and their asymptotics are given in Sect. 3.5.

Corollary 4 *Suppose* $0 < T < 1$ *and* $\alpha > -1$. *Then*

$$\int_{\mathbb{S}^d} g_4(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = T^\alpha \left(1 - \frac{1}{2} \mathrm{I}_{1-T^2}(d/2, 1/2)\right) + \frac{1}{2} A_1 \, \mathrm{I}_{1-T^2}(d/2, (\alpha+1)/2),$$

where $\mathrm{I}_z(a,b)$ *denotes the regularized incomplete beta function given in* (20) *and*

$$A_1 := \frac{\Gamma((d+1)/2)\Gamma((\alpha+1)/2)}{\sqrt{\pi}\,\Gamma((d+1+\alpha)/2)}.$$

Remark 2 We also include the following alternative forms for $0 < T < 1, \alpha > -1$:

$$\int_{\mathbb{S}^d} g_4(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{1}{2} A_1 + T^\alpha \Bigg(\frac{1}{2} + \frac{\omega_{d-1}}{\omega_d} T \,{}_2\mathrm{F}_1\Big(\begin{matrix} 1/2, 1-d/2 \\ 3/2 \end{matrix}; T^2 \Big) - \frac{1}{\alpha+1} \frac{\omega_{d-1}}{\omega_d} T \,{}_2\mathrm{F}_1\Big(\begin{matrix} (\alpha+1)/2, 1-d/2 \\ 1+(\alpha+1)/2 \end{matrix}; T^2 \Big) \Bigg).$$

Note that the hypergeometric functions, defined in (21), reduce to polynomials if d is an even dimension. The ratio ω_{d-1}/ω_d is given in (5).

2 Function Space Setting and Zonal Functions

The unit sphere $\mathbb{S}^d$ in the Euclidean space $\mathbb{R}^{d+1}$, $d \geq 1$, is provided with the normalized uniform surface area measure σ_d (i.e., $\int_{\mathbb{S}^d} \mathrm{d}\sigma_d = 1$) and has surface area ω_d:

$$\omega_0 := 2, \; \frac{\omega_d}{\omega_{d-1}} = \int_{-1}^{1} \left(1-t^2\right)^{d/2-1} \mathrm{d}t = \frac{\sqrt{\pi}\,\Gamma(d/2)}{\Gamma((d+1)/2)}, \; d \geq 1; \; C_d := \frac{1}{d} \frac{\omega_{d-1}}{\omega_d}; \tag{5}$$

2.1 Spherical Harmonics

The restriction to $\mathbb{S}^d$ of a homogeneous polynomial of exact degree ℓ defined in $\mathbb{R}^{d+1}$ is called a *spherical harmonic* $\mathrm{Y}_\ell = \mathrm{Y}_\ell^{(d)}$ *of degree* ℓ *on* $\mathbb{S}^d$. There are at most

$$Z(d,0) := 1, \qquad Z(d,\ell) := (2\ell+d-1)\,\frac{\Gamma(\ell+d-1)}{\Gamma(d)\,\Gamma(\ell+1)}, \quad \ell \geq 1, \tag{6}$$

such linearly independent spherical harmonics. The exact asymptotic behavior is

$$Z(d,\ell) \sim (2/\Gamma(d))\,\ell^{d-1} \quad \text{as } \ell \to \infty. \tag{7}$$

Every Y_ℓ is an eigenfunction of the negative Laplace-Beltrami operator $-\Delta_d^*$ for $\mathbb{S}^d$ with eigenvalue

$$\lambda_\ell := \ell\,(\ell + d - 1)\,, \qquad \ell \geq 0. \tag{8}$$

A system $\{\mathrm{Y}_{\ell,k} : k = 1, \ldots, Z(d,\ell)\}$ of $Z(d,\ell)$ linearly independent real spherical harmonics $\mathrm{Y}_{\ell,k}$, $\mathbb{L}_2$-orthonormal with respect to σ_d on $\mathbb{S}^d$, obeys the *addition theorem*

$$\sum_{k=1}^{Z(d,\ell)} \mathrm{Y}_{\ell,k}(\boldsymbol{x})\,\mathrm{Y}_{\ell,k}(\boldsymbol{y}) = Z(d,\ell)\,P_\ell^{(d)}(\boldsymbol{x}\cdot\boldsymbol{y}), \qquad \boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^d. \tag{9}$$

The normalized Gegenbauer (or ultraspherical) polynomials $P_\ell^{(d)}$ are orthogonal on $[-1,1]$ w.r.t. the weight function $(1-t^2)^{d/2-1}$ with $P_\ell^{(d)}(1) = 1$. The family

$$\{\mathrm{Y}_{\ell,k} : k = 1, \ldots, Z(d,\ell); \ell = 0, 1, \ldots\} \tag{10}$$

is a complete orthonormal basis of the Hilbert space $\mathbb{L}_2(\mathbb{S}^d)$ of square-integrable functions on $\mathbb{S}^d$ endowed with the inner product and induced norm [2]

$$(f,g)_{\mathbb{L}_2(\mathbb{S}^d)} := \int_{\mathbb{S}^d} f(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\sigma_d(\boldsymbol{x}), \qquad \|f\|_{\mathbb{L}_2(\mathbb{S}^d)} := \sqrt{(f,f)_{\mathbb{L}_2(\mathbb{S}^d)}}.$$

2.2 Normequivalent Sobolev Space Families

Sequences of positive weights $(a_\ell^{(s)})_{\ell\geq 0}$ that satisfy the relation [3]

$$a_\ell^{(s)} \asymp (1+\lambda_\ell)^{-s} \asymp (1+\ell)^{-2s} \tag{11}$$

define a family of inner products and induced equivalent norms

$$(f,g)_{(a_\ell^{(s)})} := \sum_{\ell=0}^{\infty} \frac{1}{a_\ell^{(s)}} \sum_{k=1}^{Z(d,\ell)} \widehat{f}_{n,k}\,\widehat{g}_{n,k}, \qquad \|f\|_{(a_\ell^{(s)})} := \sqrt{(f,f)_{(a_\ell^{(s)})}}$$

[2] The system (10) is also a complete orthogonal system for the class of continuous functions $C(\mathbb{S}^d)$, the set of k-times continuously differentiable functions $C^k(\mathbb{S}^d)$, the family of smooth functions $C^\infty(\mathbb{S}^d)$, and the Banach space $\mathbb{L}_p(\mathbb{S}^d)$, $1 \leq p < \infty$, provided with the usual p-norm. For more details, we refer the reader to [3, 29].

[3] We write $a_\ell \asymp b_\ell$ to mean that there exist $c_1, c_2 > 0$ independent of ℓ such that $c_1 a_\ell \leq b_\ell \leq c_2 a_\ell$ for all ℓ.

on $\mathbb{L}_2(\mathbb{S}^d)$ in terms of the Laplace-Fourier coefficients

$$\widehat{f}_{\ell,k} := \widehat{f}_{\ell,k}^{(d)} := (f, \mathrm{Y}_{\ell,k})_{\mathbb{L}_2(\mathbb{S}^d)} = \int_{\mathbb{S}^d} f(\boldsymbol{x})\, \mathrm{Y}_{\ell,k}(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}). \tag{12}$$

The *Sobolev space* $\mathbb{H}^s(\mathbb{S}^d)$ *over* $\mathbb{S}^d$ with smoothness index s is then the set of all $\mathbb{L}_2$-functions on $\mathbb{S}^d$ with finite *Sobolev norm* $\|f\|_{\mathbb{H}^s} := \|f\|_{(a_\ell^{(s)})}$; i.e.,

$$\mathbb{H}^s(\mathbb{S}^d) := \left\{ f \in \mathbb{L}_2(\mathbb{S}^d) : \|f\|_{\mathbb{H}^s} < \infty \right\}. \tag{13}$$

It is known that $\mathbb{H}^s(\mathbb{S}^d) \subset \mathbb{H}^{s'}(\mathbb{S}^d)$ whenever $s > s'$, and that $\mathbb{H}^s(\mathbb{S}^d)$ is embedded in the space of k-times continuously differentiable functions $C^k(\mathbb{S}^d)$ if $s > k + d/2$.

2.3 Zonal Functions

Let $g(\boldsymbol{x}) := h(\boldsymbol{x} \cdot \boldsymbol{p})$ be a zonal function; i.e., g depends only on the inner product of $\boldsymbol{x}$ with a fixed $\boldsymbol{p} \in \mathbb{S}^d$. Then (12) and the *Funk-Hecke formula* (see Müller [29])

$$\int_{\mathbb{S}^d} h(\boldsymbol{x} \cdot \boldsymbol{p})\, \mathrm{Y}_\ell(\boldsymbol{x})\, \mathrm{d}\sigma_d(\boldsymbol{x}) = \alpha_\ell[h]\, \mathrm{Y}_\ell(\boldsymbol{p}),$$

where

$$\alpha_\ell[h] := \frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} h(t)\, P_\ell^{(d)}(t) \left(1 - t^2\right)^{d/2-1} \mathrm{d}t,$$

holding for any spherical harmonic Y_ℓ of degree ℓ on $\mathbb{S}^d$, yield

$$\widehat{g}_{\ell,k} = (g, \mathrm{Y}_{\ell,k})_{\mathbb{L}_2(\mathbb{S}^d)} = \alpha_\ell[h]\, \mathrm{Y}_{\ell,k}(\boldsymbol{p}).$$

Application of the addition theorem gives

$$\begin{aligned} g(\boldsymbol{x}) &= \sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(d,\ell)} \widehat{g}_{\ell,k}\, \mathrm{Y}_{\ell,k}(\boldsymbol{x}) \\ &= \sum_{\ell=0}^{\infty} \alpha_\ell[h] \sum_{k=1}^{Z(d,\ell)} \mathrm{Y}_{\ell,k}(\boldsymbol{p})\, \mathrm{Y}_{\ell,k}(\boldsymbol{x}) \\ &= \sum_{\ell=0}^{\infty} \alpha_\ell[h]\, Z(d,\ell)\, P_\ell^{(d)}(\boldsymbol{x} \cdot \boldsymbol{p}). \end{aligned}$$

On the other hand, the function h can be expanded (formally) w.r.t. the orthogonal system of normalized ultraspherical polynomials (cf. [40]) by means of

$$\begin{aligned} h(t) &= \sum_{\ell=0}^{\infty} \widehat{h}_\ell \, P_\ell^{(d)}(t), \text{ where} \\ \frac{\widehat{h}_\ell}{Z(d,\ell)} &= \tfrac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} h(t) P_\ell^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t. \end{aligned} \tag{14}$$

(Note that $\frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} P_\ell^{(d)}(t) P_\ell^{(d)}(t)(1-t^2)^{d/2-1} \mathrm{d}t = 1/Z(d,\ell)$; cf., e.g., [30].) Hence, the connecting formula relating the Laplace-Fourier coefficients $\widehat{g}_{\ell,1}, \dots, \widehat{g}_{\ell,Z(d,\ell)}$ of g with the coefficient $\widehat{h}_\ell$ in the ultraspherical expansion of h is

$$\begin{aligned} \widehat{g}_{\ell,k} &= \alpha_\ell[h] \, \mathrm{Y}_{\ell,k}(\boldsymbol{p}) \\ &= \frac{\widehat{h}_\ell}{Z(d,\ell)} \, \mathrm{Y}_{\ell,k}(\boldsymbol{p}), \; k = 1, \dots, Z(d,\ell); \ell = 0, 1, 2, \dots. \end{aligned} \tag{15}$$

Thus, the squared Sobolev norm of the zonal function g is then given by

$$\begin{aligned} \|g\|_{\mathbb{H}^s}^2 &= \sum_{\ell=0}^{\infty} \frac{1}{a_\ell^{(s)}} \sum_{k=1}^{Z(d,\ell)} [\widehat{g}_{\ell,k}]^2 = \sum_{\ell=0}^{\infty} \frac{1}{a_\ell^{(s)}} \sum_{k=1}^{Z(d,\ell)} \left[\frac{\widehat{h}_\ell}{Z(d,\ell)} \mathrm{Y}_{\ell,k}(\boldsymbol{p}) \right]^2 \\ &= \sum_{\ell=0}^{\infty} \frac{Z(d,\ell)}{a_\ell^{(s)}} \left[\frac{\widehat{h}_\ell}{Z(d,\ell)} \right]^2, \end{aligned} \tag{16}$$

where in the last step the addition theorem (9) is used. Clearly, the zonal function g is in $\mathbb{H}^s(\mathbb{S}^d)$ if and only if the last infinite series converges.

Furthermore, using that $\mathrm{Y}_{0,1} \equiv 1$, we have

$$\begin{aligned} \int_{\mathbb{S}^d} g(\boldsymbol{x}) \, \mathrm{d}\sigma_d(\boldsymbol{x}) &= \int_{\mathbb{S}^d} h(\boldsymbol{x} \cdot \boldsymbol{p}) \, \mathrm{d}\sigma_d(\boldsymbol{x}) \\ &= \frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} h(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t = \widehat{h}_0. \end{aligned} \tag{17}$$

We conclude this section with the observation that due to the Funk-Hecke[4] formula the zonal function g satisfies the following relation for $p \in [1, \infty)$:

$$\|g\|_{\mathbb{L}_p(\mathbb{S}^d)}^p = \int_{\mathbb{S}^d} |h(\boldsymbol{x} \cdot \boldsymbol{p})|^p \, \mathrm{d}\sigma_d(\boldsymbol{x}) = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} |h(t)|^p \left(1-t^2\right)^{d/2-1} \mathrm{d}t. \tag{18}$$

[4] The Funk-Hecke formula holds for $\mathbb{L}_1$ functions h; see [3].

3 Proofs

3.1 Preliminaries[5]

We use the Pochhammer symbol defined by $(a)_0 = 1$, $(a)_{n+1} = (a)_n(n-1+a)$ for $n = 0, 1, \dots$, which can be written in terms of the gamma function Γ:

$$(a)_n = \frac{\Gamma(n+a)}{\Gamma(a)}, \qquad (a)_{-n} = \frac{(-1)^n}{(1-a)_n} = \frac{\Gamma(a-n)}{\Gamma(a)}.$$

The incomplete beta function $\mathrm{B}_z(a,b)$, its regularized form $\mathrm{I}_z(a,b)$, and the beta function $\mathrm{B}(a,b)$ are given by ($\operatorname{Re} a, \operatorname{Re} b > 0$)

$$\mathrm{B}_z(a,b) = \int_0^z u^{a-1}\,(1-u)^{b-1}\,\mathrm{d}u, \tag{19}$$

$$\mathrm{I}_z(a,b) = \frac{\mathrm{B}_z(a,b)}{\mathrm{B}(a,b)}, \qquad \mathrm{B}(a,b) = \mathrm{B}_1(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{20}$$

The Gauss hypergeometric function, its regularized form, and the generalized hypergeometric ${}_3\mathrm{F}_2$-function are the analytic continuations of the power series

$${}_2\mathrm{F}_1\left(\begin{matrix}a, b\\ c\end{matrix}; z\right) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k k!} z^k, \qquad {}_2\widetilde{F}_1\left(\begin{matrix}a, b\\ c\end{matrix}; z\right) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{\Gamma(k+c) k!} z^k, \tag{21}$$

$${}_3\mathrm{F}_2\left(\begin{matrix}a_1, a_2, a_3\\ b_1, b_2\end{matrix}; z\right) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k (a_2)_k}{(b_1)_k (b_2)_k} \frac{z^k}{k!}.$$

The Ferrers function of the first kind is defined as

$$\mathrm{P}_\nu^\mu(x) = \left(\frac{1+x}{1-x}\right)^{\mu/2} {}_2\widetilde{F}_1\left(\begin{matrix}-\nu, 1+\nu\\ 1-\mu\end{matrix}; \frac{1-x}{2}\right), \quad \mu, \nu \in \mathbb{R}, -1 < x < 1. \tag{22}$$

Ultraspherical and classical Gegenbauer polynomials $\mathrm{C}_\ell^{(\lambda)}$ are related: For $d = 1$,

$$Z(1,\ell) P_\ell^{(1)}(t) = \lim_{\lambda \to 0} \frac{\ell + \lambda}{\lambda} \mathrm{C}_\ell^{(\lambda)}(t), \qquad P_\ell^{(1)}(t) = \mathrm{T}_\ell(t), \tag{23}$$

where T_ℓ is the *Chebyshev polynomial of the first kind*, and for $d \geq 2$,

$$Z(d,\ell) P_\ell^{(d)}(t) = \frac{\ell + \lambda}{\lambda} \mathrm{C}_\ell^{(\lambda)}(t), \qquad \lambda = \frac{d-1}{2}. \tag{24}$$

[5] For details, see the standard reference NIST Digital Library of Mathematical Functions [30].

We shall, in particular, use that $\mathrm{C}_\ell^{(\lambda)}(1) = (2\lambda)_\ell/\ell!$ and

$$P_0^{(d)} \equiv 1, \quad P_\ell^{(d)}(-t) = (-1)^\ell P_\ell(t), \quad \ell \geq 1, \quad |P_\ell^{(d)}(t)| \leq P_\ell^{(d)}(1) = 1. \tag{25}$$

For the asymptotic analysis part, we collect the following auxiliary results.

The gamma function satisfies the asymptotic relation [30, Eq. 5.11.12]

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} \sim z^{a-b} \qquad \text{as } z \to \infty \text{ in the sector } |\arg z| \leq \pi - \delta(< \pi). \tag{26}$$

Ursell [41] gives the following asymptotic expansion for Gegenbauer polynomials with large degree n in terms of Bessel functions of the first kind,

$$\begin{aligned}\mathrm{C}_n^{(\alpha)}(\cos\theta) \sim{}& \frac{\Gamma(\alpha+1/2)}{\Gamma(2\alpha)} \frac{\mathrm{J}_{\alpha-1/2}(N\theta)}{(N\theta/2)^{\alpha-1/2}} \\ &\times N^{2\alpha-1}\left\{a_0(\theta^2) + \frac{1}{N^2}a_1(\theta^2) + \cdots\right\} \\ &- \frac{\Gamma(3-2\alpha)\Gamma(\alpha-1/2)}{\Gamma(1-2\alpha)\Gamma(2\alpha)} \frac{\mathrm{J}_{\alpha-3/2}(N\theta)}{(N\theta/2)^{\alpha-3/2}} \\ &\times N^{2\alpha-3}\left\{b_0(\theta^2) + \frac{1}{N^2}b_1(\theta^2) + \cdots\right\},\end{aligned} \tag{27}$$

where $N = n + \alpha$ and $\alpha \neq 1, 1/2, 0, -1, -2, \ldots$, the functions $a_p(\theta^2)$ and $b_p(\theta^2)$ are analytic in θ^2, and successive terms in the asymptotic series decay in powers of N^{-2}. This asymptotic expansion is valid in a disc $|\theta| < \pi$ excluding small discs around $\pm\pi$. In particular, one has

$$a_0(\theta^2) := \left(\frac{\theta}{\sin\theta}\right)^\alpha, \quad b_0(\theta^2) := \frac{\alpha}{4}\left(\frac{\theta}{\sin\theta}\right)^\alpha \frac{\sin\theta - \theta\cos\theta}{\theta^2\sin\theta}. \tag{28}$$

We also need the asymptotic expansion of $\mathrm{C}_n^{(1)}(\cos\theta)$ as $n \to \infty$ which can be obtained from the above asymptotic expansion by letting $\alpha \to 1$ (cf. [41, Section 6]).

The Jacobi polynomials admit the uniform asymptotic expansion [16]

$$\begin{aligned}&\left(\sin\frac{\theta}{2}\right)^\alpha \left(\cos\frac{\theta}{2}\right)^\beta P_n^{(\alpha,\beta)}(\cos\theta) \\ &= \frac{\Gamma(n+\alpha+1)}{n!}\left(\frac{\theta}{\sin\theta}\right)^{1/2} \left\{\sum_{\ell=0}^{m-1} A_\ell(\theta)\frac{\mathrm{J}_{\alpha+\ell}(N\theta)}{N^{\alpha+\ell}} + \theta^\alpha \mathscr{O}(N^{-m})\right\}\end{aligned} \tag{29}$$

as $n \to \infty$, where $\alpha > -1/2, \alpha - \beta > -2m, \alpha + \beta \geq -1$, and $N = n + (\alpha + \beta + 1)/2$. The coefficients $A_\ell(\theta)$ are analytic functions for $0 \leq \theta < \pi$ and

$$A_0(\theta) = 1, \qquad A_1(\theta) = \left(\alpha^2 - 1/4\right) \frac{\sin\theta - \theta\cos\theta}{2\theta\sin\theta} - \frac{\alpha^2 - \beta^2}{4} \tan\frac{\theta}{2}. \tag{30}$$

The $\mathscr{O}$-term is uniform with respect to $\theta \in [0, \pi - \delta]$, $\delta > 0$.

For the Jacobi function (α, β, ν real and $\nu + \alpha \neq -1, -2, \dots$)

$$P_\nu^{(\alpha,\beta)}(x) = \frac{\Gamma(\nu + \alpha + 1)}{\Gamma(\nu + 1)} \, {}_2\widetilde{F}_1\left(\begin{matrix} -\nu, \nu + \alpha + \beta + 1 \\ \alpha + 1 \end{matrix}; \frac{1 - x}{2}\right), \tag{31}$$

which reduces to a Jacobi polynomial for ν a non-negative integer (and $\alpha, \beta > -1$), one has the following asymptotic expansion (see [20])

$$\begin{aligned} &(\sin\theta)^\alpha \, (\cos\theta)^\beta \, P_\nu^{(\alpha,\beta)}(\cos(2\theta)) \\ &\quad = \frac{2^{2\nu+\alpha+\beta+3/2} \mathrm{B}(\nu + \alpha + 1, \nu + \beta + 1)}{\pi} \left[\sin(2\theta)\right]^{-1/2} \\ &\qquad \times \left\{ \sum_{\ell=0}^{p-1} \frac{f_\ell(\theta)}{2^\ell (2\nu + \alpha + \beta + 2)_\ell} + \mathscr{O}(\nu^{-p}) \right\} \quad \text{as } \nu \to \infty, \end{aligned} \tag{32}$$

where α, β are real and bounded and ν is real. The functions $f_\ell(\theta)$ are given by

$$\begin{aligned} f_\ell(\theta) = \sum_{k=0}^{\ell} & \frac{(1/2 + \alpha)_k (1/2 - \alpha)_k (1/2 + \beta)_{\ell-k} (1/2 - \alpha)_{\ell-k}}{k! (\ell - k)!} \\ & \times \frac{\cos[(2\nu + \alpha + \beta + \ell + 1)\theta - (\alpha + k + 1/2)\pi/2]}{(\sin\theta)^k \, (\cos\theta)^{\ell-k}}. \end{aligned}$$

The $\mathscr{O}$-term is uniform w.r.t. $\theta \in [\theta_1, \theta_2]$ with $0 < \theta_1 < \theta_2 < \pi/2$. In particular,

$$P_\nu^{(\alpha,\beta)}(\cos(2\theta)) \sim \frac{1}{\sqrt{\pi}} \frac{\cos[(2\nu + \alpha + \beta + 1)\theta - (\alpha + 1/2)\pi/2]}{(\sin\theta)^{\alpha+1/2} \, (\cos\theta)^{\beta+1/2}} \, \nu^{-1/2}. \tag{33}$$

The Bessel functions of the first kind have the asymptotic behaviour [30, Eq. 10.17.3]

$$\mathrm{J}_\nu(z) \sim \left(\frac{2}{\pi z}\right)^{1/2} \left\{ \cos\omega \sum_{k=0}^{\infty} (-1)^k \frac{a_{2k}(\nu)}{z^{2k}} - \sin\omega \sum_{k=0}^{\infty} (-1)^k \frac{a_{2k+1}(\nu)}{z^{2k+1}} \right\} \tag{34}$$

as $z \to \infty$ for $|\arg z| \leq \pi - \delta$, where $\omega = z - \nu\pi/2 - \pi/4$ and the coefficients $a_k(\nu)$ are given in [30, Eq. 10.17.1]. This asymptotic form motivates the definition of the (asymptotically) normalized Bessel function of the first kind

$$\mathfrak{J}_\nu(z) := \left(\frac{2}{\pi z}\right)^{-1/2} \mathrm{J}_\nu(z), \qquad |\arg z| \leq \pi - \delta. \tag{35}$$

Regarding a remainder estimate, we will use for $\nu > -1/2$ (cf. Olver [31, Sec. 8.11.3]),

$$\mathrm{J}_\nu(x) = \left(\frac{2}{\pi x}\right)^{1/2} \left\{\cos\left(x - \frac{2\nu+1}{2}\frac{\pi}{2}\right) + \mathscr{O}(x^{-1})\right\} \quad \text{as } x \to \infty (x \text{ real}). \tag{36}$$

The Ferrers function of the first kind has the asymptotic behavior [30, Eq. 14.15.11]

$$\begin{aligned} \mathrm{P}_\nu^{-\mu}(\cos\theta) &= \frac{1}{\nu^\mu}\left(\frac{\theta}{\sin\theta}\right)^{1/2} \\ &\times \left\{\mathrm{J}_\mu\left((\nu+\frac{1}{2})\theta\right) + \mathscr{O}\left(\frac{1}{\nu}\right) \operatorname{env} \mathrm{J}_\mu\left((\nu+\frac{1}{2})\theta\right)\right\} \end{aligned} \tag{37}$$

as $\nu \to \infty$ and μ (≥ 0) fixed, where the remainder term is represented by means of the envelope function associated with J_μ. The convergence is uniform for $\theta \in (0, \pi - \delta]$.

3.2 *The Trial Functions of Type (1)*

Set $h \equiv f_1$. Let $T = 0$. For $d \geq 1$, the Laplace-Fourier coefficients take the form

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d} \int_0^1 t^\alpha P_\ell^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t, \qquad \ell = 0, 1, 2, \dots. \tag{38}$$

The integral is finite for all $\ell = 0, 1, 2, \dots$ if and only if $\alpha > -1$ as can be seen from the behavior of the integrand near $t = 0$.

Lemma 1 *Let $d \geq 1$, $T = 0$, and $\alpha > -1$. Then for $\ell = 2m + \varepsilon$ ($\varepsilon = 0$ or $\varepsilon = 1$),*

$$\begin{aligned} \frac{\widehat{h}_{2m+\varepsilon}}{Z(d, 2m+\varepsilon)} &= \frac{(-1)^m}{2} \frac{\Gamma((d+1)/2)\Gamma((\varepsilon+\alpha+1)/2)}{\sqrt{\pi}} \\ &\times \frac{((\varepsilon-\alpha)/2)_m}{\Gamma(m+(d+1+\alpha+\varepsilon)/2)}. \end{aligned} \tag{39}$$

Proof Let $T = 0$. Let $d \geq 2$. By (24), the integral in (38) is a special case of [33, Eq. 2.21.2.5] valid for $\alpha > -1 - \varepsilon$. Let $d = 1$. Because of (23), the integral in (38) is a special case of [33, Eq. 2.18.1.1]. We get (39) but valid for $\alpha > -1$. □

We remark that if $\alpha = 0, 1, 2, \dots$, then $\widehat{h}_\ell = 0$ for $\ell = \alpha + 2, \alpha + 4, \dots$, since $\widehat{h}_{2m+\varepsilon}$ vanishes whenever $\alpha - \varepsilon = 0, 2, 4, \dots$ and $2m + \varepsilon > \alpha$.

Lemma 2 *Let $d \geq 1$, $T = 0$, and $\alpha > -1$. Then for $\alpha - \varepsilon \neq 0, 2, 4, 6, \dots$ (as $m \to \infty$),*

$$\frac{\widehat{h}_{2m+\varepsilon}}{Z(d, 2m+\varepsilon)} \sim \frac{(-1)^m}{2} \frac{\Gamma((d+1)/2)\Gamma((\varepsilon+\alpha+1)/2)}{\sqrt{\pi}\,\Gamma((\varepsilon-\alpha)/2)} m^{-(d+1)/2-\alpha}. \tag{40}$$

Proof We expand $((\varepsilon - \alpha)/2)_m$ in (39) and use (26) to obtain (40). □

Let $0 < |T| < 1$. For $d \geq 1$, the Laplace-Fourier coefficients take the form

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d} \int_T^1 (t-T)^\alpha P_\ell^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t, \qquad \ell = 0, 1, 2, \dots. \tag{41}$$

The integral is well-defined and finite for all $\ell = 0, 1, 2, \dots$ if and only if $\alpha > -1$ as can be seen from the behaviour of the integrand near the critical point $t = T$.

Lemma 3 *Let $d \geq 1$, $0 < |T| < 1$, and $\alpha > -1$. Then for $\ell = 0, 1, 2, \dots$,*

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = 2^{d/2-1} \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\sqrt{\pi}} \left(1-T^2\right)^{(\alpha+d/2)/2} \mathrm{P}_{\ell+d/2-1}^{-(\alpha+d/2)}(T)$$

in terms of the Ferrers function defined in (22).

Proof Let $d \geq 2$ and $0 < T < 1$. Set $\lambda = (d-1)/2$. By (24), the integral in (41) is a special case of [33, Eq. 2.21.4.10]. Thus, using (41) and (5), we get

$$\begin{aligned} \widehat{h}_\ell = {} & \frac{\Gamma(\lambda+1)}{\sqrt{\pi}\,\Gamma(\lambda+1/2)} \frac{\ell+\lambda}{\lambda} \frac{(2\lambda)_\ell}{\ell!} \\ & \times \mathrm{B}(\alpha+1, \lambda+1/2) 2^{\lambda-1/2} (1-T)^{\alpha+\lambda+1/2} \\ & \times {}_3\mathrm{F}_2\left(\begin{matrix} \lambda+1/2, 1/2-\lambda-\ell, \ell+\lambda+1/2 \\ \alpha+\lambda+3/2, \lambda+1/2 \end{matrix} ; \frac{1-T}{2} \right), \end{aligned} \tag{42}$$

provided $\alpha > -1$ and $\lambda > -1/2$ and with the understanding that, by (23) and (25),

$$\lim_{\lambda \to 0} \frac{\ell+\lambda}{\lambda} \frac{(2\lambda)_\ell}{\ell!} = \lim_{\lambda \to 0} \frac{\ell+\lambda}{\lambda} \mathrm{C}_\ell^{(\lambda)}(1) = Z(1,\ell)\, P_\ell^{(d)}(1) = Z(1,\ell). \tag{43}$$

Since the hypergeometric function reduces to a regularized Gauss hypergeometric function, expanding the beta function and simplifying expressions yields

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{2^{1-d/2}\sqrt{\pi}} (1-T)^{\alpha+d/2} {}_2\widetilde{F}_1\left(\begin{matrix} 1-d/2-\ell, \ell+d/2 \\ \alpha+1+d/2 \end{matrix} ; \frac{1-T}{2} \right).$$

By (43), this formula holds for all $d \geq 1$. The hypergeometric function can be expressed in terms of a Ferrers function (cf. (22)) and we obtain the desired result for $0 < T < 1$. If $-1 < T < 0$, then relation (42) follows from [33, Eq. 2.21.4.6]. □

Lemma 4 *Let $d \geq 1$, $0 < |T| < 1$ with $T =: \cos\theta_T$ for $0 < \theta_T < \pi$, and $\alpha > -1$. Then (as $\ell \to \infty$)*

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} \sim 2^{(d-1)/2}\frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\pi} \times (\sin\theta_T)^{\alpha+(d-1)/2}\,\frac{\mathfrak{J}_{\alpha+d/2}\big((\ell+(d-1)/2)\theta_T\big)}{\ell^{(d+1)/2+\alpha}}. \tag{44}$$

Proof We express (37) in terms of the normalized Bessel function,

$$\mathrm{P}_\nu^{-\mu}(\cos\theta_T) \sim \left(\frac{2}{\pi}\right)^{1/2} \nu^{-\mu-1/2}\,(\sin\theta_T)^{-1/2}\,\mathfrak{J}_\mu\big((\nu+1/2)\theta_T\big) \qquad \text{as } \nu \to \infty,$$

and (44) follows from Lemma 3. □

Proof of Theorem 1 Let $T = 0$. Using the asymptotics (7), (11), and (40), we obtain for $\ell = 2m + \varepsilon$ with $\varepsilon \in \{0, 1\}$ and $\alpha - \varepsilon \neq 0, 2, 4, 6, \ldots$,

$$\frac{Z(d,\ell)}{a_\ell^{(s)}}\left[\frac{\widehat{h}_\ell}{Z(d,\ell)}\right]^2 \asymp \ell^{2s}\,\ell^{d-1}\left[m^{-(d+1)/2-\alpha}\right]^2 \asymp \ell^{2s-2\alpha-2} \quad \text{as } \ell \to \infty.$$

Hence, the series in (16) converges if and only if $2s - 2\alpha - 2 < -1$; i.e., the zonal trial function g_1 belongs to $\mathbb{H}^s(\mathbb{S}^d)$, $s > 0$, if and only if $0 < s < \alpha + 1/2$.

Let $0 < |T| < 1$. Then (7), (11), and (44) imply $\frac{Z(d,\ell)}{a_\ell^{(s)}}\left[\frac{\widehat{h}_\ell}{Z(d,\ell)}\right]^2 \asymp \ell^{2s-2\alpha-2}$ as $\ell \to \infty$. Hence, g_1 belongs to $\mathbb{H}^s(\mathbb{S}^d)$, $s > 0$, if and only if $0 < s < \alpha + 1/2$. □

3.3 The Trial Functions of Type (2)

Set $h \equiv f_2$. Let $T = 0$. For $d \geq 1$, the Laplace-Fourier coefficients take the form

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d}\int_{-1}^{1} |t|^{2\alpha}\,P_\ell^{(d)}(t)\,\big(1-t^2\big)^{d/2-1}\,\mathrm{d}t, \qquad \ell = 0, 1, 2, \ldots. \tag{45}$$

This integral is finite for all $\ell = 0, 1, 2, \ldots$ if and only if $\alpha > -1/2$ as can be seen from the behavior of the integrand near $t = 0$.

Lemma 5 *Let $d \geq 1$, $T = 0$, and $\alpha > -1/2$. Then $\widehat{h}_{2m+1} = 0$ and*

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = (-1)^m \frac{\Gamma((d+1)/2)\Gamma(\alpha+1/2)}{\sqrt{\pi}\,\Gamma(-\alpha)} \frac{\Gamma(m-\alpha)}{\Gamma(m+(d+1)/2+\alpha)}, \tag{46}$$

where $\Gamma(m-\alpha)/\Gamma(-\alpha)$ is interpreted as $(-\alpha)_m$ if α is a non-negative integer.

Proof By (25), $\widehat{h}_\ell = 0$ for $\ell = 1, 3, \dots$ and

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = 2\,\frac{\omega_{d-1}}{\omega_d} \int_0^1 t^{2\alpha}\, P_{2m}^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t, \qquad m = 0, 1, 2, \dots.$$

The integral has essentially been dealt with in the proof of Lemma 1. □

We remark that if $\alpha = 0, 1, 2, \dots$, then the coefficient $\widehat{h}_\ell$ vanishes for $\ell \geq 2\alpha + 1$.

Lemma 6 *Let $d \geq 1$, $T = 0$, and $\alpha > -1/2$ not an integer. Then*

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} \sim (-1)^m \frac{\Gamma((d+1)/2)\Gamma(\alpha+1/2)}{\sqrt{\pi}\,\Gamma(-\alpha)}\, m^{-(d+1)/2-2\alpha} \quad \text{as } m \to \infty. \tag{47}$$

Proof We apply to (46) the asymptotic expansion (26) and obtain (47). □

Let $0 < T < 1$. For $d \geq 1$, the Laplace-Fourier coefficients take the form

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^1 \left(t^2 - T^2\right)_+^\alpha P_\ell^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t, \quad \ell = 0, 1, 2, \dots. \tag{48}$$

The integral is well-defined and finite for all $\ell = 0, 1, 2, \dots$ if and only if $\alpha > -1$ as can be seen from the behaviour of the integrand near the critical points $t = \pm T$.

Lemma 7 *Let $d \geq 1$, $0 < T < 1$, and $\alpha > -1$. Then $\widehat{h}_{2m+1} = 0$ and*

$$\begin{aligned} \frac{\widehat{h}_{2m}}{Z(d,2m)} &= \frac{\Gamma((d+1)/2)}{\sqrt{\pi}} \frac{\Gamma(\alpha+1)\Gamma(m+1/2)}{\Gamma(m+\alpha+(d+1)/2)} \\ &\quad \times \left(1-T^2\right)^{\alpha+d/2} P_{m-1/2}^{(\alpha+d/2,-\alpha-1/2)}(2T^2-1). \end{aligned}$$

Proof By (25), $\widehat{h}_\ell = 0$ for $\ell = 1, 3, \dots$ and

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = 2\,\frac{\omega_{d-1}}{\omega_d} \int_T^1 \left(t^2 - T^2\right)^\alpha P_{2m}^{(d)}(t) \left(1-t^2\right)^{d/2-1} \mathrm{d}t, \quad m = 0, 1, 2, \dots.$$

Let $d \geq 2$. Using (5) and (24), we have $\widehat{h}_{2m} = 2H_{2m}((d-1)/2, \alpha; T)$, where

$$H_{2m}(\lambda, \alpha; T) := \frac{\Gamma(\lambda+1)}{\sqrt{\pi}\,\Gamma(\lambda+1/2)} \frac{2m+\lambda}{\lambda} \int_T^1 \left(t^2 - T^2\right)^\alpha \mathrm{C}_{2m}^{(\lambda)}(t) \left(1-t^2\right)^{\lambda-1/2} \mathrm{d}t.$$

The above integral is a special case of [33, Eq. 2.21.4.3]. Transformations like [30, Eqs. 15.8.1 and 15.8.4] in the arising Gauss hypergeometric functions yield

$$H_{2m}(\lambda,\alpha;T) = \frac{1}{2}\frac{\Gamma(\lambda+1)\Gamma(\alpha+1)}{\sqrt{\pi}}\frac{2m+\lambda}{\lambda}\frac{(2\lambda)_{2m}}{(2m)!}$$
$$\times\left(1-T^2\right)^{\alpha+\lambda+1/2}\,{}_2\widetilde{F}_1\left(\begin{matrix}1/2-m, m+\lambda+1/2\\ \alpha+\lambda+3/2\end{matrix};1-T^2\right),$$

valid for $0<T<1$, $\lambda>-1/2$ with $\lambda\neq 0$, and $\alpha>-1$ with $\alpha\neq -1/2, 1/2, 3/2,\dots$ which also extends to the case when $1/2+\alpha$ is an integer. The formula for $d=1$ follows by taking the limit as $\lambda\to 0$. Hence, for all $d\geq 1$, $0<T<1$, and $\alpha>-1$,

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\sqrt{\pi}}\left(1-T^2\right)^{\alpha+d/2}\,{}_2\widetilde{F}_1\left(\begin{matrix}1/2-m, m+d/2\\ d/2+\alpha+1\end{matrix};1-T^2\right).$$

The result follows by changing to Jacobi functions (cf. (31)). □

Lemma 8 *Let $d\geq 1$, $0<T=:\cos\theta_T<1$, and $\alpha>-1$. Then (as $m\to\infty$)*

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} \sim \frac{\Gamma((d+1)/2)}{\pi}\Gamma(\alpha+1)\,(\sin\theta_T)^{\alpha+(d-1)/2}$$
$$\times(\cos\theta_T)^{\alpha}\cos\left[(2m+\frac{d-1}{2})\theta_T-(\alpha+\frac{d+1}{2})\pi/2\right]m^{-\alpha-(d+1)/2}.$$

Proof The result follows from Lemma 7 using the asymptotics (33) and (26). □

Proof of Theorem 2 Let $T=0$ and $\alpha>-1/2$ and α not an integer. Using the asymptotics (7), (11), and (47), we obtain for the even terms in (16),

$$\frac{Z(d,2m)}{a_{2m}^{(s)}}\left[\frac{\widehat{h}_{2m}}{Z(d,2m)}\right]^2 \asymp m^{2s}\,m^{d-1}\left[m^{-(d+1)/2-2\alpha}\right]^2 = m^{2s-4\alpha-2} \qquad \text{as } m\to\infty;$$

i.e., the series in (16) converges iff $2s-4\alpha-2<-1$. Thus, $g_2\in\mathbb{H}^s(\mathbb{S}^d)$, $s>0$, iff $0<s<2\alpha+1/2$. If $\alpha=0,1,2,\dots$, then the Laplace-Fourier expansion of g_2 has only finitely many terms and $g_2\in\mathbb{H}^s(\mathbb{S}^d)$ for all $s\geq 0$.

Let $0<T<1$ and $\alpha>-1$. Then (7), (11), and Lemma 8 imply $\frac{Z(d,2m)}{a_{2m}^{(s)}}\left[\frac{\widehat{h}_{2m}}{Z(d,2m)}\right]^2 \asymp m^{2s-2\alpha-2}$ as $m\to\infty$ and $g_2\in\mathbb{H}^s(\mathbb{S}^d)$, $s>0$, iff $0<s<\alpha+1/2$. □

3.4 The Trial Functions of Type (3)

Set $h \equiv f_3$. Let $0 < T < 1$. For $d \geq 1$, the Laplace-Fourier coefficients take the form

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} \left(T^2 - t^2\right)_+^\alpha P_\ell^{(d)}(t) \left(1 - t^2\right)^{d/2-1} \mathrm{d}t, \quad \ell = 0, 1, 2, \dots. \tag{49}$$

The integral is well-defined and finite for all $\ell = 0, 1, 2, \dots$ if and only if $\alpha > -1$ as can be seen from the behaviour of the integrand near the critical points $t = \pm T$.

Lemma 9 *Let $d \geq 1$, $0 < T = \cos\theta_T < 1$, and $\alpha > -1$. Then $\widehat{h}_{2m+1} = 0$ and*

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = (-1)^m \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\sqrt{\pi}} \frac{\Gamma(m+1/2)}{\Gamma(m+(d+1)/2+\alpha)} \\ \times (\cos\theta_T)^{1+2\alpha} P_{m+d/2-1}^{(\alpha+1/2,-\alpha-d/2)}(\cos(2(\pi/2-\theta_T))).$$

Proof By (25), $\widehat{h}_\ell = 0$ for $\ell = 1, 3, 5, \dots$ and $\widehat{h}_{2m} = 2H_{2m}((d-1)/2, \alpha; T)$, where

$$H_{2m}(\lambda, \alpha; T) := \frac{\Gamma(\lambda+1)}{\sqrt{\pi}\,\Gamma(\lambda+1/2)} \frac{2m+\lambda}{\lambda} \int_0^T \left(T^2 - t^2\right)^\alpha \mathrm{C}_{2m}^{(\lambda)}(t) \left(1 - t^2\right)^{\lambda-1/2} \mathrm{d}t.$$

The above integral is a special case of [33, Eq. 2.21.4.1] valid for $0 < T < 1$, $\alpha > -1$, and $\lambda > -1/2$. The formula for $d = 1$ follows by taking the limit as $\lambda \to 0$. Hence,

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} = (-1)^m \frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\Gamma(d/2)} \frac{(1/2)_m}{(d/2)_m} \\ \times T^{2\alpha+1} {}_2\widetilde{F}_1\left(\begin{matrix} 1-d/2-m, m+1/2 \\ \alpha+3/2 \end{matrix}; T^2\right).$$

The result follows by changing to a Jacobi function (cf. (31)) and using the substitution $T = \cos\theta_T$. (Note that the argument of the Jacobi function is $1 - 2T^2$.) □

Lemma 10 *Let $d \geq 1$, $0 < T = \cos\theta_T < 1$, and $\alpha > -1$. Then (as $m \to \infty$)*

$$\frac{\widehat{h}_{2m}}{Z(d,2m)} \sim -\frac{\Gamma((d+1)/2)\Gamma(\alpha+1)}{\pi} (\cos\theta_T)^\alpha (\sin\theta_T)^{\alpha+(d-1)/2} \\ \times \sin[(2m+(d-1)/2)\theta_T + (\alpha-(d-1)/2)\pi/2]\, m^{-\alpha-(d+1)/2}.$$

Proof The result follows from Lemma 9 using (32) and (26). □

Proof of Theorem 3 Using (7) and (11), Lemma 10 implies $\frac{Z(d,2m)}{a_{2m}^{(s)}}\left[\frac{\widehat{h}_{2m}}{Z(d,2m)}\right]^2 \asymp m^{2s-2\alpha-2}$ as $m \to \infty$ and $g_3 \in \mathbb{H}^s(\mathbb{S}^d)$, $s > 0$, iff $0 < s < \alpha + 1/2$. □

3.5 The Trial Functions of Type (4)

Let $h = f_4$. The Laplace-Fourier coefficients can be written as

$$\frac{\widehat{h}_\ell}{Z(d,\ell)} = \frac{\omega_{d-1}}{\omega_d}\int_{-1}^{1}[\max\{t,T\}]^\alpha\, P_\ell^{(d)}(t)\left(1-t^2\right)^{d/2-1}\mathrm{d}t = T^\alpha \mathscr{A}_\ell + \mathscr{D}_\ell, \tag{50}$$

where the integrals

$$\mathscr{A}_\ell := \frac{\omega_{d-1}}{\omega_d}\int_{-1}^{T} P_\ell^{(d)}(t)\left(1-t^2\right)^{d/2-1}\mathrm{d}t,\ \ \mathscr{D}_\ell := \frac{\omega_{d-1}}{\omega_d}\int_{T}^{1} t^\alpha\, P_\ell^{(d)}(t)\left(1-t^2\right)^{d/2-1}\mathrm{d}t$$

are well-defined and finite for all $\ell = 0, 1, 2, \dots$, $T > 0$, and $\alpha \in \mathbb{R}$. Set

$$A_1 := \frac{\mathrm{B}(d/2,(\alpha+1)/2)}{\mathrm{B}(d/2,1/2)} = \frac{\Gamma((d+1)/2)\Gamma((\alpha+1)/2)}{\sqrt{\pi}\,\Gamma((d+1+\alpha)/2)}.$$

Lemma 11 *Let $d \geq 1$, $0 < T < 1$, and $\alpha > -1$. Then*

$$\widehat{h}_0 = T^\alpha\left(1 - \frac{1}{2}\,\mathrm{I}_{1-T^2}(d/2,1/2)\right) + \frac{1}{2}A_1\,\mathrm{I}_{1-T^2}(d/2,(\alpha+1)/2)$$

and for $\ell = 2m + \varepsilon \geq 1$ ($\varepsilon = 0$ or $\varepsilon = 1$),

$$\begin{aligned}\frac{\widehat{h}_{2m+\varepsilon}}{Z(d,2m+\varepsilon)} &= \frac{(-1)^m}{2}\,\frac{\Gamma((d+1)/2)\Gamma((\varepsilon+\alpha+1)/2)}{\sqrt{\pi}}\,\frac{((\varepsilon-\alpha)/2)_m}{\Gamma(m+(d+1+\alpha+\varepsilon)/2)}\\ &\quad - (-1)^{m+\varepsilon-1}T^{\alpha+1-\varepsilon}\,\frac{2^{d-\varepsilon}\alpha\Gamma((d+1)/2)}{\pi(\alpha+1-\varepsilon)}\,\frac{\Gamma(m+(d+1)/2)\Gamma(2m+\varepsilon)}{\Gamma(m+\varepsilon)\Gamma(2m+\varepsilon+d)}\\ &\quad \times {}_3\mathrm{F}_2\!\left(\begin{matrix}1-d/2-m-\varepsilon, m+1/2, (\alpha+1-\varepsilon)/2\\ 3/2-\varepsilon, 1+(\alpha+1-\varepsilon)/2\end{matrix}; T^2\right).\end{aligned}$$

Proof Let $\ell = 0$. Then $P_0^{(d)} \equiv 1$ and $Z(d,0) = 1$. Hence, (50) yields

$$\widehat{h}_0 = T^\alpha\,\mathscr{A}_0 + \mathscr{D}_0.$$

The integral $\mathscr{A}_0$ is the σ_d-measure of the spherical cap $\{\boldsymbol{x} \in \mathbb{S}^d : \boldsymbol{x}\cdot\boldsymbol{p} \leq T\}$:

$$\mathscr{A}_0 = \mathrm{I}_{(1+T)/2}(d/2,d/2) = 1 - \frac{1}{2}\,\mathrm{I}_{1-T^2}(d/2,1/2).$$

Since $T > 0$, the substitution $u = 1 - t^2$ reduces $\mathscr{D}_0$ to an incomplete beta function,

$$\mathscr{D}_0 = \frac{1}{2} \frac{\omega_{d-1}}{\omega_d} \mathrm{B}_{1-T^2}(d/2, (\alpha + 1)/2).$$

The final form of the coefficient $\widehat{h}_0$ follows then by using (5) and (20).

Let $\ell \geq 1$. By orthogonality of the ultraspherical polynomials, $\mathscr{A}_\ell$ in (50) becomes

$$\mathscr{A}_\ell = -\frac{\omega_{d-1}}{\omega_d} \int_T^1 P_\ell^{(d)}(t) \left(1 - t^2\right)^{d/2-1} \mathrm{d}t.$$

Inserting (5) and (24) into (50), we get $\widehat{h}_\ell = H_\ell((d-1)/2, \alpha; T)$ for $d \geq 2$, where

$$H_\ell(\lambda, \alpha; T) = \frac{\ell + \lambda}{\lambda} \frac{\Gamma(\lambda + 1)}{\sqrt{\pi}\, \Gamma(\lambda + 1/2)} \int_T^1 (t^\alpha - T^\alpha)\, \mathrm{C}_\ell^{(\lambda)}(t) \left(1 - t^2\right)^{\lambda - 1/2} \mathrm{d}t.$$

Integration by parts, using relation [30, Eq. 18.9.20] yields

$$\begin{aligned} H_\ell(\lambda, \alpha; T) = & \frac{2\alpha\, \Gamma(\lambda + 1)}{\sqrt{\pi}\, \Gamma(\lambda + 1/2)} \frac{\ell + \lambda}{\ell\, (\ell + 2\lambda)} \\ & \times \int_T^1 t^{\alpha - 1}\, \mathrm{C}_{\ell-1}^{(\lambda+1)}(t) \left(1 - t^2\right)^{\lambda + 1/2} \mathrm{d}t. \end{aligned} \tag{51}$$

This formula holds, in particular, in the limit as $\lambda \to 0$ which corresponds to $d = 1$. The above integral is a special case of [33, Eq. 2.21.4.3]. The result follows. □

Lemma 12 *Let* $d \geq 1$, $0 < T = \cos\theta_T < 1$, *and* $\alpha > -1$. *Then*

$$\begin{aligned} \frac{\widehat{h}_\ell}{Z(d, \ell)} \sim & \, \alpha 2^{(d-1)/2} \frac{\Gamma((d + 1)/2)}{\pi} (\cos\theta_T)^{\alpha - 1} (\sin\theta_T)^{(d+1)/2} \\ & \times \frac{\mathfrak{J}_{d/2+1}\Big((\ell + (d - 1)/2)\, \theta_T \Big)}{\ell^{(d-1)/2+2}} \qquad \textit{as } \ell \to \infty. \end{aligned}$$

Proof Let $d \geq 2$. Set $\lambda = (d - 1)/2$ and $T = \cos\theta_T$. Then (51) turns into

$$\frac{\widehat{h}_n}{Z(d, n)} = \frac{\alpha}{2\lambda + 1} \frac{\omega_{d-1}}{\omega_d} \int_0^{\theta_T} (\cos\theta)^{\alpha - 1} \frac{\mathrm{C}_{n-1}^{(\lambda+1)}(\cos\theta)}{\mathrm{C}_{n-1}^{(\lambda+1)}(1)} (\sin\theta)^{2\lambda + 2} \mathrm{d}\theta.$$

Since $\lambda + 1 \neq 1, 1/2, 0, -1, -2, \ldots$, by Ursell's result (27), and exploiting integration by parts and properties of the Bessel functions, we arrive at

$$\frac{\widehat{h}_n}{Z(d,n)} \sim \alpha 2^{\lambda} \frac{\Gamma(\lambda+1)}{\pi} \frac{\Gamma(n)}{\Gamma(n+2\lambda+1)} N^{\lambda-1} (\cos\theta_T)^{\alpha-1} (\sin\theta_T)^{\lambda+1} \mathfrak{J}_{\lambda+3/2}(N\theta_T)$$
$$+ \frac{\Gamma(n)}{\Gamma(n+2\lambda+1)} \mathcal{O}\big(N^{\lambda-3/2}\big) \qquad \text{as } N := n+\lambda \to \infty.$$

The result follows by using (26) and substituting $N = n + \lambda$ and $\lambda = (d-1)/2$.

The result for $d = 1$ follows in a similar way by taking into account that Ursell's asymptotic expansion also holds as $\lambda \to 0$. □

Proof of Theorem 4 Using asymptotics (7) and (11), by Lemma 12,

$$\frac{Z(d,\ell)}{a_\ell^{(s)}} \left[\frac{f_\ell}{Z(d,\ell)} \right]^2 \asymp \ell^{2s} \, \ell^{d-1} \left[\ell^{-(d-1)/2-2} \right]^2 = \ell^{2s-4} \qquad \text{as } \ell \to \infty.$$

Hence, $g_4 \in \mathbb{H}^s(\mathbb{S}^d)$, $s > 0$, iff $0 < s < 3/2$. □

Acknowledgements The research of this author was supported, in part, by the Austrian Science Fund FWF project F5510 (part of the Special Research Program (SFB) "Quasi-Monte Carlo Methods: Theory and Applications") and M2030 Lise Meitner Programm "Self organization by local interaction".

References

1. An, C., Chen, X., Sloan, I.H., Womersley, R.S.: Well conditioned spherical designs for integration and interpolation on the two-sphere. SIAM J. Numer. Anal. **48**(6), 2135–2157 (2010)
2. An, C., Chen, X., Sloan, I.H., Womersley, R.S.: Regularized least squares approximations on the sphere using spherical designs. SIAM J. Numer. Anal. **50**(3), 1513–1534 (2012)
3. Berens, H., Butzer, P.L., Pawelke, S.: Limitierungsverfahren von Reihen mehrdimensionaler Kugelfunktionen und deren Saturationsverhalten. Publ. Res. Inst. Math. Sci. Ser. A **4**, 201–268 (1968/1969)
4. Bilyk, D., Lacey, T.: One bit sensing, discrepancy, and Stolarsky principle. Mat. Sb. **208**(6), 4–25 (2017)
5. Bondarenko, A., Radchenko, D., Viazovska, M.: Optimal asymptotic bounds for spherical designs. Ann. Math. (2) **178**(2), 443–452 (2013)
6. Brauchart, J.S., Dick, J.: Quasi-Monte Carlo rules for numerical integration over the unit sphere $\mathbb{S}^2$. Numer. Math. **121**(3), 473–502 (2012)
7. Brauchart, J.S., Dick, J.: A characterization of Sobolev spaces on the sphere and an extension of Stolarsky's invariance principle to arbitrary smoothness. Constr. Approx. **38**(3), 397–445 (2013)
8. Brauchart, J.S., Dick, J.: A simple proof of Stolarsky's invariance principle. Proc. Am. Math. Soc. **141**(6), 2085–2096 (2013)

9. Brauchart, J.S., Hesse, K.: Numerical integration over spheres of arbitrary dimension. Constr. Approx. **25**(1), 41–71 (2007)
10. Brauchart, J.S., Saff, E.B., Sloan, I.H., Womersley, R.S.: QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere. Math. Comput. **83**(290), 2821–2851 (2014)
11. Brauchart, J.S., Dick, J., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces. J. Math. Anal. Appl. **431**(2), 782–811 (2015)
12. Brauchart, J.S., Dick, J., Fang, L.: Spatial low-discrepancy sequences, spherical cone discrepancy, and applications in financial modeling. J. Comput. Appl. Math. **286**, 28–53 (2015)
13. Chen, X., Womersley, R.S.: Existence of solutions to systems of underdetermined equations and spherical designs. SIAM J. Numer. Anal. **44**(6), 2326–2341 (electronic) (2006)
14. Chen, X., Frommer, A., Lang, B.: Computational existence proofs for spherical t-designs. Numer. Math. **117**(2), 289–305 (2011)
15. Delsarte, P., Goethals, J.M., Seidel, J.J.: Spherical codes and designs. Geom. Dedicata **6**(3), 363–388 (1977)
16. Frenzen, C.L., Wong, R.: A uniform asymptotic expansion of the Jacobi polynomials with error bounds. Can. J. Math. **37**(5), 979–1007 (1985)
17. Genz, A.: Testing multidimensional integration routines. In: Proceedings of International Conference on Tools, Methods and Languages for Scientific and Engineering Computation, pp. 81–94. Elsevier North-Holland, Inc., New York, NY (1984)
18. Genz, A.: Fully symmetric interpolatory rules for multiple integrals over hyper-spherical surfaces. J. Comput. Appl. Math. **157**(1), 187–195 (2003)
19. Grabner, P.J., Klinger, B., Tichy, R.F.: Discrepancies of point sequences on the sphere and numerical integration. In: Multivariate Approximation (Witten-Bommerholz, 1996). Mathematical Research, vol. 101, pp. 95–112. Akademie Verlag, Berlin (1997)
20. Hahn, E.: Asymptotik bei Jacobi-Polynomen und Jacobi-Funktionen. Math. Z. **171**(3), 201–226 (1980)
21. Hesse, K.: A lower bound for the worst-case cubature error on spheres of arbitrary dimension. Numer. Math. **103**(3), 413–433 (2006)
22. Hesse, K., Sloan, I.H.: Optimal lower bounds for cubature error on the sphere S^2. J. Complexity **21**(6), 790–803 (2005)
23. Hesse, K., Sloan, I.H.: Worst-case errors in a Sobolev space setting for cubature over the sphere S^2. Bull. Aust. Math. Soc. **71**(1), 81–105 (2005)
24. Hesse, K., Sloan, I.H.: Cubature over the sphere S^2 in Sobolev spaces of arbitrary order. J. Approx. Theory **141**(2), 118–133 (2006)
25. Hesse, K., Sloan, I.H.: Hyperinterpolation on the sphere. In: Frontiers in Interpolation and Approximation. Pure and Applied Mathematics (Boca Raton), vol. 282, pp. 213–248. Chapman & Hall/CRC, Boca Raton, FL (2007)
26. Hesse, K., Sloan, I.H., Womersley, R.S.: Numerical Integration on the Sphere, pp. 2671–2710. Springer, Berlin (2015)
27. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**(221), 299–322 (1998)
28. Mhaskar, H.N., Narcowich, F.J., Ward, J.D.: Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature. Math. Comput. **70**(235), 1113–1130 (2001)
29. Müller, C.: Spherical Harmonics. Lecture Notes in Mathematics, vol. 17. Springer, Berlin (1966)
30. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.10 of 2015-08-07
31. Olver, F.W.J.: Asymptotics and Special Functions. Computer Science and Applied Mathematics. Academic [A Subsidiary of Harcourt Brace Jovanovich, Publishers], New York (1974)
32. Owen, A.B.: The dimension distribution and quadrature test functions. Stat. Sin. **13**(1), 1–17 (2003)

33. Prudnikov, A.P., Brychkov, Y.A., Marichev, O.I.: Integrals and Series. Special functions, vol. 2. Gordon & Breach Science Publishers, New York (1986). Translated from the Russian by N.M. Queen
34. Renka, R.J.: Multivariate interpolation of large sets of scattered data. ACM Trans. Math. Softw. **14**, 139–148 (1988)
35. Sloan, I.H., Womersley, R.S.: Extremal systems of points and numerical integration on the sphere. Adv. Comput. Math. **21**(1–2), 107–125 (2004)
36. Sloan, I.H., Womersley, R.S.: A variational characterisation of spherical designs. J. Approx. Theory **159**(2), 308–318 (2009)
37. Sloan, I.H., Womersley, R.S.: Filtered hyperinterpolation: a constructive polynomial approximation on the sphere. GEM Int. J. Geomath. **3**(1), 95–117 (2012)
38. Stolarsky, K.B.: Sums of distances between points on a sphere. II. Proc. Am. Math. Soc. **41**, 575–582 (1973)
39. Surjanovic, S., Bingham, D.: Virtual library of simulation experiments: test functions and datasets. Retrieved September 20, 2016, from http://www.sfu.ca/~ssurjano
40. Szegő, G.: Orthogonal Polynomials. Colloquium Publications, vol. XXIII, 4th edn. American Mathematical Society, Providence (1975)
41. Ursell, F.: Integrals with nearly coincident branch points: Gegenbauer polynomials of large degree. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **463**(2079), 697–710 (2007)
42. Wang, Y.: Filtered polynomial approximation on the sphere. Bull. Aust. Math. Soc. **93**(1), 162–163 (2016)
43. Wang, H., Wang, K.: Optimal recovery of Besov classes of generalized smoothness and Sobolev classes on the sphere. J. Complexity **32**(1), 40–52 (2016)

Logarithmic and Riesz Equilibrium for Multiple Sources on the Sphere: The Exceptional Case

Johann S. Brauchart, Peter D. Dragnev, Edward B. Saff, and Robert S. Womersley

To Ian Sloan, an outstanding mathematician, mentor, and colleague, with much appreciation for his insights, guidance, and friendship.

Abstract We consider the minimal discrete and continuous energy problems on the unit sphere $\mathbb{S}^d$ in the Euclidean space $\mathbb{R}^{d+1}$ in the presence of an external field due to finitely many localized charge distributions on $\mathbb{S}^d$, where the energy arises from the Riesz potential $1/r^s$ (r is the Euclidean distance) for the critical Riesz parameter $s = d - 2$ if $d \geq 3$ and the logarithmic potential $\log(1/r)$ if $d = 2$. Individually, a localized charge distribution is either a point charge or assumed to be rotationally symmetric. The extremal measure solving the continuous external field problem for weak fields is shown to be the uniform measure on the sphere but restricted to the exterior of spherical caps surrounding the localized charge distributions. The radii are determined by the relative strengths of the generating charges. Furthermore, we show that the minimal energy points solving the related discrete external field problem are confined to this support. For $d - 2 \leq s < d$, we show that for point sources on the sphere, the equilibrium measure has support in the complement of

J. S. Brauchart
Institute of Analysis and Number Theory, Graz University of Technology, Graz, Austria
e-mail: j.brauchart@tugraz.at

P. D. Dragnev
Department of Mathematical Sciences, Indiana University - Purdue University, Fort Wayne, IN, USA
e-mail: dragnevp@ipfw.edu

E. B. Saff (✉)
Center for Constructive Approximation, Department of Mathematics, Vanderbilt University, Nashville, TN, USA
e-mail: edward.b.saff@vanderbilt.edu

R. S. Womersley
School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: r.womersley@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_10

the union of specified spherical caps about the sources. Numerical examples are provided to illustrate our results.

1 Introduction

Let $\mathbb{S}^d := \{\mathbf{x} \in \mathbb{R}^{d+1} : |\mathbf{x}| = 1\}$ be the unit sphere in $\mathbb{R}^{d+1}$, where $|\cdot|$ denotes the Euclidean norm. Given a compact set $E \subset \mathbb{S}^d$, consider the class $\mathscr{M}(E)$ of unit positive Borel measures supported on E. For $0 < s < d$ the *Riesz s-potential* and *Riesz s-energy* of a measure $\mu \in \mathscr{M}(E)$ are given, respectively, by

$$U_s^\mu(\mathbf{x}) := \int k_s(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu(\mathbf{y}), \quad \mathbf{x} \in \mathbb{R}^{d+1}, \quad \mathscr{I}_s(\mu) := \int\int k_s(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{y}),$$

where $k_s(\mathbf{x}, \mathbf{y}) := |\mathbf{x} - \mathbf{y}|^{-s}$ for $s > 0$ is the so-called *Riesz kernel*. For the case $s = 0$ we use the logarithmic kernel $k_0(\mathbf{x}, \mathbf{y}) := \log(1/|\mathbf{x} - \mathbf{y}|)$. The *s-capacity* of E is then defined as $C_s(E) := 1/W_s(E)$ for $s > 0$ and $C_0(E) = \exp(-W_0(E))$, where $W_s(E) := \inf\{\mathscr{I}_s(\mu) : \mu \in \mathscr{M}(E)\}$. A property is said to hold *quasi-everywhere (q.e.)* if the exceptional set has s-capacity zero. When $C_s(E) > 0$, there exists a unique minimizer $\mu_E = \mu_{s,E}$, called the *s-equilibrium measure on E*, such that $\mathscr{I}_s(\mu_E) = W_s(E)$. The s-equilibrium measure is just the normalized surface area measure on $\mathbb{S}^d$ which we denote with σ_d. For more details see [7, Chapter II].

We remind the reader that the s-energy of $\mathbb{S}^d$ is given by,

$$U_s^{\sigma_d}(\mathbf{x}) = \mathscr{I}_s(\sigma_d) = W_s(\mathbb{S}^d) = \frac{\Gamma(d)\Gamma((d-s)/2)}{2^s\Gamma(d/2)\Gamma(d-s/2)}, \qquad 0 < s < d, \text{ for } \mathbf{x} \in \mathbb{S}^d, \tag{1}$$

and the logarithmic energy of $\mathbb{S}^d$ is given by

$$U_0^{\sigma_d}(\mathbf{x}) = \mathscr{I}_0(\sigma_d) = W_0(\mathbb{S}^d) = \left.\frac{\mathrm{d}W_s(\mathbb{S}^d)}{\mathrm{d}s}\right|_{s=0} = -\log(2) + \frac{1}{2}\left(\psi(d) - \psi(d/2)\right),$$

where $\psi(s) := \Gamma'(s)/\Gamma(s)$ is the digamma function. Using cylindrical coordinates

$$\mathbf{x} = (\sqrt{1-u^2}\,\overline{\mathbf{x}}, u), \qquad -1 \le u \le 1, \ \overline{\mathbf{x}} \in \mathbb{S}^{d-1}, \tag{2}$$

we can write the decomposition

$$\mathrm{d}\sigma_d(\mathbf{x}) = \frac{\omega_{d-1}}{\omega_d}\left(1-u^2\right)^{d/2-1} \mathrm{d}u\, \mathrm{d}\sigma_{d-1}(\overline{\mathbf{x}}). \tag{3}$$

Here ω_d is the surface area of $\mathbb{S}^d$ and the ratio of these areas can be evaluated as

$$\omega_0 = 2, \quad \frac{\omega_d}{\omega_{d-1}} = \int_{-1}^{1} \left(1 - u^2\right)^{d/2-1} \mathrm{d}u = \frac{\sqrt{\pi}\,\Gamma(d/2)}{\Gamma((d+1)/2)} \\ = 2^{d-1} \frac{[\Gamma(d/2)]^2}{\Gamma(d)}. \tag{4}$$

We shall refer to a non-negative lower semi-continuous function $Q : \mathbb{S}^d \to [0, \infty]$ such that $Q(\mathbf{x}) < \infty$ on a set of positive Lebesgue surface area measure as an *external field*. The weighted energy associated with Q is then given by

$$I_Q(\mu) := \mathscr{I}_s(\mu) + 2 \int Q(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}). \tag{5}$$

Definition 1 The minimal energy problem on the sphere in the presence of the external field Q refers to the quantity

$$V_Q := \inf\left\{ I_Q(\mu) : \mu \in \mathscr{M}(\mathbb{S}^d)\right\}. \tag{6}$$

A measure $\mu_Q = \mu_{Q,s} \in \mathscr{M}(\mathbb{S}^d)$ such that $I_Q(\mu_Q) = V_Q$ is called an *s-extremal* (or *s-equilibrium*) *measure associated with Q*.

The discretized version of the minimal s-energy problem is also of interest. The associated optimal point configurations have a variety of possible applications, such as for generating radial basis functions on the sphere that are used in the numerical solutions to PDEs (see, e.g., [8, 9]).

Given a positive integer N, we consider the optimization problem

$$\mathscr{E}_{Q,N} := \min_{\{\mathbf{x}_1,\ldots,\mathbf{x}_N\} \subset \mathbb{S}^d} \sum_{1 \le i \ne j \le N} \left[k_s(\mathbf{x}_i, \mathbf{x}_j) + Q(\mathbf{x}_i) + Q(\mathbf{x}_j) \right]. \tag{7}$$

A system that minimizes the discrete energy is called an *optimal (minimal) s-energy N-point configuration w.r.t. Q*. The field-free case $Q \equiv 0$ is particularly important.

The following Frostman-type result as stated in [6] summarizes the existence and uniqueness properties for s-equilibrium measures on $\mathbb{S}^d$ in the presence of external fields (see also [12, Theorem I.1.3] for the complex plane case and [14] for more general spaces).

Proposition 1 *Let* $0 \le s < d$. *For the minimal s-energy problem on* $\mathbb{S}^d$ *with external field Q the following properties hold:*

(a) V_Q *is finite.*
(b) *There exists a unique s-equilibrium measure* $\mu_Q = \mu_{Q,s} \in \mathscr{M}(\mathbb{S}^d)$ *associated with Q. Moreover, the support* S_Q *of this measure is contained in the compact set* $E_M := \{\mathbf{x} \in \mathbb{S}^d : Q(\mathbf{x}) \le M\}$ *for some* $M > 0$.

(c) *The measure μ_Q satisfies the variational inequalities*

$$U_s^{\mu_Q}(\mathbf{x}) + Q(\mathbf{x}) \geq F_Q \quad \text{q.e. on } \mathbb{S}^d, \tag{8}$$

$$U_s^{\mu_Q}(\mathbf{x}) + Q(\mathbf{x}) \leq F_Q \quad \text{for all } \mathbf{x} \in S_Q, \tag{9}$$

where

$$F_Q := V_Q - \int Q(\mathbf{x}) \mathrm{d}\mu_Q(\mathbf{x}). \tag{10}$$

(d) *Inequalities (8) and (9) completely characterize the extremal measure μ_Q in the sense that if $\nu \in \mathscr{M}(\mathbb{S}^d)$ is a measure with finite s-energy such that*

$$U_s^{\nu}(\mathbf{x}) + Q(\mathbf{x}) \geq C \quad \text{q.e. on } \mathbb{S}^d, \tag{11}$$

$$U_s^{\nu}(\mathbf{x}) + Q(\mathbf{x}) \leq C \quad \text{for all } \mathbf{x} \in \operatorname{supp}(\nu) \tag{12}$$

for some constant C, we have then $\mu_Q = \nu$ and $F_Q = C$.

Remark 1 We note that a similar statement holds true when $\mathbb{S}^d$ is replaced by any compact subset $K \subset \mathbb{S}^d$ of positive s-capacity.

The explicit determination of s-equilibrium measures or their support is not an easy task. In [6] an external field exerted by a single point mass on the sphere was applied to establish that, in the field-free case, minimal s-energy N-point systems on $\mathbb{S}^d$, as defined in (7), are "well-separated" for $d - 2 < s < d$. Axis-supported external fields were studied in [4] and rotationally invariant external fields on $\mathbb{S}^2$ in [3]. The separation of minimal s-energy N-point configurations for more general external fields, namely Riesz s-potentials of signed measures with negative charge outside the unit sphere, was established in [5].

Here we shall focus primarily on the exceptional case when $s = d - 2$ and Q is the external field exerted by finitely many localized charge distributions. Let $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m \in \mathbb{S}^d$ be m fixed points with associated positive charges $q_1, q_2, \dots, q_m$. Then the external field is given by

$$Q(\mathbf{x}) := \sum_{i=1}^{m} q_i \, k_{d-2}(\mathbf{a}_i, \mathbf{x}). \tag{13}$$

For sufficiently small charges $q_1, \dots, q_m$ we completely characterize the $(d-2)$-equilibrium measure for the external field (13).

The outline of the paper is as follows. In Sect. 2, we introduce some notion from potential theory utilized in our analysis. In Sect. 3, we present the important case of the unit sphere in the 3-dimensional space and logarithmic interactions. An interesting corollary in its own right for discrete external fields in the complex plane is exhibited as well. The situation when $d \geq 3$, considered in Sect. 4, is more involved as there is a loss of mass in the balayage process. Finally, in Sect. 5, we derive a result on regions free of optimal points and formulate an open problem.

2 Signed Equilibria, Mhaskar-Saff $\mathcal{F}$-Functional, and Balayage

A significant role in our analysis is played by the so-called *signed equilibrium* (see [4, 5]).

Definition 2 Given a compact subset $E \subset \mathbb{R}^p$, $p \geq 3$, and an external field Q, we call a signed measure $\eta_{E,Q} = \eta_{E,Q,s}$ supported on E and of total charge $\eta_{E,Q}(E) = 1$ *a signed s-equilibrium on E associated with Q* if its weighted Riesz s-potential is constant on E:

$$U_s^{\eta_{E,Q}}(\mathbf{x}) + Q(\mathbf{x}) = F_{E,Q} \qquad \text{for all } \mathbf{x} \in E. \tag{14}$$

We note that if the signed equilibrium exists, it is unique (see [4, Lemma 23]). In view of (8) and (9), the signed equilibrium on S_Q is actually a non-negative measure and coincides with the s-extremal measure associated with Q, and hence can be obtained by solving a singular integral equation on S_Q. Moreover, for the equilibrium support we have that $S_Q \subset \operatorname{supp}(\eta_{E,Q}^+)$ whenever $S_Q \subset E \subset \mathbb{S}^d$ (see [5, Theorem 9]).

An important tool in our analysis is the Riesz analog of the *Mhaskar-Saff F-functional* from classical logarithmic potential theory in the plane (see [10] and [12, Chapter IV, p. 194]).

Definition 3 The *$\mathcal{F}_s$-functional* of a compact subset $K \subset \mathbb{S}^d$ of positive s-capacity is defined as

$$\mathcal{F}_s(K) := W_s(K) + \int Q(\mathbf{x})\, \mathrm{d}\mu_K(\mathbf{x}), \tag{15}$$

where $W_s(K)$ is the s-energy of K and μ_K is the s-equilibrium measure on K.

Remark 2 As pointed out in [4, 5], when $d - 2 \leq s < d$, a relationship exists between the signed s-equilibrium constant in (14) and the $\mathcal{F}_s$-functional (15), namely $\mathcal{F}_s(K) = F_{K,Q}$. Moreover, the equilibrium support minimizes the $\mathcal{F}_s$-functional; i.e., if $d - 2 \leq s < d$ and Q is an external field on $\mathbb{S}^d$, then the $\mathcal{F}_s$-functional is minimized for $S_Q = \operatorname{supp}(\mu_Q)$ (see [4, Theorem 9]).

A tool we use extensively is the *Riesz s-balayage measure* (see [7, Section 4.5]). Given a measure ν supported on $\mathbb{S}^d$ and a compact subset $K \subset \mathbb{S}^d$, the measure $\widehat{\nu} := \mathrm{Bal}_s(\nu, K)$ is called the *Riesz s-balayage* of ν onto K, $d - 2 \leq s < d$, if $\widehat{\nu}$ is supported on K and

$$\begin{aligned} U_s^{\widehat{\nu}}(\mathbf{x}) &= U_s^{\nu}(\mathbf{x}) \qquad \text{on } K, \\ U_s^{\widehat{\nu}}(\mathbf{x}) &\leq U_s^{\nu}(\mathbf{x}) \qquad \text{on } \mathbb{S}^d. \end{aligned} \tag{16}$$

In general, there is some loss of mass, namely $\widehat{\nu}(\mathbb{S}^d) < \nu(\mathbb{S}^d)$. However, in the logarithmic interaction case $s = 0$ and $d = 2$, the mass of the balayage measures is preserved, but as in the classical complex plane potential theory we have equality of potentials up to a constant term

$$\begin{aligned} U_0^{\widehat{\nu}}(\mathbf{x}) &= U_0^{\nu}(\mathbf{x}) + C \qquad \text{on } K, \\ U_0^{\widehat{\nu}}(\mathbf{x}) &\leq U_0^{\nu}(\mathbf{x}) + C \qquad \text{on } \mathbb{S}^2. \end{aligned} \tag{17}$$

Balayage of a signed measure η is achieved by taking separately the balayage of its positive and its negative part in the Jordan decomposition $\eta = \eta^+ - \eta^-$. An important property is that we can take balayage in steps: if $F \subset K \subset \mathbb{S}^d$, then

$$\mathrm{Bal}_s(\nu, F) = \mathrm{Bal}_s(\mathrm{Bal}_s(\nu, K), F). \tag{18}$$

We also use the well-known relation

$$\mathrm{Bal}_s(\nu, K) = \nu_{|_K} + \mathrm{Bal}_s(\nu_{|_{\mathbb{S}^d \setminus K}}, K). \tag{19}$$

3 Logarithmic Interactions on $\mathbb{S}^2$

We first state and prove our main theorem for the case of logarithmic interactions on $\mathbb{S}^2$. We associate with Q in (13) (or equivalently with $\{\mathbf{a}_i\}$ and $\{q_i\}$) the total charge

$$q := q_1 + \cdots + q_m,$$

the vector

$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m), \qquad \epsilon_i := 2\sqrt{\frac{q_i}{1+q}}, \quad i = 1, \ldots, m, \tag{20}$$

and the set

$$\Sigma_{\boldsymbol{\epsilon}} = \bigcap_{i=1}^{m} \Sigma_{i,\epsilon_i}, \quad \Sigma_{i,\epsilon} := \left\{\mathbf{x} \in \mathbb{S}^2 : |\mathbf{x} - \mathbf{a}_i| \geq \epsilon\right\}, \quad i = 1, \ldots, m, \; \epsilon \geq 0. \tag{21}$$

More generally, with any vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ with non-negative components we associate the set $\Sigma_{\boldsymbol{\gamma}} = \bigcap_{i=1}^m \Sigma_{i,\gamma_i}$. Note: if $\boldsymbol{\gamma} \leq \boldsymbol{\epsilon}$ (i.e., $\gamma_i \leq \epsilon_i$, $1 \leq i \leq m$), then $\Sigma_{\boldsymbol{\epsilon}} \subset \Sigma_{\boldsymbol{\gamma}}$.

Theorem 1 *Let $d = 2$ and $s = 0$. Let Q, $\boldsymbol{\epsilon}$, and $\Sigma_{\boldsymbol{\epsilon}}$ be defined by* (13), (20), *and* (21). *Suppose that $\Sigma_{i,\epsilon_i}^c \cap \Sigma_{j,\epsilon_j}^c = \emptyset$, $1 \leq i < j \leq m$ (K^c denotes the complement* *of K relative to the sphere). Then the logarithmic extremal measure associated with Q is* $\mu_Q = (1+q)\,\sigma_2|_{\Sigma_\epsilon}$ *and the extremal support is* $S_Q = \Sigma_\epsilon$.

Remark 3 The theorem has the following electrostatics interpretation. As positively charged particles $\mathbf{a}_i$ are introduced on a positively pre-charged unit sphere, they create charge-free regions which we call *regions of electrostatic influence*. The theorem then states that if the potential interaction is logarithmic and the charges of the particles are sufficiently small (so that the regions of influence do not overlap), then these regions are perfect spherical caps Σ_{i,ϵ_i}^c whose radii depend only on the amount of charge and the position of the particles. In Sect. 5, we partially investigate what happens when the q_i's increase beyond the critical values imposed by the non-overlapping conditions $\Sigma_{i,\epsilon_i}^c \cap \Sigma_{i,\epsilon_j}^c = \emptyset$, $1 \le i < j \le m$.

Proof Let $m = 1$. This case has already been solved in [4]. By [4, Theorem 17], the signed equilibrium on Σ_γ associated with $Q(\mathbf{x}) := q \log \frac{1}{|\mathbf{x}-\mathbf{a}|}$, $\mathbf{a} \in \mathbb{S}^2$, is given by

$$\begin{aligned}\eta_{\Sigma_\gamma,Q} &= (1+q)\,\mathrm{Bal}_0(\sigma_2, \Sigma_\gamma) - q\,\mathrm{Bal}_0(\delta_{\mathbf{a}}, \Sigma_\gamma)\\ &= (1+q)\,\sigma_2|_{\Sigma_\gamma} + (1+q)\left(\frac{\gamma^2}{4} - \frac{q}{1+q}\right)\beta,\end{aligned} \tag{22}$$

where β is the normalized Lebesgue measure on the boundary circle of Σ_γ.

The logarithmic extremal measure on $\mathbb{S}^2$ associated with Q is then given by

$$\mu_Q = (1+q)\,\sigma_2|_{\Sigma_\epsilon}, \qquad \text{where } \epsilon = 2\sqrt{\frac{q}{1+q}}.$$

Let $\xi := \langle \mathbf{x}, \mathbf{a}\rangle$ and $\gamma^2 = 2(1-t)$, where t is the projection of the boundary circle $\partial\Sigma_\gamma$ onto the $\mathbf{a}$-axis. For future reference, by [4, Lemmas 39 and 41] we have

$$U_0^{\mathrm{Bal}_0(\sigma_2,\Sigma_\gamma)}(\mathbf{x}) = \begin{cases} W_0(\Sigma_\gamma), & \mathbf{x} \in \Sigma_\gamma,\\ W_0(\Sigma_\gamma) + \dfrac{1}{2}\log\dfrac{1+t}{1+\xi}, & \mathbf{x} \in \Sigma_\gamma^c \end{cases} \tag{23}$$

and

$$U_0^{\mathrm{Bal}_0(\delta_{\mathbf{a}},\Sigma_\gamma)}(\mathbf{x}) = U_0^{\delta_{\mathbf{a}}}(\mathbf{x}) + \begin{cases} \dfrac{1}{2}\log\dfrac{1-t}{1+t} - \dfrac{1}{2}\log\dfrac{1-t}{2}, & \mathbf{x} \in \Sigma_\gamma,\\ \dfrac{1}{2}\log\dfrac{1-\xi}{1+\xi} - \dfrac{1}{2}\log\dfrac{1-t}{2}, & \mathbf{x} \in \Sigma_\gamma^c. \end{cases} \tag{24}$$

Moreover, $\widehat{\nu} = \mathrm{Bal}_0(\sigma_2|_{\Sigma_\gamma^c}, \Sigma_\gamma)$ and $\widehat{\delta_{\mathbf{a}}} = \mathrm{Bal}_0(\delta_{\mathbf{a}}, \Sigma_\gamma)$ are multiples of β:

$$\widehat{\nu} = \sigma_2(\Sigma_\gamma^c)\,\beta = \frac{\gamma^2}{4}\beta = \frac{1-t}{2}\beta, \qquad \widehat{\delta_{\mathbf{a}}} = \beta. \tag{25}$$

Let $m \geq 2$. First, we determine the signed equilibrium on the set $\Sigma_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \leq \boldsymbol{\epsilon}$, associated with Q. We consider the signed measure

$$\tau := (1+q)\,\mathrm{Bal}_0(\sigma_2, \Sigma_{\boldsymbol{\gamma}}) - \mathrm{Bal}_0(q_1\,\delta_{\mathbf{a}_1} + \cdots + q_m\,\delta_{\mathbf{a}_m}, \Sigma_{\boldsymbol{\gamma}}).$$

As balayage under logarithmic interaction is linear and preserves mass, we have[1]

$$\|\tau\| = (1+q)\,\|\sigma_2\| - \sum_{i=1}^{m} q_i\,\|\delta_{\mathbf{a}_i}\| = 1 + q - \sum_{i=1}^{m} q_i = 1.$$

The hypotheses on $\Sigma_{\boldsymbol{\epsilon}}$ and the fact that $\Sigma_{\boldsymbol{\epsilon}} \subset \Sigma_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \leq \boldsymbol{\epsilon}$, imply the non-overlapping conditions

$$\Sigma_{i,\gamma_i}^c \cap \Sigma_{j,\gamma_j}^c = \emptyset, \qquad 1 \leq i < j \leq m.$$

For $i = 1, \dots, m$ let

$$\nu_i := \sigma_2|_{\Sigma_{i,\gamma_i}^c}, \qquad \widehat{\nu_i} := \mathrm{Bal}_0(\nu_i, \Sigma_{i,\gamma_i}), \qquad \widehat{\delta_{\mathbf{a}_i}} := \mathrm{Bal}_0(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}). \tag{26}$$

Since $\Sigma_{i,\gamma_i} \supset \Sigma_{\boldsymbol{\gamma}}$, balayage in steps (cf. (18)) yields

$$\mathrm{Bal}_0(\nu_i, \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_0(\mathrm{Bal}_0(\nu_i, \Sigma_{i,\gamma_i}), \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_0(\nu_i, \Sigma_{i,\gamma_i}) = \widehat{\nu_i}.$$

The second step follows because $\widehat{\nu_i}$ is supported on $\partial\Sigma_{i,\gamma_i}$ which is included in $\partial\Sigma_{\boldsymbol{\gamma}}$. Hence

$$\begin{aligned}\mathrm{Bal}_0(\sigma_2, \Sigma_{\boldsymbol{\gamma}}) &= \mathrm{Bal}_0(\sigma_2|_{\Sigma_{\boldsymbol{\gamma}}} + \nu_1 + \cdots + \nu_m, \Sigma_{\boldsymbol{\gamma}}) \\ &= \sigma_2|_{\Sigma_{\boldsymbol{\gamma}}} + \sum_{i=1}^{m} \mathrm{Bal}_0(\nu_i, \Sigma_{\boldsymbol{\gamma}}) \\ &= \sigma_2|_{\Sigma_{\boldsymbol{\gamma}}} + \sum_{i=1}^{m} \widehat{\nu_i}.\end{aligned}$$

Likewise,

$$\mathrm{Bal}_0(\delta_{\mathbf{a}_i}, \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_0(\mathrm{Bal}_0(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}), \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_0(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}) = \widehat{\delta_{\mathbf{a}_i}}.$$

[1] The mass of a signed measure μ is defined as $\|\mu\| := \int \mathrm{d}\mu$.

Hence, we obtain the following representation of τ:

$$\tau = (1+q)\,\sigma_{2|_{\Sigma_{\boldsymbol{\gamma}}}} + (1+q)\sum_{i=1}^{m}\widehat{\nu_i} - \sum_{i=1}^{m} q_i\,\widehat{\delta_{\mathbf{a}_i}}. \tag{27}$$

We show that the weighted logarithmic potential of τ satisfies (14). Let $\mathbf{x} \in \Sigma_{\boldsymbol{\gamma}}$. Then $\mathbf{x} \in \Sigma_{i,\gamma_i}$ for every $1 \le i \le m$ and, by (17) and (24), for every $1 \le i \le m$

$$U_0^{\widehat{\nu_i}}(\mathbf{x}) = U_0^{\nu_i}(\mathbf{x}) + C_i \qquad \text{on } \Sigma_{i,\gamma_i},$$
$$U_0^{\widehat{\delta_{\mathbf{a}_i}}}(\mathbf{x}) = U_0^{\delta_{\mathbf{a}_i}}(\mathbf{x}) - \frac{1}{2}\log\frac{1+t_i}{2} \qquad \text{on } \Sigma_{i,\gamma_i}.$$

Hence, computing the logarithmic potential of τ in (27) yields, after simplification,

$$U_0^{\tau}(\mathbf{x}) + Q(\mathbf{x}) = (1+q)\,U_0^{\sigma_2}(\mathbf{x}) + \sum_{i=1}^{m}\left((1+q)\,C_i + \frac{q_i}{2}\log\frac{1+t_i}{2}\right), \qquad \mathbf{x} \in \Sigma_{\boldsymbol{\gamma}}.$$

Since $U_0^{\sigma_2}(\mathbf{x}) = W_0(\mathbb{S}^2)$, the weighted potential of τ is constant on $\Sigma_{\boldsymbol{\gamma}}$; i.e., τ is a signed equilibrium on $\Sigma_{\boldsymbol{\gamma}}$ associated with Q and, by uniqueness, $\eta_{\Sigma_{\boldsymbol{\gamma}},Q} = \tau$ and

$$F_{\Sigma_{\boldsymbol{\gamma}},Q} = (1+q)\,W_0(\mathbb{S}^2) + \sum_{i=1}^{m}\left((1+q)\,C_i + \frac{q_i}{2}\log\frac{1+t_i}{2}\right).$$

Let $\mathbf{x} \in \Sigma_{\boldsymbol{\gamma}}^c$. Then $\mathbf{x} \in \Sigma_{i_0,\gamma_{i_0}}^c$ for some $i_0 \in \{1,\dots,m\}$ and $\mathbf{x} \in \Sigma_{i,\gamma_i}$ for $i \neq i_0$. Using (27), (23), and (24),

$$\begin{aligned} U_0^{\eta_{\Sigma_{\boldsymbol{\gamma}},Q}}(\mathbf{x}) + Q(\mathbf{x}) = F_{\Sigma_{\boldsymbol{\gamma}},Q} &+ (1+q)\left[U_0^{\widehat{\nu_{i_0}}}(\mathbf{x}) - \left(U_0^{\nu_{i_0}}(\mathbf{x}) + C_{i_0}\right)\right] \\ &+ \frac{q_{i_0}}{2}\log\frac{(1+\xi_{i_0})\,(1-t_{i_0})}{(1-\xi_{i_0})\,(1+t_{i_0})}. \end{aligned} \tag{28}$$

Observe that the square-bracketed expression is ≤ 0 by (17). Because of $t_{i_0} < \xi_{i_0} < 1$, the ratio under the logarithm is > 1 and the logarithm tends to zero as ξ_{i_0} goes to t_{i_0} and the logarithm tends to $+\infty$ as ξ_{i_0} approaches 1 from below. Using (23) again, we derive

$$\begin{aligned} U_0^{\eta_{\Sigma_{\boldsymbol{\gamma}},Q}}(\mathbf{x}) + Q(\mathbf{x}) = (1+q)\,W_0(\Sigma_{i_0,\gamma_{i_0}}) &+ \sum_{\substack{i=1,\\ i\neq i_0}}^{m}\left((1+q)\,C_i + \frac{q_i}{2}\log\frac{1+t_i}{2}\right) \\ &+ \frac{q_{i_0}}{2}\log\frac{1+t_{i_0}}{2} + f(t_{i_0}) - f(\xi_{i_0}), \end{aligned} \tag{29}$$

where

$$f(u) := \frac{1+q-q_{i_0}}{2}\log(1+u) + \frac{q_{i_0}}{2}\log(1-u), \qquad -1<u<1. \tag{30}$$

The function f has a unique maximum at $u^* = 1 - \frac{2q_{i_0}}{1+q}$ in the interval $(-1,1)$ for $0 < q_{i_0} < 1+q$. Assuming that $-1 < t_{i_0} < u < 1, f'(u) < 0$ if and only if

$$\max\left\{t_{i_0}, 1-\frac{2q_{i_0}}{1+q}\right\} < u < 1 \;\Leftrightarrow\; 0 < 2(1-u) < \min\left\{\frac{4q_{i_0}}{1+q}, \gamma_{i_0}^2\right\} = \min\left\{\epsilon_{i_0}^2, \gamma_{i_0}^2\right\}.$$

By assumption, $\gamma_{i_0} \leq \epsilon_{i_0}$. Hence, the infimum of the weighted potential of $\eta_{\Sigma_\gamma,Q}$ in the set $\Sigma_{i_0,\gamma_{i_0}}^c$ is assumed on its boundary. Continuity of the potentials in (28) yields

$$U_0^{\eta_{\Sigma_\gamma,Q}}(\mathbf{x}) + Q(\mathbf{x}) \geq F_{\Sigma_\gamma,Q} \qquad \text{on } \Sigma_{i_0,\gamma_{i_0}}^c.$$

As i_0 was determined by $\mathbf{x} \in \Sigma_{\boldsymbol{\gamma}}^c$, we deduce that the last relation holds on $\Sigma_{\boldsymbol{\gamma}}^c$.

Summarizing, for each $\boldsymbol{\gamma} \leq \boldsymbol{\epsilon}$

$$U_0^{\eta_{\Sigma_\gamma,Q}}(\mathbf{x}) + Q(\mathbf{x}) \geq F_{\Sigma_\gamma,Q} \qquad \text{on } \Sigma_{\boldsymbol{\gamma}}^c, \tag{31}$$

$$U_0^{\eta_{\Sigma_\gamma,Q}}(\mathbf{x}) + Q(\mathbf{x}) = F_{\Sigma_\gamma,Q} \qquad \text{on } \Sigma_{\boldsymbol{\gamma}} \tag{32}$$

and from (27) and (25),

$$\eta_{\Sigma_\gamma,Q} = (1+q)\,\sigma_{2|_{\Sigma_\gamma}} + (1+q)\sum_{i=1}^{m}\left(\frac{\gamma_i^2}{4} - \frac{\epsilon_i^2}{4}\right)\beta_i.$$

It is not difficult to see that the signed equilibrium $\eta_{\Sigma_\gamma,Q}$ becomes a positive measure, and at the same time satisfies the characterization inequalities (11) and (12), if and only if $\boldsymbol{\gamma} = \boldsymbol{\epsilon}$. By Proposition 1(d), $\mu_Q = \eta_{\Sigma_\epsilon,Q} = (1+q)\,\sigma_{2|\Sigma_\epsilon}$. □

Theorem 1 and [5, Corollary 13] yield the following result.

Corollary 1 *Under the assumptions of Theorem 1, the optimal logarithmic energy N-point configurations w.r.t. Q are contained in S_Q for every $N \geq 2$.*

Proof From [5, Corollary 13] we have that the optimal N-point configurations lie in

$$\widetilde{S}_Q = \{\mathbf{x} \,:\, U_0^{\mu_Q}(\mathbf{x}) + Q(\mathbf{x}) \leq F_Q\}.$$

The strict monotonicity of the function f in (30) yields $\widetilde{S}_Q = S_Q$. □

Remark 4 Theorem 1 and Corollary 1 are illustrated in Figs. 1 and 2 for two and three point sources, respectively. Observe, that the density of the (approximate) log-optimal configuration approaches the normalized surface area of the equilibrium support $S_Q = \Sigma_{\boldsymbol{\epsilon}}$.

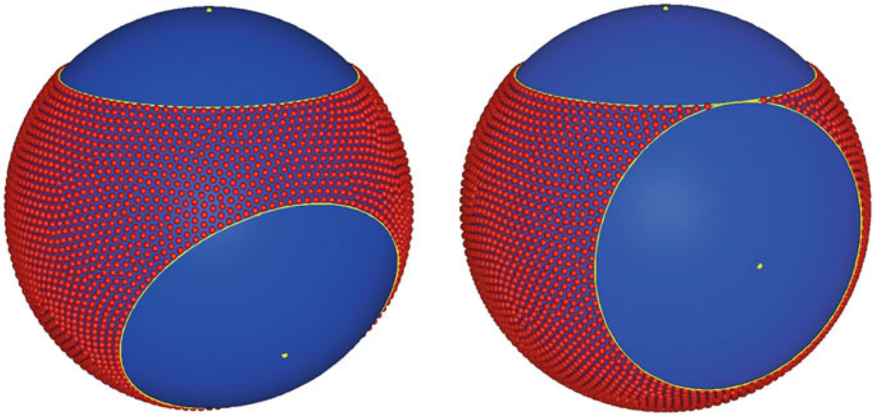


Fig. 1 Approximate log-optimal points for $m = 2, N = 4000$ with $q_1 = q_2 = \frac{1}{4}$, $\mathbf{a}_1 = (0, 0, 1)$ and $\mathbf{a}_2 = (\frac{\sqrt{91}}{10}, 0, -\frac{3}{10})$ or $\mathbf{a}_2 = (\frac{4\sqrt{5}}{9}, 0, -\frac{1}{9})$

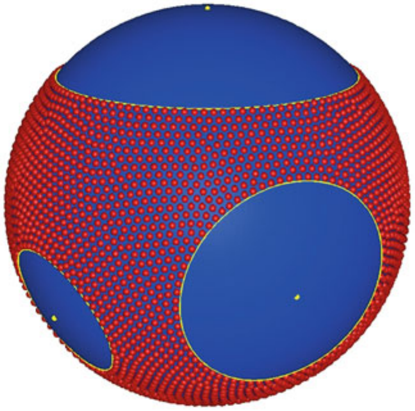


Fig. 2 Approximate log-optimal points for $m = 3, N = 4000$ with $q_1 = \frac{1}{4}$, $q_2 = \frac{1}{8}$, $q_3 = \frac{1}{20}$, $\mathbf{a}_1 = (0, 0, 1)$, $\mathbf{a}_2 = (\frac{\sqrt{91}}{10}, 0, -\frac{3}{10})$, and $\mathbf{a}_3 = (0, \frac{\sqrt{3}}{2}, -\frac{1}{2})$

Remark 5 The objective function for the optimization problem (7) with the discrete external field (13) is

$$E_{Q,N}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \sum_{1 \le i \ne j \le N} k_s(\mathbf{x}_i, \mathbf{x}_j) + 2(N-1) \sum_{i=1}^{m} q_i \sum_{j=1}^{N} k_{s_i}(\mathbf{a}_i, \mathbf{x}_j),$$

where $k(\mathbf{x}, \mathbf{y})$ is the Riesz kernel defined at the beginning of Sect. 1. The standard spherical parametrisation, $\mathbf{x}_i = (\sin(\theta_i)\cos(\phi_i),\ \sin(\theta_i)\sin(\phi_i),\ \cos(\theta_i)) \in \mathbb{S}^2$ for $\theta_i \in [0, \pi]$ and $\phi_i \in [0, 2\pi)$ is used to avoid the non-linear constraints $|\mathbf{x}_i| = 1, i = 1, \dots, N$. This introduces singularities at the poles $\theta = 0, \pi$, one of which can be avoided by using the rotational invariance of the objective function to place the first external field at the North Pole. For $\theta_i \neq 0, \pi$ the gradient of $E_{Q,N}(\theta_1, \phi_1, \dots, \theta_N, \phi_N)$ can be calculated for use in a nonlinear optimization method.

Point sets $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ that provide approximate optimal s-energy configurations were obtained using this spherical parametrisation of the points and applying a nonlinear optimization method, for example a limited memory BFGS method for bound constrained problems [13], to find a local minimum of $E_{Q,N}$. The initial point sets $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ used as starting points for the nonlinear optimization were uniformly distributed on $\mathbb{S}^2$, so did not reflect the structure of the external fields. A local perturbation of the point set achieving a local minimum was then used to generate a new starting point and the nonlinear optimization applied again. The best local minimizer found provided an approximation (upper bound) on the global minimum of $E_{Q,N}$. Different local minima arose from the fine structure of the points within their support.

The results above lend themselves to the following generalization. Given m points $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{S}^2$, for each $i = 1, \dots, m$ let ϕ_i be a radially-symmetric measure centered at $\mathbf{a}_i$ and supported on Σ_{i,ρ_i}^c for some $\rho_i > 0$ that has absolutely continuous density with respect to σ_2; i.e.,

$$\mathrm{d}\phi_i(\mathbf{x}) = f_i(\langle \mathbf{x}, \mathbf{a}_i \rangle)\, \mathrm{d}\sigma_2(\mathbf{x}), \qquad f_i(u) = 0 \quad \text{on} \left[-1, \sqrt{1 - \rho_i^2/2} \right]. \tag{33}$$

Let $q_i := \|\phi_i\| = \int \mathrm{d}\phi_i$, $1 \leq i \leq m$, and define the external field

$$Q_{\boldsymbol{\phi}}(\mathbf{x}) := \sum_{i=1}^{m} U_0^{\phi_i}(\mathbf{x}) = \sum_{i=1}^{m} \int \log \frac{1}{|\mathbf{x} - \mathbf{a}_i|}\, \mathrm{d}\phi_i(\mathbf{x}), \tag{34}$$

where $\boldsymbol{\phi} = (\phi_1, \dots, \phi_m)$. Then the following theorem holds.

Theorem 2 *Let $d = 2$ and $s = 0$. Let $Q_{\boldsymbol{\phi}}$ be defined by (34) and $\boldsymbol{\epsilon}$, $\Sigma_{\boldsymbol{\epsilon}}$ be defined by (20), (21). Suppose that $\Sigma_{i,\epsilon_i}^c \cap \Sigma_{j,\epsilon_j}^c = \emptyset$, $1 \leq i < j \leq m$. Then the logarithmic extremal measure associated with $Q_{\boldsymbol{\phi}}$ is $\mu_{Q_{\boldsymbol{\phi}}} = (1 + q)\, \sigma_2|_{\Sigma_{\boldsymbol{\epsilon}}}$ and the extremal support is $S_{Q_{\boldsymbol{\phi}}} = \Sigma_{\boldsymbol{\epsilon}}$.*

Proof The proof proceeds as in the proof of Theorem 1 with the adaption that the balayage measure of ϕ_i is given by

$$\widehat{\phi}_i = \mathrm{Bal}_0(\phi_i, \Sigma_{i,\epsilon_i}) = \|\phi\|\, \beta_i = q_i\, \beta_i,$$

which follows easily from the hypothesis $\rho_i \leq \epsilon_i$ and the uniqueness of balayage measures. □

We next formulate the analog of Theorem 1 in the complex plane $\mathbb{C}$. Let us fix one of the charges, say $\mathbf{a}_m$, at the North Pole $\mathbf{p}$, which will also serve as the center of the *Kelvin transformation* $\mathscr{K}$ (stereographic projection, or equivalently, inversion about the center $\mathbf{p}$) with radius $\sqrt{2}$ onto the equatorial plane. Set $w_i := \mathscr{K}(\mathbf{a}_i)$, $1 \leq i \leq m$. The image of $\mathbf{a}_m$ under the Kelvin transformation is the "point at infinity" in $\mathbb{C}$. Letting $z = \mathscr{K}(\mathbf{x})$, $\mathbf{x} \in \mathbb{S}^2$, we can utilize the following formulas

$$|\mathbf{x} - \mathbf{p}| = \frac{2}{\sqrt{1 + |z|^2}}, \qquad |\mathbf{x} - \mathbf{a}_i| = \frac{2|z - w_i|}{\sqrt{1 + |z|^2}\,\sqrt{1 + |w_i|^2}}, \quad 1 \leq i \leq m - 1$$

to convert the continuous minimal energy problem (cf. (6)) and the discrete minimal energy problem (cf. (7)) on the sphere to their analogous forms in the complex plane $\mathbb{C}$. Neglecting a constant term, we obtain in the complex plane the external field

$$\widetilde{Q}(z) := \sum_{i=1}^{m-1} q_i \log \frac{1}{|z - w_i|} + (1 + q) \log \sqrt{1 + |z|^2}, \qquad z \in \mathbb{C}. \tag{35}$$

This external field is admissible in the sense of Saff-Totik [12], since

$$\lim_{|z| \to \infty} \big(\widetilde{Q}(z) - \log |z|\big) = \lim_{|z| \to \infty} q_m \log |z| = \infty.$$

Therefore, there is a unique equilibrium measure $\mu_{\widetilde{Q}}$ characterized by variational inequalities similar to the ones in Proposition 1(d). The following theorem giving the extremal support $S_{\widetilde{Q}}$ and the extremal measure $\mu_{\widetilde{Q}}$ associated with the external field $\widetilde{Q}$ in (35) for sufficiently small q_i's is a direct consequence of Theorem 1.

Theorem 3 *Let $w_1, \dots, w_{m-1} \in \mathbb{C}$ be fixed and $q_1, \dots, q_m$ be positive real numbers with $q = q_1 + \cdots + q_m$ and $\widetilde{Q}$ be the corresponding external field given in (35). Further, let $\mathbf{a}_1, \dots, \mathbf{a}_{m-1} \in \mathbb{S}^2$ be the pre-images under the Kelvin transformation $\mathscr{K}$, i.e., $w_i = \mathscr{K}(\mathbf{a}_i)$, $1 \leq i \leq m - 1$, and $a_m = \mathbf{p}$. If the q_i's are sufficiently small so that $\Sigma_{i,\epsilon_i}^c \cap \Sigma_{j,\epsilon_j}^c = \emptyset$, $1 \leq i < j \leq m$, where the spherical caps Σ_{i,ϵ_i} are defined in (21), then there are open discs $D_1, \dots, D_{m-1}$ in $\mathbb{C}$ with $w_i \in D_i = \mathscr{K}(\Sigma_{i,\epsilon_i})$, $1 \leq i \leq m - 1$, such that*

$$S_{\widetilde{Q}} = \left\{ z \in \mathbb{C} \,:\, |z| \leq \sqrt{\frac{1 + q - q_m}{q_m}} \right\} \setminus \bigcup_{i=1}^{m-1} D_i. \tag{36}$$

The extremal measure $\mu_{\widetilde{Q}}$ associated with $\widetilde{Q}$ is given by

$$\mathrm{d}\mu_{\widetilde{Q}}(z) = \frac{1 + q}{\pi \left(1 + |z|^2\right)^2}\, \mathrm{d}A(z), \tag{37}$$

where $\mathrm{d}A$ denotes the Lebesgue area measure in the complex plane.

Proof The proof follows by a straight forward application of the Kelvin transformation to the weighted potential $U_0^{\mu_Q}(\mathbf{x}) + Q(\mathbf{x})$ and using the identity relating the regular (not normalized) Lebesgue measure on the sphere and the area measure on the complex plane

$$\frac{4\pi}{|\mathbf{x}-\mathbf{p}|^2}\,\mathrm{d}\sigma_2(\mathbf{x}) = \frac{1}{1+|z|^2}\,\mathrm{d}A(z).$$

This change of variables yields the identity

$$U_0^{\mu_Q}(\mathbf{x}) + Q(\mathbf{x}) = U_0^{\mu_{\widetilde{Q}}}(z) + \widetilde{Q}(z) + const$$

from which, utilizing (31) and (32), one derives

$$U_0^{\mu_{\widetilde{Q}}}(z) + \widetilde{Q}(z) \geq C \quad \text{in } \mathbb{C}, \tag{38}$$

$$U_0^{\mu_{\widetilde{Q}}}(z) + \widetilde{Q}(z) = C \quad \text{on } S_{\widetilde{Q}}, \tag{39}$$

which implies that $\mu_{\widetilde{Q}}$ is the equilibrium measure by Saff and Totik [12, Theorem 1.3]. □

Remark 6 At first it seems like a surprising fact that the equilibrium measure in Theorem 2 is uniform on S_Q (i.e. has constant density). However, this can be easily seen alternatively from the planar version Theorem 3. Once we derive that the support $S_{\widetilde{Q}}$ is given by (36), we can recover the measure $\mu_{\widetilde{Q}}$ by applying Gauss' theorem (cf. [12, Theorem II.1.3]), namely on any subregion of $S_{\widetilde{Q}}$ we have

$$\mathrm{d}\mu_{\widetilde{Q}} = -\frac{1}{2\pi}\Delta U^{\mu_{\widetilde{Q}}}\mathrm{d}A(z) = \frac{1}{2\pi}\Delta\widetilde{Q}(z) = \frac{1+q}{\pi(1+|z|^2)^2}\,\mathrm{d}A(z).$$

Recall that on this subregion $\log|z-w_i|$ is harmonic for all $i = 1,\dots,m-1$. As $\mathrm{d}\sigma_2(\mathbf{x}) = \mathrm{d}A(z)/[\pi(1+|z|^2)^2]$, we get that μ_Q is the normalized Lebesgue surface measure on S_Q. Observe, that the same argument will apply to the setting of Theorem 5 ($d = 2,\ s = 0$), from which we derive $\mu_Q = (1+q)\,\sigma_2|_{S_Q}$ even in the case when $\Sigma_{\epsilon_i}^c$ are not disjoint. Of course, we don't know the equilibrium support S_Q in this case. For related results see [1, 2].

4 Riesz $(d-2)$-Energy Interactions on $\mathbb{S}^d$, $d \geq 3$

The case of $(d-2)$-energy interactions on $\mathbb{S}^d$, $d \geq 3$, and an external field Q given by (13) is considerably more involved as the balayage measures utilized to determine the signed equilibrium on Σ_γ diminish their masses. This phenomenon yields an implicit nonlinear system for the critical values of the radii $\epsilon_1,\dots,\epsilon_m$ (see (54) and (55)) characterizing the regions of electrostatic influence.

Let $d \geq 3$ and $0 < d-2 \leq s < d$. Let $\Phi_s(t_i) := \mathscr{F}_s(\Sigma_{i,\gamma_i})$ be the Mhaskar-Saff $\mathscr{F}_s$-functional associated with the external field $Q_i(\mathbf{x}) := q_i\,|\mathbf{x}-\mathbf{a}_i|^{-s}$ evaluated for the spherical cap Σ_{i,γ_i}. Then the signed s-equilibrium measure $\eta_{i,s} := \eta_{\Sigma_{i,\gamma_i},Q_i,s}$ on Σ_{i,γ_i} associated with Q_i is given by (see [4, Theorem 11 and 15])

$$\eta_{i,s} = \frac{\Phi_s(t_i)}{W_s(\mathbb{S}^d)} \operatorname{Bal}_s(\sigma_d, \Sigma_{i,\gamma_i}) - q_i \operatorname{Bal}_s(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}). \tag{40}$$

For $d-2 < s < d$ this signed measure is absolutely continuous

$$\mathrm{d}\eta_{i,s}(\mathbf{x}) = \frac{\omega_{d-1}}{\omega_d} \eta_{i,s}'(u) \left(1-u^2\right)^{d/2-1} \mathrm{d}u \, \mathrm{d}\sigma_{d-1}(\overline{\mathbf{x}}), \quad \mathbf{x} = (\sqrt{1-u^2}\, \overline{\mathbf{x}}, u) \in \Sigma_{i,\gamma_i},$$

with density function

$$\begin{aligned} \eta_{i,s}'(u) = \frac{1}{W_s(\mathbb{S}^d)} \frac{\Gamma(d/2)}{\Gamma(d-s/2)} \left(\frac{1-t_i}{1-u}\right)^{d/2} \left(\frac{t_i-u}{1-t_i}\right)^{(s-d)/2} \\ \times \left\{ \Phi_s(t_i) \, {}_2\mathbf{F}_1\left(\begin{matrix} 1, d/2 \\ 1-(d-s)/2 \end{matrix}; \frac{t_i-u}{1-u} \right) - \frac{q_i \, 2^{d-s}}{\gamma_i{}^d} \right\}. \end{aligned} \tag{41}$$

For the ratio $\frac{\omega_{d-1}}{\omega_d}$ see (4), a formula for the Riesz s-energy $W_0(\mathbb{S}^d)$ is given in (1), and ${}_2\mathbf{F}_1$ denotes Olver's regularized ${}_2F_1$-hypergeometric function [11, Eq. 15.2.2]. For $s = d-2$ the signed $(d-2)$-equilibrium

$$\eta_{i,d-2} = \frac{\Phi_{d-2}(t_i)}{W_{d-2}(\mathbb{S}^d)} \sigma_{d}{}_{|\Sigma_{i,\gamma_i}} + \frac{1-t_i}{2} \left(1-t_i^2\right)^{d/2-1} \left[\Phi_{d-2}(t_i) - \frac{4q_i}{\gamma_i^d} \right] \beta_i \tag{42}$$

has, like in the logarithmic case (see (22)), a boundary-supported component β_i, which is the normalized Lebesgue measure on the boundary circle of $\Sigma_{i,\gamma}$. Observe that in either case the signed equilibrium has a negative component if and only if

$$\Phi_s(t_i) - \frac{2^{d-s} q_i}{\gamma_i^d} < 0, \qquad \text{where } 2(1-t_i) = \gamma_i^2. \tag{43}$$

The weighted s-potential of $\eta_{i,s}$, $d-2 < s < d$, satisfies [4, Theorem 11]

$$U_s^{\eta_{i,s}}(\mathbf{z}) + Q_i(\mathbf{z}) = \Phi_s(t_i), \qquad \mathbf{z} \in \Sigma_{i,\gamma_i}, \tag{44}$$

$$\begin{aligned} U_s^{\eta_{i,s}}(\mathbf{z}) + Q_i(\mathbf{z}) = \Phi_s(t_i) + \frac{q_i}{[2(1-\xi_i)]^{s/2}} \mathrm{I}\left(\frac{2}{1-t_i} \frac{\xi_i - t_i}{1+\xi_i}; \frac{d-s}{2}, \frac{s}{2} \right) \\ - \Phi_s(t_i) \, \mathrm{I}\left(\frac{\xi_i - t_i}{1+\xi_i}; \frac{d-s}{2}, \frac{s}{2} \right), \qquad \mathbf{z} \in \mathbb{S}^d \setminus \Sigma_{i,\gamma_i}, \end{aligned} \tag{45}$$

where $\mathbf{z} = (\sqrt{1-\xi_i^2}\,\overline{\mathbf{z}}, \xi_i) \in \mathbb{S}^d$, $-1 \leq \xi_i \leq 1$ and $\overline{\mathbf{z}} \in \mathbb{S}^{d-1}$, and

$$\mathrm{I}(x;a,b) := \frac{\mathrm{B}(x;a,b)}{\mathrm{B}(a,b)}, \quad \mathrm{B}(a,b) := \mathrm{B}(1;a,b), \quad \mathrm{B}(x;a,b) := \int_0^x u^{a-1}(1-u)^{b-1}du$$

are the regularized incomplete beta function, the beta function, and the incomplete beta function [11, Ch. 5 and 8]; whereas [4, Lemmas 33 and 36]

$$U_{d-2}^{\eta_{i,d-2}}(\mathbf{z}) + Q_i(\mathbf{z}) = \Phi_{d-2}(t_i), \qquad \mathbf{z} \in \Sigma_{i,\gamma_i}, \tag{46}$$

$$\begin{aligned} U_{d-2}^{\eta_{i,d-2}}(\mathbf{z}) + Q_i(\mathbf{z}) = \Phi_{d-2}(t_i)\left(\frac{1+t_i}{1+\xi_i}\right)^{d/2-1} &+ \frac{q_i}{(2(1-\xi_i))^{d/2-1}} \\ - \frac{q_i}{\gamma_i^{d-2}}\left(\frac{1+t_i}{1+\xi_i}\right)^{d/2-1}&, \qquad \mathbf{z} \in \mathbb{S}^d \setminus \Sigma_{i,\gamma_i}. \end{aligned} \tag{47}$$

The last relation follow from (45) if s is changed to $d-2$.

In the proof of our main result for $s = d-2$, $d \geq 3$, we need the analog of (31), which we derive from a similar result for the weighted potential (45). As this is of independent interest, we state and prove the following lemma for $d-2 \leq s < d$.

Lemma 1 *Let $d \geq 3$ and $d-2 \leq s < d$. If (43) is satisfied, then the weighted s-potential of the signed s-equilibrium $\eta_{i,s}$ satisfies the variational inequalities*

$$U_s^{\eta_{i,s}}(\mathbf{z}) + Q_i(\mathbf{z}) = \Phi_s(t_i), \qquad \mathbf{z} \in \Sigma_{i,\gamma_i}, \tag{48}$$

$$U_s^{\eta_{i,s}}(\mathbf{z}) + Q_i(\mathbf{z}) > \Phi_s(t_i), \qquad \mathbf{z} \in \mathbb{S}^d \setminus \Sigma_{i,\gamma_i}. \tag{49}$$

Furthermore, both relations remain valid if equality is allowed in (43).

Proof The first equality (48) was established in [4, Theorems 11 and 15].

Let $d \geq 3$ and $d-2 \leq s < d$. The right-hand side of (45) is a function of ξ_i with $t_i < \xi_i \leq 1$. We denote it by $G(\xi_i)$. Using the integral form of the incomplete regularized beta function, we get

$$\begin{aligned} \mathrm{B}\Big(\frac{d-s}{2}, \frac{s}{2}\Big)\Big(G(\xi_i) - \Phi_s(t_i)\Big) &= \left(\frac{1-t_i}{1-\xi_i}\right)^{s/2} \frac{2^{d-s} q_i}{\gamma_i^d} \\ \times \int_0^{\frac{\xi_i - t_i}{1+\xi_i}} u^{\frac{d-s}{2}-1}\left(1 - \frac{2\,u}{1-t_i}\right)^{\frac{s}{2}-1} \mathrm{d}u &- \Phi_s(t_i) \int_0^{\frac{\xi_i - t_i}{1+\xi_i}} u^{\frac{d-s}{2}-1}\,(1-u)^{\frac{s}{2}-1}\,\mathrm{d}u. \end{aligned}$$

Let (43) be satisfied. Then

$$\mathrm{B}\Big(\frac{d-s}{2},\frac{s}{2}\Big)\,\frac{G(\xi_i)-\Phi_s(t_i)}{\Phi_s(t_i)} > \left[\frac{1-t_i}{1-\xi_i}\right]^{s/2}\int_0^{\frac{\xi_i-t_i}{1+\xi_i}} u^{\frac{d-s}{2}-1}\left(1-\frac{2}{1-t_i}\,u\right)^{\frac{s}{2}-1}\mathrm{d}u$$

$$-\int_0^{\frac{\xi_i-t_i}{1+\xi_i}} u^{\frac{d-s}{2}-1}\,(1-u)^{\frac{s}{2}-1}\,\mathrm{d}u.$$

The square-bracketed expression is > 1 for $-1 < t_i < \xi_i \leq 1$. Since $\frac{2}{1-t_i} > 1$, the first integrand is bounded from below by the second integrand if $\frac{s}{2}-1 \leq 0$. In the case $\frac{s}{2}-1 > 0$, we observe that for $0 \leq u \leq \frac{\xi_i-t_i}{1+\xi_i}$,

$$\begin{aligned}\left[\frac{1-t_i}{1-\xi_i}\right]^{s/2}\left(1-\frac{2}{1-t_i}\,u\right)^{\frac{s}{2}-1} &= \frac{1-t_i}{1-\xi_i}\left(\frac{1-t_i}{1-\xi_i}-\frac{2}{1-\xi_i}\,u\right)^{\frac{s}{2}-1}\\ &> \left(\frac{1-t_i}{1-\xi_i}-\frac{2}{1-\xi_i}\,u\right)^{\frac{s}{2}-1}\\ &\geq (1-u)^{\frac{s}{2}-1}\,.\end{aligned}$$

The estimates are strict in both cases, which yields (49). Moreover, examining these inequalities shows that (49) still holds when (43) is an equality. □

We are now ready to state and prove the second main result.

Theorem 4 *Let $d \geq 3$ and $s = d-2$. Let Q be defined by (13). Suppose the positive charges $q_1,\dots,q_m$ are sufficiently small. Then there exists a critical $\boldsymbol{\epsilon} = (\epsilon_1,\dots,\epsilon_m)$, uniquely defined by these charges, such that $\Sigma_{\epsilon_i}^c \cap \Sigma_{\epsilon_j}^c = \emptyset$, $1 \leq i < j \leq m$, and the $(d-2)$-extremal measure associated with Q is $\mu_Q = C\,\sigma_d|_{\Sigma_{\boldsymbol{\epsilon}}}$ for a uniquely defined normalization constant $C > 1$ and the extremal support is $S_Q = \Sigma_{\boldsymbol{\epsilon}}$.*

Furthermore, an optimal $(d-2)$-energy N-point configuration w.r.t. Q is contained in S_Q for every $N \geq 2$.

Proof Let $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_m)$ be a vector of m positive numbers such that $\Sigma_{i,\gamma_i}^c \cap \Sigma_{i,\gamma_j}^c = \emptyset$, $1 \leq i < j \leq m$. We consider the signed measure

$$\tau := C\,\mathrm{Bal}_{d-2}(\sigma_d,\Sigma_{\boldsymbol{\gamma}}) - \mathrm{Bal}_{d-2}(q_1\,\delta_{\mathbf{a}_1}+\cdots+q_m\,\delta_{\mathbf{a}_m},\Sigma_{\boldsymbol{\gamma}}).$$

As balayage under Riesz $(d-2)$-kernel interactions satisfies (16), we have

$$\begin{aligned} U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) &= C\,U_{d-2}^{\sigma_d}(\mathbf{z}) = C\,W_{d-2}(\mathbb{S}^d), && \mathbf{z}\in\Sigma_{\boldsymbol{\gamma}},\\ U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) &\geq C\,U_{d-2}^{\mathrm{Bal}_{d-2}(\sigma_d,\Sigma_{\gamma})}(\mathbf{z}), && \mathbf{z}\in\mathbb{S}^d\setminus\Sigma_{\boldsymbol{\gamma}}.\end{aligned}$$

If the normalization constant $C = C(\boldsymbol{\gamma})$ is chosen such that $\|\tau\| = \tau(\Sigma_{\boldsymbol{\gamma}}) = 1$, then τ is a signed $(d-2)$-equilibrium measure on $\Sigma_{\boldsymbol{\gamma}}$ associated with Q and, by uniqueness, $\eta_{\Sigma_{\boldsymbol{\gamma}},Q} = \tau$ and $F_{\Sigma_{\boldsymbol{\gamma}},Q} = \mathscr{F}_s(\Sigma_{\boldsymbol{\gamma}})$ with $C = \mathscr{F}_s(\Sigma_{\boldsymbol{\gamma}})/W_{d-2}(\mathbb{S}^d)$.

We show the variational inequality for $\mathbf{z} \in \mathbb{S}^d \setminus \Sigma_{\boldsymbol{\gamma}}$ and proceed in a similar fashion as in the proof of Theorem 1. For $i = 1, \dots, m$ let

$$\nu_i := \sigma_d|_{\Sigma^c_{i,\gamma_i}}, \qquad \widehat{\nu_i} := \mathrm{Bal}_{d-2}(\nu_i, \Sigma_{i,\gamma_i}), \qquad \widehat{\delta_{\mathbf{a}_i}} := \mathrm{Bal}_{d-2}(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}), \tag{50}$$

where t_i is the projection of the boundary circle $\partial\Sigma_{i,\gamma_i}$ onto the $\mathbf{a}_i$-axis; recall that $2(1-t_i) = \gamma_i^2$. As the open spherical caps Σ^c_{i,γ_i}, $1 \leq i \leq m$, do not intersect for $i \neq j$, we have $\partial\Sigma_{\gamma_i} \subset \partial\Sigma_{\boldsymbol{\gamma}}$. Balayage in steps yields

$$\mathrm{Bal}_{d-2}(\nu_i, \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_{d-2}(\nu_i, \Sigma_{i,\gamma_i}) = \widehat{\nu_i} = W_{d-2}(\mathbb{S}^d)\, \frac{1-t_i}{2} \left(1-t_i^2\right)^{d/2-1} \beta_i,$$

$$\mathrm{Bal}_{d-2}(\delta_{\mathbf{a}_i}, \Sigma_{\boldsymbol{\gamma}}) = \mathrm{Bal}_{d-2}(\delta_{\mathbf{a}_i}, \Sigma_{i,\gamma_i}) = \widehat{\delta_{\mathbf{a}_i}} = \frac{4}{\gamma_i^d}\, \frac{1-t_i}{2} \left(1-t_i^2\right)^{d/2-1} \beta_i,$$

where the respective last step follow from [4, Lemmas 33 and 36] and it is crucial that $\widehat{\nu_i}$ and $\widehat{\delta_{\mathbf{a}_i}}$ are supported on $\partial\Sigma_{i,\gamma_i}$ and thus $\partial\Sigma_{\boldsymbol{\gamma}}$, so that

$$\tau = C\, \sigma_d|_{\Sigma_{\boldsymbol{\gamma}}} + C \sum_{i=1}^{m} \widehat{\nu_i} - \sum_{i=1}^{m} q_i\, \widehat{\delta_{\mathbf{a}_i}} \tag{51}$$

$$= C\, \sigma_d|_{\Sigma_{\boldsymbol{\gamma}}} + \sum_{i=1}^{m} \left(C\, W_{d-2}(\mathbb{S}^d) - \frac{4q_i}{\gamma_i^d} \right) \frac{1-t_i}{2} \left(1-t_i^2\right)^{d/2-1} \beta_i. \tag{52}$$

Observe, the signed measure τ has a negative component if and only if

$$C\, W_{d-2}(\mathbb{S}^d) - \frac{4q_i}{\gamma_i^d} < 0 \qquad \text{for at least one } i \in \{1, \dots, m\}.$$

Let $\mathbf{z} \in \mathbb{S}^d \setminus \Sigma_{\boldsymbol{\gamma}}$. Then $\mathbf{z} \in \Sigma^c_{i_0,\gamma_{i_0}}$ for some $i_0 \in \{1, \dots, m\}$ and $\mathbf{z} \in \Sigma_{i,\gamma_i}$ for all $i \neq i_0$. Hence,

$$U^{\tau}_{d-2}(\mathbf{z}) + Q(\mathbf{z}) = C\, W_{d-2}(\mathbb{S}^d) + C \left(U^{\widehat{\nu_{i_0}}}_{d-2}(\mathbf{z}) - U^{\nu_{i_0}}_{d-2}(\mathbf{z}) \right)$$
$$- q_{i_0} \left(U^{\widehat{\delta_{\mathbf{a}_{i_0}}}}_{d-2}(\mathbf{z}) - U^{\delta_{\mathbf{a}_{i_0}}}_{d-2}(\mathbf{z}) \right).$$

Using (19), from [4, Lemmas 33]

$$U^{\widehat{\nu_{i_0}}}_{d-2}(\mathbf{z}) - U^{\nu_{i_0}}_{d-2}(\mathbf{z}) = W_{d-2}(\mathbb{S}^d) \left(\frac{1+t_{i_0}}{1+\xi_{i_0}} \right)^{d/2-1} - W_{d-2}(\mathbb{S}^d) < 0$$

and from [4, Lemmas 36],

$$U_{d-2}^{\widehat{\delta_{\mathbf{a}_{i_0}}}}(\mathbf{z}) - U_{d-2}^{\delta_{\mathbf{a}_{i_0}}}(\mathbf{z}) = \frac{1}{\gamma_{i_0}^{d-2}} \left(\frac{1+t_{i_0}}{1+\xi_{i_0}} \right)^{d/2-1} - \frac{1}{(2(1-\xi_{i_0}))^{d/2-1}} < 0;$$

hence

$$U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) = C\, W_{d-2}(\mathbb{S}^d) \left(\frac{1+t_{i_0}}{1+\xi_{i_0}} \right)^{d/2-1} - \frac{q_{i_0}}{\gamma_{i_0}^{d-2}} \left(\frac{1+t_{i_0}}{1+\xi_{i_0}} \right)^{d/2-1} + \frac{q_{i_0}}{(2(1-\xi_{i_0}))^{d/2-1}}.$$

Observe the similarity to (47). Essentially the same argument as in the proof of Lemma 1 shows that

$$U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) > C\, W_{d-2}(\mathbb{S}^d), \qquad \mathbf{z} \in \Sigma_{i_0,\gamma_{i_0}}$$

in the case when

$$C\, W_{d-2}(\mathbb{S}^d) - \frac{4 q_{i_0}}{\gamma_{i_0}^{d}} \leq 0, \qquad i = 1, \dots, m. \tag{53}$$

It is not difficult to see that near $\partial \Sigma_{i_0,\gamma_{i_0}}$ the following asymptotics holds:

$$\begin{aligned} U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) &= C\, W_{d-2}(\mathbb{S}^d) + \left(\frac{d}{2} - 1 \right) \left(\frac{4 q_{i_0}}{\gamma_{i_0}^{d}} - C\, W_0(\mathbb{S}^d) \right) \frac{\xi_{i_0} - t_{i_0}}{1 + t_{i_0}} \\ &\quad + \frac{1}{2} \left(\frac{d}{2} - 1 \right) \frac{d}{2} \left(\frac{4 q_{i_0}}{\gamma_{i_0}^{d}} \frac{2 t_{i_0}}{1 + t_{i_0}} + C\, W_0(\mathbb{S}^d) \right) \left(\frac{\xi_{i_0} - t_{i_0}}{1 + t_{i_0}} \right)^2 \\ &\quad + \mathcal{O}\Big(\left(\frac{\xi_{i_0} - t_{i_0}}{1 + t_{i_0}} \right)^3 \Big) \qquad \text{as } \xi_{i_0} \to t_{i_0}^{+}; \end{aligned}$$

i.e., the weighted $(d-2)$-potential of τ will be negative sufficiently close to $\partial \Sigma_{i_0,\gamma_{i_0}}$ if (53) does not hold. Hence, if the necessary conditions (53) are satisfied, then

$$U_{d-2}^{\tau}(\mathbf{z}) + Q(\mathbf{z}) > C\, W_{d-2}(\mathbb{S}^d), \qquad \mathbf{z} \in \Sigma_{\boldsymbol{\gamma}}^{c}.$$

Suppose, the system

$$C\, W_{d-2}(\mathbb{S}^d) = \frac{4 q_i}{\gamma_i^{d}}, \qquad i = 1, \dots, m, \tag{54}$$

$$C\, \sigma_d(\Sigma_{\boldsymbol{\gamma}}) = 1, \tag{55}$$

subject to the geometric side conditions

$$\Sigma_{i,\gamma_i} \cap \Sigma_{j,\gamma_j} = \emptyset, \qquad 1 \leq i < j \leq m, \tag{56}$$

has a solution $(\boldsymbol{\gamma}, C)$ with $\boldsymbol{\gamma} = \boldsymbol{\gamma}(C) \in (0,2)^m$ and $C > 0$, then $\eta_{\Sigma_{\boldsymbol{\gamma}},Q} = \tau = C\,\sigma_{d|_{\Sigma_{\boldsymbol{\gamma}}}}$ with $F_{\Sigma_{\boldsymbol{\gamma}},Q} = C\,W_{d-2}(\mathbb{S}^d)$ satisfies the variational inequalities

$$\begin{aligned} U_{d-2}^{\eta_{\Sigma_{\boldsymbol{\gamma}},Q}}(\mathbf{z}) + Q(\mathbf{z}) &= F_{\Sigma_{\boldsymbol{\gamma}},Q}, \qquad \mathbf{z} \in \Sigma_{\boldsymbol{\gamma}}, \\ U_{d-2}^{\eta_{\Sigma_{\boldsymbol{\gamma}},Q}}(\mathbf{z}) + Q(\mathbf{z}) &> F_{\Sigma_{\boldsymbol{\gamma}},Q}, \qquad \mathbf{z} \in \Sigma_{\boldsymbol{\gamma}}^c, \end{aligned} \tag{57}$$

and thus, by Proposition 1(d), $\mu_Q = \eta_{\Sigma_{\boldsymbol{\gamma}},Q} = C\,\sigma_{d|_{\Sigma_{\boldsymbol{\gamma}}}}$ and $S_Q = \Sigma_{\boldsymbol{\gamma}}$. Observe that, given a collection of pairwise different points $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{S}^d$, for sufficiently small charges $q_1, \dots, q_m$, there always exists such a solution. In particular, this is the case if (55) holds for $\gamma_i = [4q_i/W_{d-2}(\mathbb{S}^d)]^{1/d}$.

To determine the parameter C, denote $g(C) := C\sigma_d(\Sigma_{\boldsymbol{\gamma}})$, where

$$\gamma_i := \gamma_i(C) = \left[\frac{4q_i}{CW_{d-2}(\mathbb{S}^d)}\right]^{1/d}, \qquad i = 1, \dots, m.$$

As $\gamma_i = \gamma_i(C)$ are decreasing and continuous functions for all $i = 1, \dots, m$, we derive that $\sigma_d(\Sigma_{\boldsymbol{\gamma}})$ is an increasing and continuous function of C and so is $g(C)$. Also, note that $g(1) = \sigma_d(\Sigma_{\boldsymbol{\gamma}}) < 1$, and $\lim_{C\to\infty} g(C) = \infty$. Therefore, there exists a unique solution C^* of the equation

$$C\sigma_d\left(\bigcap_{i=1}^{m} \Sigma_{i,\gamma_i}\right) = 1,$$

where the γ_i's are defined by (54).

Finally, we invoke [5, Corollary 13] and (57) to conclude that an optimal $(d-2)$-energy N-point configuration w.r.t. Q is contained in S_Q. □

5 Regions of Electrostatic Influence and Optimal $(d-2)$-Energy Points

In this section we consider what happens when the regions of electrostatic influence (see Remark 3 after Theorem 1) have intersecting interiors. We are going to utilize the techniques in the proofs of [5, Theorem 14 and Corollary 15] to show that the support of the $(d-2)$-equilibrium measure associated with the external field (13) satisfies $S_Q \subset \Sigma_\epsilon$, and hence the optimal $(d-2)$-energy points stay away from Σ_ϵ^c. We are going to prove our result for s in the range $d-2 \leq s < d$.

Let $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{S}^d$ be m fixed points with associated positive charges $q_1, \dots, q_m$. We define for $d-2 \leq s < d$ the external field

$$Q_s(\mathbf{x}) := \sum_{i=1}^{m} q_i \, k_s(\mathbf{a}_i, \mathbf{x}), \qquad \mathbf{x} \in \mathbb{S}^d. \tag{58}$$

We introduce the reduced charges

$$\overline{q}_i := \frac{q_i}{1+q-q_i}, \qquad 1 \leq i \leq m.$$

Let $\overline{\Phi_s}(t_i)$ be the Mhaskar-Saff $\mathscr{F}_s$-functional associated with the external field $\overline{q}_i \, k_s(\mathbf{a}_i, \cdot)$ evaluated for the spherical cap Σ_{i,γ_i} (cf. Sect. 4) where it is used that t_i and γ_i are related by $2(1-t_i) = \gamma_i^2$. Let $\overline{\gamma}_i$ denote the unique solution of the equation

$$\overline{\Phi}_s(t_i) = \frac{2^{d-s}\overline{q}_i}{\gamma_i^d}, \qquad 1 \leq i \leq m. \tag{59}$$

Theorem 5 *Let $d-2 \leq s < d$, $d \geq 2$, and let $\overline{\boldsymbol{\gamma}} = (\overline{\gamma}_1, \dots, \overline{\gamma}_m)$ be the vector of solutions of (59). Then the support S_{Q_s} of the s-extremal measure μ_{Q_s} associated with the external field Q_s defined in (58) is contained in the set $\Sigma_{\overline{\boldsymbol{\gamma}}} = \bigcap_{i=1}^{m} \Sigma_{i,\overline{\gamma}_i}$. If $d=2$ and $s=0$, then $\overline{\gamma}_i = \epsilon_i$, $1 \leq i \leq m$, where ϵ_i is defined in (20).*

Furthermore, no point of an optimal N-point configuration w.r.t. Q_s lies in $\Sigma_{i,\overline{\gamma}_i}$, $1 \leq i \leq m$.

Proof First, we consider the case $d-2 < s < d$. Let i be fixed. Since the external field (58) has a singularity at $\mathbf{a}_i$, it is true that $S_{Q_s} \subset \Sigma_{i,\rho}$ for some $\rho > 0$. Moreover, as noted after Definition 2, $S_{Q_s} \subset \operatorname{supp}(\eta^{+}_{\Sigma_{i,\gamma},Q_s})$ for all γ such that $S_{Q_s} \subset \Sigma_{i,\gamma}$. It is easy to see that the signed equilibrium on $\Sigma_{i,\gamma}$ associated with Q_s is given by

$$\eta_{\Sigma_{i,\gamma},Q_s} = \frac{1+\sum_{j=1}^{m} q_j \, \|\widehat{\delta_{\mathbf{a}_j}}\|}{\|\widehat{\nu_i}\|} \, \widehat{\nu_i} - \sum_{j=1}^{m} q_j \, \widehat{\delta_{\mathbf{a}_j}}, \tag{60}$$

where

$$\widehat{\nu_i} = \operatorname{Bal}_s(\sigma_d, \Sigma_{i,\gamma}), \qquad \widehat{\delta_{\mathbf{a}_j}} = \operatorname{Bal}_s(\delta_{\mathbf{a}_j}, \Sigma_{i,\gamma}).$$

Observe, that if $\mathbf{a}_j \in \Sigma_{i,\gamma}$ then $\widehat{\delta_{\mathbf{a}_j}} = \delta_{\mathbf{a}_j}$. We will show that for all $\rho < \gamma < \overline{\gamma}_i$ the signed s-equilibrium measure in (60) will be negative near the boundary $\partial\Sigma_{i,\gamma}$. Indeed, with the convention that the inequality between two signed measures

$\nu_1 \le \nu_2$ means that $\nu_2 - \nu_1$ is a non-negative measure, we have

$$\begin{aligned}\eta_{\Sigma_{i,\gamma},Q_s} &\le \frac{1+\sum_{j=1}^{m} q_j \,\|\widehat{\delta_{\mathbf{a}_j}}\|}{\|\widehat{\nu_i}\|}\,\widehat{\nu_i} - q_i\,\widehat{\delta_{\mathbf{a}_i}} \\ &\le (1+q-q_i)\left(\frac{1+\overline{q}_i\,\|\widehat{\delta_{\mathbf{a}_i}}\|}{\|\widehat{\nu_i}\|}\,\widehat{\nu_i} - \overline{q}_i\,\widehat{\delta_{\mathbf{a}_i}}\right) \\ &= (1+q-q_i)\left[\frac{\overline{\Phi}_s(t)}{W_s(\mathbb{S}^d)}\widehat{\nu_i} - \overline{q}_i\,\widehat{\delta_{\mathbf{a}_i}}\right],\end{aligned} \tag{61}$$

where $2(1-t) = \gamma^2$. The square-bracketed part is the signed equilibrium measure on $\Sigma_{i,\gamma}$ associated with the external field $\overline{q}_i\,k_s(\mathbf{a}_i,\cdot)$ and has a negative component near the boundary $\partial\Sigma_{i,\gamma}$ if and only if $\overline{\Phi}_s(t) - \frac{2^{d-s}\overline{q}_i}{\gamma^d} < 0$ as noted after (42). This inequality holds whenever $\rho < \gamma < \overline{\gamma}_i$ and the inclusion relation $S_{Q_s} \subset \Sigma_{i,\gamma}$ for all $\rho < \gamma < \overline{\gamma}_i$ can now be easily deduced. As i was arbitrarily fixed, we derive $S_{Q_s} \subset \Sigma_{\overline{\boldsymbol{\gamma}}}$. As an optimal N-point configuration w.r.t. Q_s is confined to S_{Q_s}, no point of such a configuration lies in $\Sigma_{i,\overline{\gamma}_i}^c$, $1 \le i \le m$.

In order to obtain the result of the theorem for $d = 2$ and $s = 0$, we use that balayage under logarithmic interaction preserves mass. Hence

$$\eta_{\Sigma_{i,\gamma},Q_0} = \overline{\Phi}_0(t)\,\widehat{\nu_i} - \sum_{j=1}^{m} q_j\,\widehat{\delta_{\mathbf{a}_j}} \le \overline{\Phi}_0(t)\,\widehat{\nu_i} - \overline{q}_i\,\widehat{\delta_{\mathbf{a}_i}}, \qquad \overline{\Phi}_0(t) := 1+q$$

and the characteristic equation $\overline{\Phi}_0(t) = \frac{4\overline{q}_i}{\overline{\gamma}^2}$ reduces to $\overline{\gamma}^2 = \frac{4\overline{q}_i}{1+\overline{q}_i}$. As before, no point of an optimal N-point configuration w.r.t. Q_0 lies in $\Sigma_{i,\overline{\gamma}_i}^c$, $1 \le i \le m$ (as illustrated in Fig. 3). This completes the proof. □

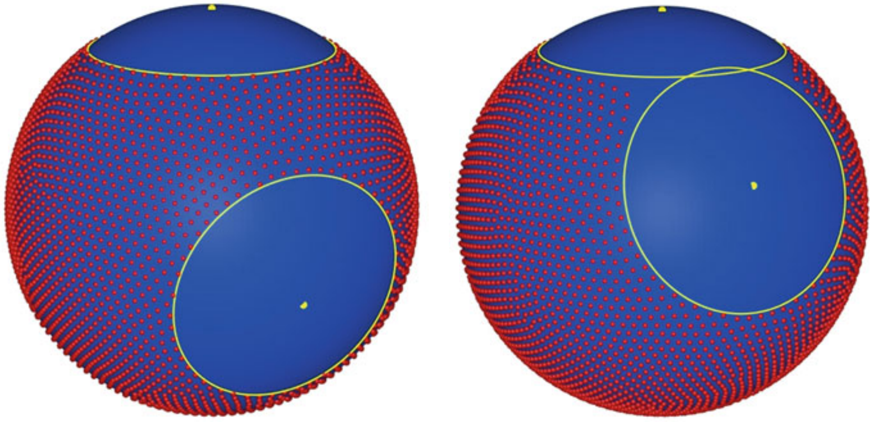


Fig. 3 Approximate Coulomb-optimal points for $m = 2, N = 4000, q_1 = q_2 = \frac{1}{4}, \mathbf{a}_1 = (0,0,1)$ and $\mathbf{a}_2 = (0, \frac{\sqrt{91}}{10}, -\frac{3}{10})$ or $\mathbf{a}_2 = (0, \frac{\sqrt{91}}{10}, \frac{3}{10})$

Example 1 Observe, that if the charges $q_1, \ldots, q_m$ are selected sufficiently small so that for all i we have $\mathbf{a}_j \in \Sigma_{\gamma_i}$, then close to the boundary $\partial\Sigma_{\gamma_i}$ equality holds in (61). So, the critical γ_i can be determined by solving the equation

$$\overline{\Phi}_s(t_i) - 2^{d-s}\overline{q}_i/\gamma_i^d = 0,$$

where

$$\overline{\Phi}_s(t_i) = W_s(\mathbb{S}^d)\frac{1 + \overline{q}_i\|\widehat{\delta}_{t_i,s}\|}{\|\mathrm{Bal}_s(\sigma_2, \Sigma_{\gamma_i})\|}. \tag{62}$$

Motivated by this, we consider the important case of Coulomb interaction potential, namely when $d = 2$ and $s = 1$. We find that (see [4, Lemmas 29 and 30])

$$W_1(\mathbb{S}^2) = 1, \quad \|\widehat{\delta}_{t_i,1}\| = \frac{\arcsin t_i}{\pi} + \frac{1}{2}, \quad \|\mathrm{Bal}_s(\sigma_2, \Sigma_{\gamma_i})\| = \frac{\sqrt{1 - t_i^2} + \arcsin t_i}{\pi} + \frac{1}{2}.$$

Maximizing the Mhaskar-Saff $\mathscr{F}_1$-functional $\overline{\Phi}_1(t)$ is equivalent to solving the equation

$$\frac{\pi(1 + \overline{q}_i/2) + \overline{q}_i \arcsin t_i}{\sqrt{1 - t_i^2} + \arcsin t_i + \pi/2} = \frac{\overline{q}_i}{1 - t_i}.$$

An equivalent equation in term of the geodesic radius α_i of the cap $\Sigma_{\epsilon_i}^c$ of electrostatic influence, so $t_i = \cos(\alpha_i)$ is

$$(\bar{q}_i + 1)\pi\cos(\alpha) - \bar{q}_i\alpha\cos(\alpha) + \bar{q}_i\sin(\alpha) - \pi = 0.$$

Problem 1 The two images in Fig. 4 compare approximate log-optimal configurations with 4000 and 8000 points. The two circles are the boundaries of Σ_{1,ϵ_1} and Σ_{2,ϵ_2}. It is evident that optimal log-energy points stay away from the caps of electrostatic influence Σ_{1,ϵ_1}^c and Σ_{2,ϵ_2}^c of the two charges. In the limit, the log-optimal points approach the log-equilibrium support, which seems to be a smooth region excluding these caps of electrostatic influence. We conclude this section by posing as an open problem, the precise determination of the support in such a case.

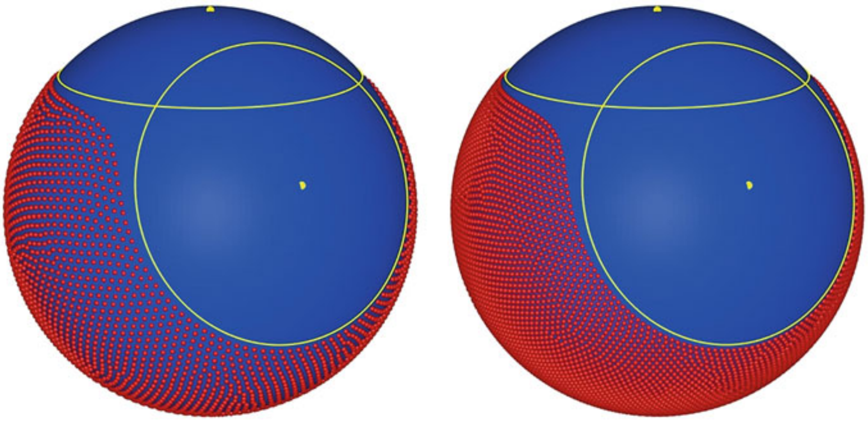


Fig. 4 Approximate log-optimal points for $m = 2$, $N = 4000$ (left) and $N = 8000$ (right), $q_1 = q_2 = \frac{1}{4}$, $\mathbf{a}_1 = (0, 0, 1)$, $\mathbf{a}_2 = (\frac{\sqrt{91}}{10}, 0, \frac{3}{10})$

Acknowledgements The research of Johann S. Brauchart was supported, in part, by the Austrian Science Fund FWF project F5510 (part of the Special Research Program (SFB) "Quasi-Monte Carlo Methods: Theory and Applications") and was also supported by the Meitner-Programme M2030 "Self organization by local interaction" funded by the Austrian Science Fund FWF. The research of Peter D. Dragnev was supported by the Simon's Foundation grant no. 282207. The research of Edward B. Saff was supported by U.S. National Science Foundation grant DMS-1516400. The research of Robert S. Womersley was supported by IPFW Scholar-in-Residence program. All the authors acknowledge the support of the Erwin Schrödinger Institute in Vienna, where part of the work was carried out. This research includes computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW Sydney.

References

1. Beltrán, C.: Harmonic properties of the logarithmic potential and the computability of elliptic Fekete points. Constr. Approx. **37**(1), 135–165 (2013)
2. Beltrán, C.: A facility location formulation for stable polynomials and elliptic Fekete points. Found. Comput. Math. **15**(1), 125–157 (2015)
3. Bilogliadov, M.: Weighted energy problem on the unit sphere. Anal. Math. Phys. **6**(4), 403–424 (2016)
4. Brauchart, J.S., Dragnev, P.D., Saff, E.B.: Riesz extremal measures on the sphere for axis-supported external fields. J. Math. Anal. Appl. **356**(2), 769–792 (2009)
5. Brauchart, J.S., Dragnev, P.D., Saff, E.B.: Riesz external field problems on the hypersphere and optimal point separation. Potential Anal. **41**(3), 647–678 (2014)
6. Dragnev, P.D., Saff, E.B.: Riesz spherical potentials with external fields and minimal energy points separation. Potential Anal. **26**(2), 139–162 (2007)
7. Landkof, N.S.: Foundations of Modern Potential Theory. Springer, New York, Heidelberg (1972). Translated from the Russian by A.P. Doohovskoy, Die Grundlehren der mathematischen Wissenschaften, Band 180

8. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. Appl. Comput. Harmon. Anal. **32**(3), 401–412 (2012)
9. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Multiscale RBF collocation for solving PDEs on spheres. Numer. Math. **121**(1), 99–125 (2012)
10. Mhaskar, H.N., Saff, E.B.: Where does the sup norm of a weighted polynomial live? (A generalization of incomplete polynomials). Constr. Approx. **1**(1), 71–91 (1985)
11. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.14 of 2016-12-21. Online companion to F.W.J. Olver, D.W. Lozier, R.F. Boisvert, C.W. Clark: NIST Handbook of Mathematical Functions. Cambridge University Press, New York, NY (2010)
12. Saff, E.B., Totik, V.: Logarithmic Potentials with External Fields, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 316. Springer, Berlin (1997). Appendix B by Thomas Bloom
13. Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans. Math. Software **23**(4), 550–560 (1997)
14. Zoriĭ, N.V.: Equilibrium potentials with external fields. Ukraïn. Mat. Zh. **55**(9), 1178–1195 (2003)

Numerical Analysis and Computational Solution of Integro-Differential Equations

Hermann Brunner

Dedicated to Ian Sloan, with my best wishes on your 80th birthday.

Abstract The aim of this paper is to describe the current state of the numerical analysis and the computational solution of non-standard integro-differential equations of Volterra and Fredholm types that arise in various applications. In order to do so, we first give a brief review of recent results concerning the numerical analysis of standard (ordinary and partial) Volterra and Fredholm integro-differential equations, with the focus being on collocation and (continuous and discontinuous) Galerkin methods. In the second part of the paper we look at the extension of these results to various classes of non-standard integro-differential equations type that arise as mathematical models in applications. We shall see that in addition to numerous open problems in the numerical analysis of such equations, many challenges in the computational solution of non-standard Volterra and Fredholm integro-differential equations are waiting to be addressed.

1 Introduction

In applications integro-differential equations (IDEs) of Volterra or Fredholm type often arise in 'non-standard' form. While the numerical analysis and the computational solution of standard IDEs are now well understood, this is largely not true for many of their non-standard versions. Thus, the aim of this paper is to present first a concise overview of collocation and Galerkin methods for standard

H. Brunner (✉)
Department of Mathematics, Hong Kong Baptist University, Hong Kong SAR, China

Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada
e-mail: hbrunner@math.hkbu.edu.hk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_11

Volterra and Fredholm IDEs (with the focus being on the former class of equations) and then to describe various classes of non-standard IDEs where the analysis and the implementation of those numerical schemes is rather incomplete. Owing to limitation of space we will deal with partial (e.g. parabolic) IDEs only in passing, as the spatial discretization of such IDEs leads to a (usually large) system of IDEs in time.

1.1 Standard Volterra Integro-Differential Equations

The generic (standard) forms of linear and nonlinear first-order Volterra integro-differential equations (VIDEs) are respectively given by

$$u'(t) = a(t)u(t) + f(t) + \int_0^t (t-s)^{\alpha-1} K(t,s)u(s)\,ds, \;\; t \in I := [0,T], \tag{1}$$

and

$$u'(t) = F(t,u(t)) + \int_0^t (t-s)^{\alpha-1} k(t,s,u(s))\,ds, \;\; t \in I \;\; (0 < \alpha \le 1), \tag{2}$$

with $0 < \alpha \le 1$ and complemented by an initial condition $u(0) = u_0$. The kernels $K = K(t,s)$ and $k = k(t,s,u)$ are assumed to be continuous on their respective domains $D := \{(t,s) : \;\; 0 \le s \le t \le T\}$ and $D \times \mathbb{R}$. If $0 < \alpha < 1$ we will refer to (1) and (2) as *weakly singular* VIDEs.

In applications, nonlinear VIDEs usually occur in *Hammerstein* form; that is, the function $k(t,s,u)$ in (2) is

$$k(t,s,u) = K(t,s)G(s,u), \tag{3}$$

where $G = G(s,u)$ is a smooth function in s and u.

The spatial discretization (by, e.g., finite element or finite difference techniques) of parabolic partial VIDEs for $u = u(t,x)$, for example

$$\frac{\partial u}{\partial t} = \mathscr{A}u + \int_0^t (t-s)^{\alpha-1} K(t,s)(\mathscr{B}u)(s,\cdot)\,ds, \;\; x \in \Omega \subset \mathbb{R}^d, \; t \in I, \tag{4}$$

where $\mathscr{A}$ denotes a linear, uniformly elliptic spatial differential operator (e.g. $\mathscr{A} = \Delta$, the spatial Laplace operator) and $\mathscr{B}$ is a spatial differential operator of order not exceeding 2, leads to a (usually very large) system of VIDEs (1). If the partial VIDE is of hyperbolic type, e.g.

$$\frac{\partial^2 u}{\partial t^2} = \mathscr{A}u + \int_0^t (t-s)^{\alpha-1} K(t,s)(\mathscr{B}u)(s,\cdot)\,ds, \tag{5}$$

spatial discretization yields a (large) system of ordinary second-order VIDEs that is the matrix analogue of the VIDE

$$u^{(r)}(t) = \sum_{j=0}^{r-1} a_j(t)u^{(j)}(t) + f(t) + \int_0^t \sum_{j=0}^{r} K_j(t,s)u^{(j)}(s)\,ds, \quad t \in I, \tag{6}$$

with $r = 2$.

1.2 Non-standard Integro-Differential Equations

VIDEs arising in the mathematical modelling of physical or biological phenomena (for example, in materials with memory, in population dynamics, and in chemical reaction-diffusion processes) often occur in 'non-standard' form. Typical examples are

$$u'(t) = F(t,u(t)) + \int_0^t k(t-s)G(u(t),u(s))\,ds, \quad u(0) = u_0; \tag{7}$$

$$\varepsilon u'(t) = F(t,u(t)) + \int_{-\infty}^t k(t-s)G(u(t),u(s))\,ds, \quad t > 0 \ \ (0 < \varepsilon \ll 1) \tag{8}$$

(*singularly perturbed* VIDE), with $u(t) = \phi(t)$, $(t \le 0)$; and

$$u'(t) = a(t)u(t) + f(t) + \int_0^t K(t,s)G(u(t-s))G(u(s))\,ds, \quad u(0) = u_0 \tag{9}$$

(generalized *auto-convolution* VIDE). A representative example corresponds to $G(u) = u^{\beta}$ $(\beta > 0)$.

In these VIDEs the nonlinearity under the integral sign does not only depend on $u(s)$ but also on $u(t)$ or on $u(t-s)$. We illustrate this, also for further reference, by means of six representative examples. They show that these equations may also depend on a (constant or variable) delay $\tau > 0$.

Example 1 (Volterrra [95–98]; see also Brunner [17]) The system of VIDEs

$$\frac{dN_1(t)}{dt} = N_1(t)\Big(\varepsilon_1 - \gamma_1 N_2(t) - \int_{t-\tau}^t F_1(t-s)N_1(s)\,ds\Big), \tag{10}$$

$$\frac{dN_2(t)}{dt} = N_2(t)\Big(-\varepsilon_2 + \gamma_2 N_1(t) - \int_{t-\tau}^t F_2(t-s)N_2(s)\,ds\Big), \tag{11}$$

where the ε_i and γ_i are given positive parameters, is a mathematical model describing the size of the populations $N_1(t)$ and $N_2(t)$ of interacting predators and preys. The integral operators describing these VIDEs contain a constant delay $\tau > 0$. (See also Cushing [38, Ch. 4] for related population growth models of interacting species.)

Example 2 (Markowich and Renardy [65, 66]) The non-standard VIDE

$$\mu u'(t) = b(t)u^{\beta}(t) + \int_{-\infty}^{t} k(t-s)\Big(\frac{u^3(t)}{u^2(s)} - u(s)\Big)\,ds, \ \ t \geq 0, \tag{12}$$

is a mathematical model for the stretching (and recovery) of a filament or a sheet of a certain molten polymer under a prescribed force. The constant $\mu \geq 0$ is related to the Newtonian contribution to viscosity, and $\beta = 2$ (polymeric filament) or $\beta = 1/2$ (polymeric sheet). For small parameters $0 < \mu \ll 1$ this is a *singularly perturbed* VIDE. (See also Lodge et al. [58] and Jordan [48] for related mathematical models.)

Example 3 (Janno and von Wolfersdorf [47], von Wolfersdorf and Janno [99]) A particular case of the VIDE of auto-convolution type,

$$u'(t) = a(t)u(t) + f(t) + \int_{0}^{t} K(t,s)u(t-s)u(s)\,ds, \ \ t \geq 0, \tag{13}$$

namely,

$$u'(t) + \frac{\gamma}{t}u(t) = a(t)\int_{0}^{t} u(t-s)u(s)\,ds,\ t > 0, \tag{14}$$

arises in the theory of Burgers' turbulence. Details on the physical background of the model can be found in the above papers and their references.

Example 4 (Burns et al. [32]) The mathematical modelling of the elastic motions of a 3-degree of freedom airfoil with flap in a 2-dimensional incompressible flow leads to a system of neutral Volterra functional differential equations of the form

$$\frac{d}{dt}\Big(A_0u(t) - \int_{-\tau}^{0} A_1(s)u(t+s)\,ds\Big) = B_0u(t) + B_1u(t-\tau) + f(t), \ \ t > 0 \tag{15}$$

($\tau > 0$), with A_0, $A_1(\cdot)$ and B_0, B_1 denoting square matrices in $\mathbb{R}^{d\times d}$ (where $d = 8$). The matrix A_0 is singular ($\det A_0 = 0$ but $\operatorname{rank} A_0 \geq 1$); typically, its last row consists of zeros. Compare also Ito and Turi [46] for details and additional references.

Example 5 (Doležal [39, Ch. 5]) The related system of integro-differential algebraic equations (IDAEs)

$$A(t)u'(t) + B(t)u(t) = f(t) + \int_{0}^{t} (t-s)^{\alpha-1}K(t,s)u(s)\,ds, \ \ t \geq 0, \tag{16}$$

with $\alpha = 1$, arises in the theory of electrical networks. Here, $A(\cdot)$, $B(\cdot)$ and $K(\cdot,\cdot)$ are square matrices in $\mathbb{R}^{d\times d}$ $(d \geq 2)$, with $\det A(t) = 0$ for all $t \geq 0$ and $\operatorname{rank} A(t) > 0$. A similar system of IDAEs occurs in the mathematical modelling of a hydraulic circuit that feeds on a combustion process (cf. Nassirharand [78]). The paper by Bulatov et al. [31] is concerned with the theory of IDAEs (16).

Example 6 *Population growth models* (Cushing [38]):

$$u'(t) = u(t)\Big(1 - \int_{-\infty}^{t} k(t-s)u(s)\,ds\Big),\ \ t > 0. \tag{17}$$

See also Aves et al. [4] and its references.

Example 7 *Thermal behavior for a confined reactive gas* (Bebernes and Kassoy [8], Bebernes and Bressan [6]):

$$u_t - \Delta u = \delta \mathrm{e}^u + \big((\gamma-1)/\gamma\big)\big(1/\mathrm{vol}(\Omega)\big)\int_{\Omega} u_t(\cdot,y)\,dy, \tag{18}$$

with $u(t,x) = 0$ $(x \in \partial\Omega,\ t > 0)$, $u(0,x) = u_0(x)$ (where $\Omega \in \mathbb{R}^n$ is bounded with boundary $\partial\Omega$), and $\delta > 0,\ \gamma = 1$ or $\gamma > 1$. Note that this (Fredholm) integro-differential equation is *implicit* in u_t. The monograph by Bebernes and Eberly [7] conveys the general framework of such combustion problems.

Example 8 *Local chemical reaction-diffusion processes* (Chadam et al. [35], Chadam and Yin [34]): The VIDE

$$u_t - \Delta u = \int_{\Omega} H(u(\cdot,y))\,dy,\ \ t > 0,\ x \in \Omega \subset \mathbb{R}^d, \tag{19}$$

complemented by homogeneous Dirichlet or Neumann boundary conditions on $\partial\Omega$, represents a mathematical model of chemical reaction-diffusion processes in which, owing to the effects of a catalyst, the reaction occurs only at some local sites. A typical example corresponds to $H(u) = \mathrm{e}^u$. For certain (large) initial data $u(0,x) \geq 0$ the solution blows up in finite time.

Example 9 *Non-local reaction-diffusion equations with finite-time blow-up* (Souplet [89], Quittner and Souplet [83, Ch. V]):

$$u_t - \Delta u = a\int_0^t u^p(s)\,ds - bu^q,\ \ x \in \Omega \subset \mathbb{R}^d \tag{20}$$

$(a > 0,\ b > 0,\ p,q \geq 1)$, with $u(0,x) = u_0(x) \geq 0$ $(x \in \Omega)$, $u(t,x) = 0$ $(x \in \partial\Omega,\ t \geq 0)$. For u_0 with $u_0(x) \not\equiv 0$, the solution blows up in finite time T_b (i.e. $\|u(t,\cdot)\|_\infty \to \infty$ (as $t \to T_b^-$) whenever $p > q$. For $p \leq q$ the solution exists for all $t > 0$ but is unbounded: $\limsup_{t\to\infty} |u(t,\cdot)| = \infty$.

An analogous result holds for a similar partial VIDE containing integrals over both time and space,

$$u_t - \Delta u = a \int_0^t \int_\Omega k(s) u^p(s,y)\, dy\, ds - b u^q(t,\cdot). \tag{21}$$

Example 10 Dynamics of price adjustment in a single commodity market i (Bélair and Mackey [9]): If $D(\cdot)$ and $S(\cdot)$ denote, respectively, the demand and supply functions for a particular commodity, and P_D and P_S are the demand and supply prices, then a model for the relative variations in market price $P = P(t)$ is

$$\frac{1}{P(t)} P'(t) = F\big(D(P_D), S(P_S)\big), \;\; t \geq 0, \tag{22}$$

subject to some appropriate initial condition. The function $F = F(D,S)$ is the price range function (a simple example is $F(D,S) = D - S$). The demand price

$$P_D(t) = \int_{-\infty}^t K_D(t-v) P(v)\, dv$$

is the weighted average of all the past prices, where $K_D(t-s)$ is a weight attached by the consumer to a past market price $P(s)$ $(-\infty \leq s \leq t)$. The weighting function K_D (the demand price kernel) is assumed to be normalized so that $\int_0^\infty K_D(s) ds = 1$. An analogous expression exists for the supply price P_S: it is

$$P_S(t) = \int_{-\infty}^{t-T_{\min}} K_S(t - T_{\min} - v) P(v)\, dv,$$

where $T_{\min}$ denotes the minimum time which must elapse before a decision to alter production is translated into an actual change of supply.

Example 11 Evolution of a spherical flame initiated by a point source energy input and subject to heat loss (Audounet et al. [3], Rouzaud [84]):

$$\pi^{-1/2} \int_0^t (t-s)^{-1/2} u'(s)\, ds = u(t) \log\big(u(t)\big) + Eq(t) - \lambda u^3(t), \tag{23}$$

with $u(0) = 0$. Here, the constant $E > 0$ and the function $q(t) \geq 0$ are given; $Eq(t)$ is a point source energy input. It can be shown that there exists a 'critical value' $\lambda = \lambda^*$ for which the flame always quenches.

The monographs by Prüss [82] and by Appell et al. [2] contain, in addition to the theory of VIDEs, additional applications of (ordinary and partial) VIDEs.

We note that in many of the above examples the numerical analysis and computational treatment of the respective integro-differential equations are as yet little understood.

The paper is organized as follows. In Sect. 2 we give a concise review of results on the optimal (global and local) order of convergence of collocation and (continuous or discontinuous) Galerkin solutions for standard linear VIDEs. The extension of the convergence analysis for collocation and Galerkin-type solutions for various classes of non-standard VIDEs, including equations with delay arguments, integro-algebraic equations and fractional evolution equations, is the subject of Sect. 3. There we describe theoretical and computational issues that are waiting to be addressed. In Sect. 4 we turn our attention to Volterra and Fredholm IDEs whose solutions blow up in finite time. Owing to limitation of space we only briefly discuss partial VIDEs. However, since the first step in the discretization of such problems usually consists in the approximation (e.g. by finite difference of finite element Galerkin techniques) of the spatial derivatives of the solution, the numerical schemes described in this paper can be employed for the subsequent temporal discretization (time-stepping). The aim of the list of references is to guide the reader to papers that reflect the current 'state of the art' of the numerical analysis and the computational solution of VIDEs, as well as to papers on integro-differential equations not treated in the present paper.

2 Numerical Analysis of Ordinary VIDEs

In his paper [94] of 1909 (the first paper on applications of partial VIDEs) Volterra makes the following observation:

> The problem of solving integro-differential equations constitutes a problem which is fundamentally different from the problems of solving differential equations and integral equations.

As we shall see in the following sections, this comment remains true for the numerical analysis and computational solution of VIDEs.

We first present a brief overview of prominent time-stepping schemes for VIDEs. They include collocation methods, continuous and discontinuous Galerkin methods, and convolution quadrature methods.

2.1 *Collocation and Galerkin Spaces*

Suppose we want to approximate the solution of an ordinary VIDE (1.1) (we use the notation (1.1) to refer to Eq. (1) in Sect. 1, etc.) on the time interval $I := [0, T]$ $(T < \infty)$, and let $I_h := \{t_n :\ 0 = t_0 < t_1 < \cdots < t_N = T\}$ be a (not necessarily uniform) mesh for I (i.e. an h-discretization of I), with

$$h_n := t_{n+1} - t_n, \quad e_n := (t_n, t_{n+1}], \quad h := \max\{h_n :\ 0 \le n \le N-1\}.$$

The approximating space will be either

$$S_m^{(0)}(I_h) := \{v \in C(I) :\ v\big|_{e_n} \in \mathscr{P}_m\ (0 \le n \le N-1)\},$$

the space of globally continuous piecewise polynomials of (fixed) degree $m \ge 1$ for all e_n ($\mathscr{P}_m = \mathscr{P}_m(e_n)$ denotes the set of (real-valued) polynomials on e_n of degree not exceeding m), or

$$S_m^{(-1)}(I_h) := \{v :\ v\big|_{e_n} \in \mathscr{P}_m\ (0 \le n \le N-1)\},$$

the space of piecewise polynomials of degree $m \ge 0$ that may possess finite jump discontinuities at the interior mesh points of I_h. The dimensions of these linear spaces are given respectively by

$$\dim S_m^{(0)}(I_h) = Nm + 1 \qquad \text{and} \qquad \dim S_m^{(-1)}(I_h) = N(m+1).$$

It is often advantageous (especially when approximating non-smooth solutions of (1.1)) not to use the same polynomial degree m on each subinterval e_n. Thus, an *hp*-discretization of I is defined as follows: for the given mesh and given nonnegative integers m_i $(i = 0, 1, \dots, N-1)$ we consider the *degree vector* $\underline{m} := (m_0, m_1, \dots, m_{N-1})$, with $|\underline{m}| := \max\{m_n :\ 0 \le n \le N-1\}$. For $d \in \{-1, 0\}$ the corresponding piecewise polynomial spaces are then

$$S_{\underline{m}}^{(d)}(I_h) := \{v \in C^d(I) :\ v\big|_{e_n} \in \mathscr{P}_{m_n}\ (0 \le n \le N-1)\}.$$

(If $d = -1$ the elements of $S_{\underline{m}}^{(-1)}(I_h)$ are in general not continuous at the interior points of I_h.) It is easily seen that we have

$$\dim S_{\underline{m}}^{(0)}(I_h) = \sum_{n=0}^{N-1} m_n + 1 \qquad \text{and} \qquad \dim S_{\underline{m}}^{(-1)}(I_h) = \sum_{n=0}^{N-1} m_n + N.$$

In order to obtain high-order collocation or Galerkin approximations to VIDEs with weakly singular kernels whose solutions typically have unbounded second derivatives at $t = 0$, one will choose a mesh on $I = [0, T]$ that is *locally refined* near $t = 0$. Such meshes, denoted by $I_h(r, \sigma)$, correspond to a *grading parameter* $\sigma \in (0, 1)$ and $r \ge 1$ *levels of refinement* and are defined by the mesh points

$$\{t_{0,0} := 0,\ t_{0,\mu} := \sigma^{r-\mu} t_1\ (1 \le \mu \le r),\ t_n\ (1 \le n \le N)\}. \tag{24}$$

We associate with each subinterval $e_{0,\mu} := (t_{0,\mu}, t_{0,\mu+1}]$ $(0 \le \mu \le r-1)$ a nonnegative integer $m_{0,\mu}$; these integers define the initial degree vector $\underline{m}_0 := (m_{0,0}, \dots, m_{0,r-1})$. For $d \in \{-1, 0\}$ the corresponding piecewise polynomial spaces are defined by

$$\begin{aligned} &S^{(d)}_{\underline{m}}(I_h(r,\sigma)) \\ &\quad := \{v \in C^d(I) :\ v\big|_{e_{0,\mu}} \in \mathscr{P}_{m_{0,\mu}}\ (0 \le \mu \le r-1);\ v\big|_{e_n} \in \mathscr{P}_{m_n}\}, \end{aligned} \tag{25}$$

where $(m_1, \dots, m_{N-1})$ is the degree vector for the intervals $e_1, \dots, e_{N-1}$. The dimensions of these linear spaces are

$$\dim S^{(-1)}_{\underline{m}}(I_h(r,\sigma)) = \sum_{k=0}^{r-1} m_{0,k} + \sum_{n=1}^{N-1} m_n + r + N - 1 \tag{26}$$

and

$$\dim S^{(0)}_{\underline{m}}(I_h(r,\sigma)) = \sum_{k=0}^{r-1} m_{0,k} + \sum_{n=1}^{N-1} m_n + 1,$$

respectively. These spaces will be used below in the formulation of the *hp*-versions of collocation and (continuous and discontinuous) Galerkin methods.

2.2 Collocation Time-Stepping

We first recall time-stepping schemes based on collocation in the piecewise polynomial space $S^{(0)}_m(I_h)$ for the VIDE

$$u'(t) = a(t)u(t) + f(t) + (\mathscr{V}_\alpha u)(t),\ \ t \in I := [0,T],\ \ u(0) = u_0, \tag{27}$$

where the Volterra integral operator $\mathscr{V}_\alpha :\ C(I) \to C(I)$ is

$$(\mathscr{V}_\alpha u)(t) := \int_0^t (t-s)^{\alpha-1} K(t,s) u(s)\, ds$$

$(0 < \alpha \le 1)$, with $K \in C(D)$ $(D := \{(t,s) :\ 0 \le s \le t \le T\})$. Since $\dim S^{(0)}_m(I_h) = Nm + 1$, we choose the set of collocation points

$$X_h := \{t_{n,i} := t_n + c_i h_n :\ i = 1, \dots, m\ (0 \le n \le N-1)\} \tag{28}$$

of cardinality $|X_h| = Nm$ and defined by $m \ge 1$ prescribed collocation parameters $\{c_i :\ 0 < c_1 < \cdots < c_m \le 1\}$. The collocation equation defining the collocation solution $u_h \in S_m^{(0)}(I_h)$ for (4) is then given by

$$u_h'(t) = a(t)u_h(t) + f(t) + (\mathcal{V}_\alpha u)(t), \;\; t \in X_h, \tag{29}$$

and complemented by the initial condition $u_h(0) = u_0$.

The local (time-stepping) version of (4) (for $t = t_n + vh_n \in e_n$) has the form

$$\begin{aligned} u'(t_n + vh_n) &= a(t_n + vh_n)u(t_n + vh_n) + f(t_n + vh_n) + H_n(t_n + vh_n) \\ &\quad + h_n^\alpha \int_0^v (v - s)^{\alpha - 1} K(t_n + vh_n, t_n + sh_n)u(t_n + sh_n)\, ds \end{aligned} \tag{30}$$

($v \in (0, 1]$), with the history term

$$\begin{aligned} H_n(t) &:= \int_0^{t_n} (t - s)^{\alpha - 1} K(t, s)u(s)\, ds \\ &= \sum_{\ell=0}^{n-1} h_\ell \int_0^1 (t - (t_\ell + sh_\ell))^{\alpha - 1} K(t, t_\ell + sh_\ell)u(t_\ell + sh_\ell)\, ds \end{aligned} \tag{31}$$

($t = t_n + vh_n \in e_n$). Thus, the corresponding (local) collocation equation for $t_{n,i} \in e_n$ is

$$\begin{aligned} & u_h'(t_{n,i}) - a(t_{n,i})u_h(t_{n,i}) - h_n^\alpha \int_0^{c_i} (c_i - s)^{\alpha - 1} K(t_{n,i}, t_n + sh_n)u_h(t_n + sh_n)\, ds \\ &\quad = f(t_{n,i}) + \hat{H}_n(t_{n,i}) \end{aligned} \tag{32}$$

($i = 1, \ldots, m$), where the approximation $\hat{H}_n(t)$ to the history term $H_n(t)$ in (8) is

$$\hat{H}_n(t) := \int_0^{t_n} (t - s)^{\alpha - 1} K(t, s)u_h(s)\, ds \;\; (t \in e_n).$$

In order to obtain the computational form of (9) we set

$$Y_{n,i} := u_h'(t_{n,i}), \quad L_j(v) := \prod_{k=1, k \neq j}^{m} \frac{v - c_k}{c_j - c_k}, \quad b_j(v) := \int_0^v L_j(s)\, ds \;\; (v \in [0, 1]).$$

Since u_h' on e_n is a polynomial of degree $m - 1$ we may write

$$u_h'(t_n + vh_n) = \sum_{j=1}^{m} L_j(v)Y_{n,j} \;\; (v \in (0, 1]).$$

This implies that on e_n the collocation approximation u_h has the local representation

$$u_h(t_n + vh_n) = u_n + h_n \sum_{j=1}^{m} b_j(v) Y_{n,j} \ \ (v \in [0, 1]), \tag{33}$$

where $u_n := u_h(t_n)$. It allows us to write the local collocation Eq. (9) in the form

$$Y_{n,i} - h_n a(t_{n,i}) \sum_{j=1}^{m} b_{i,j} Y_{n,j} - h_n^2 \sum_{j=1}^{m} \int_0^{c_i} (c_i - s)^{\alpha-1} K(t_{n,i}, t_n + sh_n) b_j(s) ds \cdot Y_{n,j}$$

$$= f(t_{n,i}) + \hat{H}(t_{n,i}) + \Big(a(t_{n,i}) + h_n \int_0^{c_i} (c_i - s)^{\alpha-1} K(t_{n,i}, t_n + sh_n) ds\Big) u_n, \tag{34}$$

with $b_{i,j} := b_j(c_i)$. This is a system of m linear algebraic equations for the vector $Y_n := (Y_{n,1}, \ldots, Y_{n,m})^T$. For $a, f \in C(I)$, $K \in C(D)$ and $0 < \alpha \leq 1$, it possesses a unique solution $Y_n \in \mathbb{R}^m$ for $0 \leq n \leq N - 1$ and for all meshes I_h with sufficiently small mesh diameter $h > 0$.

Remark 1 Since the integrals in (11) cannot, in general, be evaluated analytically a further discretization step consisting in approximating these integrals by appropriate numerical quadrature schemes, e.g. m-point interpolatory quadrature formulas with abscissas coinciding with the collocation points will be necessary (cf. Brunner [16] and Remark 2 below).

The attainable order of convergence of the collocation solution $u_h \in S_m^{(0)}(I_h)$ depends strongly on the regularity of the solution u of the VIDE (3). If $\alpha = 1$, then u essentially inherits the regularity of the data: C^m data a, f and K imply that $u \in C^{m+1}(I)$. For $0 < \alpha < 1$ this is no longer true: for such C^m data we obtain in general only $u \in C^1(I) \cap C^{m+1}(0, T]$: its second derivative behaves like $u''(t) \sim t^{\alpha-1}$ as $t \to 0^+$. We summarize these observations in the following theorems (see for example Brunner [16, Ch. 3]).

Theorem 1 *Assume that $a, f \in C^d(I)$, $K \in C^d(D)$ $(d \geq m)$, $\alpha = 1$, and let $u_h \in S_m^{(0)}(I_h)$ be the collocation solution defined by the collocation Eq. (6), with I_h being (quasi-)uniform.*

(a) If $d \geq m$ and the collocation parameters $\{c_i\}$ are chosen arbitrarily, there holds $\|u - u_h\|_\infty \leq Ch^m$. The exponent m can in general not be replaced by $m + 1$.

(b) If $d \geq m + 1$ and if the collocation parameters satisfy the orthogonality condition

$$\int_0^1 \prod_{i=1}^{m} (s - c_i)\, ds = 0, \tag{35}$$

the attainable order of convergence of u_h is given by $\|u - u_h\|_\infty \leq Ch^{m+1}$. This holds in particular when the c_i are the (shifted) Gauss-Legendre points (i.e. the *zeros of the Legendre polynomial $P_m(2s-1)$) or the (shifted) Radau II points (the zeros of $P_m(2s-1) - P_{m-1}(2s-1)$, with $c_m = 1$).*

For sufficiently regular solutions and very special choices of the collocation parameters $\{c_i\}$ the collocation solution $u_h \in S_m^{(0)}(I_h)$ exhibits a higher order of (local) superconvergence at the mesh points I_h.

Theorem 2 *Let the assumptions of Theorem 1 hold and assume that the collocation parameters are such that*

$$\int_0^1 s^\nu \prod_{i=1}^m (s - c_i)\, ds = 0 \quad \textit{for} \quad \nu = 0, \ldots, \kappa - 1 \;\; (\kappa \le m).$$

Then the optimal order of (local) convergence of $u_h \in S_m^{(0)}(I_h)$ at the points of I_h is

$$\max_{1 \le n \le N} |u(t_n) - u_h(t_n)| \le C_d h^{m+\kappa}.$$

Important special cases are the m Gauss-Legendre points (corresponding to $\kappa = m$) and the Radau II points (corresponding to $\kappa = m - 1$, with $c_m = 1$).

Remark 2 The local superconvergence results on I_h remain valid (with different, usually larger, error constants C_d) if the integrals in the collocation Eq. (11) are approximated by m-point *interpolatory* quadrature formulas whose abscissas are the collocation points. The resulting 'fully discretized' collocation equation represents an implicit m-stage Volterra-Runge-Kutta method (Brunner [16, Section 3.2.2]; see also Brunner and van der Houwen [21]).

The general theory of (explicit and implicit) Runge-Kutta is due to Lubich [60] (compare also Brunner and van der Houwen [21, Section 4.2]). Implicit Runge-Kutta time discretization (and their asymptotic stability properties) were studied by Brunner et al. [24]. See also Kauthen [49] on implicit Runge-Kutta methods for singularly perturbed VIDEs.

For VIDEs (4) with weakly singular kernels (corresponding to $0 < \alpha < 1$) the above results on the attainable order of convergence of the collocation solution $u_h \in S_m^{(0)}(I_h)$ are no longer valid, owing to the low regularity of the solution u at $t = 0$. The following theorem is due to Brunner [12], Tang [91] (see also Brunner et al. [25] and Brunner [16, Section 7.2]).

Theorem 3 *Let the functions a, f, K in (4) be subject to the assumptions of Theorem 1, with $0 < \alpha < 1$. Then the collocation solution $u_h \in S_m^{(0)}(I_h)$ defined by (6), (7) possesses the following convergence properties:*

(a) If the mesh I_h is (quasi-)uniform, then $\|u - u_h\|_\infty \le C_\alpha h^{1+\alpha}$ for any $m \ge 2$.
(b) If I_h is a (globally) graded mesh whose points are given by $t_n = (n/N)^r T$, with $r \ge (m + \alpha)/(1 + \alpha)$, then the attainable order of convergence of u_h on I is

described by

$$\max_{1\le n\le N} |u(t_n) - u_h(t_n)| \le C_\alpha N^{-(m+\alpha)} \ \ (m \ge 2),$$

provided the collocation parameters $\{c_i\}$ are such that (12) holds.

While the use of globally graded meshes restores the higher order of convergence of collocation solutions for VIDEs (1.1) with $0 < \alpha < 1$, it has the drawback that h_n, the size of the subinterval e_n becomes very large (compared to h_0) as n tends to $N-1$. There are a number of approaches that avoid this problem.

(a) *Piecewise non-polynomial spline collocation:* For a given (uniform) mesh I_h the piecewise polynomial space $S_m^{(0)}(I_h)$ is augmented by an appropriate number (depending on m and α) of non-polynomial basis functions that, on the initial interval e_0, reflect the singular behaviour of higher derivatives of the solution u (cf. Brunner [11]).

(b) *Hybrid collocation:* This approach combines non-polynomial spline collocation near the initial point $t = 0$ and piecewise polynomial spline collocation on e_n with $n \ge 1$. It was analyzed for weakly singular Volterra integral equations in Cao et al. [33]; it seems that for weakly singular VIDEs this has not yet been studied.

(c) *hp-collocation with local mesh refinement:* As we shall see at the end of Sect. 2.3, piecewise polynomial collocation in $S_m^{(0)}(I_h)$ for the VIDE (4) is closely related to *discretized* cG and dG methods in $S_m^{(0)}(I_h)$ and $S_{m-1}^{(-1)}(I_h)$, respectively. Thus, the convergence analysis for the latter approximations to the solution of (4) with $0 < \alpha \le 1$ can be employed to derive optimal convergence results for *hp*-collocation methods. This analysis is currently being carried out.

2.3 Continuous and Discontinuous Galerkin Time-Stepping

Based on the variational form of the VIDE (4) the exact *continuous* Galerkin (cG) equation for $u_h \in S_m^{(0)}(I_h)$ has the form

$$\langle u_h' - au_h, \phi\rangle = \langle f, \phi\rangle + \langle \mathcal{V}_\alpha u_h, \phi\rangle \quad \text{for all} \quad \phi \in S_m^{(0)}(I_h), \tag{36}$$

where the (global) inner product of g and h is given by $\langle g, h\rangle := \int_I g(s)h(s)ds$. (We use the terminology 'exact Galerkin equation' to indicate that the inner products are evaluated exactly.) In analogy to the collocation Eq. (6) in $S_m^{(0)}(I_h)$ the cG Eq. (13) is complemented by the initial condition $u_h(0) = u_0$.

The (exact) *discontinuous* Galerkin (dG) equation in $S_m^{(-1)}(I_h)$ for (4) is

$$\langle u_h' - au_h, \phi\rangle = \langle f, \phi\rangle + \langle \mathcal{V}_\alpha u_h, \phi\rangle \quad \text{for all} \quad \phi \in S_m^{(-1)}(I_h). \tag{37}$$

The above cG and dG equations can be written in local ‘time-stepping’ form where the inner products are now taken over the subintervals e_n. We will do this first for the dG Eq. (14) where we have to take into account the jump discontinuities of the test functions $\phi \in S_m^{(-1)}(I_h)$ across the interior points of the mesh I_h. It is readily verified that on e_n the dG equation assumes the form

$$\begin{aligned}\int_{e_n} u_h'(t)\phi(t)\,dt - U_n^+\phi_n^+ = \int_{e_n} a(t)u_h(t)\phi(t)\,dt \\ + \int_{e_n}\Big(\int_{t_n}^{t}(t-s)^{\alpha-1}K(t,s)u_h(s)ds\Big)\phi(t)\,dt \\ + U_n^-\phi_n^+ + \int_{e_n} f(t)\phi(t)dt + \int_{e_n}\hat{H}_n(t)\phi(t)\,dt\end{aligned} \tag{38}$$

for all $\phi \in \mathscr{P}_m(e_n)$ and $0 \le n \le N-1$ (see also Brunner and Schötzau [20]). Here, we have set

$$U_n^+ := u_h(t_n^+),\quad U_n^- := u_h(t_n^-),\quad \phi_n^+ := \phi(t_n^+),$$

and $\hat{H}_n(t)$ is as in (9).

An equation analogous to (15) holds for the cG Eq. (13), except that now there are no jump discontinuity terms (since $u_h \in C(I)$):

$$\begin{aligned}\int_{e_n} u_h'(t)\phi(t)\,dt = \int_{e_n} a(t)u_h(t)\phi(t)\,dt \\ + \int_{e_n}\Big(\int_{t_n}^{t}(t-s)^{\alpha-1}K(t,s)u(s)ds\Big)\phi(t)\,dt \\ + \int_{e_n} f(t)\phi(t)\,ds + \int_{e_n}\hat{H}_n(t)\phi(t)\,dt\end{aligned} \tag{39}$$

for all $\phi \in \mathscr{P}_m(e_n)$ $(0 \le n \le N-1)$.

We cite two representative results on the attainable order of convergence of *hp*-dG approximations $u_h \in S_{\underline{m}}^{(-1)}(I_h)$ and $u_h \in S_{\underline{m}}^{(-1)}(I_h(r,\sigma))$. The underlying VIDE (1.1) is assumed to be parabolic (that is, $a \in C(I)$ satisfies $\underline{a} \le -a(t) \le \bar{a}$ $(t \in I)$ for some constants $\underline{a} \le \bar{a} < \infty$), as well as subject to some additional technical assumptions (see Brunner and Schötzau [20] and Mustapha et al. [76] for details).

(1) In the (atypical) case where the solution u of the VIDE (1.1) with $0 < \alpha < 1$ is analytic on I there holds

$$\|u - u_h\|_\infty \le C\mathrm{e}^{-b|\underline{m}|},$$

where the constants C and b are independent of the degree vector $\underline{m}$.

(2) If the data af, and K are analytic on I and D, respectively, (implying that u is not analytic on I) then there exist degree vectors $\underline{m}_0$ (on e_0) and $\underline{m}$ (on $[t_1, T]$) so that for the locally geometrically refined mesh $I_h(r, \sigma)$ the dG solution $u_h \in S_{\underline{m}}^{(-1)}(I_h(r, \sigma))$ satisfies

$$\|u - u_h\|_\infty \le C\mathrm{e}^{-bM_m^{-1/2}},$$

where

$$M_m := \dim S_{\underline{m}}^{(-1)}(I_h(r, \sigma)) = \sum_{k=0}^{r-1} m_{0,k} + \sum_{n=1}^{N-1} m_n + r + N - 1,$$

(cf. (3)), with constants C and b not depending on the degree vectors.

Remark 3

(i) While 'good' values of the grading parameter $\sigma \in (0, 1)$ can be determined numerically (see Brunner and Schötzau [20, pp. 242–243] for a discussion), the analysis of how to select an optimal grading parameter σ remains to be carried out.

(ii) Superconvergence results for dG solutions (h-version) for weakly singular VIDEs (1.1) can be found in Mustapha [72]. Analogous results for cG solutions do not seem to have been derived yet.

(iii) An interesting alternative to hp-dG methods for VIDEs with weakly singular kernels are hp-Petrov-Galerkin methods: here, the approximate solution is sought in the space $S_m^{(0)}(I_h)$ while the test space is a space of discontinuous piecewise polynomials. This extension of the hp-dG methods analyzed in [20] can be found in Yi and Guo [100], together with results on the attainable order of convergence of such Petrov-Galerkin solutions.

Since the (local) integrals (inner products) in the Galerkin equations (15) and (16) can in general not be found analytically, they need to be approximated by appropriate quadrature schemes in order to obtain the computational form of these equations. For the dG Eq. (15) the obvious choice are m-point *interpolatory* (product) quadrature formulas with abscissas $0 \le d_0 < d_1 < \cdots < d_m \le 1$. For the approximation of the first integral on the right-hand side of (15) this results in

$$\int_{e_n} a(t)u_h(t)\phi(t)\,dt = h_n \int_0^1 a(t_n + vh_n)u_h(t_n + vh_n)\phi(t_n + vh_n)\,dv$$

$$\approx h_n \sum_{j=0}^{m} w_j a(t_n + d_j h_n)u_h(t_n + d_j h_n)\phi(t_n + d_j h_n).$$

The resulting discretized dG equation is related to (but, owing to the finite jump terms, not identical with) the collocation Eq. (9) for $u_h \in S_{m+1}^{(0)}(I_h)$ with the

$\{d_i\}$ as collocation parameters. (This is an extension of Lasaint and Raviart [51] where this relationship was explored for ODEs; see also [20].) On the other hand, the discretized cG Eq. (16) coincides with the collocation Eq. (9) if m-point interpolatory quadrature with abscissas $d_i = c_i$ is used.

Remark 4

(i) For long-time integration and very large values of N the re-evaluation of the history terms (i.e. the integrals over $[0, t_n]$) in (15) and (16) for each new interval e_n will become very expensive. In such situations the use of 'sparse quadrature' may reduce the computational effort; see for example Sloan and Thomeé [88] or Adolfssson et al. [1].

(ii) For certain partial VIDEs with *convolution kernels*, for example

$$\frac{\partial u}{\partial t} + \int_0^t (t-s)^{\alpha-1} \mathscr{A} u(s)\, ds = f(t), \;\; t \in I, \;\; u(0) = u_0 \tag{40}$$

where $0 < \alpha < 1$ and $\mathscr{A}$ is an elliptic (spatial) differential operator, convolution quadrature based on Laplace transform techniques leads to efficient time-stepping schemes (see for example McLean and Thomée [71], Schädle et al. [85], López-Fernández et al. [59], Mustapha and McLean [73], as well as Cuesta et al. [37] for more general versions of (17) and the use of modified convolution quadrature techniques for time-stepping). A particular example (Fujita [40]) is the VIDE

$$u_t = f + \int_0^t k(t-s) \Delta u(s, \cdot)\, ds \;\; :$$

it 'interpolates' between heat equation (corresponding to $k(t-s) = \delta(t-s)$) and the wave equation ($k(t-s) \equiv 1$).

2.4 Collocation and Galerkin Methods for FIDEs

A comprehensive analysis of piecewise polynomial collocation solutions for boundary-value problems for nonlinear Fredholm integro-differential equations

$$u^{(r)}(t) = F\big(t, u(t), \ldots, u^{(r-1)}(t)\big) + \int_a^b k\big(t, s, u(s), \ldots, u^{(r)}(s)\big) ds, \;\; t \in [a, b], \tag{41}$$

with $r \geq 1$, is due to Hangelbroek et al. [45]. In particular, they derived optimal local superconvergence results at the mesh points I_h. Similar local superconvergence results hold for initial-value problems for analogous rth-order VIDEs (cf. Brunner

[13]). A boundary-value problem for the nonlinear second-order nonlinear FIDE

$$u''(t) + \int_0^1 k(t-s)u^4(s)\,ds = f(t), \;\; t \in [0,1],$$

arises when studying a coupled system of integro-differential-algebraic equations that models exothermic catalytic combustion in a cylinder. Its numerical treatment by orthogonal collocation methods and the derivation of optimal convergence results are discussed in Ganesh and Spence [42]. An alternative numerical scheme, namely a Petrov-Galerkin method, is analyzed in Ganesh and Sloan [41].

An analysis of projection methods, and in particular of cG methods, for FIDEs (18) can be found in Volk [92, 93]. The second paper also contains superconvergence results for cG solutions.

Parts, Pedas and Tamme [79] and Pedas and Tamme [80] established a comprehensive theory on the regularity of solutions of linear, weakly singular FIDEs

$$u'(t) = a(t)u(t) + f(t) + \int_0^T K(t,s)u(s)\,ds, \;\; t \in [0,T],$$

where $K(t,s)$ contains weak algebraic or logarithmic singularities, or is bounded but has unbounded derivatives. This is complemented by an equally comprehensive analysis of the order of optimal convergence of piecewise polynomial collocation solutions (see also [81]).

Large systems of FIDEs with dense matrices are encountered in the spatial discretization of linear parabolic FIDEs $u_t + \mathscr{A}u = 0$ where $\mathscr{A}$ is the sum of a second-order elliptic (spatial) differential operator and a linear Fredholm integral operator over some bounded domain $\Omega \subset \mathbb{R}^d$. Such FIDEs arise in the mathematical modelling of stochastic processes in financial mathematics (e.g. in option pricing). Matache et al. [67] proposed a numerical scheme, based on wavelet discretization in space and dG time discretization, in which the (large) dense matrix is replaced by using wavelet compression techniques. The complexity of such schemes is analyzed in Matache et al. [68].

3 Numerical Analysis of Non-standard VIDEs

3.1 Auto-Convolution VIDEs

It was shown in Brunner [14] that for the non-standard VIDE

$$u'(t) = a(t)u(t) + f(t) + \int_0^t K(t,s)G(u(t),u(s))\,ds, \;\; t \in I, \tag{42}$$

the optimal orders of (global and local) convergence of collocation solutions $u_h \in S_m^{(0)}(I_h)$ described in Theorems 1 and 2 remain valid (see also Brunner et al. [26] for a study of similar time-stepping for analogous partial VIDEs). The proof of these results proceeds along the lines of the one for standard nonlinear VIDEs (cf. Brunner [16]). Discontinuous Galerkin methods for (1) were analyzed in Ma and Brunner [62]; the paper includes the derivation of a posteriori error bounds for the piecewise polynomial spaces $S_{m-1}^{(-1)}(I_h)$ with $m = 1$ and $m = 2$.

Consider now the (generalized) auto-convolution VIDE

$$u'(t) = a(t)u(t) + f(t) + \int_0^t K_\alpha(t,s)G(u(t-s))G(u(s))\,ds, \quad t \in I, \tag{43}$$

with $a \in C(I), f \in C(I)$ and $K_\alpha(t,s) := (t-s)^{\alpha-1}K(t,s)$ $(0 < \alpha \leq 1,\ K \in C(D))$. If $G(u) = u$ the analysis of its solvability differs significantly from one for the non-standard VIDE (1). It follows by a fixed-point argument similar to the one used in Zhang et al. [102] for auto-convolution VIEs that there exists a (small) $\delta_0 > 0$ (depending on $\bar{a} := \|a\|_\infty, \bar{f} := \|f\|_\infty$ and $\bar{K} := \|K\|_\infty$) so that (1) possesses a unique (local) solution $w_0 \in C^1[0,\delta_0]$. For $t \in [\delta_0, 2\delta_0]$ we may write (1) in the form

$$\begin{aligned} u'(t) &= a(t)u(t) + f(t) + \int_0^{\delta_0} K_\alpha(t,s)u(t-s)u(s)ds + \int_{\delta_0}^t K_\alpha(t,s)u(t-s)u(s)ds \\ &= a(t)u(t) + f(t) + \int_{t-\delta_0}^{\delta_0} K_\alpha(t,s)w_0(t-s)w_0(s)\,ds \\ &\quad + \int_{\delta_0}^t \big(K_\alpha(t,t-s) + K_\alpha(t,s)\big)u(t-s)u(s)\,ds. \end{aligned} \tag{44}$$

We see that in (3), $t - s \in [0, \delta_0]$. Since $u(t-s) = w_0(t-s)$ is known, (3) is *linear* in u. This process can be continued to the entire (bounded) interval $[\delta_0, T]$ because there exists an integer $\bar{N}$ so that $T \in [(M-1)\delta_0, M\delta_0]$.

The above observation implies that the results on the attainable orders of superconvergence of Theorems 1 and 2 are also valid for the auto-convolution VIDE (2) with $G(u) = u$. (A different, though rather sketchy, convergence analysis for implicit, collocation-based Runge-Kutta methods for (2) with $\alpha = 1$ and $G(u) = u$ was given in Yuan and Tang [101].)

For more general (nonlinear) functions G in (2), for example $G(u) = u^\beta$ with $\beta > 1$, the analysis of the optimal order of (global or local) superconvergence of collocation solutions $u_h \in S_m^{(0)}(I_h)$ remains open.

3.2 VIDEs with Delay Arguments

The generic form of a linear Volterra functional integro-differential equation (VFIDE) with (real-valued) delay function θ is

$$u'(t) = a(t)u(t)+b(t)u(\theta(t))+\int_{\theta(t)}^{t}(t-s)^{\alpha-1}K(t,s)u(s)\,ds,\ \ t\in I=[0,T], \tag{45}$$

where $0<\alpha\leq 1$, and $\theta(t):=t-\tau(t)$ is either a vanishing delay ($\tau(0)=0$, $0<\theta(t)<t$ if $t>0$) or a non-vanishing delay ($\tau(t)\geq\tau_0>0$, $t\in I$). Regularity results for the solutions of weakly singular VFIDES (4) with non-vanishing delays can be found in Brunner and Ma [19]. For (4) with $\alpha = 1$, optimal (super-)convergence results analogous to the ones in Theorems 1 and 2 were established in Brunner [15]. Shakourifar and Enright [86] studied continuous implicit Runge-Kutta methods for such VFIDEs; an alternative to collocation, using (explicit) continuous Volterra-Runge-Kutta methods together with C^1 Hermite interpolants at non-mesh points is described in Shakourifar and Enright [87] (compare also [86]). These methods are then used to solve Volterra's predator-prey system (1.10), (1.11). A very general theoretical framework for the analysis of Runge-Kutta methods for Volterra functional differential equations is due to Lin [52] and Li and Li [53].

If a VFIDE is of the form

$$\begin{aligned} u'(t) &= a(t)u(t)+b(t)u(\theta(t))+c(t)u'(\theta(t)) \\ &\quad +\int_{\theta(t)}^{t}(t-s)^{\alpha-1}\big(K(t,s)u(s)+K_1(t,s)u'(s)\big)ds, \end{aligned} \tag{46}$$

it is said to be of *neutral type*; it may be viewed as the nonlocal analogue of a neutral delay differential equation. The terminology 'neutral' VIDE or VFIDE is also used for equations like

$$\frac{d}{dt}\Big(u(t)-\int_{\theta(t)}^{t}(t-s)^{\alpha-1}K(t,s)u(s)ds\Big)=a(t)u(t)+f(t),\ \ t\in I, \tag{47}$$

with, respectively, $\theta(t)\equiv 0$ and $\theta(t)=t-\tau(t)$. We have encountered a closely related system of such neutral VFIDEs in Example 4. That system of VFIDEs is also closely related to a system of integral-algebraic equations (cf. following section). The numerical analysis and computational solution of VFIDEs (6) was studied by, e.g., Brunner and Vermiglio [22] ($\theta(t)\equiv 0$ and $\alpha=1$) and, for $0<\alpha<1$), by Ito and Turi [46] (using a semigroup framework) and by Brunner [18]. (The latter two papers also contain numerous additional references.)

3.3 Volterra Integro-Differential-Algebraic Equations

The system

$$A(t)u'(t) + B(t)u(t) = f(t) + (\mathcal{V}_\alpha u)(t),\;\; t \in I = [0,T], \tag{48}$$

with

$$(\mathcal{V}_\alpha u)(t) := \int_0^t (t-s)^{\alpha-1} K(t,s)u(s)\,ds\;\; (0 < \alpha \leq 1),$$

and $A(\cdot),\ B(\cdot),\ K(\cdot,\cdot) \in \mathbb{R}^{d\times d}$ $(d \geq 2)$ and $0 < \alpha \leq 1$, is called a system of Volterra integro-differential-algebraic equations (IDAEs). It may be viewed as a nonlocal extension of the system of differential-algebraic equations (DAEs)

$$A(t)u'(t) + B(t)u(t) = f(t),\;\; t \in I. \tag{49}$$

The numerical analysis of systems of DAEs is now well understood (see for example Lamour et al. [50] and its references), and this is to a somewhat lesser extent also true for systems of integral-algebraic equations (IAEs) (that is, (7) with $A(t) \equiv 0$; cf. Liang and Brunner [54, 55]). The extension of the optimal convergence results for collocation methods from IAEs (which used an adaptation of the projection techniques of [50]) to systems of IDAEs is currently being studied by Liang and Brunner [56]. Owing to the non-local character of IAEs and IDAEs, the analysis becomes much more complex than the one for DAEs because it not only requires an appropriate understanding of the (tractability) index of the IDAE system but has also to take into account the degree of ill-posedness of the inherent system of first-kind Volterra integral equations. However, the analysis of collocation methods for IAEs and IDAEs with weakly singular kernels remains open.

3.4 Time-Fractional Evolution Equations

An equation of the form

$$\left({}^{C}D_t^{\alpha} u\right)(t) = a(t)u(t) + f(t),\;\; t \in I, \tag{50}$$

is a basic example of a (time-)fractional VIDE. For $0 < \alpha < 1$,

$$\left({}^{C}D_t^{\alpha} u\right)(t) := \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{(t-s)^{1-\alpha-1}}{\Gamma(1-\alpha)} \frac{du(s)}{ds}\,ds$$

is the *Caputo* fractional derivative of order α of $u(t)$. It is related to the *Riemann-Liouville* fractional derivative,

$$\left({}^{RL}D_t^{\alpha}u\right)(t) := \frac{1}{\Gamma(1-\alpha)}\frac{d}{dt}\int_0^t (t-s)^{-\alpha}u(s)\,ds,$$

via

$$\left({}^{RL}D_t^{\alpha}u\right)(t) = \frac{t^{-\alpha}}{\Gamma(1-\alpha)}u(0) + \left({}^{C}D_t^{\alpha}u\right)(t).$$

Using the inverse (fractional time-integration) operator corresponding to ${}^{C}D_t^{\alpha}$ the fractional VIDE (9) can be written as an equivalent first-order VIDE or a VIE with weakly singular kernel (see for example Ma and Huang [63] where this is used as the basis for a numerical scheme). Although the numerical treatment of time-fractional VIDEs (and more general time-fraction evolution equations) has by now become a substantial 'industry', many issues are still waiting to be addressed. These include a detailed (analytical and numerical) comparison of computational schemes for fractional diffusion equations based on either the Caputo or the Riemann-Liouville fractional derivative (and the relationship between the respective schemes), as well as a thorough analysis of the merits of solving (9) directly, rather than its corresponding VIDE or VIE version.

Owing to limitation of space, and the sheer mass of recent papers on fractional diffusion equations, we will have to restrict this section to pointing the reader to a selection recent contributions relevant to the topics treated in the present paper. The 2010 monograph by Mainardi [64] contains, in addition to an introduction to fractional calculus, numerous applications of fractional diffusion-wave equations. The regularity of solutions to fractional diffusion is analyzed in McLean [69] (see also Clément and Londen [36] and its references). Various aspects (including a maximum principle) of discretizing such problems are treated in Mustapha and McLean [73], Brunner et al. [27], Mustapha and McLean [74], Ling and Yamamoto [57], Mustapha and Schötzau [75], Mustapha et al. [77], McLean and Mustapha [70], and Brunner et al. [28, 29]. Most of these papers contain extensive references.

4 Computational Challenges and Open Problems

4.1 Semilinear VIDEs with Blow-Up Solutions

For certain functions a, f, smooth or weakly singular kernels k, and (smooth) G the solution of the semilinear VIDE

$$u'(t) = a(t)u(t) + f(t) + \int_0^t k(t-s)G(u(s))\,ds, \quad t \geq 0 \tag{51}$$

(with $a(t) \leq 0$) may blow up in finite time. For VIDEs (1) whose solution behaves monotonically the blow-up analysis of nonlinear VIEs developed in Brunner and Yang [23] can be used to derive necessary and sufficient conditions for finite-time blow-up. (Sufficient conditions for very special case of (1) were derived in Ma [61].) However, the blow-up theory for general VIDEs (1) whose solutions do (typically) not exhibit a monotone behavior remains to be established.

The finite-time blow-up of solutions of semilinear parabolic VIDEs

$$u_t - \Delta u = f + \int_0^t k(t-s)G(u(s,\cdot))\,ds, \;\; x \in \Omega \subset \mathbb{R}^d, \;\; t \geq 0, \tag{52}$$

with typical nonlinearities $G(u) = (u+\lambda)^p$ $(p > 1, \lambda > 0)$ or $G(u) = \mathrm{e}^{\beta u}$ ($\beta > 0$ was studied by Bellout [10] (see also Souplet [90]), under the assumption that Ω is bounded and has a smooth boundary $\partial\Omega$. Blow-up results for different classes of semilinear parabolic VIDEs, including VIDEs of the form

$$u_t - \Delta u = \mu \int_0^t u^p(s,\cdot)\,ds - au^q \;\; (p,\, q \geq 1), \tag{53}$$

where μ is Hölder continuous, with $\mu \geq 0\,(\mu \not\equiv 0)$, and $a > 0$, and analogous partial VIDEs of Fredholm type, can be found in Souplet [89] and in Chapter V of Quittner and Souplet [83]. The blow-up of solutions for IDEs whose right-hand sides contain the composition of temporal and spatial integrals are also studied. The analysis is again based on the assumption that the spatial domain Ω possesses a smooth boundary. It appears that the blow-up theory for (2) and (3) with $d = 2$ and rectangular Ω remains to be established (in contrast to semilinear parabolic PDEs; cf. Bandle and Brunner [5] and its references).

The computational solution of parabolic VIDEs (2) on *unbounded* spatial domains Ω was studied in, e.g., Han et al. [44] and Brunner et al. [30] (see also for additional references). It is based on the choice of an appropriate bounded computational domain $\bar{\Omega}$ and the construction of corresponding artificial boundary conditions for $\bar{\Omega}$. (Compare also the monograph by Han and Wu [43] on the underlying theory of artificial boundary conditions for various classes of PDEs.)

On the other hand, the numerical analysis of parabolic VIDEs with finite-time blow-up, in particular the derivation of a posteriori error bounds for the numerical blow-up time, remains open.

4.2 Semilinear FIDEs with Blow-Up Solutions

As we have seen in Sect. 1.2, semilinear Fredholm integro-differential equations with nonlocal reaction term,

$$u_t - \Delta u = f + \int_\Omega H(u(\cdot, y))\,dy, \;\; t > 0, \;\; x \in \Omega \in \mathbb{R}^d, \tag{54}$$

where Ω is bounded with smooth boundary, occur in chemical reaction-diffusion processes. It was shown in Chadam et al. [35] and Chadam and Yin [34] that for typical nonlinearities like $H(u) = \mathrm{e}^u$ the solution of (4) may blow up in finite time. While the theory of such FIDEs is well understood, this is not true of the numerical analysis and the efficient computational solution of these problems. The key difference between the spatial semidiscretization of the parabolic VIDE (2) and the parabolic FIDE (4) is that the approximation of the spatial integral in (4) leads to a large, *dense* system of semilinear FIDEs. It would be of interest to see if a discretization scheme similar to the one described at the end of Section 2.4 [67] can be used in the efficient computational solution of (4).

Acknowledgements The paper is an extended and updated version of an invited plenary talk presented during the 2013 Biennial Conference on Numerical Analysis at the University of Strathclyde in Glasgow (Scotland). The research was supported by the Hong Kong Research Grants Council (GRF Grant No. HKBU 200210 and HKBU 12300014) and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant No. 9406).
I thank the two reviewers for their careful reading of the original manuscript and for their valuable comments.

References

1. Adolfsson, K., Enelund, M., Larsson, S.: Adaptive discretization on an integro-differential equation with a weakly singular kernel. Comput. Methods Appl. Mech. Eng. **192**, 5285–5304 (2003)
2. Appell, J.M., Kalitvin, A.S., Zabrejko, P.P.: Partial Integral Operators and Integro-Differential Equations. Marcel Dekker, New York (2000)
3. Audounet, J., Roquejoffre, J.M., Rouzaud, H.: Numerical simulation of a point-source initiated flame ball with heat loss. M2AN Math. Mod. Numer. Anal. **36**, 273–291 (2002)
4. Aves, M.A., Davies, P.J., Higham, D.J.: The effect of quadrature on the dynamics of a discretized nonlinear integro-differential equation. Appl. Numer. Math. **32**, 1–20 (2000)
5. Bandle, C., Brunner, H.: Blowup in diffusion equations: a survey. J. Comput. Appl. Math. **97**, 3–32 (1998)
6. Bebernes, J., Bressan, A.: Thermal behavior for a confined reactive gas. J. Differ. Equ. **44**, 118–133 (1982)
7. Bebernes, J., Eberly, D.: Mathematical Problems from Combustion Theory. Springer, New York (1989)
8. Bebernes, J., Kassoy, D.R.: A mathematical analysis of blow-up for thermal reactions – the spatially nonhomogeneous case. SIAM J. Appl. Math. **40**, 476–484 (1981)
9. Bélair, J., Mackey, M.C.: Consumer memory and price fluctuations in commodity markets: an integrodifferentiation model. J. Dynam. Differ. Equ. **1**, 299–325 (1989)
10. Bellout, H.: Blow-up of solutions of parabolic equations with nonlinear memory. J. Differ. Equ. **70**, 42–68 (1987)
11. Brunner, H.: Nonpolynomial spline collocation for Volterra equations with weakly singular kernels. SIAM J. Numer. Anal. **20**, 1106–1119 (1983)
12. Brunner, H.: Polynomial spline collocation methods for Volterra integro-differential equations with weakly singular kernels. IMA J. Numer. Anal. **6**, 221–239 (1986)
13. Brunner, H.: The approximate solution of initial-value problems for general Volterra integro-differential equations. Computing **40**, 125–137 (1988)

14. Brunner, H.: Collocation methods for nonlinear Volterra integro-differential equations with infinite delay. Math. Comput. **53**, 571–587 (1989)
15. Brunner, H.: The numerical solution of neutral Volterra integro-differential equations with delay arguments. Ann. Numer. Math. **1**, 309–322 (1994)
16. Brunner, H.: Collocation Methods for Volterra Integral and Related Functional Differential Equations. Cambridge University Press, Cambridge (2004)
17. Brunner, H.: The numerical analysis of functional integral and integro-differential equations of Volterra type. Acta Numer. **13**, 55–145 (2004)
18. Brunner, H.: The numerical solution of weakly singular Volterra functional integro-differential equations with variable delays. Commun. Pure Appl. Anal. **5**, 261–276 (2006)
19. Brunner, H., Ma, J.T.: On the regularity of solutions to Volterra functional integro-differential equations with weakly singular kernels. J. Integr. Equ. Appl. **18**, 143–167 (2006)
20. Brunner, H., Schötzau, D.: *hp*-discontinuous Galerkin time-stepping for Volterra integro-differential equations. SIAM J. Numer. Anal. **44**, 224–245 (2006)
21. Brunner, H., van der Houwen, P.J.: The Numerical Solution of Volterra Equations. CWI Monographs, vol. 3. North-Holland, Amsterdam (1986)
22. Brunner, H., Vermiglio, R.: Stability of solutions of delay functional integro-differential equations and their discretizations. Computing **71**, 229–245 (2003)
23. Brunner, H., Yang, Z.W.: Blow-up behavior of Hammerstein-type Volterra integral equations. J. Integr. Equ. Appl. **24**, 487–512 (2012)
24. Brunner, H., Kauthen, J.-P., Ostermann, A.: Runge-Kutta time discretization of parabolic integro-differential equations. J. Integr. Equ. Appl. **7**, 1–16 (1995)
25. Brunner, H., Pedas, A., Vainikko, G.: Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels. SIAM J. Numer. Anal. **39**, 957–982 (2001)
26. Brunner, H., van der Houwen, P.J., Sommeijer, B.P.: Splitting methods for partial Volterra integro-differential equations. In: Lu, Y., Sun, W., Tang, T. (eds.) Advances in Scientific Computing and Applications (Hong Kong 2003), pp. 68–81. Science Press, Beijing/New York (2004)
27. Brunner, H., Ling, L., Yamamoto, M.: Numerical simulation of 2D fractional subdiffusion problems. J. Comput. Phys. **229**, 6613–6622 (2010)
28. Brunner, H., Han, H.D., Yin, D.S: Artificial boundary conditions and finite difference approximations for a time-fractional diffusion-wave equation on a two-dimensional unbounded domain. J. Comput. Phys. **276**, 541–562 (2014)
29. Brunner, H., Han, H.D., Yin, D.S.: The maximum principle for time-fractional diffusion equations and its applications. Numer. Funct. Anal. Optim. **36**, 1307–1321 (2015)
30. Brunner, H., Tang, T., Zhang, J.W.: Numerical blow-up of nonlinear parabolic integro-differential equations on unbounded domain. J. Sci. Comput. **68**, 1281–1298 (2016)
31. Bulatov, M.V., Lima, P., Weinmüller, E.B.: Existence and uniqueness of solutions to weakly singular integral-algebraic and integro-differential equations. Cent. Eur. J. Math. **12**, 308–321 (2014)
32. Burns, J.A., Cliffs, E.M., Herdman, T.L.: A state-space model for an aeroelastic system. In: 22nd IEEE Conference on Decision and Control, vol. 3, pp. 1074–1077 (1983)
33. Cao, Y.Z., Herdman, T.L., Xu, Y.S.: A hybrid collocation method for Volterra integral equations with weakly singular kernels. SIAM J. Numer. Anal. **41**, 364–381 (2003)
34. Chadam, J.M., Yin, H.M.: A diffusion equation with localized chemical reactions. Proc. Edinb. Math Soc. (2) **37**, 101–118 (1994)
35. Chadam, J.M., Peirce, A., Yin, H.M.: The blowup property of solutions to some diffusion equations with localized nonlinear reactions. J. Math. Anal. Appl. **169**, 313–328 (1992)
36. Clément, P., Londen, S.-O.: Regularity aspects of fractional evolution equations. Rend. Ist. Mat. Univ. Trieste **31**, 19–30 (2000)
37. Cuesta, E., Lubich, Ch., Palencia, C.: Convolution quadrature time discretization of fractional diffusion-wave equations. Math. Comput. **75**, 673–696 (2006)

38. Cushing, J.M.: Integrodifferential Equations and Delay Models in Population Dynamics. Lecture Notes in Biomathematics, vol. 20. Springer, Berlin/Heidelberg (1977)
39. Doležal, V.: Dynamics of Linear Systems. Publishing House of the Czechoslovak Academy of Sciences, Prague (1964)
40. Fujita, Y.: Integrodifferential equation which interpolates the heat equation and the wave equation (I). Osaka J. Math. **27**, 309–321 (1990); (II) **27**, 797–804 (1990)
41. Ganesh, M., Sloan, I.H.: Optimal order spline methods for nonlinear differential and integro-differential equations. Appl. Numer. Math. **29**, 445–478 (1999)
42. Ganesh, M., Spence, A.: Orthogonal collocation for nonlinear integro-differential equations. IMA J. Numer. Anal. **18**, 191–206 (1998)
43. Han, H.D., Wu, X.N.: Artificial Boundary Method. Springer, Heidelberg/Tsinghua University Press, Beijing (2013)
44. Han, H.D., Zhu, L., Brunner, H., Ma, J.T.: Artificial boundary conditions for parabolic Volterra integro-differential equations on unbounded two-dimensional domains. J. Comput. Appl. Math. **197**, 406–420 (2006)
45. Hangelbroek, R.J., Kaper, H.G., Leaf, G.K.: Collocation methods for integro-differential equations. SIAM J. Numer. Anal. **14**, 377–390 (1977)
46. Ito, K., Turi, J.: Numerical methods for a class of singular integro-differential equations based on semigroup approximation. SIAM J. Numer. Anal. **28**, 1698–1722 (1991)
47. Janno, J., von Wolfersdorf, L.: Integro-differential equations of first order with autoconvolution integral. J. Integr. Equ. Appl. **21**, 39–75 (2009)
48. Jordan, G.S.: A nonlinear singularly perturbed Volterra integrodifferential equation of nonconvolution type. Proc. R. Soc. Edinb. Sect. A **80**, 235–277 (1978)
49. Kauthen, J.-P.: Implicit Runge-Kutta methods for singularly perturbed integro-differential systems. Appl. Numer. Math. **18**, 201–210 (1995)
50. Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Springer, Berlin/Heidelberg (2013)
51. Lasaint, P., Raviart, P.A.: On a finite element method for solving the neutron transport equation. In: de Boor, C. (ed.) Mathematical Aspects of Finite Elements in Partial Differential Equations, pp. 89–145. Academic, New York (1974)
52. Li, S.F.: High order contractive Runge-Kutta methods for Volterra functional differential equations. SIAM J. Numer. Anal. **47**, 4290–4325 (2010)
53. Li, S.F., Li, Y.F.: *B*-convergence theory of Runge-Kutta methods for stiff Volterra functional differential equations with infinite interval of integration. SIAM J. Numer. Anal. **53**, 2570–2583 (2015)
54. Liang, H., Brunner, H.: Integral-algebraic equations: theory of collocation methods I. SIAM J. Numer. Anal. **51**, 2238–2259 (2013)
55. Liang, H., Brunner, H.: Integral-algebraic equations: theory of collocation methods II. SIAM J. Numer. Anal. **54**, 2640–2663 (2016)
56. Liang, H., Brunner, H.: Collocation methods for integro-differential-algebraic equations with index 1. IMA J. Numer. Anal. (submitted)
57. Ling, L., Yamamoto, M.: Numerical simulations for space-time fractional diffusion equations. Int. J. Comput. Methods **10**, 13 pp. (2013)
58. Lodge, A.S., McLeod, J.B., Nohel, J.A.: A nonlinear singularly perturbed Volterra integrodifferential equation occurring in polymer rheology. Proc. R. Soc. Edinb. Sect. A **80**, 99–137 (1978)
59. López-Fernández, M., Lubich, Ch., Schädle, A.: Adaptive, fast, and oblivious convolution in evolution equations with memory. SIAM J. Sci. Comput. **30**, 1015–1037 (2008)
60. Lubich, Ch.: Runge-Kutta theory for Volterra integrodifferential equations. Numer. Math. **40**, 119–135 (1982)
61. Ma, J.T.: Blow-up solutions of nonlinear Volterra integro-differential equations. Math. Comput. Model. **54**, 2551–2559 (2011)
62. Ma, J.T., Brunner, H.: A posteriori error estimates of discontinuous Galerkin methods for non-standard Volterra integro-differential equations. IMA J. Numer. Anal. **26**, 78–95 (2006)

63. Ma, X.H., Huang, C.M.: Numerical solution of fractional integro-differential equations by a hybrid collocation method. Appl. Math. Comput. **219**, 6750–6760 (2013)
64. Mainardi, F.: Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models. Imperial College Press, London (2010)
65. Markowich, P., Renardy, M.: A nonlinear Volterra integro-differential equation describing the stretching of polymer liquids. SIAM J. Math. Anal. **14**, 66–97 (1983)
66. Markowich, P., Renardy, M.: The numerical solution of a class of quasilinear parabolic Volterra equations arising in polymer rheology. SIAM J. Numer. Anal. **20**, 890–908 (1983)
67. Matache, A.-M., Schwab, C., Wihler, T.P.: Fast numerical solution of parabolic integro-differential equations with applications in finance. SIAM J. Sci. Comput. **27**, 369–393 (2005)
68. Matache, A.-M., Schwab, C., Wihler, T.P.: Linear complexity of parabolic integro-differential equations. Numer. Math. **104**, 69–102 (2006)
69. McLean, W.: Regularity of solutions to a time-fractional diffusion equation. ANZIAM J. **52**, 123–138 (2010)
70. McLean, W., Mustapha, K.: Time-stepping error bounds for fractional diffusion problems with non-smooth initial data. J. Comput. Phys. **293**, 201–217 (2015)
71. McLean, W., Thomée, V.: Time discretization of an evolution equation via Laplace transforms. IMA J. Numer. Anal. **24**, 439–463 (2004)
72. Mustapha, K.: A superconvergent discontinuous Galerkin method for Volterra integro-differential equations. Math. Comput. **82**, 1987–2005 (2013)
73. Mustapha, K., McLean, W.: Discontinuous Galerkin method for an evolution equation with a memory term of positive type. Math. Comput. **78**, 1975–1995 (2009)
74. Mustapha, K., McLean, W.: Uniform convergence for a discontinuous Galerkin, time-stepping method applied to a fractional differential equation. IMA J. Numer. Anal. **32**, 906–925 (2012)
75. Mustapha, K., Schötzau, D.: Well-posedness of *hp*-version discontinuous Galerkin methods for fractional diffusion wave equations. IMA J. Numer. Anal. **34**, 1426–1446 (2014)
76. Mustapha, K., Brunner, H., Mustapha, H., Schötzau, D.: An *hp*-version discontinuous Galerkin method for integro-differential equations of parabolic type. SIAM J. Numer. Anal. **49**, 1369–1396 (2011)
77. Mustapha, K., Abdallah, B., Furati, K.M.: A discontinuous Petrov-Galerkin method for time-fractional diffusion equations. SIAM J. Numer. Anal. **52**, 2512–2529 (2014)
78. Nassirharand, A.: A new technique for solving sets of coupled nonlinear algebraic and integro-differential equations. Int. J. Contemp. Math. Sci. **3**, 1611–1617 (2008)
79. Parts, I., Pedas, A., Tamme, E.: Piecewise polynomial collocation for Fredholm integro-differential equations with weakly singular kernels. SIAM J. Numer. Anal. **43**, 1897–1911 (2005)
80. Pedas, A., Tamme, E.: Spline collocation method for integro-differential equations with weakly singular kernels. J. Comput. Appl. Math. **197**, 253–269 (2006)
81. Pedas, A., Tamme, E.: A discrete collocation method for Fredholm integro-differential equations with weakly singular kernels. Appl. Numer. Math. **61**, 738–751 (2011)
82. Prüss, J.: Evolutionary Integral Equations and Applications. Birkhäuser, Basel (1993)/Reprint (2012)
83. Quittner, P., Souplet, P.: Superlinear Parabolic Problems. Birkhäuser, Basel (2007)
84. Rouzaud, H.: Long-time dynamics of an integro-differential equation describing the evolution of a spherical flame. Rev. Mat. Complut. **16**, 207–232 (2003)
85. Schädle, A., López-Fernández, M., Lubich, Ch.: Fast and oblivious convolution quadrature. SIAM J. Sci. Comput. **28**, 421–438 (2006)
86. Shakourifar, M., Enright, W.H.: Reliable approximate solution of systems of Volterra integro-differential equations with time-dependent delays. SIAM J. Sci. Comput. **33**, 1134–1158 (2011)
87. Shakourifar, M., Enright, W.H.: Superconvergent interpolants for collocation methods applied to Volterra integro-differential equations with delay. BIT Numer. Math. **52**, 725–740 (2012)
88. Sloan, I.H., Thomée, V.: Time discretization of an integro-differential equation of parabolic type. SIAM J. Numer. Anal. **23**, 1052–1061 (1986)

89. Souplet, P.: Blow-up in nonlocal reaction-diffusion equations. SIAM J. Math. Anal. **29**, 1301–1334 (1998)
90. Souplet, P.: Monotonicity of solutions and blow-up for semilinear parabolic equations with nonlinear memory. Z. Angew. Math. Phys. **55**, 28–31 (2004)
91. Tang, T.: Superconvergence of numerical solutions to weakly singular Volterra integro-differential equations. Numer. Math. **61**, 373–382 (1992)
92. Volk, W.: The numerical solution of linear integro-differential equations by projection methods. J. Integr. Equ. **9**, 171–190 (1985)
93. Volk, W.: The iterated Galerkin method for linear integro-differential equations. J. Comput. Appl. Math. **21**, 63–74 (1988)
94. Volterra, V.: Sulle equazioni integro-differenziali. Rend. Accad. Lincei Ser. 5 **XVIII**, 167–174 (1909)
95. Volterra, V.: Sur les équations intégro-différentielles et leurs applications. Acta Math. **35**, 295–356 (1912)
96. Volterra, V.: Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem. R. Com. Talass. Ital. **CXXXI**, 142 pp. (1927)
97. Volterra, V.: Theory of Functionals and of Integral and Integro-Differential Equations (1927/1930). Dover, New York (1959)
98. Volterra, V.: Leçons sur la théorie mathématique de la lutte pour la vie [Lessons on the Mathematical Theory of the Struggle for Survival] (reprint of the original 1931 Gauthier-Villars edition). Éditions Jacques Gabay, Sceaux (1990)
99. von Wolfersdorf, L., Janno, J.: Integro-differential equations of first order with autoconvolution integral II. J. Integr. Equ. Appl. **23**, 331–349 (2011)
100. Yi, L.J., Guo, B.Q.: An $h-p$ version of the continuous Petrov-Galerkin finite element method for Volterra integro-differential equations with smooth and nonsmooth solutions. SIAM J. Numer. Anal. **53**, 2677–2704 (2015)
101. Yuan, W., Tang, T.: The numerical analysis of implicit Runge-Kutta methods for a certain integro-differential equation. Math. Comput. **54**, 155–168 (1990)
102. Zhang, R., Liang, H., Brunner, H.: Analysis of collocation methods for auto-convolution Volterra integral equations. SIAM J. Numer. Anal. **54**, 899–920 (2016)

Multivariate Approximation in Downward Closed Polynomial Spaces

Albert Cohen and Giovanni Migliorati

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The task of approximating a function of d variables from its evaluations at a given number of points is ubiquitous in numerical analysis and engineering applications. When d is large, this task is challenged by the so-called *curse of dimensionality*. As a typical example, standard polynomial spaces, such as those of total degree type, are often uneffective to reach a prescribed accuracy unless a prohibitive number of evaluations is invested. In recent years it has been shown that, for certain relevant applications, there are substantial advantages in using certain *sparse* polynomial spaces having anisotropic features with respect to the different variables. These applications include in particular the numerical approximation of high-dimensional parametric and stochastic partial differential equations. We start by surveying several results in this direction, with an emphasis on the numerical algorithms that are available for the construction of the approximation, in particular through interpolation or discrete least-squares fitting. All such algorithms rely on the assumption that the set of multi-indices associated with the polynomial space is *downward closed*. In the present paper we introduce some tools for the study of approximation in multivariate spaces under this assumption, and use them in the derivation of error bounds, sometimes independent of the dimension d, and in the development of adaptive strategies.

The original version of this chapter was revised. The correction to this chapter can be found at https://doi.org/10.1007/978-3-319-72456-0_59.

A. Cohen · G. Migliorati (✉)
Laboratoire Jacques-Louis Lions, Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7598, Paris, France
e-mail: cohen@ljll.math.upmc.fr; migliorati@ljll.math.upmc.fr

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_12

1 Introduction

The mathematical modeling of complex physical phenomena often demands for functions that depend on a large number of variables. One typical instance occurs when a quantity of interest u is given as the solution to an equation written in general form as

$$\mathscr{P}(u, y) = 0, \tag{1}$$

where $y = (y_j)_{j=1,\dots,d} \in \mathbb{R}^d$ is a vector that concatenates various physical parameters which have an influence on u.

Supposing that we are able to solve the above problem, either exactly or approximately by numerical methods, for any y in some domain of interest $U \subset \mathbb{R}^d$ we thus have access to the *parameter-to-solution* map

$$y \mapsto u(y). \tag{2}$$

The quantity $u(y)$ may be of various forms, namely:

1. a real number, that is, $u(y) \in \mathbb{R}$;
2. a function in some Banach space, for example when (1) is a partial differential equation (PDE);
3. a vector of eventually large dimension, in particular when (1) is a PDE whose solution is numerically approximated using some numerical method with fixed discretization parameters.

In all three cases, the above maps act from U to some finite- or infinite-dimensional Banach space which we shall generically denote by V.

As a guiding example which will be further discussed in this paper, consider the elliptic diffusion equation

$$-\operatorname{div}(a\nabla u) = f, \tag{3}$$

set on a given bounded Lipschitz domain $D \subset \mathbb{R}^k$ (say with $k = 1, 2$ or 3), for some fixed right-hand side $f \in L^2(D)$, homogeneous Dirichlet boundary conditions $u_{|\partial D} = 0$, and where a has the general form

$$a = a(y) = \overline{a} + \sum_{j\geq 1} y_j \psi_j. \tag{4}$$

Here, $\overline{a}$ and ψ_j are given functions in $L^\infty(D)$, and the y_j range in finite intervals that, up to renormalization, can all be assumed to be $[-1, 1]$. In this example $y = (y_j)_{j\geq 1}$ is countably infinite-dimensional, that is, $d = \infty$. The standard weak formulation of (3) in $H_0^1(D)$,

$$\int_D a\nabla u \nabla v = \int_D f v, \quad v \in H_0^1(D),$$

is ensured to be well-posed for all such a under the so-called *uniform ellipticity assumption*

$$\sum_{j\geq 1} |\psi_j| \leq \overline{a} - r, \quad \text{a.e. on } D, \tag{5}$$

for some $r > 0$. In this case, the map $y \mapsto u(y)$ acts from $U = [-1,1]^{\mathbb{N}}$ to $H_0^1(D)$. However, if we consider the discretization of (3) in some finite element space $V_h \subset H_0^1(D)$, where h refers to the corresponding mesh size, using for instance the Galerkin method, then the resulting map

$$y \mapsto u_h(y),$$

acts from $U = [-1,1]^{\mathbb{N}}$ to V_h. Likewise, if we consider a quantity of interest such as the flux $q(u) = \int_{\Sigma} a\nabla u \cdot \sigma$ over a given interface $\Sigma \subset D$ with σ being the outward pointing normal vector, then the resulting map

$$y \mapsto q(y) = q(u(y)),$$

acts from $U = [-1,1]^{\mathbb{N}}$ to $\mathbb{R}$. In all three cases, the above maps act from U to the finite- or infinite-dimensional Banach space V, which is either H_0^1, V_h or $\mathbb{R}$.

In the previous instances, the functional dependence between the input parameters y and the output $u(y)$ is described in clear mathematical terms by Eq. (1). In other practical instances, the output $u(y)$ can be the outcome of a complex physical experiment or numerical simulation with input parameter y. However the dependence on y might not be given in such clear mathematical terms.

In all the abovementioned cases, we assume that we are able to query the map (2) at any given parameter value $y \in U$, eventually up to some uncertainty. Such uncertainty may arise due to:

1. measurement errors, when $y \mapsto u(y)$ is obtained by a physical experiment, or
2. computational errors, when $y \mapsto u(y)$ is obtained by a numerical computation.

The second type of errors may result from the spatial discretization when solving a PDE with a given method, and from the round-off errors when solving the associated discrete systems.

One common way of modeling such errors is by assuming that we observe u at the selected points y up to a an additive noise η which may depend on y, that is, we evaluate

$$y \mapsto u(y) + \eta(y), \tag{6}$$

where η satisfies a uniform bound

$$\|\eta\|_{L^\infty(U,V)} := \sup_{y\in U} \|\eta(y)\|_V \leq \varepsilon, \tag{7}$$

for some $\varepsilon > 0$ representing the noise level.

Queries of the exact $u(y)$ or of the noisy $u(y) + \eta(y)$ are often expensive since they require numerically solving a PDE, or setting up a physical experiment, or running a time-consuming simulation algorithm. A natural objective is therefore to approximate the map (2) from some fixed number m of such queries at points $\{y^1, \ldots, y^m\} \in U$. Such approximations $y \mapsto \widetilde{u}(y)$ are sometimes called *surrogate or reduced models*.

Let us note that approximation of the map (2) is sometimes a preliminary task for solving other eventually more complicated problems, such as:

1. **Optimization and Control**, i.e. find a y which minimizes a certain criterion depending on $u(y)$. In many situations, the criterion takes the form of a convex functional of $u(y)$, and the minimization is subject to feasibility constraints. See e.g. the monographs [3, 30] and references therein for an overview of classical formulations and numerical methods for optimization problems.
2. **Inverse Problems**, i.e. find an estimate y from some data depending on the output $u(y)$. Typically, we face an ill-posed problem, where the parameter-to-solution map does not admit a global and stable inverse. Nonetheless, developing efficient numerical methods for approximating the parameter-to-solution map, i.e. solving the so-called direct problem, is a first step towards the construction of numerical methods for solving the more complex inverse problem, see e.g. [36].
3. **Uncertainty Quantification**, i.e. describe the stochastic properties of the solution $u(y)$ in the case where the parameter y is modeled by a random variable distributed according to a given probability density. We may for instance be interested in computing the expectation or variance of the V-valued random variable $u(y)$. Note that this task amounts in computing multivariate integrals over the domain U with respect to the given probability measure. This area also embraces, among others, optimization and inverse problems whenever affected by uncertainty in the data. We refer to e.g. [24] for the application of polynomial approximation to uncertainty quantification, and to [35] for the Bayesian approach to inverse problems.

There exist many approaches for approximating an unknown function of one or several variables from its evaluations at given points. One of the most classical approaches consists in picking the approximant in a given suitable n-dimensional space of elementary functions, such that

$$n \leq m. \tag{8}$$

Here, by "suitable" we mean that the space should have the ability to approximate the target function to some prescribed accuracy, taking for instance advantage of its smoothness properties. By "elementary" we mean that such functions should have simple explicit form which can be efficiently exploited in numerical computations. The simplest type of such functions are obviously polynomials in the variables y_j. As a classical example, we may decide to use, for some given $k \in \mathbb{N}_0$, the total degree polynomial space of order k, namely

$$\mathbb{P}_k := \operatorname{span}\left\{ y \mapsto y^\nu \; : \; |\nu| := \|\nu\|_1 = \sum_{j=1}^{d} \nu_j \le k \right\},$$

with the standard notation

$$y^\nu := \prod_{j=1}^{d} y_j^{\nu_j}, \quad \nu = (\nu_j)_{j=1,\dots,d}.$$

Note that since $u(y)$ is V-valued, this means that we actually use the V-valued polynomial space

$$\mathbb{V}_k := V \otimes \mathbb{P}_k = \left\{ y \mapsto \sum_{|\nu| \le k} w_\nu y^\nu \; : \; w_\nu \in V \right\}.$$

Another classical example is the polynomial space of degree k in each variable, namely

$$\mathbb{Q}_k := \operatorname{span}\left\{ y \mapsto y^\nu \; : \; \|\nu\|_\infty = \max_{j=1,\dots,d} \nu_j \le k \right\}.$$

A critical issue encountered by choosing such spaces is the fact that, for a fixed value of k, the dimension of $\mathbb{P}_k$ grows with d like d^k, and that of $\mathbb{Q}_k$ grows like k^d, that is, exponentially in d. Since capturing the fine structure of the map (2) typically requires a large polynomial degree k in some coordinates, we expect in view of (8) that the number of needed evaluations m becomes prohibitive as the number of variables becomes large. This state of affairs is a manifestation of the so-called *curse of dimensionality*. From an approximation theoretic or information-based complexity point of view, the curse of dimensionality is expressed by the fact that functions in standard smoothness classes such as $C^s(U)$ cannot be approximated in $L^\infty(U)$ with better rate then $n^{-s/d}$ by any method using n degrees of freedom or n evaluations, see e.g. [17, 31].

Therefore, in high dimension, one is enforced to give up on classical polynomial spaces of the above form, and instead consider more general spaces of the general form

$$\mathbb{P}_\Lambda := \text{span}\{y \mapsto y^\nu \ : \ \nu \in \Lambda\}, \tag{9}$$

where Λ is a subset of $\mathbb{N}_0^d$ with a given cardinality $n := \#(\Lambda)$. In the case of infinitely many variables $d = \infty$, we replace $\mathbb{N}_0^d$ by the set

$$\mathscr{F} := \ell^0(\mathbb{N}, \mathbb{N}_0) := \{\nu = (\nu_j)_{j \geq 1} \ : \ \#(\text{supp}(\nu)) < \infty\},$$

of finitely supported sequences of nonnegative integers. For V-valued functions, we thus use the space

$$\mathbb{V}_\Lambda = V \otimes \mathbb{P}_\Lambda := \left\{ y \mapsto \sum_{\nu \in \Lambda} w_\nu y^\nu \ : \ w_\nu \in V \right\}.$$

Note that $\mathbb{V}_\Lambda = \mathbb{P}_\Lambda$ in the particular case where $V = \mathbb{R}$.

The main objective when approximating the map (2) is to maintain a reasonable trade-off between accuracy measured in a given error norm and complexity measured by m or number of degrees of freedom measured by n, exploiting the different importance of each variable. Intuitively, large polynomial degrees should only be allocated to the most important variables. In this sense, if d is the dimension and k is the largest polynomial degree in any variable appearing in Λ, we view Λ as a very *sparse* subset of $\{0, \dots, k\}^d$.

As generally defined by (9), the space $\mathbb{P}_\Lambda$ does not satisfy some natural properties of usual polynomial spaces such as closure under differentiation in any variable, or invariance by a change of basis when replacing the monomials y^ν by other tensorized basis functions of the form

$$\phi_\nu(y) = \prod_{j \geq 1} \phi_{\nu_j}(y_j),$$

where the univariate functions $\{\phi_0, \dots, \phi_k\}$ form a basis of $\mathbb{P}_k$ for any $k \geq 0$, for example with the Legendre or Chebyshev polynomials. In order to fulfill these requirements, we ask that the set Λ has the following natural property.

Definition 1 A set $\Lambda \subset \mathbb{N}_0^d$ or $\Lambda \subset \mathscr{F}$ is *downward closed* if and only if

$$\nu \in \Lambda \text{ and } \widetilde{\nu} \leq \nu \implies \widetilde{\nu} \in \Lambda,$$

where $\widetilde{\nu} \leq \nu$ means that $\widetilde{\nu}_j \leq \nu_j$ for all j.

Downward closed sets are also called lower sets. We sometimes use the terminology of downward closed polynomial spaces for the corresponding $\mathbb{P}_\Lambda$. To our knowledge, such spaces have been first considered in [23] in the bivariate case $d = 2$ and referred to as *polynômes pleins*. Their study in general dimension d

has been pursued in [25] and [16]. The objective of the present paper is to give a survey of recent advances on the use of downward closed polynomial spaces for high-dimensional approximation.

The outline is the following. We review in Sect. 2 several polynomial approximation results obtained in [1, 2] in which the use of well-chosen index sets allows one to *break the curse of dimensionality* for relevant classes of functions defined on $U = [-1, 1]^d$, e.g. such as those occurring when solving the elliptic PDE (3) with parametric diffusion coefficient (4). Indeed, we obtain an algebraic convergence rate n^{-s}, where s is independent of d in the sense that such a rate may even hold when $d = \infty$. Here, we consider the error between the map (2) and its approximant in either norms $L^\infty(U, V) = L^\infty(U, V, d\mu)$ or $L^2(U, V) = L^2(U, V, d\mu)$, where $d\mu$ is the uniform probability measure,

$$d\mu := \bigotimes_{j\geq 1} \frac{dy_j}{2}.$$

We also consider the case of lognormal diffusion coefficients of the form

$$a = \exp(b), \quad b = b(y) = \sum_{j\geq 1} y_j \psi_j, \tag{10}$$

where the y_j are i.i.d. standard Gaussian random variables. In this case, we have $U = \mathbb{R}^d$ and the error is measured in $L^2(U, V, d\gamma)$ where

$$d\gamma := \bigotimes_{j\geq 1} g(y_j)dy_j, \quad g(t) := \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \tag{11}$$

is the tensorized Gaussian probability measure. The above approximation results are established by using n-term truncations of polynomial expansions, such as Taylor, Legendre or Hermite, which do not necessarily result in downward closed index sets. In the present paper we provide a general approach to establish similar convergence rates with downward closed polynomial spaces.

The coefficients in the polynomial expansions cannot be computed exactly from a finite number of point evaluations of (2). One first numerical procedure that builds a polynomial approximation from point evaluations is interpolation. In this case the number m of samples is exactly equal to the dimension n of the polynomial space. We discuss in Sect. 3 a general strategy to choose evaluation points and compute the interpolant in arbitrarily high dimension. One of its useful features is that the evaluations and interpolants are updated in a sequential manner as the polynomial space is enriched, exploiting in a crucial way the downward closed structure. We study the stability of this process and its ability to achieve the same convergence rates in L^∞ established in Sect. 2.

A second numerical procedure for building a polynomial approximation is the least-squares method, which applies to the overdetermined case $m > n$. To keep the presentation concise, we confine to results obtained in the analysis of this method only for the case of evaluations at random points. In Sect. 4 we discuss standard least squares, both in the noisy and noiseless cases, and in particular explain under which circumstances the method is stable and compares favorably with the best approximation error in L^2. Afterwards we discuss the more general method of weighted least squares, which allows one to optimize the relation between the dimension of the polynomial space n and the number of evaluation points m that warrants stability and optimal accuracy.

The success of interpolation and least squares is critically tied to the choice of proper downward closed sets $(\Lambda_n)_{n\geq 1}$ with $\#(\Lambda_n) = n$. Ideally we would like to choose the set Λ_n^* that minimizes the best approximation error

$$e(u, \Lambda) := \min_{v \in V_\Lambda} \|u - v\|, \tag{12}$$

in some norm $\| \cdot \|$ of interest, among all possible downward closed sets Λ of cardinality n. In addition to be generally nonunique, such a set Λ_n^* is often not accessible. In practice we need to rely on some a-priori analysis to select "suboptimal yet good" sets. An alternative strategy is to select the sequence $(\Lambda_n)_{n\geq 1}$ in an adaptive manner, that is, make use of the computation of the approximation for Λ_{n-1} in order to choose Λ_n.

We discuss in Sect. 5 several adaptive and nonadaptive strategies which make critical use of the downward closed structure of such sets. While our paper is presented in the framework of polynomial approximation, the concept of downward closed set may serve to define multivariate approximation procedures in other nonpolynomial frameworks. At the end of the paper we give some remarks on this possible extension, including, as a particular example, approximation by sparse piecewise polynomial spaces using hierarchical bases, such as sparse grid spaces.

Let us finally mention that another class of frequently used methods in high-dimensional approximation is based on Reproducing Kernel Hilbert Space (RKHS) or equivalently Gaussian process regression, also known as *kriging*. In such methods, for a given Mercer kernel $K(\cdot, \cdot)$ the approximant is typically searched by minimizing the associated RKHS norm among all functions agreeing with the data at the evaluation points, or equivalently by computing the expectation of a Gaussian process with covariance function K conditional to the observed data. Albeit natural competitors, these methods do not fall in the category discussed in the present paper, in the sense that the space where the approximation is picked varies with the evaluation points. It is not clear under which circumstances they may also break the curse of dimensionality.

2 Sparse Approximation in Downward Closed Polynomial Spaces

2.1 Truncated Polynomial Expansions

As outlined in the previous section, we are interested in deriving polynomial approximations of the map (2) acting from $U = [-1, 1]^d$ with $d \in \mathbb{N}$ or $d = \infty$ to the Banach space V. Our first vehicle to derive such approximations, together with precise error bounds for relevant classes of maps, consists in truncating certain polynomial expansions of (2) written in general form as

$$\sum_{\nu \in \mathscr{F}} u_\nu \phi_\nu, \tag{13}$$

where for each $\nu = (\nu_j)_{j \geq 1} \in \mathscr{F}$ the function $\phi_\nu : U \to \mathbb{R}$ has the tensor product form

$$\phi_\nu(y) = \prod_{j \geq 1} \phi_{\nu_j}(y_j),$$

and $u_\nu \in V$. Here we assume that $(\phi_k)_{k \geq 0}$ is a sequence of univariate polynomials such that $\phi_0 \equiv 1$ and the degree of ϕ_k is equal to k. This implies that $\{\phi_0, \dots, \phi_k\}$ is a basis of $\mathbb{P}_k$ and that the above product only involves a finite number of factors, even in the case where $d = \infty$. Thus, we obtain polynomial approximations of (2) by fixing some sets $\Lambda_n \subset \mathscr{F}$ with $\#(\Lambda_n) = n$ and defining

$$u_{\Lambda_n} := \sum_{\nu \in \Lambda_n} u_\nu \phi_\nu. \tag{14}$$

Before discussing specific examples, let us make some general remarks on the truncation of countable expansions with V-valued coefficients, not necessarily of tensor product or polynomial type.

Definition 2 The series (13) is said to converge *conditionally* with limit u in a given norm $\| \cdot \|$ if there exists an exhaustion $(\Lambda_n)_{n \geq 1}$ of $\mathscr{F}$ (which means that for any $\nu \in \mathscr{F}$ there exists n_0 such that $\nu \in \Lambda_n$ for all $n \geq n_0$), with the convergence property

$$\lim_{n \to \infty} \|u - u_{\Lambda_n}\| = 0. \tag{15}$$

The series (13) is said to converge *unconditionally* towards u in the same norm, if and only if (15) holds for every exhaustion $(\Lambda_n)_{n \geq 1}$ of $\mathscr{F}$.

As already mentioned in the introduction, we confine our attention to the error norms $L^\infty(U, V)$ or $L^2(U, V)$ with respect to the uniform probability measure $d\mu$.

We are interested in establishing unconditional convergence, as well as estimates of the error between u and its truncated expansion, for both norms.

In the case of the L^2 norm, unconditional convergence can be established when $(\phi_\nu)_{\nu\in\mathscr{F}}$ is an orthonormal basis of $L^2(U)$. In this case we know from standard Hilbert space theory that if (2) belongs to $L^2(U,V)$ then the inner products

$$u_\nu := \int_U u(y)\phi_\nu(y)\,d\mu, \quad \nu \in \mathscr{F},$$

are elements of V, and the series (13) converges unconditionally towards u in $L^2(U,V)$. In addition, the error is given by

$$\|u - u_{\Lambda_n}\|_{L^2(U,V)} = \left(\sum_{\nu\notin\Lambda_n} \|u_\nu\|_V^2\right)^{1/2}, \tag{16}$$

for any exhaustion $(\Lambda_n)_{n\geq 1}$. Let us observe that, since $d\mu$ is a probability measure, the $L^\infty(U,V)$ norm controls the $L^2(U,V)$ norm, and thus the above holds whenever the map u is uniformly bounded over U.

For the L^∞ norms, consider an expansion (13) where the functions $\phi_\nu : U \mapsto \mathbb{R}$ are normalized such that $\|\phi_\nu\|_{L^\infty(U)} = 1$, for all $\nu \in \mathscr{F}$. Then $(\|u_\nu\|_V)_{\nu\in\mathscr{F}} \in \ell^1(\mathscr{F})$, and it is easily checked that, whenever the expansion (13) converges conditionally to a function u in $L^\infty(U,V)$, it also converges unconditionally to u in $L^\infty(U,V)$. In addition, for any exhaustion $(\Lambda_n)_{n\geq 1}$, we have the error estimate

$$\|u - u_{\Lambda_n}\|_{L^\infty(U,V)} \leq \sum_{\nu\notin\Lambda_n} \|u_\nu\|_V. \tag{17}$$

The above estimate is simply obtained by triangle inequality, and therefore generally it is not as sharp as (16). One particular situation is when $(\phi_\nu)_{\nu\in\mathscr{F}}$ is an orthogonal basis of $L^2(U)$ normalized in L^∞. Then, if $u \in L^2(U,V)$ and if the

$$u_\nu := \frac{1}{\|\phi_\nu\|_{L^2(U,V)}^2} \int_U u(y)\phi_\nu(y)\,d\mu, \quad \nu \in \mathscr{F},$$

satisfy $(\|u_\nu\|_V)_{\nu\in\mathscr{F}} \in \ell^1(\mathscr{F})$, we find on the one hand that (13) converges unconditionally to a limit in $L^\infty(U,V)$ and in turn in $L^2(U,V)$. On the other hand, we know that it converges toward $u \in L^2(U,V)$. Therefore, its limit in $L^\infty(U,V)$ is also u.

A crucial issue is the choice of the sets Λ_n that we decide to use when defining the n-term truncation (14). Ideally, we would like to use the set Λ_n which minimizes the truncation error in some given norm $\|\cdot\|$ among all sets of cardinality n.

In the case of the L^2 error, if $(\phi_\nu)_{\nu\in\mathscr{F}}$ is an orthonormal basis of $L^2(U)$, the estimate (16) shows that the optimal Λ_n is the set of indices corresponding to the

n largest $\|u_\nu\|_V$. This set is not necessarily unique, in which case any realization of Λ_n is optimal.

In the case of the L^∞ error, there is generally no simple description of the optimal Λ_n. However, when the ϕ_ν are normalized in $L^\infty(U)$, the right-hand side in the estimate (17) provides an upper bound for the truncation error. This bound is minimized by again taking for Λ_n the set of indices corresponding to the n largest $\|u_\nu\|_V$, with the error now bounded by the ℓ^1 tail of the sequence $(\|u_\nu\|_V)_{\nu\in\mathscr{F}}$, in contrast to the ℓ^2 tail which appears in (16).

The properties of a given sequence $(c_\nu)_{\nu\in\mathscr{F}}$ which ensure a certain rate of decay n^{-s} of its ℓ^q tail after one retains its n largest entries are well understood. Here, we use the following result, see [12], originally invoked by Stechkin in the particular case $q = 2$. This result says that the rate of decay is governed by the ℓ^p summability of the sequence for values of p smaller than q.

Lemma 1 *Let $0 < p < q < \infty$ and let $(c_\nu)_{\nu\in\mathscr{F}} \in \ell^p(\mathscr{F})$ be a sequence of nonnegative numbers. Then, if Λ_n is a set of indices which corresponds to the n largest c_ν, one has*

$$\left(\sum_{\nu\notin\Lambda_n} c_\nu^q\right)^{1/q} \le C(n+1)^{-s}, \quad C := \|(c_\nu)_{\nu\in\mathscr{F}}\|_{\ell^p}, \quad s := \frac{1}{p} - \frac{1}{q}.$$

In view of (16) or (17), application of the above result shows that ℓ^p summability of the sequence $(\|u_\nu\|_V)_{\nu\in\mathscr{F}}$ implies a convergence rate n^{-s} when retaining the terms corresponding to the n largest $\|u_\nu\|_V$ in (13). From (16), when $(\phi_\nu)_{\nu\in\mathscr{F}}$ is an orthonormal basis, we obtain $s = \frac{1}{p} - \frac{1}{2}$ if $p < 2$. From (17), when the ϕ_ν are normalized in $L^\infty(U)$, we obtain $s = \frac{1}{p} - 1$ if $p < 1$.

In the present setting of polynomial approximation, we mainly consider four types of series corresponding to four different choices of the univariate functions ϕ_k:

- Taylor (or power) series of the form

$$\sum_{\nu\in\mathscr{F}} t_\nu y^\nu, \quad t_\nu := \frac{1}{\nu!}\partial^\nu u(y=0), \quad \nu! := \prod_{j\ge 1} \nu_j!, \tag{18}$$

  with the convention that $0! = 1$.
- Legendre series of the form

$$\sum_{\nu\in\mathscr{F}} w_\nu L_\nu(y), \quad L_\nu(y) = \prod_{j\ge 1} L_{\nu_j}(y_j), \quad w_\nu := \int_U u(y)L_\nu(y)\,d\mu, \tag{19}$$

  where $(L_k)_{k\ge 0}$ is the sequence of Legendre polynomials on $[-1,1]$ normalized with respect to the uniform measure $\int_{-1}^1 |L_k(t)|^2 \frac{dt}{2} = 1$, so that $(L_\nu)_{\nu\in\mathscr{F}}$ is an orthonormal basis of $L^2(U, d\mu)$.

- Renormalized Legendre series of the form

$$\sum_{\nu\in\mathcal{F}} \widetilde{w}_\nu \widetilde{L}_\nu(y), \quad \widetilde{L}_\nu(y) = \prod_{j\geq 1} \widetilde{L}_{\nu_j}(y_j), \quad \widetilde{w}_\nu := \left(\prod_{j\geq 1}(1+2\nu_j)\right)^{1/2} w_\nu, \tag{20}$$

  where $(\widetilde{L}_k)_{k\geq 0}$ is the sequence of Legendre polynomials on $[-1,1]$ with the standard normalization $\|\widetilde{L}_k\|_{L^\infty([-1,1])} = \widetilde{L}_k(1) = 1$, so that $\widetilde{L}_k = (1+2k)^{-1/2}L_k$.
- Hermite series of the form

$$\sum_{\nu\in\mathcal{F}} h_\nu H_\nu(y), \quad H_\nu(y) = \prod_{j\geq 1} H_{\nu_j}(y_j), \quad h_\nu := \int_U u(y)H_\nu(y)\,d\gamma, \tag{21}$$

  with $(H_k)_{k\geq 0}$ being the sequence of Hermite polynomials normalized according to $\int_{\mathbb{R}} |H_k(t)|^2 g(t)dt = 1$, and $d\gamma$ given by (11). In this case $U = \mathbb{R}^d$ and $(H_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of $L^2(U, d\gamma)$.

We may therefore estimate the L^2 error resulting from the truncation of the Legendre series (19) by application of (16), or the L^∞ error resulting from the truncation of the Taylor series (18) or renormalized Legendre series (20) by application of (17). According to Lemma 1, we derive convergence rates that depend on the value of p such that the coefficient sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|w_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|\widetilde{w}_\nu\|_V)_{\nu\in\mathcal{F}}$ or $(\|h_\nu\|_V)_{\nu\in\mathcal{F}}$ belong to $\ell^p(\mathcal{F})$.

In a series of recent papers such summability results have been obtained for various types of parametric PDEs. We refer in particular to [1, 13] for the elliptic PDE (3) with affine parameter dependence (4), to [2, 20] for the lognormal dependence (10), and to [9] for more general PDEs and parameter dependence. One specific feature is that these conditions can be fulfilled in the infinite-dimensional framework. We thus obtain convergence rates that are immune to the curse of dimensionality, in the sense that they hold with $d = \infty$. Here, we mainly discuss the results established in [1, 2] which have the specificity of taking into account the support properties of the functions ψ_j.

One problem with this approach is that the sets Λ_n associated to the n largest values in these sequences are generally not downward closed. In the next sections, we revisit these results in order to establish similar convergence rates for approximation in downward closed polynomial spaces.

2.2 Summability Results

The summability results in [1, 2] are based on certain weighted ℓ^2 estimates which can be established for the previously defined coefficient sequences under various relevant conditions for the elliptic PDE (3). We first report below these weighted estimates. The first one from [1] concerns the affine parametrization (4). Here, we have $V = H_0^1(D)$ and V' denotes its dual $H^{-1}(D)$.

Theorem 1 *Assume that $\rho = (\rho_j)_{j\geq 1}$ is a sequence of positive numbers such that*

$$\sum_{j\geq 1} \rho_j |\psi_j(x)| \leq \overline{a}(x) - \widetilde{r}, \quad x \in D, \tag{22}$$

for some fixed number $\widetilde{r} > 0$. Then, one has

$$\sum_{\nu\in\mathscr{F}} (\rho^\nu \|t_\nu\|_V)^2 < \infty, \quad \rho^\nu = \prod_{j\geq 1} \rho_j^{\nu_j}, \tag{23}$$

as well as

$$\sum_{\nu\in\mathscr{F}} \left(\beta(\nu)^{-1}\rho^\nu \|w_\nu\|_V\right)^2 = \sum_{\nu\in\mathscr{F}} \left(\beta(\nu)^{-2}\rho^\nu \|\widetilde{w}_\nu\|_V\right)^2 < \infty, \tag{24}$$

with

$$\beta(\nu) := \prod_{j\geq 1} (1 + 2\nu_j)^{1/2}.$$

The constants bounding these sums depend on $\widetilde{r}$, $\|f\|_{V'}$, $\overline{a}_{\min}$ and $\|\overline{a}\|_{L^\infty}$.

A few words are in order concerning the proof of these estimates. The first estimate (23) is established by first proving that the uniform ellipticity assumption (5) implies the ℓ^2 summability of the Taylor sequence $(\|t_\nu\|_V)_{\nu\in\mathscr{F}}$. Since the assumption (22) means that (5) holds with the ψ_j replaced by $\rho_j\psi_j$, this gives the ℓ^2 summability of the Taylor sequence for the renormalized map

$$y \mapsto u(\rho y), \quad \rho y = (\rho_j y_j)_{j\geq 1},$$

which is equivalent to (23). The second estimate is established by first showing that $\sum_{j\geq 1} \rho_j|\psi_j| \leq \overline{a} - \widetilde{r}$ implies finiteness of the weighted Sobolev-type norm

$$\sum_{\nu\in\mathscr{F}} \frac{\rho^{2\nu}}{\nu!} \int_U \|\partial^\nu u(y)\|_V^2 \prod_{j\geq 1} (1 - |y_j|)^{2\nu_j}\, d\mu < \infty.$$

Then, one uses the Rodrigues formula $L_k(t) = \left(\frac{d}{dt}\right)^k \left(\frac{\sqrt{2k+1}}{k!\,2^k}(t^2-1)^k\right)$ in each variable y_j to bound the weighted ℓ^2 sum in (24) by this norm.

Remark 1 As shown in [1], the above result remains valid for more general classes of orthogonal polynomials of Jacobi type, such as the Chebyshev polynomials which are associated with the univariate measure $\frac{dt}{2\pi\sqrt{1-t^2}}$.

The second weighted ℓ^2 estimate from [2] concerns the lognormal parametrization (10).

Theorem 2 *Let $r \geq 0$ be an integer. Assume that there exists a positive sequence $\rho = (\rho_j)_{j\geq 1}$ such that $\sum_{j\geq 1} \exp(-\rho_j^2) < \infty$ and such that*

$$\sum_{j\geq 1} \rho_j |\psi_j(x)| = K < C_r := \frac{\ln 2}{\sqrt{r}}, \quad x \in D. \tag{25}$$

Then, one has

$$\sum_{\nu\in\mathcal{F}} \xi_\nu \|h_\nu\|_V^2 < \infty, \tag{26}$$

where

$$\xi_\nu := \sum_{\|\widetilde{\nu}\|_{\ell^\infty}\leq r} \binom{\nu}{\widetilde{\nu}} \rho^{2\widetilde{\nu}} = \prod_{j\geq 1}\left(\sum_{l=0}^{r}\binom{\nu_j}{l}\rho_j^{2l}\right), \quad \binom{\nu}{\widetilde{\nu}} := \prod_{j\geq 1}\binom{\nu_j}{\widetilde{\nu}_j},$$

with the convention that $\binom{k}{l} = 0$ when $l > k$. The constant bounding this sum depends on $\|f\|_{V'}$, $\sum_{j\geq 1}\exp(-\rho_j^2)$ and on the difference $C_r - K$.

Similar to the weighted ℓ^2 estimate (24) for the Legendre coefficients, the proof of (26) follows by first establishing finiteness of a weighed Sobolev-type norm

$$\sum_{\|\nu\|_{\ell^\infty}\leq r} \frac{\rho^{2\nu}}{\nu!} \int_U \|\partial^\nu u(y)\|_V^2 \, d\gamma < \infty,$$

under the assumption (25) in the above theorem. Then one uses the Rodrigues formula $H_k(t) = \frac{(-1)^k}{\sqrt{k!}}\frac{g^{(k)}(t)}{g(t)}$, with g given by (11), in each variable y_j to bound the weighted ℓ^2 sum in (26) by this norm.

In summary, the various estimates expressed in the above theorems all take the form

$$\sum_{\nu\in\mathcal{F}} (\omega_\nu c_\nu)^2 < \infty,$$

where

$$c_\nu \in \{\|t_\nu\|_V, \|w_\nu\|_V, \|\widetilde{w}_\nu\|_V, \|h_\nu\|_V\},$$

or equivalently

$$\omega_\nu \in \{\rho^\nu, \rho^\nu\beta(\nu)^{-1}, \rho^\nu\beta(\nu)^{-2}, \xi_\nu^{1/2}\}.$$

Then, one natural strategy for establishing ℓ^p summability of the sequence $(c_\nu)_{\nu\in\mathcal{F}}$ is to invoke Hölder's inequality, which gives, for all $0 < p < 2$,

$$\left(\sum_{\nu\in\mathcal{F}} |c_\nu|^p\right)^{1/p} \le \left(\sum_{\nu\in\mathcal{F}} (\omega_\nu c_\nu)^2\right)^{1/2} \left(\sum_{\nu\in\mathcal{F}} |\kappa_\nu|^q\right)^{1/q} < \infty, \quad \frac{1}{q} := \frac{1}{p} - \frac{1}{2},$$

where the sequence $(\kappa_\nu)_{\nu\in\mathcal{F}}$ is defined by

$$\kappa_\nu := \omega_\nu^{-1}. \tag{27}$$

Therefore ℓ^p summability of $(c_\nu)_{\nu\in\mathcal{F}}$ follows from ℓ^q summability of $(\kappa_\nu)_{\nu\in\mathcal{F}}$ with $0 < q < \infty$ such that $\frac{1}{q} = \frac{1}{p} - \frac{1}{2}$. This ℓ^q summability can be related to that of the univariate sequence

$$b = (b_j)_{j\ge 1}, \quad b_j := \rho_j^{-1}.$$

Indeed, from the factorization

$$\sum_{\nu\in\mathcal{F}} b^{q\nu} = \prod_{j\ge 1}\sum_{n\ge 0} b_j^{nq},$$

one readily obtains the following elementary result, see [12] for more details.

Lemma 2 *For any* $0 < q < \infty$*, one has*

$$b \in \ell^q(\mathbb{N}) \quad \text{and} \quad \|b\|_{\ell^\infty} < 1 \iff (b^\nu)_{\nu\in\mathcal{F}} \in \ell^q(\mathcal{F}).$$

In the case $\omega_\nu = \rho^\nu$, i.e. $\kappa_\nu = b^\nu$, this shows that the ℓ^p summability of the Taylor coefficients $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ follows if the assumption (22) holds with $b = (\rho_j^{-1})_{j\ge 1} \in \ell^q(\mathbb{N})$ and $\rho_j > 1$ for all j. By a similar factorization, it is also easily checked that for any algebraic factor of the form $\alpha(\nu) := \prod_{j\ge 1}(1 + c_1\nu_j)^{c_2}$ with $c_1, c_2 \ge 0$, one has

$$b \in \ell^q(\mathbb{N}) \quad \text{and} \quad \|b\|_{\ell^\infty} < 1 \iff (\alpha(\nu) b^\nu)_{\nu\in\mathcal{F}} \in \ell^q(\mathcal{F}).$$

This allows us to reach a similar conclusion in the cases $\omega_\nu = \beta(\nu)^{-1}\rho^\nu$ or $\omega_\nu = \beta(\nu)^{-2}\rho^\nu$, which correspond to the Legendre coefficients $(\|w_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|\widetilde{w}_\nu\|_V)_{\nu\in\mathcal{F}}$, in view of (24).

Likewise, in the case where $\omega_\nu = \xi_\nu^{1/2}$, using the factorization

$$\sum_{\nu\in\mathcal{F}} \kappa_\nu^q = \prod_{j\ge 1}\sum_{n\ge 0}\left(\sum_{l=0}^{r}\binom{n}{l}\rho_j^{2l}\right)^{-q/2},$$

it is shown in [2] that the sum on the left converges if $b \in \ell^q$, provided that r was chosen large enough such that $q > \frac{2}{r}$. This shows that the ℓ^p summability of the Hermite coefficients $(\|h_\nu\|_V)_{\nu\in\mathscr{F}}$ follows if the assumption (25) holds with $b = (\rho_j^{-1})_{j\geq 1} \in \ell^q(\mathbb{N})$. Note that, since the sequence b can be renormalized, we may replace (25) by the condition

$$\sup_{x\in D}\sum_{j\geq 1}\rho_j|\psi_j(x)| < \infty, \tag{28}$$

without a specific bound.

2.3 Approximation by Downward Closed Polynomials

The above results, combined with Lemma 1, allow us to build polynomial approximations u_{Λ_n} with provable convergence rates n^{-s} in L^∞ or L^2 by n-term truncation of the various polynomial expansions. However, we would like to obtain such convergence rates with sets Λ_n that are in addition downward closed.

Notice that if a sequence $(\kappa_\nu)_{\nu\in\mathscr{F}}$ of nonnegative numbers is monotone nonincreasing, that is

$$\nu \leq \widetilde{\nu} \implies \kappa_{\widetilde{\nu}} \leq \kappa_\nu,$$

then the set Λ_n corresponding to the n largest values of κ_ν (up to a specific selection in case of equal values) is downward closed. More generally, there exists a sequence $(\Lambda_n)_{n\geq 1}$ of downward closed realizations of such sets which is nested, i.e. $\Lambda_1 \subset \Lambda_2 \ldots$, with $\Lambda_1 = 0_{\mathscr{F}} := (0, 0, \ldots)$.

Since the general sequences $(\kappa_\nu)_{\nu\in\mathscr{F}}$ that are defined through (27) may not always be monotone nonincreasing, we introduce the following notion: for any sequence $(\kappa_\nu)_{\nu\in\mathscr{F}}$ tending to 0, in the sense that $\#\{\nu \;:\; |\kappa_\nu| > \delta\} < \infty$ for all $\delta > 0$, we introduce its *monotone majorant* $(\widehat{\kappa}_\nu)_{\nu\in\mathscr{F}}$ defined by

$$\widehat{\kappa}_\nu := \max_{\widetilde{\nu}\geq\nu} |\kappa_{\widetilde{\nu}}|,$$

that is the smallest monotone nonincreasing sequence that dominates $(\kappa_\nu)_{\nu\in\mathscr{F}}$. In order to study best n-term approximations using downward closed sets, we adapt the ℓ^q spaces as follows.

Definition 3 For $0 < q < \infty$, we say that $(\kappa_\nu)_{\nu\in\mathscr{F}} \in \ell^\infty(\mathscr{F})$ belongs to $\ell^q_m(\mathscr{F})$ if and only if its monotone majorant $(\widehat{\kappa}_\nu)_{\nu\in\mathscr{F}}$ belongs to $\ell^q(\mathscr{F})$.

We are now in position to state a general theorem that gives a condition for approximation using downward closed sets in terms of weighted ℓ^2 summability.

Theorem 3 *Let* $(c_\nu)_{\nu\in\mathcal{F}}$ *and* $(\omega_\nu)_{\nu\in\mathcal{F}}$ *be positive sequences such that*

$$\sum_{\nu\in\mathcal{F}}(\omega_\nu c_\nu)^2<\infty,$$

and such that $(\kappa_\nu)_{\nu\in\mathcal{F}}\in\ell^q_m(\mathcal{F})$ *for some* $0<q<\infty$ *with* $\kappa_\nu=\omega_\nu^{-1}$*. Then, for any* $0<r\le 2$ *such that* $\frac{1}{q}>\frac{1}{r}-\frac{1}{2}$*, there exists a nested sequence* $(\Lambda_n)_{n\ge1}$ *of downward closed sets such that* $\#(\Lambda_n)=n$ *and*

$$\left(\sum_{\nu\notin\Lambda_n}c_\nu^r\right)^{1/r}\le Cn^{-s},\quad s:=\frac{1}{q}+\frac{1}{2}-\frac{1}{r}>0. \tag{29}$$

Proof With $(\widehat{\kappa}_\nu)_{\nu\in\mathcal{F}}$ being the monotone majorant of $(\kappa_\nu)_{\nu\in\mathcal{F}}$, we observe that

$$A^2:=\sum_{\nu\in\mathcal{F}}(\widehat{\kappa}_\nu^{-1}c_\nu)^2\le\sum_{\nu\in\mathcal{F}}(\kappa_\nu^{-1}c_\nu)^2=\sum_{\nu\in\mathcal{F}}(\omega_\nu c_\nu)^2<\infty.$$

We pick a nested sequence $(\Lambda_n)_{n\ge1}$ of downward closed sets, such that Λ_n consists of the indices corresponding to the n largest $\widehat{\kappa}_\nu$. Denoting by $(\widehat{\kappa}_n)_{n\ge1}$ the decreasing rearrangement of $(\widehat{\kappa}_\nu)_{\nu\in\mathcal{F}}$, we observe that

$$n\widehat{\kappa}_n^q\le\sum_{j=1}^n\widehat{\kappa}_j^q\le B^q,\quad B:=\|(\widehat{\kappa}_\nu)_{\nu\in\mathcal{F}}\|_{\ell^q}<\infty.$$

With p such that $\frac{1}{p}=\frac{1}{r}-\frac{1}{2}$, we find that

$$\begin{aligned}\left(\sum_{\nu\notin\Lambda_n}c_\nu^r\right)^{1/r}&\le\left(\sum_{\nu\notin\Lambda_n}(\widehat{\kappa}_\nu^{-1}c_\nu)^2\right)^{1/2}\left(\sum_{\nu\notin\Lambda_n}\widehat{\kappa}_\nu^p\right)^{1/p}\\&\le A\left(\widehat{\kappa}_{n+1}^{p-q}\sum_{\nu\notin\Lambda_n}\widehat{\kappa}_\nu^q\right)^{1/p}\\&\le AB(n+1)^{1/p-1/q},\end{aligned}$$

where we have used Hölder's inequality and the properties of $(\widehat{\kappa}_n)_{n\ge1}$. This gives (29) with $C:=AB$. □

We now would like to apply the above result with $c_\nu\in\{\|t_\nu\|_V,\|w_\nu\|_V,\|\widetilde{w}_\nu\|_V,\|h_\nu\|_V\}$, and the corresponding weight sequences $\omega_\nu\in\{\rho^\nu,\rho^\nu\beta(\nu)^{-1},\rho^\nu\beta(\nu)^{-2},\xi_\nu^{1/2}\}$, or equivalently $\kappa_\nu\in\{b^\nu,b^\nu\beta(\nu),b^\nu\beta(\nu)^2,\xi_\nu^{-1/2}\}$. In the case of the Taylor series, where $\kappa_\nu=b^\nu$, we readily see that if $b_j<1$ for all $j\ge1$, then the sequence

$(\kappa_\nu)_{\nu\in\mathscr{F}}$ is monotone nonincreasing, and therefore Lemma 2 shows that $b \in \ell^q$ implies $(\kappa_\nu)_{\nu\in\mathscr{F}} \in \ell^q_m(\mathscr{F})$. By application of Theorem 3 with the value $r = 1$, this leads to the following result.

Theorem 4 *If* (22) *holds with* $(\rho_j^{-1})_{j\geq 1} \in \ell^q(\mathbb{N})$ *for some* $0 < q < 2$ *and* $\rho_j > 1$ *for all j, then*

$$\|u - u_{\Lambda_n}\|_{L^\infty(U,V)} \leq Cn^{-s}, \quad s := \frac{1}{q} - \frac{1}{2},$$

where u_{Λ_n} *is the truncated Taylor series and* Λ_n *is any downward closed set corresponding to the n largest* κ_ν*.*

In the case of the Legendre series, the weight $\kappa_\nu = b^\nu\beta(\nu)$ is not monotone nonincreasing due to the presence of the algebraic factor $\beta(\nu)$. However, the following result holds.

Lemma 3 *For any* $0 < q < \infty$ *and for any algebraic factor of the form* $\alpha(\nu) := \prod_{j\geq 1}(1 + c_1\nu_j)^{c_2}$ *with* $c_1, c_2 \geq 0$*, one has*

$$b \in \ell^q(\mathbb{N}) \quad \text{and} \quad \|b\|_{\ell^\infty} < 1 \iff (\alpha(\nu)b^\nu)_{\nu\in\mathscr{F}} \in \ell^q_m(\mathscr{F}).$$

Proof The implication from right to left is a consequence of Lemma 2, and so we concentrate on the implication from left to right. For this it suffices to find a majorant $\widetilde{\kappa}_\nu$ of $\kappa_\nu := \alpha(\nu)b^\nu$ which is monotone nonincreasing and such that $(\widetilde{\kappa}_\nu)_{\nu\in\mathscr{F}} \in \ell^q(\mathscr{F})$. We notice that for any $\tau > 1$, there exists $C = C(\tau, c_1, c_2) \geq 1$ such that

$$(1 + c_1 n)^{c_2} \leq C\tau^n, \quad n \geq 0.$$

For some $J \geq 1$ and τ to be fixed further, we may thus write

$$\kappa_\nu \leq \widetilde{\kappa}_\nu := C^J \prod_{j=1}^{J}(\tau b_j)^{\nu_j} \prod_{j>J}(1 + c_1\nu_j)^{c_2} b_j^{\nu_j}.$$

Since $\|b\|_{\ell^\infty} < 1$ we can take $\tau > 1$ such that $\theta := \tau\|b\|_{\ell^\infty} < 1$. By factorization, we find that

$$\sum_{\nu\in\mathscr{F}} \widetilde{\kappa}_\nu^q = C^{Jq}\left(\prod_{j=1}^{J}\left(\sum_{n\geq 0}\theta^{qn}\right)\right)\left(\prod_{j>J}\left(\sum_{n\geq 0}(1 + c_1 n)^{qc_2} b_j^{nq}\right)\right).$$

The first product is bounded by $(1 - \theta^q)^{-J}$. Each factor in the second product is a converging series which is bounded by $1 + cb_j^q$ for some $c > 0$ that depends on c_1, c_2 and $\|b\|_{\ell^\infty}$. It follows that this second product converges. Therefore $(\widetilde{\kappa}_\nu)_{\nu\in\mathscr{F}}$ belongs to $\ell^q(\mathscr{F})$.

Finally, we show that $\widetilde{\kappa}_\nu$ is monotone nonincreasing provided that J is chosen large enough. It suffices to show that $\widetilde{\kappa}_{\nu+e_j} \le \widetilde{\kappa}_\nu$ for all $\nu \in \mathcal{F}$ and for all $j \ge 1$ where

$$e_j := (0, \dots, 0, 1, 0, \dots),$$

is the Kronecker sequence of index j. When $j \le J$ this is obvious since $\widetilde{\kappa}_{\nu+e_j} = \tau b_j \widetilde{\kappa}_\nu \le \theta \widetilde{\kappa}_\nu \le \widetilde{\kappa}_\nu$. When $j > J$, we have

$$\widetilde{\kappa}_{\nu+e_j} \widetilde{\kappa}_\nu^{-1} = b_j \left(\frac{1 + c_1(\nu_j + 1)}{1 + c_1 \nu_j} \right)^{c_2}.$$

Noticing that the sequence $a_n := \left(\frac{1+c_1(n+1)}{1+c_1 n}\right)^{c_2}$ converges toward 1 and is therefore bounded, and that b_j tends to 0 as $j \to \infty$, we find that for J sufficiently large, the right-hand side in the above equation is bounded by 1 for all ν and $j > J$.

□

From Lemma 3, by applying Theorem 3 with $r = 1$ or $r = 2$, we obtain the following result.

Theorem 5 *If (22) holds with $(\rho_j^{-1})_{j \ge 1} \in \ell^q(\mathbb{N})$ for some $0 < q < \infty$ and $\rho_j > 1$ for all j, then*

$$\|u - u_{\Lambda_n}\|_{L^2(U,V)} \le C n^{-s}, \quad s := \frac{1}{q},$$

where u_{Λ_n} is the truncated Legendre series and Λ_n is any downward closed set corresponding to the n largest $\widehat{\kappa}_\nu$ where $\kappa_\nu := b^\nu \beta(\nu)$. If $q < 2$, we also have

$$\|u - u_{\Lambda_n}\|_{L^\infty(U,V)} \le C n^{-s}, \quad s := \frac{1}{q} - \frac{1}{2},$$

with Λ_n any downward closed set corresponding to the n largest $\widehat{\kappa}_\nu$ where $\kappa_\nu := b^\nu \beta(\nu)^2$, with $b := \rho_j^{-1})_{j \ge 1}$.

Finally, in the case of the Hermite coefficients, which corresponds to the weight

$$\kappa_\nu := \prod_{j \ge 1} \left(\sum_{l=0}^{r} \binom{\nu_j}{l} b_j^{-2l} \right)^{-1/2}, \tag{30}$$

we can establish a similar summability result.

Lemma 4 *For any $0 < q < \infty$ and any integer $r \ge 1$ such that $q > \frac{2}{r}$, we have*

$$b \in \ell^q(\mathbb{N}) \implies (\kappa_\nu)_{\nu \in \mathcal{F}} \in \ell^q(\mathcal{F}),$$

where κ_ν is given by (30). *In addition, for any integer $r \geq 0$, the sequence $(\kappa_\nu)_{\nu\in\mathscr{F}}$ is monotone nonincreasing.*

Proof For any $\nu \in \mathscr{F}$ and any $k \geq 1$ we have

$$\begin{aligned}\kappa_{\nu+e_k} &= \left(\sum_{l=0}^{r}\binom{\nu_k+1}{l}b_k^{-2l}\right)^{-1/2}\prod_{\substack{j\geq 1\\ j\neq k}}\left(\sum_{l=0}^{r}\binom{\nu_j}{l}b_j^{-2l}\right)^{-1/2}\\ &\leq \left(\sum_{l=0}^{r}\binom{\nu_k}{l}b_k^{-2l}\right)^{-1/2}\prod_{\substack{j\geq 1\\ j\neq k}}\left(\sum_{l=0}^{r}\binom{\nu_j}{l}b_j^{-2l}\right)^{-1/2} = \kappa_\nu,\end{aligned}$$

and therefore the sequence $(\kappa_\nu)_{\nu\in\mathscr{F}}$ is monotone nonincreasing.

Now we check that $(\kappa_\nu)_{\nu\in\mathscr{F}} \in \ell^q(\mathscr{F})$, using the factorization

$$\sum_{\nu\in\mathscr{F}}\kappa_\nu^q = \prod_{j\geq 1}\sum_{n\geq 0}\left(\sum_{l=0}^{r}\binom{n}{l}b_j^{-2l}\right)^{-q/2} \leq \prod_{j\geq 1}\sum_{n\geq 0}\binom{n}{r\wedge n}^{-q/2} b_j^{q(r\wedge n)}. \tag{31}$$

where the inequality follows from the fact that the value $l = n \wedge r := \min\{n, r\}$ is contained in the sum.

The j-th factor F_j in the rightmost product in (31) may be written as

$$F_j = 1 + b_j^q + \cdots + b_j^{(r-1)q} + C_{r,q}b_j^{rq},$$

where

$$C_{r,q} := \sum_{n\geq r}\binom{n}{r}^{-q/2} = (r!)^{q/2}\sum_{n\geq 0}\big[(n+1)\cdots(n+r)\big]^{-q/2} < \infty, \tag{32}$$

since we have assumed that $q > 2/r$. This shows that each F_j is finite. If $b \in \ell^q(\mathbb{N})$, there exists an integer $J \geq 0$ such that $b_j < 1$ for all $j > J$. For such j, we can bound F_j by $1 + (C_{r,q} + r - 1)b_j^q$, which shows that the product converges. □

From this lemma, and by application of Theorem 3 with the value $r = 2$, we obtain the following result for the Hermite series.

Theorem 6 *If* (28) *holds with $(\rho_j^{-1})_{j\geq 1} \in \ell^q(\mathbb{N})$ for some $0 < q < \infty$, then*

$$\|u - u_{\Lambda_n}\|_{L^2(U,V)} \leq Cn^{-s}, \quad s := \frac{1}{q},$$

where u_{Λ_n} is the truncated Hermite series and Λ_n is a downward closed set corresponding to the n largest κ_ν given by (30).

In summary, we have established convergence rates for approximation by downward closed polynomial spaces of the solution map (2) associated to the elliptic PDE (3) with affine or lognormal parametrization. The conditions are stated in terms of the control on the L^∞ norm of $\sum_{j\geq 1}\rho_j|\psi_j|$, where the ρ_j have a certain growth measured by the ℓ^q summability of the sequence $b=(b_j)_{j\geq 1}=(\rho_j^{-1})_{j\geq 1}$. This is a way to quantify the decay of the size of the ψ_j, also taking their support properties into account, and in turn to quantify the anisotropic dependence of $u(y)$ on the various coordinates y_j. Other similar results have been obtained with different PDE models, see in particular [12]. In the above results, the polynomial approximants are constructed by truncation of infinite series. The remainder of the paper addresses the construction of downward closed polynomial approximants from evaluations of the solution map at m points $\{y^1,\dots,y^m\}\in U$, and discusses the accuracy of these approximants.

3 Interpolation

3.1 Sparse Interpolation by Downward Closed Polynomials

Interpolation is one of the most standard processes for constructing polynomial approximations based on pointwise evaluations. Given a downward closed set $\Lambda\subset\mathcal{F}$ of finite cardinality, and a set of points

$$\Gamma\subset U,\quad \#(\Gamma)=\#(\Lambda),$$

we would like to build an interpolation operator I_Λ, that is, $I_\Lambda u\in\mathbb{V}_\Lambda$ is uniquely characterized by

$$I_\Lambda u(y)=u(y),\quad y\in\Gamma,$$

for any V-valued function u defined on U.

In the univariate case, it is well known that such an operator exists if and only if Γ is a set of pairwise distinct points, and that additional conditions are needed in the multivariate case. Moreover, since the set Λ may come from a nested sequence $(\Lambda_n)_{n\geq 1}$ as discussed in Sect. 2, we are interested in having similar nestedness properties for the corresponding sequence $(\Gamma_n)_{n\geq 1}$, where

$$\#(\Gamma_n)=\#(\Lambda_n)=n.$$

Such a nestedness property allows us to recycle the n evaluations of u which have been used in the computation of $I_{\Lambda_n}u$, and use only one additional evaluation for the next computation of $I_{\Lambda_{n+1}}u$.

It turns out that such hierarchical interpolants can be constructed in a natural manner by making use of the downward closed structure of the index sets Λ_n. This construction is detailed in [7] but its main principles can be traced from [23]. In order to describe it, we assume that the parameter domain is of either form

$$U = [-1,1]^d \quad \text{or} \quad [-1,1]^{\mathbb{N}},$$

with the convention that $d = \infty$ in the second case. However, it is easily checked that the construction can be generalized in a straightforward manner to any domain with Cartesian product form

$$U = \mathop{\times}_{k\geq 1} J_k,$$

where the J_k are finite or infinite intervals.

We start from a sequence of pairwise distinct points

$$T = (t_k)_{k\geq 0} \subset [-1,1].$$

We denote by I_k the univariate interpolation operator on the space $\mathbb{V}_k := V \otimes \mathbb{P}_k$ associated with the k-section $\{t_0,\dots,t_k\}$ of this sequence, that is,

$$I_k u(t_i) = u(t_i), \quad i = 0,\dots,k,$$

for any V-valued function u defined on $[-1,1]$. We express I_k in the Newton form

$$I_k u = I_0 u + \sum_{l=1}^{k} \Delta_l u, \quad \Delta_l := I_l - I_{l-1}, \tag{33}$$

and set $I_{-1} = 0$ so that we can also write

$$I_k u = \sum_{l=0}^{k} \Delta_l u.$$

Obviously the difference operator Δ_k annihilates the elements of $\mathbb{V}_{k-1}$. In addition, since $\Delta_k u(t_j) = 0$ for $j = 0,\dots,k-1$, we have

$$\Delta_k u(t) = \alpha_k B_k(t),$$

where

$$B_k(t) := \prod_{l=0}^{k-1} \frac{t - t_l}{t_k - t_l}.$$

The coefficient $\alpha_k \in V$ can be computed inductively, since it is given by

$$\alpha_k = \alpha_k(u) := u(t_k) - I_{k-1}u(t_k),$$

that is, the interpolation error at t_k when using I_{k-1}. Setting

$$B_0(t) := 1,$$

we observe that the system $\{B_0, \dots, B_k\}$ is a basis for $\mathbb{P}_k$. It is sometimes called a *hierarchical basis*.

In the multivariate setting, we tensorize the grid T, by defining

$$y_\nu := (t_{\nu_j})_{j\geq 1} \in U, \quad \nu \in \mathscr{F}.$$

We first introduce the tensorized operator

$$I_\nu := \bigotimes_{j\geq 1} I_{\nu_j},$$

recalling that the application of a tensorized operator $\otimes_{j\geq 1}A_j$ to a multivariate function amounts in applying each univariate operator A_j by freezing all variables except the jth one, and then applying A_j to the unfrozen variable. It is readily seen that I_ν is the interpolation operator on the tensor product polynomial space

$$\mathbb{V}_\nu = V \otimes \mathbb{P}_\nu, \quad \mathbb{P}_\nu := \bigotimes_{j\geq 1} \mathbb{P}_{\nu_j},$$

associated to the grid of points

$$\Gamma_\nu = \bigtimes_{j\geq 1} \{t_0, \dots, t_{\nu_j}\}.$$

This polynomial space corresponds to the particular downward closed index set of rectangular shape

$$\Lambda = R_\nu := \{\widetilde{\nu} \,:\, \widetilde{\nu} \leq \nu\}.$$

Defining in a similar manner the tensorized difference operators

$$\Delta_\nu := \bigotimes_{j\geq 1} \Delta_{\nu_j},$$

we observe that

$$I_\nu = \bigotimes_{j\geq 1} I_{\nu_j} = \bigotimes_{j\geq 1} (\sum_{l=0}^{\nu_j} \Delta_l) = \sum_{\widetilde{\nu}\in R_\nu} \Delta_{\widetilde{\nu}}.$$

The following result from [7] shows that the above formula can be generalized to *any* downward closed set in order to define an interpolation operator. We recall its proof for sake of completeness.

Theorem 7 *Let $\Lambda \subset \mathscr{F}$ be a finite downward closed set, and define the grid*

$$\Gamma_\Lambda := \{y_\nu \ : \ \nu \in \Lambda\}.$$

Then, the interpolation operator onto $\mathbb{V}_\Lambda$ for this grid is defined by

$$I_\Lambda := \sum_{\nu\in\Lambda} \Delta_\nu. \tag{34}$$

Proof From the downward closed set property, $\mathbb{V}_\nu \subset \mathbb{V}_\Lambda$ for all $\nu \in \Lambda$. Hence the image of I_Λ is contained in $\mathbb{V}_\Lambda$. With I_Λ defined by (34), we may write

$$I_\Lambda u = I_\nu u + \sum_{\widetilde{\nu}\in\Lambda, \widetilde{\nu}\not\leq\nu} \Delta_{\widetilde{\nu}} u,$$

for any $\nu \in \Lambda$. Since $y_\nu \in \Gamma_\nu$, we know that

$$I_\nu u(y_\nu) = u(y_\nu).$$

On the other hand, if $\widetilde{\nu} \not\leq \nu$, this means that there exists a $j \geq 1$ such that $\widetilde{\nu}_j > \nu_j$. For this j we thus have $\Delta_{\widetilde{\nu}} u(y) = 0$ for all $y \in U$ with the jth coordinate equal to t_{ν_j} by application of Δ_{ν_j} in the jth variable, so that

$$\Delta_{\widetilde{\nu}} u(y_\nu) = 0.$$

The interpolation property $I_\Lambda u(y_\nu) = u(y_\nu)$ thus holds, for all $\nu \in \Lambda$. □

The decomposition (34) should be viewed as a generalization of the Newton form (33). In a similar way, its terms can be computed inductively: if $\Lambda = \widetilde{\Lambda} \cup \{\nu\}$ where $\widetilde{\Lambda}$ is a downward closed set, we have

$$\Delta_\nu u = \alpha_\nu B_\nu,$$

where

$$B_\nu(y) := \prod_{j\geq 1} B_{\nu_j}(y_j),$$

and

$$\alpha_\nu = \alpha_\nu(u) := u(y_\nu) - I_{\widetilde{\Lambda}} u(y_\nu).$$

Therefore, if $(\Lambda_n)_{n\geq 1}$ is any nested sequence of downward closed index sets, we can compute I_{Λ_n} by n iterations of

$$I_{\Lambda_i} u = I_{\Lambda_{i-1}} u + \alpha_{\nu^i} B_{\nu^i},$$

where $\nu^i \in \Lambda_i$ is such that $\Lambda_i = \Lambda_{i-1} \cup \{\nu^i\}$.

Note that $(B_\nu)_{\nu\in\Lambda}$ is a basis of $\mathbb{P}_\Lambda$ and that any $f \in \mathbb{V}_\Lambda$ has the unique decomposition

$$f = \sum_{\nu\in\Lambda} \alpha_\nu B_\nu,$$

where the coefficients $\alpha_\nu = \alpha_\nu(f) \in V$ are defined by the above procedure. Also note that $\alpha_\nu(f)$ does not depend on the choice of Λ but only on ν and f.

3.2 Stability and Error Estimates

The pointwise evaluations of the function u could be affected by errors, as modeled by (6) and (7). The stability of the interpolation operator with respect to such perturbations is quantified by the *Lebesgue constant*, which is defined by

$$\mathbb{L}_\Lambda := \sup \frac{\|I_\Lambda f\|_{L^\infty(U,V)}}{\|f\|_{L^\infty(U,V)}},$$

where the supremum is taken over the set of all V-valued functions f defined everywhere and uniformly bounded over U. It is easily seen that this supremum is in fact independent of the space V, so that we may also write

$$\mathbb{L}_\Lambda := \sup \frac{\|I_\Lambda f\|_{L^\infty(U)}}{\|f\|_{L^\infty(U)}},$$

where the supremum is now taken over real-valued functions. Obviously, we have

$$\|u - I_\Lambda(u+\eta)\|_{L^\infty(U,V)} \leq \|u - I_\Lambda u\|_{L^\infty(U,V)} + \mathbb{L}_\Lambda \varepsilon,$$

where ε is the noise level from (7).

The Lebesgue constant also allows us to estimate the error of interpolation $\|u - I_\Lambda u\|_{L^\infty(U,V)}$ for the noiseless solution map in terms of the best approximation error in the L^∞ norm: for any $u \in L^\infty(U,V)$ and any $\widetilde{u} \in \mathbb{V}_\Lambda$ we have

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq \|u - \widetilde{u}\|_{L^\infty(U,V)} + \|I_\Lambda \widetilde{u} - I_\Lambda u\|_{L^\infty(U,V)},$$

which by infimizing over $\widetilde{u} \in \mathbb{V}_\Lambda$ yields

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq (1 + \mathbb{L}_\Lambda) \inf_{\widetilde{u} \in \mathbb{V}_\Lambda} \|u - \widetilde{u}\|_{L^\infty(U,V)}.$$

We have seen in Sect. 2 that for relevant classes of solution maps $y \mapsto u(y)$, there exist sequences of downward closed sets $(\Lambda_n)_{n\geq 1}$ with $\#(\Lambda_n) = n$, such that

$$\inf_{\widetilde{u} \in \mathbb{V}_{\Lambda_n}} \|u - \widetilde{u}\|_{L^\infty(U,V)} \leq C n^{-s}, \quad n \geq 1,$$

for some $s > 0$. For such sets, we thus have

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq C(1 + \mathbb{L}_{\Lambda_n}) n^{-s}. \tag{35}$$

This motivates the study of the growth of $\mathbb{L}_{\Lambda_n}$ as $n \to \infty$.

For this purpose, we introduce the univariate Lebesgue constants

$$\mathbb{L}_k := \sup \frac{\|I_k f\|_{L^\infty([-1,1])}}{\|f\|_{L^\infty([-1,1])}}.$$

Note that $\mathbb{L}_0 = 1$. We also define an analog quantity for the difference operator

$$\mathbb{D}_k := \sup \frac{\|\Delta_k f\|_{L^\infty([-1,1])}}{\|f\|_{L^\infty([-1,1])}}.$$

In the particular case of the rectangular downward closed sets $\Lambda = R_\nu$, since $I_\Lambda = I_\nu = \otimes_{j\geq 1} I_{\nu_j}$, we have

$$\mathbb{L}_{R_\nu} = \prod_{j\geq 1} \mathbb{L}_{\nu_j}.$$

Therefore, if the sequence $T = (t_k)_{k\geq 0}$ is such that

$$\mathbb{L}_k \leq (1 + k)^\theta, \quad k \geq 0, \tag{36}$$

for some $\theta \geq 1$, we find that

$$\mathbb{L}_{R_\nu} \leq \prod_{j\geq 1}(1+\nu_j)^\theta = (\#(R_\nu))^\theta,$$

for all $\nu \in \mathscr{F}$.

For arbitrary downward closed sets Λ, the expression of I_Λ shows that

$$\mathbb{L}_\Lambda \leq \sum_{\nu\in\Lambda}\prod_{j\geq 1}\mathbb{D}_{\nu_j}.$$

Therefore, if the sequence $T = (t_k)_{k\geq 0}$ is such that

$$\mathbb{D}_k \leq (1+k)^\theta, \quad k \geq 0, \tag{37}$$

we find that

$$\mathbb{L}_\Lambda \leq \sum_{\nu\in\Lambda}\prod_{j\geq 1}(1+\nu_j)^\theta = \sum_{\nu\in\Lambda}(\#(R_\nu))^\theta \leq \sum_{\nu\in\Lambda}(\#(\Lambda))^\theta = (\#(\Lambda))^{\theta+1}.$$

The following result from [7] shows that this general estimate is also valid under the assumption (36) on the growth of $\mathbb{L}_k$.

Theorem 8 *If the sequence $T = (t_k)_{k\geq 0}$ is such that* (36) *or* (37) *holds for some $\theta \geq 1$, then*

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^{\theta+1},$$

for all downward closed sets Λ.

One noticeable feature of the above result is that the bound on $\mathbb{L}_\Lambda$ only depends on $\#(\Lambda)$, independently of the number of variables, which can be infinite, as well as of the shape of Λ.

We are therefore interested in univariate sequences $T = (t_k)_{k\geq 0}$ such that $\mathbb{L}_k$ and $\mathbb{D}_k$ have moderate growth with k. For Chebyshev or Gauss-Lobatto points, given by

$$\mathscr{C}_k := \left\{\cos\left(\frac{2l+1}{2k+2}\pi\right) \;:\; l=0,\dots,k\right\} \text{ and } \mathscr{G}_k := \left\{\cos\left(\frac{l}{k}\pi\right) \;:\; l=0,\dots,k\right\},$$

it is well known that the Lebesgue constant has logarithmic growth $\mathbb{L}_k \sim \ln(k)$, thus slower than algebraic. However these points are not the k section of a single sequence T, and therefore they are not convenient for our purposes. Two examples of univariate sequences of interest are the following.

- The *Leja points*: from an arbitrary $t_0 \in [-1, 1]$ (usually taken to be 1 or 0), this sequence is recursively defined by

$$t_k := \operatorname{argmax} \left\{ \prod_{l=0}^{k-1} |t - t_l| \ : \ t \in [-1, 1] \right\} .$$

Note that this choice yields hierarchical basis functions B_k that are uniformly bounded by 1. Numerical computations of $\mathbb{L}_k$ for the first 200 values of k indicates that the linear bound

$$\mathbb{L}_k \leq 1 + k, \tag{38}$$

holds. Proving that this bound, or any other algebraic growth bound, holds for all values of $k \geq 0$ is currently an open problem.
- The $\Re$-*Leja* points: they are the real part of the Leja points defined on the complex unit disc $\{|z| \leq 1\}$, taking for example $e_0 = 1$ and recursively setting

$$e_k := \operatorname{argmax} \left\{ \prod_{l=0}^{k-1} |e - e_l| \ : \ |e| \leq 1 \right\} .$$

These points have the property of accumulating in a regular manner on the unit circle according to the so-called Van der Corput enumeration [4]. It is proven in [5] that the linear bound (38) holds for the Lebesgue constant of the complex interpolation operator on the unit disc associated to these points. The sequence of real parts

$$t_k := \Re(e_k),$$

is defined after eliminating the possible repetitions corresponding to $e_k = \overline{e}_l$ for two different values of $k = l$. These points coincide with the Gauss-Lobatto points for values of k of the form $2^n + 1$ for $n \geq 0$. A quadratic bound

$$\mathbb{D}_k \leq (1 + k)^2,$$

is established in [6].

If we use such sequences, application of Theorem 8 gives bounds of the form

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^{1+\theta},$$

for example with $\theta = 2$ when using the $\Re$-Leja points, or $\theta = 1$ when using the Leja points provided that the conjectured bound (38) holds. Combining with (35), we obtain the convergence estimate

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq C n^{-(s-1-\theta)},$$

which reveals a serious deterioration of the convergence rate when using interpolation instead of truncated expansions.

However, for the parametric PDE models discussed in Sect. 2, it is possible to show that this deterioration actually does not occur, based on the following lemma which relates the interpolation error to the summability of coefficient sequences in general expansions of u.

Lemma 5 *Assume that u admits an expansion of the type* (13)*, where $\|\phi_\nu\|_{L^\infty(U)} \leq 1$ which is unconditionally convergent towards u in $L^\infty(U,V)$. Assume in addition that $y \mapsto u(y)$ is continuous from U equipped with the product topology toward V. If the univariate sequence $T = (t_k)_{k\geq 0}$ is such that that* (36) *or* (37) *holds for some $\theta \geq 1$, then, for any downward closed set Λ,*

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq 2 \sum_{\nu\notin\Lambda} \pi(\nu)\|u_\nu\|_V, \quad \pi(\nu) := \prod_{j\geq 1}(1+\nu_j)^{\theta+1}. \tag{39}$$

Proof The unconditional convergence of (13) and the continuity of u with respect to the product topology allow us to say that the equality in (13) holds everywhere in U. We may thus write

$$I_\Lambda u = I_\Lambda\left(\sum_{\nu\in\mathcal{F}} u_\nu \phi_\nu\right) = \sum_{\nu\in\mathcal{F}} u_\nu I_\Lambda \phi_\nu = \sum_{\nu\in\Lambda} u_\nu\phi_\nu + \sum_{\nu\notin\Lambda} u_\nu I_\Lambda\phi_\nu,$$

where we have used that $I_\Lambda\phi_\nu = \phi_\nu$ for every $\nu \in \Lambda$ since $\phi_\nu \in \mathbb{P}_\Lambda$. For the second sum on the right-hand side, we observe that for each $\nu \notin \Lambda$,

$$I_\Lambda\phi_\nu = \sum_{\widetilde{\nu}\in\Lambda} \Delta_{\widetilde{\nu}}\phi_\nu = \sum_{\widetilde{\nu}\in\Lambda\cap R_\nu} \Delta_{\widetilde{\nu}}\phi_\nu = I_{\Lambda\cap R_\nu}\phi_\nu,$$

since $\Delta_{\widetilde{\nu}}$ annihilates $\mathbb{P}_\nu$ whenever $\widetilde{\nu} \not\leq \nu$. Therefore

$$u - I_\Lambda u = \sum_{\nu\notin\Lambda} u_\nu (I - I_{\Lambda\cap R_\nu})\phi_\nu,$$

where I stands for the identity operator. This implies

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq \sum_{\nu\notin\Lambda}(1 + \mathbb{L}_{\Lambda\cap R_\nu})\|u_\nu\|_V \leq 2\sum_{\nu\notin\Lambda} \mathbb{L}_{\Lambda\cap R_\nu}\|u_\nu\|_V \ .$$

Since (36) or (37) holds, we obtain from Theorem 8 that

$$\mathbb{L}_{\Lambda\cap R_\nu} \leq (\#(\Lambda \cap R_\nu))^{\theta+1} \leq (\#(R_\nu))^{\theta+1} = \pi(\nu),$$

which yields (39). □

We can apply the above lemma with the Taylor series (18) or the renormalized Legendre series (20). This leads us to analyze the ℓ^1 tail of the sequence $(c_\nu)_{\nu\in\mathcal{F}}$ where c_ν is either $\pi(\nu)\|t_\nu\|_V$ or $\pi(\nu)\|\widetilde{w}_\nu\|_V$. If (22) holds, we know from Theorem 1 that this sequence satisfies the bound

$$\sum_{\nu\in\mathcal{F}} (\omega_\nu c_\nu)^2 < \infty,$$

where ω_ν is either $\pi(\nu)^{-1}\rho^\nu$ or $\pi(\nu)^{-1}\beta(\nu)^{-2}\rho^\nu$. Since $\pi(\nu)$ has algebraic growth similar to $\beta(\nu)$, application of Lemma 3 and of Theorem 3 with the value $r = 1$, leads to the following result.

Theorem 9 *If (22) holds with $(\rho_j^{-1})_{j\geq 1} \in \ell^q(\mathbb{N})$ for some $0 < q < 2$ and $\rho_j > 1$ for all j, then*

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq C n^{-s}, \quad s := \frac{1}{q} - \frac{1}{2},$$

where Λ_n is any downward closed set corresponding to the n largest $\widehat{\kappa}_\nu$ where κ_ν is either $\pi(\nu)b^\nu$ or $\pi(\nu)\beta(\nu)^2 b^\nu$, where $b := (\rho_j^{-1})_{j\geq 1}$.

4 Discrete Least Squares Approximations

4.1 Discrete Least Squares on V-Valued Linear Spaces

Least-squares fitting is an alternative approach to interpolation for building a polynomial approximation of u from $\mathbb{V}_\Lambda$. In this approach we are given m observations $u^1, \dots, u^m$ of u at points $y^1, \dots, y^m \in U \subseteq \mathbb{R}^d$ where $m \geq n = \#(\Lambda)$.

We first discuss the least-squares method in the more general setting of V-valued linear spaces,

$$\mathbb{V}_n := V \otimes \mathbb{Y}_n,$$

where $\mathbb{Y}_n$ is the space of real-valued functions defined everywhere on U such that $\dim(\mathbb{Y}_n) = n$. In the next section, we discuss more specifically the case where $\mathbb{Y}_n = \mathbb{P}_\Lambda$. Here we study the approximation error in the $L^2(U, V, d\mu)$ norm for some given probability measure $d\mu$, when the evaluation points y^i are independent and drawn according to this probability measure. For notational simplicity we use the shorthand

$$\|\cdot\| := \|\cdot\|_{L^2(U,V,d\mu)}.$$

The *least-squares* method selects the approximant of u in the space $\mathbb{V}_n$ as

$$u_L := \underset{\widetilde{u}\in\mathbb{V}_n}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \|\widetilde{u}(y^i) - u^i\|_V^2.$$

In the noiseless case where $u^i := u(y^i)$ for any $i = 1, \dots, m$, this also writes

$$u_L = \underset{\widetilde{u}\in\mathbb{V}_\Lambda}{\operatorname{argmin}} \|u - \widetilde{u}\|_m, \tag{40}$$

where the discrete seminorm is defined by

$$\|f\|_m := \left(\frac{1}{m}\sum_{i=1}^{m} \|f(y^i)\|_V^2\right)^{1/2}.$$

Note that $\|f\|_m^2$ is an unbiased estimator of $\|f\|^2$ since we have

$$\mathbb{E}(\|f\|_m^2) = \|f\|^2.$$

Let $\{\phi_1, \dots, \phi_n\}$ denote an arbitrary $L^2(U, d\mu)$ orthonormal basis of the space $\mathbb{Y}_n$. If we expand the solution to (40) as $\sum_{j=1}^n c_j\phi_j$, with $c_j \in V$, the V-valued vector $\mathbf{c} = (c_1, \dots, c_n)^t$ is the solution to the normal equations

$$\mathbf{G}\mathbf{c} = \mathbf{d}, \tag{41}$$

where the matrix $\mathbf{G}$ has entries

$$\mathbf{G}_{j,k} = \frac{1}{m}\sum_{i=1}^{m} \phi_j(y^i)\phi_k(y^i),$$

and where the V-valued data vector $\mathbf{d} = (d_1, \dots, d_n)^t$ is given by

$$d_j := \frac{1}{m}\sum_{i=1}^{m} u^i\phi_j(y^i).$$

This linear system always has at least one solution, which is unique when $\mathbf{G}$ is nonsingular. When $\mathbf{G}$ is singular, we may define u_L as the unique minimal $\ell^2(\mathbb{R}^n, V)$ norm solution to (41).

In the subsequent analysis, we sometimes work under the assumption of a known uniform bound

$$\|u\|_{L^\infty(U,V)} \leq \tau. \tag{42}$$

We introduce the truncation operator

$$z \mapsto T_\tau(z) := \begin{cases} z, & \text{if } \|z\|_V \leq \tau, \\ \frac{z}{\|z\|_V}, & \text{if } \|z\|_V > \tau, \end{cases}$$

and notice that it is a contraction: $\|T_\tau(z) - T_\tau(\tilde{z})\|_V \leq \|z - \tilde{z}\|_V$ for any $z, \tilde{z} \in V$. The *truncated least-squares approximation* is defined by

$$u_T := T_\tau \circ u_L.$$

Note that, in view of (42), we have $\|u(y) - u_T(y)\|_V \leq \|u(y) - u_L(y)\|_V$ for any $y \in U$ and therefore

$$\|u - u_T\| \leq \|u - u_L\|.$$

Note that the random matrix $\mathbf{G}$ concentrates toward its expectation which is the identity matrix $\mathbf{I}$ as $m \to \infty$. In other words, the probability that $\mathbf{G}$ is ill-conditioned becomes very small as m increases. The truncation operator aims at avoiding instabilities which may occur when $\mathbf{G}$ is ill-conditioned. As an alternative proposed in [15], we may define for some given $A > 1$ the *conditioned least-squares approximation* by

$$u_C := u_L, \text{ if } \mathrm{cond}(\mathbf{G}) \leq A, \quad u_C := 0, \text{ otherwise},$$

where $\mathrm{cond}(\mathbf{G}) := \lambda_{\max}(\mathbf{G})/\lambda_{\min}(\mathbf{G})$ is the usual condition number.

The property that $\|\mathbf{G} - \mathbf{I}\|_2 \leq \delta$ for some $0 < \delta < 1$ amounts to the norm equivalence

$$(1-\delta)\|f\|^2 \leq \|f\|_m^2 \leq (1+\delta)\|f\|^2, \quad f \in \mathbb{V}_n.$$

It is well known that if $m \geq n$ is too close to n, least-squares methods may become unstable and inaccurate for most sampling distributions. For example, if $U = [-1, 1]$ and $\mathbb{Y}_n = \mathbb{P}_{n-1}$ is the space of algebraic polynomials of degree $n - 1$, then with $m = n$ the estimator coincides with the Lagrange polynomial interpolation which can be highly unstable and inaccurate, in particular for equispaced points. Therefore, m should be sufficiently large compared to n for the probability that $\mathbf{G}$ is ill-conditioned to be small. This trade-off between m and n has been analyzed in

[11], using the function

$$y \mapsto k_n(y) := \sum_{j=1}^{n} |\phi_j(y)|^2,$$

which is the diagonal of the integral kernel of the $L^2(U, d\mu)$ projector on $\mathbb{Y}_n$. This function depends on $d\mu$, but not on the chosen orthonormal basis. It is strictly positive in U under minimal assumptions on the orthonormal basis, for example if one element of the basis is the constant function over all U. Obviously, the function k_n satisfies

$$\int_U k_n \, d\mu = n.$$

We define

$$K_n := \|k_n\|_{L^\infty(U)} \geq n.$$

The following results for the least-squares method with noiseless evaluations were obtained in [8, 11, 15, 28] for real-valued functions, however their proof extends in a straightforward manner to the present setting of V-valued functions. They are based on a probabilistic bound for the event $\|\mathbf{G} - \mathbf{I}\|_2 > \delta$ using the particular value $\delta = \frac{1}{2}$, or equivalently the value $A = \frac{1+\delta}{1-\delta} = 3$ as a bound on the condition number of $\mathbf{G}$.

Theorem 10 *For any $r > 0$, if m and n satisfy*

$$K_n \leq \kappa \frac{m}{\ln m}, \quad \text{with} \quad \kappa := \kappa(r) = \frac{3\ln(3/2) - 1}{2 + 2r}, \tag{43}$$

then the following hold.

(i) *The matrix $\mathbf{G}$ satisfies the tail bound*

$$\Pr\left\{ \|\mathbf{G} - \mathbf{I}\|_2 > \frac{1}{2} \right\} \leq 2m^{-r}.$$

(ii) *If u satisfies (42), then the truncated least-squares estimator satisfies, in the noiseless case,*

$$\mathbb{E}(\|u - u_T\|^2) \leq (1 + \zeta(m)) \inf_{\widetilde{u} \in \mathbb{V}_n} \|u - \widetilde{u}\|^2 + 8\tau^2 m^{-r},$$

where $\zeta(m) := \frac{4\kappa}{\ln(m)} \to 0$ as $m \to \infty$, and κ is as in (43).

(iii) *The conditioned least-squares estimator satisfies, in the noiseless case,*

$$\mathbb{E}(\|u - u_C\|^2) \leq (1 + \zeta(m)) \inf_{\widetilde{u} \in \mathbb{V}_n} \|u - \widetilde{u}\|^2 + 2\|u\|^2 m^{-r},$$

where $\zeta(m)$ is as in (ii).

(iv) *If u satisfies (42), then the estimator $u_E \in \{u_L, u_T, u_C\}$ satisfies, in the noiseless case,*

$$\|u - u_E\| \leq (1 + \sqrt{2}) \inf_{\widetilde{u} \in \mathbb{V}_n} \|u - \widetilde{u}\|_{L^\infty(U,V)}, \tag{44}$$

with probability larger than $1 - 2m^{-r}$.

In the case of noisy evaluations modeled by (6)–(7), the observations are given by

$$u^i = u(y^i) + \eta(y^i). \tag{45}$$

The following result from [8] shows that (44) holds up to this additional perturbation.

Theorem 11 *For any $r > 0$, if m and n satisfy condition (43) and u satisfies (42), then the estimator $u_E \in \{u_L, u_T, u_C\}$ in the noisy case (45) satisfies*

$$\|u - u_E\| \leq (1 + \sqrt{2}) \inf_{\widetilde{u} \in \mathbb{V}_n} \|u - \widetilde{u}\|_{L^\infty(U,V)} + \sqrt{2}\varepsilon,$$

with probability larger than $1 - 2n^{-r}$, where ε is the noise level in (7).

Similar results, with more general assumptions on the type of noise, are proven in [11, 15, 29].

4.2 Downward Closed Polynomial Spaces and Weighted Least Squares

Condition (43) shows that K_n gives indications on the number m of observations required to ensure stability and accuracy of the least-squares approximation. In order to understand how demanding this condition is with respect to m, it is important to have sharp upper bounds for K_n. Such bounds have been proven when the measure $d\mu$ on $U = [-1, 1]^d$ has the form

$$d\mu = C \bigotimes_{j=1}^{d} (1 - y_j)^{\theta_1} (1 + y_j)^{\theta_2} dy_j, \tag{46}$$

where $\theta_1, \theta_2 > -1$ are real shape parameters and C is a normalization constant such that $\int_U d\mu = 1$. Sometimes (46) is called the Jacobi measure, because the Jacobi polynomials are orthonormal in $L^2(U, d\mu)$. Remarkable instances of the measure (46) are the uniform measure, when $\theta_1 = \theta_2 = 0$, and the Chebyshev measure, when $\theta_1 = \theta_2 = -\frac{1}{2}$.

When $\mathbb{Y}_n = \mathbb{P}_\Lambda$ is a multivariate polynomial space and Λ is a downward closed multi-index set with $\#(\Lambda) = n$, it is proven in [8, 27] that K_n satisfies an upper bound which only depends on n and on the choice of the measure (46) through the values of θ_1 and θ_2.

Lemma 6 *Let $d\mu$ be the measure defined in* (46)*. Then it holds*

$$K_n \leq \begin{cases} n^{\frac{\ln 3}{\ln 2}}, & \textit{if } \theta_1 = \theta_2 = -\frac{1}{2}, \\ n^{2\max\{\theta_1,\theta_2\}+2}, & \textit{if } \theta_1, \theta_2 \in \mathbb{N}_0. \end{cases} \tag{47}$$

A remarkable property of both algebraic upper bounds in (47) is that the exponent of n is independent of the dimension d, and of the shape of the downward closed set Λ. Both upper bounds are sharp in the sense that equality holds for multi-index sets of rectangular type $\Lambda = R_\nu$ corresponding to tensor product polynomial spaces.

As an immediate consequence of Theorem 10 and Lemma 6, we have the next corollary.

Corollary 1 *For any $r > 0$, with multivariate polynomial spaces $\mathbb{P}_\Lambda$ and Λ downward closed, if m and n satisfy*

$$\frac{m}{\ln m} \geq \kappa \begin{cases} n^{\frac{\ln 3}{\ln 2}}, & \textit{if } \theta_1 = \theta_2 = -\frac{1}{2}, \\ n^{2\max\{\theta_1,\theta_2\}+2}, & \textit{if } \theta_1, \theta_2 \in \mathbb{N}_0, \end{cases} \tag{48}$$

with $\kappa = \kappa(r)$ as in (43)*, then the same conclusions of Theorem 10 hold true.*

Other types of results on the accuracy of least squares have been recently established in [14], under conditions of the same type as (48).

In some situations, for example when n is very large, the conditions (48) might require a prohibitive number of observations m. It is therefore a legitimate question to ask whether there exist alternative approaches with less demanding conditions than (48) between m and n. At best, we would like that m is of order only slightly larger than n, for example by a logarithmic factor. In addition, the above analysis does not apply to situations where the basis functions ϕ_k are unbounded, such as when using Hermite polynomials in the expansion (21). It is thus desirable to ask for the development of approaches that also cover this case.

These questions have an affirmative answer by considering *weighted least-squares methods*, as proposed in [15, 18, 21]. In the following, we survey some results from [15]. For the space $\mathbb{V}_n = V \otimes \mathbb{Y}_n$, the weighted least-squares approximation is defined as

$$u_W := \operatorname*{argmin}_{\widetilde{u}\in\mathbb{V}_n} \frac{1}{m}\sum_{i=1}^{m} w^i \|\widetilde{u}(y^i) - u^i\|_V^2,$$

for some given choice of weights $w^i \geq 0$. This estimator is again computed by solving a linear system of normal equations now with the matrix $\mathbf{G}$ with entries

$$\mathbf{G}_{j,k} = \frac{1}{m}\sum_{i=1}^{m} w(y^i)\phi_j(y^i)\phi_k(y^i).$$

Of particular interest to us are weights of the form

$$w^i = w(y^i),$$

where w is some nonnegative function defined on U such that

$$\int_U w^{-1}\, d\mu = 1. \tag{49}$$

We then denote by $d\sigma$ the probability measure

$$d\sigma := w^{-1} d\mu, \tag{50}$$

and we draw the independent points $y^1, \ldots, y^m$ from $d\sigma$. The case $w \equiv 1$ and $d\sigma = d\mu$ corresponds to the previously discussed standard (unweighted) least-squares estimator u_L. As previously done for u_L, we associate to u_W a truncated estimator u_T and a conditioned estimator u_C, by replacing u_L with u_W in the corresponding definitions.

Let us introduce the function

$$y \mapsto k_{n,w}(y) := \sum_{j=1}^{n} w(y)|\phi_j(y)|^2,$$

where once again $\{\phi_1, \ldots, \phi_n\}$ is an arbitrary $L^2(U, d\mu)$ orthonormal basis of the space $\mathbb{Y}_n$. Likewise, we define

$$K_{n,w} := \|k_{n,w}\|_{L^\infty(U)}.$$

The following result, established in [15] for real-valued functions, extends Theorem 10 to this setting. Its proof in the V-valued setting is similar.

Theorem 12 *For any $r > 0$, if m and n satisfy*

$$\frac{m}{\ln m} \geq \kappa\, K_{n,w}, \quad \text{with} \quad \kappa := \kappa(r) = \frac{3\ln(3/2)-1}{2+2r},$$

then the same conclusions of Theorem 10 hold true with u_L replaced by u_W.

If we now choose

$$w(y) = \frac{n}{\sum_{j=1}^{n} |\phi_j(y)|^2}, \tag{51}$$

that satisfies condition (49) by construction, then the measure defined in (50) takes the form

$$d\sigma = \frac{\sum_{j=1}^{n} |\phi_j(y)|^2}{n} d\mu. \tag{52}$$

The choice (51) also gives

$$K_{n,w} = \|k_{n,w}\|_{L^\infty(U)} = n,$$

and leads to the next result, as a consequence of the previous theorem.

Theorem 13 *For any $r > 0$, if m and n satisfy*

$$\frac{m}{\ln m} \geq \kappa\, n, \quad \text{with} \quad \kappa := \kappa(r) = \frac{3\ln(3/2)-1}{2+2r}, \tag{53}$$

then the same conclusions of Theorem 10 hold true with u_L replaced by u_W, with w given by (51) and the weights taken as $w^i = w(y^i)$.

Remark 2 The above Theorem 13 ensures stability and accuracy of the weighted least-squares approximation, under the minimal condition that m is linearly proportional to n, up to a logarithmic factor. The fact that we may obtain near optimal approximation in L^2 with this amount of sample is remarkable and quite specific to the randomized sampling setting, as it was also observed in similar types of results obtained in the context of information based complexity [22, 32–34, 37]. For example, in the paper [37], the authors obtain the optimal L^2 approximation rate for specific classes of functions that are described by reproducing kernel Hilbert spaces. The recent results from [22] are perhaps closer to our above results since the proposed method uses the same optimal sampling measure associated to the weight (51) as in [15]. The estimates obtained in [22] compare the error of the randomized algorithm with the approximation numbers of the embedding of the RKHS in L^2, assuming a certain polynomial decay for these numbers. In Theorem 13, we do not assume any particular form of decay of the best approximation error.

Clearly the above Theorem 13 is an advantage of weighted least squares compared to standard least squares, since condition (43) is more demanding than (53) in terms of the number of observations m.

However, this advantage comes with some drawbacks that we now briefly recall, see [15] for an extensive description. In general (50) and (52) are not product measures, even if $d\mu$ is one. Therefore, the first drawback of using weighted least squares concerns the efficient generation of independent samples from multivariate probability measures, whose computational cost could be prohibitively expensive, above all when the dimension d is large. In some specific settings, for example downward closed polynomial spaces $\mathbb{Y}_n = \mathbb{P}_\Lambda$ with $\#(\Lambda) = n$, and when $d\mu$ is a product measure, this drawback can be overcome. We refer to [15], where efficient sampling algorithms have been proposed and analyzed. For any m and any downward closed set Λ, these algorithms generate m independent samples with proven bounds on the required computational cost. The dependence on the dimension d and m of these bounds is linear. For the general measure (50) the efficient generation of the sample is a nontrivial task, and remains a drawback of such an approach.

The second drawback concerns the use of weighted least squares in a hierarchical context, where we are given a nested sequence $\Lambda_1 \subset \ldots \subset \Lambda_n$ of downward closed sets, instead of a single such set Λ. Since the measure (52) depends on n, the sets $(\Lambda_n)_{n\geq 1}$ are associated to different measures $(d\sigma_n)_{n\geq 1}$. Hence, recycling samples from the previous iterations of the adaptive algorithm is not as straightforward as in the case of standard least squares.

As a final remark, let us stress that the above results of Theorems 12 and 13 hold for general approximation spaces $\mathbb{Y}_n$ other than polynomials.

5 Adaptive Algorithms and Extensions

5.1 Selection of Downward Closed Polynomial Spaces

The interpolation and least-squares methods discussed in Sects. 3 and 4 allow us to construct polynomial approximations in $\mathbb{V}_\Lambda = V \otimes \mathbb{P}_\Lambda$ of the map (2) from its pointwise evaluations, for some given downward closed set Λ. For these methods, we have given several convergence results in terms of error estimates either in $L^\infty(U, V)$ or $L^2(U, V, d\mu)$. In some cases, these estimates compare favorably with the error of best approximation $\min_{\widetilde{u}\in\mathbb{V}_\Lambda} \|u - \widetilde{u}\|$ measured in such norms.

A central issue which still needs to be addressed is the choice of the downward closed set Λ, so that this error of best approximation is well behaved, for a given map u. Ideally, for each given n, we would like to use the set

$$\Lambda_n = \operatorname*{argmin}_{\Lambda\in\mathscr{D}_n} \min_{\widetilde{u}\in\mathbb{V}_\Lambda} \|u - \widetilde{u}\|,$$

where $\mathscr{D}_n$ is the family of all downward closed sets Λ of cardinality n. However such sets Λ_n are not explicitly given to us, and in addition the resulting sequence $(\Lambda_n)_{n\geq 1}$ is generally not nested.

Concrete selection strategies aim to produce "suboptimal yet good" nested sequences $(\Lambda_n)_{n\geq 1}$ different from the above. Here, an important distinction should be made between *nonadaptive* and *adaptive* selection strategies.

In nonadaptive strategies, the selection of Λ_n is made in an a-priori manner, based on some available information on the given problem. The results from Sect. 2.3 show that, for relevant instances of solution maps associated to elliptic parametric PDEs, there exist nested sequences $(\Lambda_n)_{n\geq 1}$ of downward closed sets such that $\#(\Lambda_n) = n$ and $\min_{\widetilde{u}\in \mathbb{V}_{\Lambda_n}} \|u - \widetilde{u}\|$ decreases with a given convergence rate n^{-s} as $n \to \infty$. In addition, these results provide constructive strategies for building the sets Λ_n, since these sets are defined as the indices associated to the n largest $\widehat{\kappa}_\nu := \max_{\widetilde{\nu}\geq\nu} \kappa_{\widetilde{\nu}}$ like in Theorem 5, or directly to the n largest κ_ν like in Theorems 4 and 6, and since the κ_ν are explicitly given numbers.

In the case where we build the polynomial approximation by interpolation, Theorem 9 shows that a good choice of Λ_n is produced by taking κ_ν to be either $\pi(\nu)b^\nu$ or $\pi(\nu)\beta(\nu)^2 b^\nu$ where $b = (\rho_j^{-1})_{j\geq 1}$ is such that (22) holds. The choice of such a sequence ρ depends both on the size and support properties of the functions ψ_j. For example, when the functions ψ_j have nonoverlapping support, one natural choice is to take

$$\rho_j = \min_{x\in \operatorname{supp}(\psi_j)} \frac{\bar{a}(x) - \tilde{r}}{|\psi_j(x)|}. \tag{54}$$

We refer to [1] for the choices of sequences ρ in more general situations, for example in the case where $(\psi_j)_{j\geq 1}$ is a wavelet basis.

In the case where we build the polynomial approximation by least-squares methods, the various results from Sect. 4 show that under suitable assumptions, the error is nearly as good as that of best approximation in $L^2(U, V, d\mu)$ with respect to the relevant probability measure. In the affine case, Theorem 5 shows that a good choice of Λ_n is produced by taking κ_ν to be $b^\nu \beta(\nu)$ where $b = (\rho_j^{-1})_{j\geq 1}$ is such that (22) holds. In the lognormal case Theorem 6 shows that a good choice of Λ_n is produced by taking κ_ν to be given by (30) where $b = (\rho_j^{-1})_{j\geq 1}$ is such that (28) holds.

Let us briefly discuss the complexity of identifying the downward closed set Λ_n associated to the n largest $\widehat{\kappa}_\nu$. For this purpose, we introduce for any downward closed set Λ its set of *neighbors* defined by

$$N(\Lambda) := \{\nu \in \mathscr{F} \setminus \Lambda \text{ such that } \Lambda \cup \{\nu\} \text{ is downward closed}\}.$$

We may in principle define $\Lambda_n = \{\nu^1, \dots, \nu^n\}$ by the following induction.

- Take $\nu^1 = 0_{\mathscr{F}}$ as the null multi-index.
- Given $\Lambda_k = \{\nu^1, \dots, \nu^k\}$, choose a ν^{k+1} maximizing $\widehat{\kappa}_\nu$ over $\nu \in N(\Lambda_k)$.

In the finite-dimensional case $d < \infty$, we observe that $N(\Lambda_k)$ is contained in the union of $N(\Lambda_{k-1})$ with the set consisting of the indices

$$\nu^k + e_j, \quad j = 1, \dots, d,$$

where e_j is the Kronecker sequence with 1 at position j. As a consequence, since the values of the $\widehat{\kappa}_\nu$ have already been computed for $\nu \in N(\Lambda_{k-1})$, the step k of the induction requires at most d evaluations of $\widehat{\kappa}_\nu$, and therefore the overall computation of Λ_n requires at most nd evaluations.

In the infinite-dimensional case $d = \infty$, the above procedure cannot be practically implemented, since the set of neighbors has infinite cardinality. This difficulty can be circumvented by introducing a priority order among the variables, as done in the next definitions.

Definition 4 A monotone nonincreasing positive sequence $(c_\nu)_{\nu \in \mathscr{F}}$ is said to be *anchored* if and only if

$$l \leq j \implies c_{e_j} \leq c_{e_l}.$$

A finite downward closed set Λ is said to be *anchored* if and only if

$$e_j \in \Lambda \quad \text{and} \quad l \leq j \quad \implies \quad e_l \in \Lambda,$$

where e_l and e_j are the Kronecker sequences with 1 at position l and j, respectively.

Obviously, if $(c_\nu)_{\nu \in \mathscr{F}}$ is anchored, one of the sets Λ_n corresponding to its n largest values is anchored. It is also readily seen that all sequences $(\widehat{\kappa}_\nu)_{\nu \in \mathscr{F}}$ that are used in Theorems 4, 5, 6 or 9 for the construction of Λ_n are anchored, provided that the sequence $b = (\rho_j^{-1})_{j \geq 1}$ is monotone nonincreasing. This is always the case up to a rearrangement of the variables. For any anchored set Λ, we introduce the set of its *anchored neighbors* defined by

$$\widetilde{N}(\Lambda) := \{\nu \in N(\Lambda) \ : \ \nu_j = 0 \ \text{ if } \ j > j(\Lambda) + 1\}, \tag{55}$$

where

$$j(\Lambda) := \max\{j \ : \ \nu_j > 0 \text{ for some } \nu \in \Lambda\}.$$

We may thus modify in the following way the above induction procedure.

- Take $\nu^1 = 0_{\mathscr{F}}$ as the null multi-index.
- Given $\Lambda_k = \{\nu^1, \dots, \nu^k\}$, choose a ν^{k+1} maximizing $\widehat{\kappa}_\nu$ over $\nu \in \widetilde{N}(\Lambda_k)$.

This procedure is now feasible in infinite dimension. At each step k the number of active variables is limited by $j(\Lambda_k) \leq k-1$, and the total number of evaluations of $\widehat{\kappa}_\nu$ needed to construct Λ_n does not exceed $1 + 2 + \cdots + (n-1) \leq n^2/2$.

In adaptive strategies the sets Λ_n are not a-priori selected, but instead they are built in a recursive way, based on earlier computations. For instance, one uses the previous set Λ_{n-1} and the computed polynomial approximation $u_{\Lambda_{n-1}}$ to construct Λ_n. If we impose that the sets Λ_n are nested, this means that we should select an index $\nu^n \notin \Lambda_{n-1}$ such that

$$\Lambda_n := \Lambda_{n-1} \cup \{\nu^n\}.$$

The choice of the new index ν^n is further limited to $N(\Lambda_{n-1})$ if we impose that the constructed sets Λ_n are downward closed, or to $\widetilde{N}(\Lambda_{n-1})$ if we impose that these sets are anchored.

Adaptive methods are known to sometimes perform significantly better than their nonadaptive counterpart. In the present context, this is due to the fact that the a-priori choices of Λ_n based on the sequences κ_ν may fail to be optimal. In particular, the guaranteed rate n^{-s} based on such choices could be pessimistic, and better rates could be obtained by other choices. However, convergence analysis of adaptive methods is usually more delicate. We next give examples of possible adaptive strategies in the interpolation and least-squares frameworks.

5.2 Adaptive Selection for Interpolation

We first consider polynomial approximations obtained by interpolation as discussed in Sect. 3. The hierarchical form

$$I_\Lambda u = \sum_{\nu\in\Lambda} \alpha_\nu B_\nu, \tag{56}$$

may formally be viewed as a truncation of the expansion of u in the hierarchical basis

$$\sum_{\nu\in\mathscr{F}} \alpha_\nu B_\nu,$$

which however may not always be converging, in contrast to the series discussed in Sect. 2. Nevertheless, we could in principle take the same view, and use for Λ_n the set of indices corresponding to the n largest terms of (56) measured in some given metric $L^p(U, V, d\mu)$. This amounts in choosing the indices of the n largest $w_\nu \|\alpha_\nu\|_V$, where the weight w_ν is given by

$$w_\nu := \|B_\nu\|_{L^p(U,d\mu)}.$$

This weight is easily computable when $d\mu$ is a tensor product measure, such as the uniform measure. In the case where $p = \infty$ and if we use the Leja sequence, we know that $\|B_\nu\|_{L^\infty(U)} = 1$ and therefore this amounts to choosing the largest $\|\alpha_\nu\|_V$.

This selection strategy is not practically feasible since we cannot afford this exhaustive search over $\mathscr{F}$. However, it naturally suggests the following adaptive greedy algorithm, which has been proposed in [7].

- Initialize $\Lambda_1 := \{0_{\mathscr{F}}\}$ with the null multi-index.
- Assuming that Λ_{n-1} has been selected and that $(\alpha_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, compute α_ν for $\nu \in N(\Lambda_{n-1})$.
- Set

$$\nu^n := \operatorname{argmax}\{w_\nu \|\alpha_\nu\|_V \ : \ \nu \in N(\Lambda_{n-1})\}. \tag{57}$$

- Define $\Lambda_n := \Lambda_{n-1} \cup \{\nu^n\}$.

In the case where $p = \infty$ and if we use the Leja sequence, this strategy amounts in picking the index ν^n that maximizes the interpolation error $\|u(y_\nu) - I_{\Lambda_{n-1}} u(y_\nu)\|_V$ among all ν in $N(\Lambda_{n-1})$. By the same considerations as previously discussed for the a-priori selection of Λ_n, we find that in the finite-dimensional case, the above greedy algorithm requires at most dn evaluation after n steps. When working with infinitely many variables $(y_j)_{j\geq 1}$, we replace the infinite set $N(\Lambda_n)$ in the algorithm by the finite set of anchored neighbors $\widetilde{N}(\Lambda_n)$ defined by (55). Running n steps of the resulting greedy algorithm requires at most $n^2/2$ evaluations.

Remark 3 A very similar algorithm has been proposed in [19] in the different context of adaptive quadratures, that is, for approximating the integral of u over the domain U rather than u itself. In that case, the natural choice is to pick the new index ν^n that maximizes $|\int_U \Delta_\nu u \, d\mu|$ over $N(\Lambda_n)$ or $\widetilde{N}(\Lambda_n)$.

The main defect of the above greedy algorithm is that it may fail to converge, even if there exist sequences $(\Lambda_n)_{n\geq 1}$ such that $I_{\Lambda_n} u$ converges toward u. Indeed, if $\Delta_\nu u = 0$ for a certain ν, then no index $\widetilde{\nu} \geq \nu$ will ever be selected by the algorithm. As an example, if u is of the form

$$u(y) = u_1(y_1)u_2(y_2),$$

where u_1 and u_2 are nonpolynomial smooth functions such that $u_2(t_0) = u_2(t_1)$, then the algorithm could select sets Λ_n with indices $\nu = (k, 0)$ for $k = 0, \dots, n-1$, since the interpolation error at the point (t_k, t_1) vanishes.

One way to avoid this problem is to adopt a more conservative selection rule which ensures that all of $\mathscr{F}$ is explored, by alternatively using the rule (57), or picking the multi-index $\nu \in \widetilde{N}(\Lambda_n)$ which has appeared at the earliest stage in the neighbors of the previous sets Λ_k. This is summarized by the following algorithm.

- Initialize $\Lambda_1 := \{0_{\mathscr{F}}\}$ with the null multi-index.

- Assuming that Λ_{n-1} has been selected and that $(\alpha_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, compute α_ν for $\nu \in \widetilde{N}(\Lambda_{n-1})$.
- If n is even, set

$$\nu^n := \operatorname{argmax}\{w_\nu \|\alpha_\nu\|_V \ : \ \nu \in \widetilde{N}(\Lambda_{n-1})\}. \tag{58}$$

- If n is odd, set

$$\nu^n := \operatorname{argmin}\{k(\nu) \ : \ \nu \in \widetilde{N}(\Lambda_{n-1})\}, \quad k(\nu) := \min\{k \ : \ \nu \in \widetilde{N}(\Lambda_k)\}.$$

- Define $\Lambda_n := \Lambda_{n-1} \cup \{\nu^n\}$.

Even with such modifications, the convergence of the interpolation error produced by this algorithm is not generally guaranteed. Understanding which additional assumptions on u ensure convergence at some given rate, for a given univariate sequence T such as Leja points, is an open problem.

Remark 4 Another variant to the above algorithms consists in choosing at the iteration k more than one new index at a time within $N(\Lambda_{k-1})$ or $\widetilde{N}(\Lambda_{k-1})$. In this case, we have $n_k := \#(\Lambda_k) \geq k$. For example we may choose the smallest subset of indices that retains a fixed portion of the quantity $\sum_{\nu\in\Lambda_{k-1}} w_\nu \|\alpha_\nu\|_V$. This type of modification turns out to be particularly relevant in the least-squares setting discussed in the next section.

5.3 Adaptive Selection for Least Squares

In this section we describe adaptive selections in polynomial spaces, for the least-squares methods that have been discussed in Sect. 4. We focus on adaptive selection algorithms based on the standard (unweighted) least-squares method.

As a preliminary observation, it turns out that the most efficient available algorithms for adaptive selection of multi-indices might require the selection of more than one index at a time. Therefore, we adopt the notation that $n_k := \#(\Lambda_k) \geq k$, where the index k denotes the iteration in the adaptive algorithm.

As discussed in Sect. 4, stability and accuracy of the least-squares approximation is ensured under suitable conditions between the number of samples and the dimension of the approximation space, see e.g. condition (48). Hence, in the development of reliable iterative algorithms, such conditions need to be satisfied at each iteration. When $d\mu$ is the measure (46) with shape parameters θ_1, θ_2, condition (48) takes the form of

$$\frac{m_k}{\ln m_k} \geq \kappa \, n_k^s, \tag{59}$$

where m_k denotes the number of samples at iteration k, and

$$s = \begin{cases} \ln 3/\ln 2, & \text{if } \theta_1 = \theta_2 = -\frac{1}{2}, \\ 2\max\{\theta_1, \theta_2\} + 2, & \text{if } \theta_1, \theta_2 \in \mathbb{N}_0. \end{cases}$$

Since n_k increases with k, the minimal number of samples m_k that satisfies (59) has to increase as well at each iteration. At this point, many different strategies can be envisaged for progressively increasing m_k such that (59) remains satisfied at each iteration k. For example, one can double the number of samples by choosing $m_k = 2m_{k-1}$ whenever (59) is broken, and keep $m_k = m_{k-1}$ otherwise. The sole prescription for applying Corollary 1 is that the samples are independent and drawn from $d\mu$. Since all the samples at all iterations are drawn from the same measure $d\mu$, at the kth iteration, where m_k samples are needed, it is possible to use m_{k-1} samples from the previous iterations, thus generating only $m_k - m_{k-1}$ new samples.

We may now present a first adaptive algorithm based on standard least squares.

- Initialize $\Lambda_1 := \{0_{\mathscr{F}}\}$ with the null multi-index.
- Assuming that Λ_{k-1} has been selected, compute the least-squares approximation

$$u_L = \sum_{\nu \in \Lambda_{k-1} \cup N(\Lambda_{k-1})} c_\nu \phi_\nu$$

  of u in $\mathbb{V}_{\Lambda_{k-1} \cup N(\Lambda_{k-1})}$, using a number of samples m_k that satisfies condition (59) with $n_k = \#(\Lambda_{k-1} \cup N(\Lambda_{k-1}))$.
- Set

$$\nu^k := \underset{\nu \in N(\Lambda_{k-1})}{\operatorname{argmax}} |c_\nu|^2. \tag{60}$$

- Define $\Lambda_k := \Lambda_{k-1} \cup \{\nu^k\}$.

Similarly to the previously discussed interpolation algorithms, in the case of infinitely many variables $(y_j)_{j \geq 1}$ the set $N(\Lambda_k)$ is infinite and should be replaced by the finite set of anchored neighbors $\widetilde{N}(\Lambda_k)$ defined by (55). As for interpolation, we may define a more conservative version of this algorithm in order to ensure that all of $\mathscr{F}$ is explored. For example, when k is even, we define ν^k according to (60), and when k is odd we pick for ν^k the multi-index $\nu \in \widetilde{N}(\Lambda_k)$ which has appeared at the earliest stage in the neighbors of the previous sets Λ_k. The resulting algorithm is very similar to the one presented for interpolation, with obvious modifications due to the use of least squares.

As announced at the beginning, it can be advantageous to select more than one index at a time from $\widetilde{N}(\Lambda_{k-1})$, at each iteration k of the adaptive algorithm. For describing the multiple selection of indices from $\widetilde{N}(\Lambda_{k-1})$, we introduce the so-called *bulk chasing* procedure. Given a finite set $R \subseteq \widetilde{N}(\Lambda_{k-1})$, a nonnegative function $\mathscr{E} : R \to \mathbb{R}$ and a parameter $\alpha \in (0, 1]$, we define the procedure bulk $:=$ bulk$(R, \mathscr{E}, \alpha)$ that computes a set $F \subseteq R$ of minimal positive cardinality such that

$$\sum_{\nu \in F} \mathscr{E}(\nu) \geq \alpha \sum_{\nu \in R} \mathscr{E}(\nu).$$

A possible choice for the function $\mathscr{E}$ is

$$\mathscr{E}(\nu) = \mathscr{E}_L(\nu) := |c_\nu|^2, \quad \nu \in R,$$

where c_ν is given from an available least-squares estimator

$$u_L = \sum_{\nu \in \Lambda} c_\nu \phi_\nu,$$

that has been already computed on any downward closed set $R \subset \Lambda \subseteq \Lambda_{k-1} \cup \widetilde{N}(\Lambda_{k-1})$. Another choice for $\mathscr{E}$ is

$$\mathscr{E}(\nu) = \mathscr{E}_M(\nu) := \langle \phi_\nu, u - \widetilde{u}_L \rangle_{m_{k-1}}, \quad \nu \in R,$$

where $\widetilde{u}_L$ is the truncation to Λ_{k-1} of a least-squares estimator $u_L = \sum_{\nu \in \Lambda} c_\nu \phi_\nu$ that has been already computed on any downward closed set $\Lambda_{k-1} \subset \Lambda \subseteq \Lambda_{k-1} \cup \widetilde{N}(\Lambda_{k-1})$, using a number of samples m_{k-1} that satisfies condition (59) with $n_k = \#(\Lambda)$. The discrete norm in $\mathscr{E}_M(\nu)$ uses the same m_{k-1} evaluations of u that have been used to compute the least-squares approximation u_L on Λ.

Both $\mathscr{E}_L(\nu)$ and $\mathscr{E}_M(\nu)$ should be viewed as estimators of the coefficient $\langle u, \phi_\nu \rangle$. The estimator $\mathscr{E}_M(\nu)$ is of Monte Carlo type and computationally cheap to calculate. Combined use of the two estimators leads to the next algorithm for greedy selection with bulk chasing, that has been proposed in [26].

- Initialize $\Lambda_1 := \{0_{\mathscr{F}}\}$ with the null multi-index, and choose $\alpha_1, \alpha_2 \in (0, 1]$.
- Assuming that Λ_{k-1} has been selected, set

$$F_1 = \text{bulk}(\widetilde{N}(\Lambda_{k-1}), \mathscr{E}_M, \alpha_1), \tag{61}$$

  where $\mathscr{E}_M$ uses the least-squares approximation $u_L = \sum_{\nu \in \Lambda} c_\nu \phi_\nu$ of u in $\mathbb{V}_\Lambda$ that has been calculated at iteration $k-1$ on a downward closed set $\Lambda_{k-1} \subset \Lambda \subseteq \Lambda_{k-1} \cup \widetilde{N}(\Lambda_{k-1})$ using a number of samples m_{k-1} that satisfies (59) with $n_k = \#(\Lambda)$.
- Compute the least-squares approximation

$$u_L = \sum_{\nu \in \Lambda_{k-1} \cup F_1} c_\nu \phi_\nu \tag{62}$$

  of u on $\mathbb{V}_{\Lambda_{k-1} \cup F_1}$ using a number of samples m_k that satisfies (59) with $n_k = \#(\Lambda_{k-1} \cup F_1)$.

- Set

$$F_2 = \text{bulk}(F_1, \mathscr{E}_L, \alpha_2), \tag{63}$$

where $\mathscr{E}_L$ uses the least-squares approximation u_L computed on $\Lambda_{k-1} \cup F_1$.
- Define $\Lambda_k = \Lambda_{k-1} \cup F_2$.

The set $\widetilde{N}(\Lambda_{k-1})$ can be large, and might contain many indices that are associated to small coefficients. Discarding these indices is important in order to avoid unnecessary computational burden in the calculation of the least-squares approximation. The purpose of the bulk procedure (61) is to perform a preliminary selection of a set $F_1 \subseteq \widetilde{N}(\Lambda_{k-1})$ of indices, using the cheap estimator $\mathscr{E}_M$. At iteration k, $\mathscr{E}_M$ in (61) uses the estimator computed in (62) at iteration $k-1$ and truncated to Λ_{k-1}. Afterwards, at iteration k, the least-squares approximation in (62) is calculated on $\Lambda_{k-1} \cup F_1$, using a number of samples m_k which satisfies condition (59), with $n_k = \#(\Lambda_{k-1} \cup F_1)$. The second bulk procedure (63) selects a set F_2 of indices from F_1, using the more accurate estimator $\mathscr{E}_L$. The convergence rate of the adaptive algorithm depends on the values given to the parameters α_1 and α_2.

Finally we mention some open issues related to the development of adaptive algorithms using the weighted least-squares methods discussed in Sect. 4, instead of standard least squares. In principle the same algorithms described above can be used with the weighted least-squares estimator u_W replacing the standard least-squares estimator u_L, provided that, at each iteration k, the number of samples m_k satisfies

$$\frac{m_k}{\ln m_k} \geq \kappa \, n_k,$$

and that the samples are drawn from the optimal measure, see Theorem 13. This ensures that at each iteration k of the adaptive algorithm, the weighted least-squares approximation remains stable and accurate. However, no guarantees on stability and accuracy are ensured if the above conditions are not met, for example when the samples from previous iterations are recycled.

5.4 Approximation in Downward Closed Spaces: Beyond Polynomials

The concept of downward closed approximation spaces can be generalized beyond the polynomial setting. We start from a countable index set S equipped with a partial order $\leq$, and assume that there exists a root index $0_S \in S$ such that $0_S \leq \sigma$ for all $\sigma \in S$. We assume that $(B_\sigma)_{\sigma \in S}$ is a basis of functions defined on $[-1, 1]$ such that $B_{0_S} \equiv 1$. We then define by tensorization a basis of functions on $U = [-1, 1]^d$

when $d < \infty$, or $U = [-1, 1]^{\mathbb{N}}$ in the case of infinitely many variables, according to

$$B_\nu(y) = \prod_{j \geq 1} B_{\nu_j}(y_j), \quad \nu := (\nu_j)_{j \geq 1} \in \mathcal{F},$$

where $\mathcal{F} := S^d$ in the case $d < \infty$, or $\mathcal{F} = \ell^0(\mathbb{N}, S)$, i.e. the set of finitely supported sequences, in the case $d = \mathbb{N}$.

The set $\mathcal{F}$ is equipped with a partial order induced by its univariate counterpart: $\nu \leq \widetilde{\nu}$ if and only if $\nu_j \leq \widetilde{\nu}_j$ for all $j \geq 1$. We may then define downward closed sets $\Lambda \subset \mathcal{F}$ in the same way as in Definition 1 which corresponds to the particular case $S = \mathbb{N}$. We then define the associated downward closed approximation space by

$$\mathbb{V}_\Lambda := V \otimes \mathbb{B}_\Lambda, \quad \mathbb{B}_\Lambda := \text{span}\{B_\nu \ : \ \nu \in \Lambda\},$$

that is the space of functions of the form $\sum_{\nu \in \Lambda} u_\nu B_\nu$ with $u_\nu \in V$.

Given a sequence $T = (t_\sigma)_{\sigma \in S}$ of pairwise distinct points we say that the basis $(B_\sigma)_{\sigma \in S}$ is hierarchical when it satisfies

$$B_\sigma(t_\sigma) = 1 \text{ and } B_\sigma(t_{\widetilde{\sigma}}) = 0 \text{ if } \widetilde{\sigma} \leq \sigma \text{ and } \widetilde{\sigma} \neq \sigma.$$

We also define the tensorized grid

$$y_\nu := (t_{\nu_j})_{j \geq 1} \in U.$$

Then, if $\Lambda \subset \mathcal{F}$ is a downward closed set, we may define an interpolation operator I_Λ onto V_Λ associated to the grid

$$\Gamma_\Lambda := \{y_\nu \ : \ \nu \in \Lambda\}.$$

In a similar manner as in the polynomial case, this operator is defined inductively by

$$I_\Lambda u := I_{\widetilde{\Lambda}} u + \alpha_\nu B_\nu, \quad \alpha_\nu := \alpha_\nu(u) = u(y_\nu) - I_{\widetilde{\Lambda}} u(y_\nu),$$

where $\nu \notin \widetilde{\Lambda}$ and $\widetilde{\Lambda}$ is any downward closed set such that $\Lambda = \widetilde{\Lambda} \cup \{\nu\}$. We initialize this computation with $\Lambda_1 = \{0_{\mathcal{F}}\}$, where $0_{\mathcal{F}}$ is the null multi-index, by defining $I_{\Lambda_1} u$ as the constant function with value $u(y_{0_{\mathcal{F}}})$.

Examples of relevant hierarchical systems include the classical piecewise linear hierarchical basis functions. In this case the set S is defined by

$$S = \{\lambda_{-1}, \lambda_1, (0, 0)\} \cup \big\{(j, k) \ : \ -2^{j-1} \leq k \leq 2^{j-1} - 1, \ j = 1, 2, \dots\big\}$$

equipped with the partial order $\lambda_{-1} \leq \lambda_1 \leq (0,0)$ and

$$(j,k) \leq (j+1,2k), \qquad (j,k) \leq (j+1,2k+1), \qquad (j,k) \in S.$$

The set S is thus a binary tree where λ_{-1} is the root node, $(0,0)$ is a child of λ_1 which is itself a child of λ_{-1}, every node (j,k) has two children $(j+1,2k)$ and $(j+1,2k+1)$, and the relation $\widetilde{\lambda} \leq \lambda$ means that $\widetilde{\lambda}$ is a parent of λ. The index j corresponds to the level of refinement, i.e. the depth of the node in the binary tree. We associate with S the sequence

$$T := \{t_{\lambda_{-1}}, t_{\lambda_1}, t_{(0,0)}\} \cup \left\{ t_{(j,k)} := \frac{2k+1}{2^j} : (j,k) \in S, j \geq 1 \right\},$$

where $t_{\lambda_{-1}} = -1$, $t_{\lambda_1} = 1$ and $t_{(0,0)} = 0$. The hierarchical basis of piecewise linear functions defined over $[-1,1]$ is then given by

$$B_{\lambda_{-1}} \equiv 1, \quad B_{\lambda_1}(t) = \frac{1+t}{2}, \quad B_{(j,k)}(t) = H(2^j(t - t_{(j,k)})), \quad (j,k) \in S,$$

where

$$H(t) := \max\{0, 1 - |t|\},$$

is the usual hat function. In dimension $d = 1$, the hierarchical interpolation amounts in the following steps: start by approximating f with the constant function equal to $f(-1)$, then with the affine function that coincides with f at -1 and 1, then with the piecewise affine function that coincides with f at -1, 0 and -1; afterwards refine the approximation in further steps by interpolating f at the midpoint of an interval between two adjacents interpolation points.

Other relevant examples include piecewise polynomials, hierarchical basis functions, and more general interpolatory wavelets, see [10] for a survey.

References

1. Bachmayr, M., Cohen, A., Migliorati, G.: Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients. ESAIM:M2AN **51**, 321–339 (2017)
2. Bachmayr, M., Cohen, A., DeVore, R. Migliorati, G.: Sparse polynomial approximation of parametric elliptic PDEs. Part II: lognormal coefficients. ESAIM:M2AN **51**, 341–363 (2017)
3. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
4. Calvi, J.P., Phung, V.M.: On the Lebesgue constant of Leja sequences for the unit disk and its applications to multivariate interpolation. J. Approx. Theory **163**, 608–622 (2011)
5. Chkifa, A.: On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection. J. Approx. Theory **166**, 176–200 (2013)

6. Chkifa, A., Cohen, A.: On the stability of polynomial interpolation using hierarchical sampling. In: Sampling Theory: a Renaissance, pp. 437–458. Springer, New York (2015)
7. Chkifa, A., Cohen, A., Schwab, C.: High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. Found. Comput. Math. **14**, 601–633 (2014)
8. Chkifa, A., Cohen, A., Migliorati, G., Nobile, F., Tempone, R.: Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs. ESAIM:M2AN **49**(3), 815–837 (2015)
9. Chkifa, A., Cohen, A., Schwab, C.: Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. J. Math. Pures Appl. **103**(2), 400–428 (2015)
10. Cohen, A.: Numerical Analysis of Wavelet Methods. Elsevier, Amsterdam (2003)
11. Cohen, A., Davenport, M.A., Leviatan, D.: On the stability and accuracy of least squares approximations. Found. Comput. Math. **13**, 819–834 (2013)
12. Cohen, A., DeVore, R.: Approximation of high-dimensional parametric PDEs. Acta Numer. **24**, 1–159 (2015)
13. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic PDEs. Anal. Appl. **9**, 11–47 (2011)
14. Cohen, A., Migliorati, G., Nobile, F.: Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension. Constr. Approx. **45**, 497–519 (2017)
15. Cohen, A., Migliorati, G.: Optimal weighted least-squares methods. SMAI JCM. **3**, 181–203 (2017)
16. de Boor, C., Ron, A.: Computational aspects of polynomial interpolation in several variables. Math. Comp. **58**, 705–727 (1992)
17. DeVore, R., Howard, R., Micchelli, C.: Optimal non-linear approximation. Manuscripta Math. **63**(4), 469–478 (1989)
18. Doostan, A., Hampton, J.: Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression. Comput. Methods Appl. Mech. Eng. **290**, 73–97 (2015)
19. Gerstner, T., Griebel, M.: Dimension-adaptive tensor-product quadrature. Computing **71**, 65–87 (2003).
20. Hoang, V., Schwab, C.: n-term Wiener chaos approximation rates for elliptic PDEs with lognormal Gaussian random inputs. Math. Models Methods Appl. Sci. **24**, 797–826 (2014)
21. Jakeman, J.D., Narayan, A., Zhou, T.: A Christoffel function weighted least squares algorithm for collocation approximations. Math. Comput. **86**, 1913–1947 (2017)
22. Krieg, D.: Optimal Monte Carlo methods for L^2-approximation. Preprint. arXiv:1705.04567
23. Kuntzman, J.: Méthodes numériques - Interpolation, dérivées. Dunod, Paris (1959)
24. Le Maître, O., Knio, O.: Spectral Methods for Uncertainty Quantification. Springer, New York (2010)
25. Lorentz, G., Lorentz, R.: Solvability problems of bivariate interpolation I. Constr. Approx. **2**, 153–169 (1986)
26. Migliorati, G.: Adaptive polynomial approximation by means of random discrete least squares. In: A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso (eds.) Numerical Mathematics and Advanced Applications - ENUMATH 2013, pp. 547–554. Springer, New York (2015)
27. Migliorati, G.: Multivariate Markov-type and Nikolskii-type inequalities for polynomials associated with downward closed multi-index sets. J. Approx. Theory **189**, 137–159 (2015)
28. Migliorati, G., Nobile, F., von Schwerin, E., Tempone, R.: Analysis of discrete L^2 projection on polynomial spaces with random evaluations. Found. Comput. Math. **14**, 419–456 (2014)
29. Migliorati, G., Nobile, F., Tempone, R.: Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. J. Multivar. Anal. **142**, 167–182 (2015)
30. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006)
31. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics. Springer, New York (1988)
32. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems - Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12 (2010)

33. Novak, E., Woźniakowski, H.: On the power of function value for the approximation problem in various settings. Surv. Approx. Theory **6**, 1–23 (2011)
34. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems - Volume III: Standard Information for Operators. EMS Tracts in Mathematics, vol. 18 (2012)
35. Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)
36. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. SIAM (2005)
37. Wasilkowski, G. W., Woźniakowski, H. The power of standard information for multivariate approximation in the randomized setting. Math. Comput. **76**, 965–988 (2007)

Subperiodic Trigonometric Hyperinterpolation

Gaspare Da Fies, Alvise Sommariva, and Marco Vianello

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Using recent results on subperiodic trigonometric Gaussian quadrature and the construction of subperiodic trigonometric orthogonal bases, we extend Sloan's notion of hyperinterpolation to trigonometric spaces on subintervals of the period. The result is relevant, for example, to function approximation on spherical or toroidal rectangles.

1 Introduction

Trigonometric approximation in the absence of periodicity has been investigated along some apparently parallel paths: the theory of *Fourier Extensions* (also known as Fourier Continuations in certain applications), cf. [1, 3, 4, 8, 9, 27], the theory of *subperiodic* trigonometric interpolation and quadrature, e.g. [7, 12–15, 40], and the recent developments of *nonperiodic* trigonometric approximation [39] and of *mapped* polynomial approximation [2]. It is also worth mentioning the study of monotone trigonometric approximations, which can arise only in the subperiodic setting, cf. [29].

G. Da Fies
Department of Mathematics, Aberystwyth University, Wales, UK
e-mail: gad12@aber.ac.uk

A. Sommariva · M. Vianello (✉)
Department of Mathematics, University of Padova, Padova, Italy
e-mail: alvise@math.unipd.it; marcov@math.unipd.it

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_13

Fourier Extensions emerged in the context of trigonometric approximation of nonperiodic functions, for example as a tool for circumventing the Gibbs phenomenon. Summarizing, a smooth nonperiodic real-valued function f defined in $[-1, 1]$ is approximated by the restriction to $[-1, 1]$ of a trigonometric polynomial periodic on $[-T, T]$, $T > 1$,

$$\begin{aligned}\tau \in \mathscr{T}_n = \text{span}\{1, \cos(k\pi u/T), \sin(k\pi u/T),\ 1 \le k \le n,\\ u \in [-T, T]\ ,\ T > 1\},\end{aligned} \tag{1}$$

by solving the minimization problem

$$\min_{\tau \in \mathscr{T}_n} \|\tau - f\|_{L^2_\mu([-1,1])}, \tag{2}$$

where either μ is the Lebesgue measure on $[-1, 1]$, or a discrete measure supported at a discrete set $X \subset [-1, 1]$. In the first case one speaks of a "continuous Fourier extension", and in the second of a "discrete Fourier extension". When $\text{card}(X) = 2n + 1 = \dim(\mathscr{T}_n)$, the construction of the discrete Fourier extension becomes an interpolation problem.

A number of deep theoretical results have been obtained on Fourier extensions, concerning different aspects such as: choice of T, choice of interpolation nodes for discrete extensions, approximation power (with tight error estimates for analytic functions), resolution power on higly oscillatory functions, numerical stability; cf. e.g. [1, 4]. The main computational approach has been the solution of the least squares/interpolation linear systems working with the standard trigonometric basis (1), which leads to strongly ill-conditioned systems. Nevertheless, Fourier extensions are actually numerically stable when implemented in finite arithmetic, cf. [4] for an explanation of the phenomenon (a different approach has been recently proposed in [30]).

Let us now summarize the parallel path of *subperiodic trigonometric approximation*. In several recent papers, subperiodic trigonometric interpolation and quadrature have been studied, i.e., interpolation and quadrature formulas exact on

$$\mathbb{T}_n([-\omega, \omega]) = \text{span}\{1, \cos(k\theta), \sin(k\theta),\ 1 \le k \le n\ ,\ \ \theta \in [-\omega, \omega]\}, \tag{3}$$

where $\mathbb{T}_n([-\omega, \omega])$ denote the $(2n + 1)$-dimensional space of trigonometric polynomials restricted to the interval $[-\omega, \omega]$, $0 < \omega \le \pi$; cf. e.g. [7, 13], and [12, 15] for the construction and application of subperiodic trigonometric Gaussian formulas. All these formulas are related by the simple nonlinear transformation

$$\theta(t) = 2 \arcsin(\sin(\omega/2)t)\ ,\ \ t \in [-1, 1], \tag{4}$$

with inverse

$$t(\theta) = \frac{\sin(\theta/2)}{\sin(\omega/2)}\ , \quad \theta \in [-\omega, \omega], \tag{5}$$

to polynomial interpolation and quadrature on $[-1, 1]$, and have been called "subperiodic" since they concern subintervals of the period of trigonometric polynomials.

For example, trigonometric interpolation and quadrature by the transformed zeros of the $(2n + 1)$-th Chebyshev polynomial $T_{2n+1}(t)$ has been studied in [7]. Moreover, in [13] stability of such Chebyshev-like subperiodic trigonometric interpolation has been studied, proving that its Lebesgue constant does not depend on ω and increases logarithmically in the degree.

In this article, we apply the following subperiodic quadrature [12, 23].

Proposition 1 *Let* $\{(\xi_j, \lambda_j)\}_{1\le j\le n+1}$, *be the nodes and positive weights of the algebraic Gaussian quadrature formula on* $(-1, 1)$ *induced by the weight function*

$$W_{-1/2}(t) = \frac{2\sin(\omega/2)}{\sqrt{1 - \sin^2(\omega/2)\, t^2}}\ , \quad t \in (-1, 1). \tag{6}$$

Then

$$\int_{-\omega}^{\omega} f(\theta)\, d\theta = \sum_{j=1}^{n+1} \lambda_j f(\varphi_j)\ , \quad \forall f \in \mathbb{T}_n([-\omega, \omega])\ , \quad 0 < \omega \le \pi, \tag{7}$$

where

$$\varphi_j = 2\arcsin(\sin(\omega/2)\xi_j) \in (-\omega, \omega)\ , \quad j = 1, 2, \ldots, n + 1. \tag{8}$$

It is worth recalling that the key role played by the transformation (6) on subintervals of the period was also recognized in [6, E.3, p. 235], and more recently in [40], in the context of trigonometric polynomial inequalities. On the other hand, such a transformation (already introduced in [28]), is at the base of the recent studies on nonperiodic trigonometric approximations [39], and on "mapped" polynomial approximations [2].

The initial motivation of [7, 12] for the analysis of subperiodic interpolation and quadrature was different from that of Fourier extensions or other similar studies, since it arised from multivariate applications. The key observation is that a multivariate polynomial restricted to an arc of a circle becomes a univariate subperiodic trigonometric polynomial in the arclength. Then, multivariate polynomials on domains defined by circular arcs, such as sections of disk (circular sectors, segments, zones, lenses, lunes) and surface/solid sections of sphere, cylinder, torus (rectangles, collars, caps, slices) become in the appropriate coordinates elements of tensor-product spaces of univariate trigonometric and algebraic polynomials, where the angular variables are restricted to a subinterval of the period. This entails that approximation in polynomial spaces and in such product spaces are intimately related; cf. e.g. [12, 15, 36]. On the other hand, product approximations are simpler to construct.

Now, a closer look at Fourier extensions and subperiodic trigonometric approximation shows that we are speaking essentially of the same problem, and that the results obtained in one framework can be fruitfully adapted to the other one. Indeed, by the change of variables

$$\theta = \omega\, u\ ,\ \ \omega = \frac{\pi}{T}\ ,\ \ u \in [-1, 1], \tag{9}$$

we can immediately translate an extension problem into a subperiodic problem, and conversely.

In this paper we focus on a semi-discrete approximation in the subperiodic setting, namely *hyperinterpolation*, that is an orthogonal projection onto $\mathbb{T}_n([-\omega, \omega])$, discretized by means of the Gaussian quadrature formula of Proposition 1 for exactness degree $2n$. In view of (9), this can be seen as a kind of discrete Fourier extension.

In order to generate such orthogonal projections, we need some theoretical tools that are recalled in Sects. 2 and 3, that are: the extension of the notion of hyperinterpolation, originally introduced by Sloan in the seminal paper [35] for multivariate total-degree polynomial spaces, to a more general class of spaces, that we term *hyperinterpolation spaces*; the construction of a *subperiodic orthogonal* trigonometric basis. Moreover, in Sect. 3 we also discuss the main computational issues related to subperiodic orthogonality. Finally, in Sect. 4 we discuss the implementation of subperiodic trigonometric hyperinterpolation, together with examples and applications.

2 Hyperinterpolation Spaces

Hyperinterpolation is a powerful tool for total-degree polynomial approximation of multivariate continuous functions, introduced by Sloan in the seminal paper [35]. In brief, it corresponds to a truncated Fourier expansion in a series of orthogonal polynomials for some measure on a given multidimensional domain, where the Fourier coefficients are discretized by a positive algebraic cubature formula. Since then, theoretical as well as computational aspects of hyperinterpolation as an alternative to interpolation have attracted much interest, due to the intrinsic difficulties in finding good multivariate interpolation nodes, with special attention to the case of the sphere; cf., e.g., [16, 25, 26, 41, 42].

In order to generalize the notion of hyperinterpolation, we introduce the idea of *nested hyperinterpolation spaces*. We shall denote by $C(K)$ the space of real-valued continuous functions on a compact set $K \subset \mathbb{R}^d$.

Let $\{S_n\}$ be an increasing sequence of finite-dimensional subspaces $S_n \subset S_{n+1} \subset C(K)$, $n \geq 0$. Moreover, assume that

(i) if $u \in S_n$ and $v \in S_m$ then $uv \in S_{n+m}$;

(ii) the subalgebra $S = \bigcup_{n\geq 0} S_n$ is dense in $C(K)$ with respect to the uniform norm (by the Stone-Weierstrass theorem, if $\overline{S}$ contains the constants the latter is equivalent to the fact that $\overline{S}$ separates points in K, i.e., for every $x \in K$ there exist $u, v \in \overline{S}$ such that $u(x) \neq v(x)$; cf. e.g. [34]);

(iii) we know a sequence of positive quadrature rules $\mu_n = \{(X, \boldsymbol{w})\}$, $n \geq 0$, with nodes $X = \{x_j\}$ and weights $\boldsymbol{w} = \{w_j\}$, $1 \leq j \leq M$, that are exact in S_{2n} for a measure μ on K (for notational convenience we do not display the fact that the nodes X, the weights $\boldsymbol{w}$ and the cardinality M depend on n)

$$\int_K f(x)\, d\mu = \sum_{j=1}^{M} w_j f(x_j)\ , \quad \forall f \in S_{2n}. \tag{10}$$

Let $\{u_j\}$ be a μ-orthonormal basis of S_n, i.e. $S_n = \text{span}\{u_1, \ldots, u_N\}$, $N = N_n = \dim(S_n)$ and

$$(u_i, u_j)_\mu = \int_K u_i(x) u_j(x)\, d\mu = \delta_{ij}. \tag{11}$$

Observe that such an orthonormal basis always exists by the Gram-Schmidt process applied to a given basis of S_n, that in view of (iii) can be performed using either the scalar product $(f, g)_\mu$ or equivalently its discrete counterpart $(f, g)_{\mu_n}$ defined below in (12). It is also worth observing that a quadrature formula like (10) always exists, by a generalized version of Tchakaloff theorem on positive algebraic formulas (whose proof however is not constructive, so that in practice we have to get the formula in some other way); cf. [10] for the polynomial setting and [5, Thm. 5.1] for the extension to more general spaces.

Consider the "discrete inner product" in $C(K)$ (cf. (iii))

$$(f, g)_{\mu_n} = \sum_{j=1}^{M} w_j f(x_j) g(x_j) \tag{12}$$

together with the corresponding seminorm $\|f\|_{\ell^2_w(X)} = \sqrt{(f,f)_{\mu_n}}$, and define the discrete orthogonal projection $\mathscr{L}_n : C(K) \to S_n$

$$\mathscr{L}_n f(x) = \sum_{i=1}^{N} (f, u_i)_{\mu_n}\, u_i(x), \tag{13}$$

which solves the discrete weighted least squares problem

$$\|f - \mathscr{L}_n f\|_{\ell^2_w(X)} = \min_{u \in S_n} \|f - u\|_{\ell^2_w(X)} . \tag{14}$$

This construction was originally proposed in polynomial spaces by Sloan [35] with the name of "hyperinterpolation", namely for $S_n = \mathbb{P}^d_n(K)$ (the space of total-degree polynomials in d real variables of degree not exceeding n, restricted to a compact set or manifold K).

All the relevant properties of the hyperinterpolation operator hold true also in our more general setting. We do not give the proofs, since "mutatis mutandis" they follow exactly the lines of those in [35]. We only observe that a key fact is the coincidence of the original 2-norm and the discrete weighted 2-norm in S_n, in view of (i) and (iii), that is

$$\|u\|_{L^2_\mu(K)} = \|u\|_{\ell^2_w(X)} \ , \quad \forall u \in S_n . \tag{15}$$

It is worth collecting some of the most relevant features of hyperinterpolation in the following

Proposition 2 (cf. [35] for the Polynomial Case) *The hyperinterpolation operator $\mathscr{L}_n$ defined in (13) has the following properties:*

- *$M \geq dim\left(S_n|_{supp(\mu)}\right)$ and if the equality holds, then $\mathscr{L}_n f$ interpolates f at the quadrature nodes;*
- *for every $f \in C(K)$*

$$\|\mathscr{L}_n f\|_{L^2_\mu(K)} \leq \sqrt{\mu(K)}\, \|f\|_{L^\infty(K)} ; \tag{16}$$

- *the $L^2_\mu(K)$ error can be estimated as*

$$\|f - \mathscr{L}_n f\|_{L^2_\mu(K)} \leq 2\sqrt{\mu(K)}\, E_{S_n}(f; K), \tag{17}$$

where $E_{S_n}(f;K) = \inf_{u \in S_n} \left\{\|f - u\|_{L^\infty(K)}\right\}$.

Observe that by the density of the subalgebra $S = \bigcup_{n \geq 0} S_n$, (17) implies $L^2_\mu(K)$-convergence of the sequence of hyperinterpolants, since density is equivalent to $E_{S_n}(f; K) \to 0$, $n \to \infty$. To study $L^\infty(K)$-convergence, an estimate of uniform norm of the operator is needed, along with a Jackson-like theorem for approximation in S_n. The former can be obtained, in general, by estimating in the specific context the reciprocal Christoffel function

$$K_n(x, x) = \sum_{i=1}^{N} u_i^2(x) \ , \tag{18}$$

(i.e., the diagonal of the reproducing kernel $K_n(x,y) = \sum_{i=1}^{N} u_i(x)u_i(y)$ which does not depend on the orthonormal basis), as shown in [17] for total-degree polynomials. Indeed, one can easily prove that

$$\sup_{S_n \ni u \neq 0} \frac{\|u\|_{L^\infty(K)}}{\|u\|_{L^2_\mu(K)}} = \sqrt{\max_{x\in K} K_n(x,x)}, \tag{19}$$

and consequently by (16)

$$\|\mathscr{L}_n\| = \sup_{f\neq 0} \frac{\|\mathscr{L}_n f\|_{L^\infty(K)}}{\|f\|_{L^\infty(K)}} \le \sqrt{\max_{x\in K} K_n(x,x)}\, \sup_{f\neq 0} \frac{\|\mathscr{L}_n f\|_{L^2_\mu(K)}}{\|f\|_{L^\infty(K)}}$$

$$\le \sqrt{\mu(K)\, \max_{x\in K} K_n(x,x)}, \tag{20}$$

cf. [17, Prop. 1.1 and Cor. 1.2].

On the other hand, by the representation

$$\mathscr{L}_n f(x) = \sum_{i=1}^{N} \left(\sum_{j=1}^{M} w_j u_i(x_j) f(x_j) \right) u_i(x) = \sum_{j=1}^{M} f(x_j) w_j K_n(x,x_j), \tag{21}$$

we can obtain an explicit expression for the uniform norm of the hyperinterpolation operator

$$\|\mathscr{L}_n\| = \max_{x\in K} \sum_{j=1}^{M} w_j \left|K_n(x,x_j)\right|. \tag{22}$$

Observe that when $M = N = \dim\left(S_n|_{supp(\mu)}\right)$ and thus $\mathscr{L}_n$ is interpolant, the functions

$$\ell_j(x) = w_j K_n(x,x_j)\ ,\ \ j = 1,\dots,N, \tag{23}$$

are the cardinal functions of interpolation at X in S_n, i.e. $\ell_j(x_k) = \delta_{jk}$, and $\|\mathscr{L}_n\|$ plays the role of the Lebesgue constant of polynomial interpolation.

We stress finally that (20) is usually an overestimate of the actual norm (22). Tighter estimates can be obtained on specific geometries and functional settings, see e.g. the case of the ball [41] and the cube [42] in the polynomial framework.

3 Subperiodic Orthogonality

In this paper we wish to apply the generalized notion of hyperinterpolation of Sect. 2 in the subperiodic trigonometric framework, namely $S_n = \mathbb{T}_n([-\omega, \omega])$, $0 < \omega \leq \pi$. To this end, since the trigonometric Gaussian quadrature formula (7) is at hand (cf. (*iii*)), we need to find a subperiodic orthonormal basis.

We can now state and prove the following Proposition (see [11] for a preliminary version)

Proposition 3 *An orthonormal basis in $L^2(-\omega, \omega)$ for the $(2n+1)$-dimensional space $\mathbb{T}_n([-\omega, \omega])$ is given by the trigonometric polynomials*

$$\tau_i(\theta) = \tau_i(\theta, \omega)\ , \quad i = 0, 1, \ldots, 2n, \tag{24}$$

$$\tau_{2k}(\theta) = p_{2k}\left(\frac{\sin(\theta/2)}{\sin(\omega/2)}\right)\ , \quad k = 0, \ldots, n \tag{25}$$

where $\{p_j\}_{j\geq 0}$ are the algebraic orthonormal polynomials with respect to the weight function $W_{-1/2}(t)$ in (6) and

$$\tau_{2k-1}(\theta) = \cos(\theta/2)\, q_{2k-1}\left(\frac{\sin(\theta/2)}{\sin(\omega/2)}\right)\ , \quad k = 1, \ldots, n \tag{26}$$

where $\{q_j\}_{j\geq 0}$ are the algebraic orthonormal polynomials with respect to the weight function

$$W_{1/2}(t) = 2\sin(\omega/2)\sqrt{1 - \sin^2(\omega/2)\, t^2}\ , \quad t \in (-1, 1). \tag{27}$$

Proof First, observe that by basic trigonometric identities the functions τ_{2k} are a basis for the even trigonometric polynomials, whereas the functions τ_{2k-1} are a basis for the odd trigonometric polynomials. Indeed, an even power of a sine is a combination of cosines with frequencies up to the exponent, whereas an odd power of a sine is a combination of sines with frequencies up to the exponent, and $\cos(\theta/2)\sin(j\theta/2) = \frac{1}{2}\sin((j+1)\theta/2) + \sin((j-1)\theta/2)$ is a trigonometric polynomial of degree $(j+1)/2$ for odd j.

From the definition and the change of variable (4) it follows directly that τ_i has unit L^2-norm for even i, and that τ_i and τ_j with even i and j, $i \neq j$, are mutually orthogonal, whereas when i is even and j odd, or conversely, they are mutually orthogonal since their product is an odd function. To show that they are orthonormal also for odd i and j, write

$$\int_{-\omega}^{\omega} \tau_i(\theta)\tau_j(\theta)\, d\theta = \int_{-\omega}^{\omega} q_i\left(\frac{\sin(\theta/2)}{\sin(\omega/2)}\right) q_j\left(\frac{\sin(\theta/2)}{\sin(\omega/2)}\right) \cos^2(\theta/2)\, d\theta.$$

By the change of variable $\theta = 2\arcsin(\sin(\omega/2)t)$ we get

$$\int_{-\omega}^{\omega} \tau_i(\theta)\tau_j(\theta)\, d\theta = \int_{-1}^{1} q_i(t)q_j(t)\,(1-\sin^2(\omega/2)\,t^2)\,\frac{2\sin(\omega/2)}{\sqrt{1-\sin^2(\omega/2)\,t^2}}\, dt$$

$$= \int_{-1}^{1} q_i(t)q_j(t)\, W_{1/2}(t)\, dt = \delta_{ij}. \quad \square$$

□

We discuss now how to implement the computation of the subperiodic orthogonal basis of Proposition 3, since this is the base for an algorithm that constructs subperiodic trigonometric hyperinterpolants.

Concerning univariate algebraic orthogonal polynomials, one of the most comprehensive and reliable tools is the Matlab OPQ suite by Gautschi [22]. For example, inside OPQ one finds the routine `chebyshev`, that computes in a stable way the recurrence coefficients for orthogonal polynomials with respect to a given measure by the modified Chebyshev algorithm, as soon as the modified Chebyshev moments (i.e., the moments of the Chebyshev basis with respect to the given measure) are known. As shown in [21, 22], the modified Chebyshev algorithm computes the recurrence coefficients for the orthogonal polynomials up to degree $k+1$ using the modified Chebyshev moments up to degree $2k+1$ (in our application $k=2n$ is needed).

Our first step is then to compute the modified Chebyshev moments up to degree $2n$ for the weight functions $W_{-1/2}$ in (6) and $W_{1/2}$ in (27), namely

$$m_{2j} = 2\sin(\omega/2)\int_{-1}^{1} T_{2j}(t)\,(1-\sin^2(\omega/2)\,t^2)^{\beta}\, dt\ ,\ \ j=0,\dots,2n, \tag{28}$$

where $\beta = -1/2$ or $\beta = 1/2$; observe that only the even moments have to be computed, since the weight functions are even and thus the odd moments vanish.

More generally, in [33] it is proved that the sequence of moments

$$I_{2j}(\alpha,\beta,s) = \int_{-1}^{1} T_{2j}(t)\,(\alpha+s^2+t^2)^{\beta}\, dt, \tag{29}$$

where $\alpha \in \mathbb{C}$, $s,\beta \in \mathbb{R}$, satisfies the recurrence relation

$$\left(\frac{1}{4}+\frac{\beta+1}{2(2j+1)}\right) I_{2j+2} + \left(\frac{1}{2}+\alpha^2+s^2-\frac{\beta+1}{4j^2-1}\right) I_{2j}$$

$$+\left(\frac{1}{4}-\frac{\beta+1}{2(2j-1)}\right) I_{2j-2} = -\frac{2(1+\alpha^2+s^2)^{\beta+1}}{4j^2-1}\ ,\ \ j\geq 1. \tag{30}$$

Now, setting $\alpha = i/\sin(\omega/2)$ where i is the imaginary unit ($i^2 = -1$), $s = 0$ and $\beta = \pm 1/2$, we have that $m_{2j} = 2\sin^{2\beta+1}(\omega/2)(-1)^{-\beta} I_{2j}(i/\sin(\omega/2), \beta, 0)$, from which follows that m_{2j} satisfies the recurrence relation

$$a_j m_{2j+2} + b_j m_{2j} + c_j m_{2j-2} = d_j \ , \ \ j \geq 1, \tag{31}$$

where

$$a_j = \left(\frac{1}{4} + \frac{\beta+1}{2(2j+1)}\right) \ , \ \ b_j = \left(\frac{1}{2} - \frac{1}{\sin^2(\omega/2)} - \frac{\beta+1}{4j^2-1}\right),$$

$$c_j = \left(\frac{1}{4} - \frac{\beta+1}{2(2j-1)}\right) \ , \ \ d_j = 4\,\frac{\cos^{2\beta+2}(\omega/2)}{\sin(\omega/2)}\,\frac{1}{4j^2-1}. \tag{32}$$

Such a recurrence is however unstable, namely small errors on the starting values grow very rapidly increasing j. In order to stabilize it we have adopted the method that solves instead a linear system, with tridiagonal diagonally dominant matrix (of the recurrence coefficients) and the vector $(d_1 - c_1 m_0, d_2, \ldots, d_{2n-2}, d_{2n-1} - a_{2n-1} m_{4n})$ as right-hand side; cf. [11, 20]. We get immediately that $m_0 = 2\omega$ for $\beta = -1/2$ and $m_0 = \omega + \sin(\omega)$ for $\beta = 1/2$, whereas the last moment m_{4n} can be computed accurately in both cases by the `quadgk` Matlab function (adaptive Gauss-Kronrod quadrature).

Since the `chebyshev` routine, starting from the modified Chebyshev moments, returns the recurrence coefficients for the *monic* orthogonal polynomials, we have to modify the recurrence relation in the standard way (cf. [21, Thm. 1.29]) to get the orthonormal polynomials $\{p_{2k}\}$ for $W_{-1/2}$ and $\{q_{2k-1}\}$ for $W_{1/2}$, from which we compute the orthonormal subperiodic trigonometric basis $\{\tau_0, \ldots, \tau_{2n}\}$ as in Proposition 3.

On the other hand, it turns out numerically (in double precision) that there is a moderate loss of orthogonality when n increases. Defining the Vandermonde-like matrix

$$V = V_n(\Theta, \omega) = (v_{ij}) = (\tau_{j-1}(\varphi_i)) \ , \ \ 1 \leq i,j \leq 2n+1, \tag{33}$$

where $\Theta = \{\varphi_i\}$ are the $2n+1$ quadrature nodes for trigonometric exactness degree $2n$ (see Proposition 1), we can then measure numerical orthogonality of the basis $\{\tau_{j-1}\}$ by computing $\varepsilon_n = \|(\sqrt{\Lambda}V)^t(\sqrt{\Lambda}V) - I\|_2$, where $\Lambda = \text{diag}(\lambda_i)$ is the diagonal matrix of the quadrature weights. For example, we get $\varepsilon_{250} \approx 2 \cdot 10^{-13}$. In order to recover orthogonality at machine precision, it is sufficient to re-orthogonalize the basis by

$$\sqrt{\Lambda}\,V = QR \ , \ \ (\hat{\tau}_0(\theta), \ldots, \hat{\tau}_{2n}(\theta)) = (\tau_0(\theta), \ldots, \tau_{2n}(\theta))R^{-1}. \tag{34}$$

In such a way, we can eventually compute in a stable way the orthonormal trigonometric basis of Proposition 3. All the relevant codes are available in the

Matlab package HYPERTRIG [38]. Notice that we could not have applied the QR based orthonormalization directly to the Vandermonde matrix in the canonical trigonometric basis (3), since such a matrix turns out to be extremely ill-conditioned already at moderate values of n for $\omega < \pi$. On the contrary, the Vandermonde-like matrix $\sqrt{\Lambda}V$ in (34) being quasi-orthogonal has a condition number very close to 1, and thus $Q = \sqrt{\Lambda}\,VR^{-1}$ is orthogonal at machine precision.

4 Subperiodic (Hyper)interpolation

By the tools developed in the previous sections, we can now construct a subperiodic trigonometric hyperinterpolation operator, whose properties are an immediate consequence of Proposition 2 with $S_n = \mathbb{T}_n([-\omega,\omega])$. Notice that property (i) of hyperinterpolation spaces is immediate (n being the trigonometric degree) and concerning (ii) it is not difficult to show that subperiodic trigonometric polynomials separate points.

Corollary 1 *Consider the subperiodic trigonometric hyperinterpolation operator $\mathscr{L}_n : C([-\omega,\omega]) \to \mathbb{T}_n([-\omega,\omega])$, $0 < \omega \leq \pi$, defined as*

$$\mathscr{L}_n f(\theta) = \sum_{i=0}^{2n} (f,\tau_i)_{\mu_n}\, \tau_i(\theta), \theta \in [-\omega,\omega], (f,\tau_i)_{\mu_n} = \sum_{j=1}^{2n+1} \lambda_j f(\varphi_j)\tau_i(\varphi_j), \tag{35}$$

where $\{\tau_i\}$ is the orthonormal basis of Proposition 3 and $\{(\varphi_j,\lambda_j)\}$ are the nodes and weights of the subperiodic Gaussian formula of Proposition 1 for degree $2n$.

Then, the following properties hold

- *$\mathscr{L}_n f$ interpolates f at the $2n+1 = dim(\mathbb{T}_n([-\omega,\omega]))$ quadrature nodes*

$$\mathscr{L}_n f(\varphi_j) = f(\varphi_j)\ , \quad 1 \leq j \leq 2n+1; \tag{36}$$

- *the L^2-error can be estimated as*

$$\|f - \mathscr{L}_n f\|_{L^2([-\omega,\omega])} \leq 2\sqrt{2\omega}\, E_{\mathbb{T}_n([-\omega,\omega])}(f). \tag{37}$$

We can now give an estimate of the uniform norm of the hyperinterpolation operator, which we may call its Lebesgue constant since by Corollary 1 it is an interpolation operator.

Corollary 2 *The Lebesgue constant of the hyperinterpolation operator $\mathscr{L}_n$ of Corollary 1 can be estimated as*

$$\|\mathscr{L}_n\| \leq C_n \sim 2\sqrt{\pi}\, n + \frac{4}{\sqrt{3}} n^{3/2},$$

$$C_n = \sqrt{\pi}(2n+1) + \sqrt{\frac{(2n+1)(2n+2)(4n+3)}{3}}. \tag{38}$$

Proof In view of (20) with $K = [-\omega, \omega]$ and Proposition 3, we are reduced to estimate the trigonometric reciprocal Christoffel function. Setting $t(\theta) = \sin(\theta/2)/\sin(\omega/2) \in [-1, 1]$, we have

$$K_n(\theta,\theta) = \sum_{i=0}^{2n} \tau_i^2(\theta) = \sum_{even\ i} p_i^2(t(\theta)) + \cos^2(\theta/2) \sum_{odd\ i} q_i^2(t(\theta))$$

$$\leq \sum_{i=0}^{2n} p_i^2(t(\theta)) + \sum_{i=0}^{2n} q_i^2(t(\theta)) = \lambda_{2n}^{-1}(t(\theta); W_{-1/2}) + \lambda_{2n}^{-1}(t(\theta); W_{1/2}), \tag{39}$$

where $\lambda_m^{-1}(t; W)$ denotes the reciprocal Christoffel function for algebraic degree n and weight function $W \in L_+^1(-1, 1)$.

Now, in view of the basic property of monotonicity of Christoffel functions with respect to the underlying measure (cf., e.g., [31, Ch. 6]), we have that $\lambda_m^{-1}(t; W)$ is decreasing in W, in the sense that if $W_1 \geq W_2$ a.e., then $\lambda_m^{-1}(t; W_1) \leq \lambda_m^{-1}(t; W_2)$. Observing that $W_{1/2}(t) \geq 2\sin(\omega/2)\sqrt{1-t^2}$ and $W_{-1/2}(t) \geq 2\sin(\omega/2)$, $t \in (-1, 1)$, we get that the corresponding reciprocal Christoffel functions are bounded, up to a scaling by the factor $(2\sin(\omega/2))^{-1}$, by the reciprocal Christoffel functions of the Chebyshev measure of the second kind (sum of squares of the Chebyshev polynomials of the second kind) and of the Lebesgue measure (sum of squares of the Legendre polynomials), respectively. By well-known estimates for such reciprocal Christoffel functions (cf., e.g., [17]), we get

$$\lambda_{2n}^{-1}(t; W_{-1/2}) \leq \frac{(2n+1)^2}{4\sin(\omega/2)}, \lambda_{2n}^{-1}(t; W_{1/2}) \leq \frac{(2n+1)(2n+2)(4n+3)}{6\pi\sin(\omega/2)}, \tag{40}$$

and thus

$$\|\mathscr{L}_n\| \leq \sqrt{2\omega \max_{\theta\in[-\omega,\omega]} K_n(\theta,\theta)}$$

$$\leq 2\sqrt{\frac{\omega/2}{\sin(\omega/2)}} \left(\frac{2n+1}{\sqrt{2}} + \sqrt{\frac{(2n+1)(2n+2)(4n+3)}{6\pi}} \right), \tag{41}$$

from which (38) follows since the function $y/\sin(y)$ is increasing and bounded by $\pi/2$ for $y = \omega/2 \in [0, \pi/2]$. □ □

Even though (38) is clearly an overestimate of the actual growth, it provides a bound independent of ω and shows that the Lebesgue constant is slowly increasing with n. By (38) and the fact that $\mathscr{L}_n$ is a projection operator, we easily get the uniform error estimate

$$\|f - \mathscr{L}_n f\|_{L^\infty([-\omega,\omega])} \leq (1 + C_n)\, E_{\mathbb{T}_n([-\omega,\omega])}(f) \ , \quad \forall f \in C([-\omega,\omega]). \tag{42}$$

In Fig. 1, we have plotted the Lebesgue constant computed numerically by (22) on a fine control grid in $[-\omega, \omega]$ for some values of ω, that is

$$\Lambda_n(\omega) = \|\mathscr{L}_n\| = \max_{\theta \in [-\omega,\omega]} \sum_{j=1}^{2n+1} |\ell_j(\theta)| \ , \tag{43}$$

where

$$\ell_j(\theta) = \lambda_j K_n(\theta, \varphi_j) \ , \quad 1 \leq j \leq 2n+1 \ , \quad K_n(\theta, \phi) = \sum_{i=0}^{2n} \tau_i(\theta)\tau_i(\phi). \tag{44}$$

Observe that the Lebesgue constant appears to be decreasing in ω for fixed n and to converge to the Lebesgue constant of algebraic interpolation of degree $2n$ at the Gauss-Legendre nodes (for $\omega \to 0$), which as known is $\mathscr{O}(\sqrt{n})$ (upper solid line). This also shows that the bound (38) is a large overestimate of the

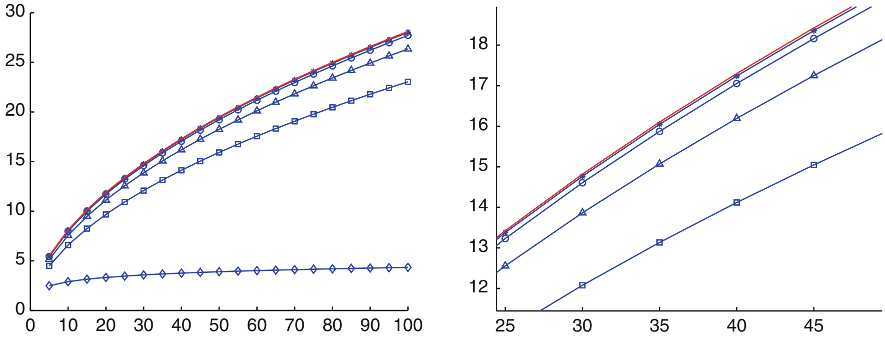


Fig. 1 The Lebesgue constant (43) for $n = 5, 10, \ldots, 95, 100$ at some values of ω: from below, $\omega = \pi$ (diamonds), $\omega = 3\pi/4$ (squares), $\omega = \pi/2$ (triangles), $\omega = \pi/4$ (circles), $\omega = \pi/8$ (asterisks); right: detail for $n = 25, \ldots, 45$. The upper solid line is the Lebesgue constant of algebraic interpolation of degree $2n$ at the Gauss-Legendre nodes

actual values. It is also worth observing that for $\omega = \pi$ the Lebesgue constant is exactly that of algebraic interpolation of degree $2n$ at the Gauss-Chebyshev nodes, that is logarithmic in n, cf. [13].

Based on these and other numerical experiments that we do not report for brevity, we can then make the following

Conjecture 1 The Lebesgue constant $\Lambda_n(\omega)$ of trigonometric hyperinterpolation in $[-\omega, \omega]$, $0 < \omega \leq \pi$, cf. (35) and (43), is a decreasing function of ω for fixed degree. Moreover, its limit for $\omega \to 0$ (that is its supremum being bounded by (38)) is the Lebesgue constant of algebraic interpolation of degree $2n$ at the Gauss-Legendre nodes.

In order to show the performance of subperiodic trigonometric hyperinterpolation, in Fig. 2 we have reported the errors in the uniform norm on six test functions

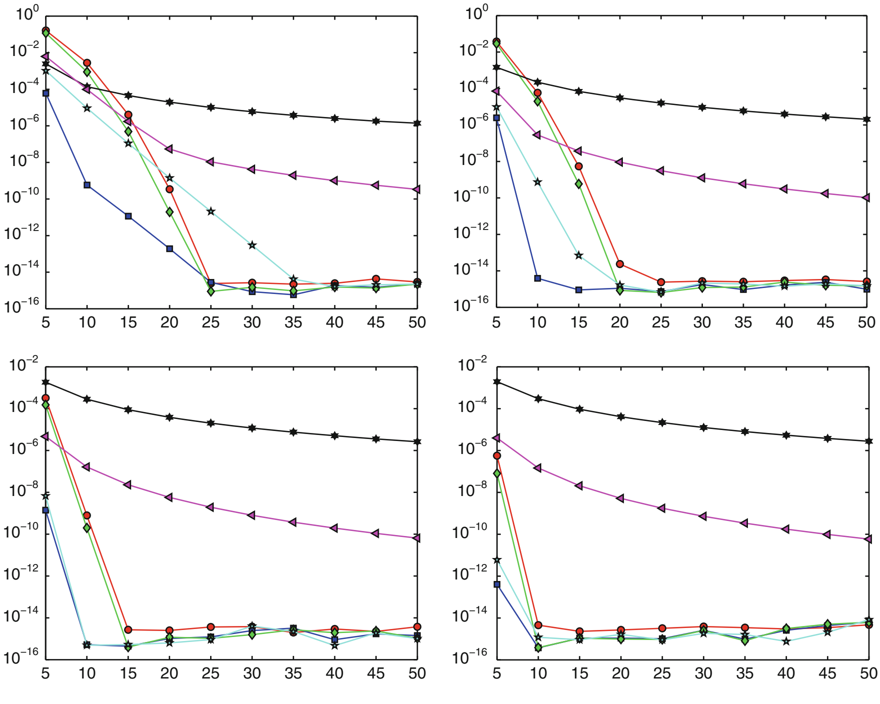


Fig. 2 Relative ℓ^2-errors of subperiodic trigonometric hyperinterpolation for degrees $n = 5, 10, \ldots, 50$ on the test functions (45) on $[-\omega, \omega]$ with $\omega = 3\pi/4$ (top-left), $\omega = \pi/2$ (top-right), $\omega = \pi/4$ (bottom-left) and $\omega = \pi/8$ (bottom-right): f_1 (bullets), f_2 (squares), f_3 (diamonds), f_4 (stars), f_5 (asterisks), f_6 (triangles)

with different regularity for some values of ω, namely

$$f_1(\theta) = (2 + \cos(\theta) + \sin(\theta))^{30}\ ,\ f_2(\theta) = \exp(-\theta^2)\ ,\ f_3(\theta) = \exp(-5\theta^2),$$

$$f_4(\theta) = \frac{1}{1 + 25(\theta/\omega)^2}\ ,\ f_5(\theta) = |\theta|^{5/2}\ ,\ f_6(\theta) = (\omega - \theta)^{5/2}, \tag{45}$$

that are a positive trigonometric polynomial, two Gaussians centered at $\theta = 0$, a Runge-like function, a function with a singularity of the third derivative at $\theta = 0$ (where the nodes do not cluster), and one with a singularity of the third derivative at $\theta = \omega$ (where the nodes cluster). The errors are measured in the relative ℓ_2-norm on a fine control grid in $[-\omega, \omega]$.

We see that convergence on the smooth functions f_1, f_2, f_3, f_4 (that are analytic) is faster by decreasing the interval length (in particular, the trigonometric polynomial f_1 is recovered at machine precision below the theoretical exactness degree $n = 30$), whereas the interval length has a substantial effect only at low degrees on the singular functions f_5, f_6 (observe that the clustering of sampling nodes at the singularity entails a faster convergence for f_6 compared to f_5).

The dependence of the convergence rate on ω for analytic functions is not surprising. Indeed, interpreting subperiodic trigonometric hyperinterpolation as a (discretized) Fourier extension, as discussed in the Introduction (cf. (2)–(9)), we may resort to deep convergence results in that theory. For example, in [1, Thm. 2.3] it is proved that the uniform convergence rate of Fourier extensions, and thus also that of the best trigonometric approximation in the relevant space if one uses (42), for "sufficiently" analytic functions (complex singularities not "too close" to the approximation interval) is (at least) exponential with order $\mathscr{O}(E(T)^{-n})$, where $E(T) = \cot^2(\pi/(4T)) = \cot^2(\omega/4)$ is a decreasing function of ω and $E(T) \to +\infty$ as $\omega \to 0^+$. This might give an explanation of the error behavior observed in Fig. 2. We do not pursue further this aspect and refer the reader to [1] for a complete discussion on the convergence features of Fourier extensions.

To conclude, we observe that we can easily extend all the constructions above to any angular interval $[\alpha, \beta] \ni \theta$, $\beta - \alpha \leq 2\pi$, by the change of variable

$$\theta' = \theta - \frac{\alpha + \beta}{2} \in [-\omega, \omega]\ ,\ \ \omega = \frac{\beta - \alpha}{2}, \tag{46}$$

namely using the orthonormal basis

$$t_i(\theta) = t_i(\theta, \alpha, \beta) = \tau_i(\theta', \omega)\ ,\ \ 0 \leq i \leq 2n, \tag{47}$$

and the subperiodic Gaussian quadrature formula with nodes and weights

$$\{(\theta_j, \lambda_j)\}\ ,\ \ \theta_j = \varphi_j + \frac{\alpha + \beta}{2}\ ,\ \ 1 \leq j \leq 2n + 1, \tag{48}$$

cf. (35).

5 Product (Hyper)interpolation on Spherical Rectangles

The general setting of Sect. 2 allows to extend immediately subperiodic trigonometric hyperinterpolation to the tensor-product case. Indeed, consider the product basis

$$\{u_i(\theta)v_j(\phi)\},\ 0 \le i,j \le 2n,\ (\theta,\phi) \in K = I_1 \times I_2 = [\alpha_1,\beta_1] \times [\alpha_2,\beta_2], \tag{49}$$

where $u_i(\theta) = t_i(\theta,\alpha_1,\beta_1)$ and $v_j(\phi) = t_j(\phi,\alpha_2,\beta_2)$, cf. (47). Clearly (49) is a L^2-orthonormal basis of the tensor-product subperiodic trigonometric space

$$S_n = \mathbb{T}_n(I_1) \otimes \mathbb{T}_n(I_2). \tag{50}$$

Then, we can construct the product hyperinterpolant of $f \in C(I_1 \times I_2)$ as

$$\mathscr{L}_n f(\theta,\phi) = \sum_{i,j=0}^{2n} c_{ij}\, u_i(\theta)v_j(\phi), \tag{51}$$

with

$$c_{ij} = \sum_{h,k=1}^{2n+1} \lambda_{h1}\lambda_{k2} f(\theta_{h1},\theta_{k2})\, u_i(\theta_{h1})v_j(\theta_{k2}), \tag{52}$$

where $\{(\theta_{is},\lambda_{is})\}$ are the angular nodes and weights of the subperiodic trigonometric Gaussian formula on I_s, $s = 1,2$, for exactness degree $2n$ (cf. Proposition 1 and (48)).

All the relevant properties of hyperinterpolation apply, in particular the hyperinterpolant is a (product) interpolant (see also [32]). Moreover, the Lebesgue constant is the product of the one-dimensional constants, $\Lambda_n(\omega_1,\omega_2) = \Lambda_n(\omega_1)\Lambda_n(\omega_2)$, and the following error estimates hold

$$\|f - \mathscr{L}_n f\|_{L^2(I_1\times I_2)} \le 4\sqrt{\omega_1\omega_2}\, E_{S_n}(f; I_1 \times I_2),$$

$$\|f - \mathscr{L}_n f\|_{L^\infty(I_1\times I_2)} \le (1 + \Lambda_n(\omega_1)\Lambda_n(\omega_2))\, E_{S_n}(f; I_1 \times I_2). \tag{53}$$

Concerning the implementation of subperiodic product hyperinterpolation, this can be constructed in a simple matrix form working on grids. Indeed, let

$$V_1 = (u_{j-1}(\theta_{i1}))\ ,\ \ V_2 = (v_{j-1}(\theta_{i2}))\ ,\ \ D_s = diag(\lambda_{is})\ ,\ \ s = 1,2,$$

$$F = (f(\theta_{i1},\theta_{j2}))\ ,\ \ 1 \le i,j \le 2n+1, \tag{54}$$

be the univariate Vandermonde-like matrices at the hyperinterpolation nodes in I_s, the diagonal matrices of the quadrature weights and the matrix of the function values at the bivariate hyperinterpolation grid $\{\theta_{i1}\} \times \{\theta_{j2}\}$, respectively. Moreover, let

$$U_1 = (u_{j-1}(\hat{\theta}_\ell)) \ , \ \ U_2 = (v_{j-1}(\hat{\phi}_t)) \ , \ \ 1 \leq \ell \leq m_1 \ , \ \ 1 \leq t \leq m_2, \tag{55}$$

be univariate Vandermonde-like matrices at the points $\{\hat{\theta}_\ell\} \subset I_1$ and $\{\hat{\phi}_t\} \subset I_2$, respectively.

Then, it is easy to check that the hyperinterpolation coefficient matrix and the values of the hyperinterpolant at the target grid $\{\hat{\theta}_\ell\} \times \{\hat{\phi}_t\}$ can be computed by matrix products as

$$C = (c_{ij}) = V_1^t D_1 F D_2 V_2 \ , \ \ L = (\mathscr{L}_n f(\hat{\theta}_\ell, \hat{\phi}_t)) = U_1 C U_2^t. \tag{56}$$

Subperiodic product hyperinterpolation via (56) has been implemented in the Matlab package [38].

Subperiodic product hyperinterpolation can be used, for example, to recover functions on spherical rectangles, in applications that require local approximation models. It is worth recalling that hyperinterpolation-like trigonometric approximation on the whole sphere, with applications in scattering theory, has been studied, e.g., in [18, 19].

Consider the spherical coordinates

$$(x, y, z) = \sigma(\theta, \phi) = (\cos(\theta)\sin(\phi), \sin(\theta)\sin(\phi), \cos(\phi)), \tag{57}$$

where θ is the azimuthal angle and ϕ the polar angle, $(\theta, \phi) \in [-\pi, \pi] \times [0, \pi]$, and a "geographic rectangle", that is

$$\Omega = \sigma(I_1 \times I_2) \ , \ \ I_1 = [\alpha_1, \beta_1] \subseteq [-\pi, \pi] \ , \ \ I_2 = [\alpha_2, \beta_2] \subseteq [0, \pi]. \tag{58}$$

Now, take a function $g \in C(\Omega)$, that we can identify with a continuous function in $I_1 \times I_2$ as $f(\theta, \phi) = g(\sigma(\theta, \phi))$. In order to estimate the hyperinterpolation errors in (53), we can observe that if $p \in \mathbb{P}_n^3(\Omega)$ then $p \circ \sigma \in S_n = \mathbb{T}_n(I_1) \otimes \mathbb{T}_n(I_2)$. Then, due to the surjectivity of the map σ, we have

$$\inf_{\psi \in S_n} \|f - \psi\|_{L^\infty(I_1 \times I_2)} \leq \inf_{p \in \mathbb{P}_n^3(\Omega)} \|f - p \circ \sigma\|_{L^\infty(I_1 \times I_2)}$$

$$= \inf_{p \in \mathbb{P}_n^3(\Omega)} \|g \circ \sigma - p \circ \sigma\|_{L^\infty(I_1 \times I_2)} = \inf_{p \in \mathbb{P}_n^3(\Omega)} \|g - p\|_{L^\infty(\Omega)},$$

that is

$$E_{S_n}(f; I_1 \times I_2) \leq E_{\mathbb{P}_n^3(\Omega)}(g; \Omega). \tag{59}$$

Moreover, it is also clear that subperiodic product trigonometric hyperinterpolation reproduces total-degree polynomials, namely

$$\mathscr{L}_n(p \circ \sigma) = p \circ \sigma \ , \ \ \forall p \in \mathbb{P}_n^3(\varOmega). \tag{60}$$

It is also worth observing that the orthonormal basis functions $\{u_i(\theta)v_j(\phi)\}$, and thus also the hyperinterpolant $\mathscr{L}_n f(\theta, \phi)$, correspond to continuous spherical functions on $\varOmega$, whenever $\varOmega$ does not contain the north or south pole (i.e., $[\alpha_2, \beta_2] \subset (0, \pi)$ or in other words $\varOmega$ is a nondegenerate rectangle).

To make an example, we have taken two geographic rectangles of the unit sphere. The first

$$\varOmega_1 = \sigma\left(\left[-\frac{125}{180}\pi, -\frac{67}{180}\pi\right] \times \left[\frac{41}{180}\pi, \frac{65}{180}\pi\right]\right), \ \omega_{11} \approx 0.506 \ , \ \omega_{12} \approx 0.209, \tag{61}$$

corresponds in standard longitude-latitude to 67°W–125°W, 25°N–49°N, a vaste rectangle approximately corresponding to the contiguous continental USA, whereas the second

$$\varOmega_2 = \sigma\left(\left[-\frac{109}{180}\pi, -\frac{102}{180}\pi\right] \times \left[\frac{49}{180}\pi, \frac{53}{180}\pi\right]\right), \ \omega_{21} \approx 0.061 \ , \ \omega_{22} \approx 0.035, \tag{62}$$

is the rectangle 102°W–109°W, 37°N–41°N, corresponding to Colorado.

In order to test the polynomial reproduction property, we have taken the positive test polynomials

$$p_n(x, y, z) = (ax + by + cz + 3)^n, \tag{63}$$

where a, b, c are random variables uniformly distributed in $[0, 1]$. In Fig. 3 we have reported the relative ℓ^2-errors (average of 100 samples) in the reconstruction of the polynomials by hyperinterpolation, computed on a 50 × 50 control grid. In particular, in Fig. 3-right we see that the reconstruction of a fixed polynomial an overprecision phenomenon occurs, more pronounced with the smaller rectangle (where near exactness is obtained already at half the polynomial degree). This may be interpreted by the dependence on ω of the convergence rate of univariate subperiodic hyperinterpolation for analytic functions, as discussed above (cf. the convergence profile for f_1 in Fig. 2).

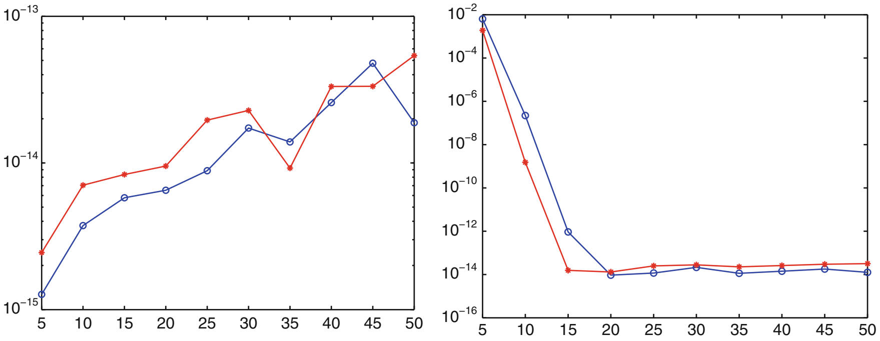


Fig. 3 Left: Average relative ℓ^2-errors of subperiodic trigonometric hyperinterpolation for degrees $n = 5, 10, \dots, 50$ on the random test polynomials p_n in (63), with spherical rectangles corresponding to USA (circles) and Colorado (asterisks). Right: Average relative ℓ^2-errors in the reconstruction of the fixed polynomial p_{30}

It is worth stressing that the fact of being on a sphere is not essential, since similar results can be obtained for example also on rectangles of the torus, with angular intervals (in the usual poloidal-toroidal coordinates) of the same length of those in (61)–(62). On the other hand, subperiodic trigonometric hyperinterpolation could be useful also to construct mixed algebraic-trigonometric product formulas for solid sections of the sphere such as (truncated) spherical sectors with rectangular base, and also for planar circular sections, such as sectors, zones, lenses, lunes (via the transformations used in [14, 15, 37]).

6 Comparison with Polynomial Hyperinterpolation

In a recent paper [24], hyperinterpolation on geographic rectangles has been studied and implemented in the usual hyperinterpolation setting of total-degree polynomials. As known, the dimension of the underlying polynomial space on the sphere is $(n + 1)^2$ for degree n. Spherical harmonics, however, are no more an orthogonal basis on a portion of the sphere, so that a costly orthonormalization process has to be applied, based on the availability of algebraic quadrature formulas exact at degree $2n$ with $(2n + 1)(2n + 2)$ nodes and positive weights. Such an orthonormalization cost is much larger than the cost of the present approach, since here orthonormalization is univariate in the components. Moreover, a substantial drawback of polynomial hyperinterpolation on geographic rectangles is that the orthonormalization process suffers from severe ill-conditioning (of the relevant Vandermonde-like matrices) already at moderate degrees, cf. [24].

A good feature of polynomial hyperinterpolation is that we work by construction with continuous spherical functions, and this allows to have rectangles containing the north or south pole and even to work with spherical polar caps (with the appropriate transformation, cf. [24]). The number of sampling points on general rectangles is slightly bigger than that of the present approach, $(2n + 1)(2n + 2)$ versus $(2n + 1)^2$, whereas the number of coefficients is smaller, namely $(n + 1)^2$ versus $(2n + 1)^2$. On the other hand, we expect smaller reconstruction errors from subperiodic product trigonometric hyperinterpolation, since we work here in a bigger space, indeed if $p \in \mathbb{P}_n^3(\Omega)$ then $p \circ \sigma \in S_n = \mathbb{T}_n(I_1) \otimes \mathbb{T}_n(I_2)$. Moreover, the hyperinterpolant is interpolant in the present context, whereas it is not in the polynomial case.

In order to make a numerical comparison of polynomial with superiodic product hyperinterpolation, we have considered the functions

$$g_1(P) = \exp\left(-5\| P - P_0\|_2^2\right)\ ,\quad g_2(P) = \| P - P_0\|_2^5\ ,\quad P = (x, y, z), \tag{64}$$

on the two rectangles above corresponding to USA (61) and Colorado (62), $P_0 = (x_0, y_0, z_0)$ being the center of the rectangle (where the sampling points do not cluster). Notice that g_1 is smooth whereas g_2 has a singularity at P_0. In Fig. 4 we have reported the relative ℓ^2-errors in the reconstruction of g_1 and g_2, computed on a 50 × 50 control grid. We see that subperiodic trigonometric hyperinterpolation is more accurate than polynomial hyperinterpolation (with essentially the same number of sampling points and a much lower computational cost).

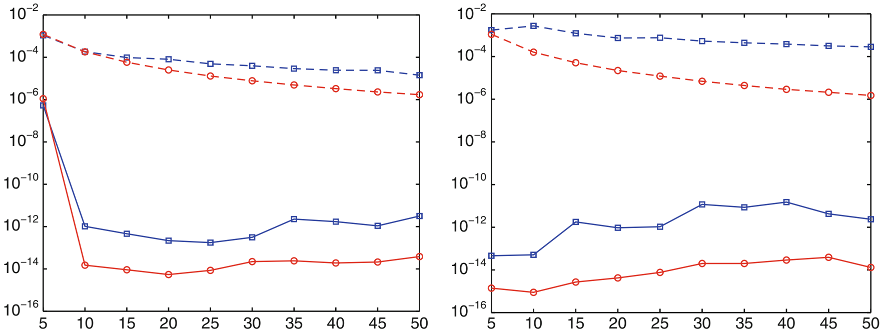


Fig. 4 Average relative ℓ^2-errors of polynomial (squares) and subperiodic trigonometric (circles) hyperinterpolation for degrees $n = 5, 10, \ldots, 50$ on the test functions g_1 (solid line) and g_2 (dashed line) in (64), with spherical rectangles corresponding to USA (left) and Colorado (right)

Acknowledgements Supported by the Horizon 2020 ERA-PLANET European project "GEOEssential", by the EU project H2020-MSCA-RISE-2014-644175-MATRIXASSAY, by the DOR funds and by the biennial projects CPDA143275 and BIRD163015 of the University of Padova, and by the GNCS-INdAM.

References

1. Adcock, B., Huybrechs, D.: On the resolution power of Fourier extensions for oscillatory functions. J. Comput. Appl. Math. **260**, 312–336 (2014)
2. Adcock, B., Platte, R.: A mapped polynomial method for high-accuracy approximations on arbitrary grids. SIAM J. Numer. Anal. **54**, 2256–2281 (2016)
3. Adcock, B., Ruan, J.: Parameter selection and numerical approximation properties of Fourier extensions from fixed data. J. Comput. Phys. **273**, 453–471 (2014)
4. Adcock, B., Huybrechs, D., Vaquero, J.M.: On the numerical stability of Fourier extensions. Found. Comput. Math. **14**, 635–687 (2014)
5. Berschneider, G., Sasvri, Z.: On a theorem of Karhunen and related moment problems and quadrature formulae, Spectral theory, mathematical system theory, evolution equations, differential and difference equations. Oper. Theory Adv. Appl. **221**, 173–187 (2012)
6. Borwein, P., Erdélyi, T.: Polynomials and Polynomial Inequalities. Springer, New York (1995)
7. Bos, L., Vianello, M.: Subperiodic trigonometric interpolation and quadrature. Appl. Math. Comput. **218**, 10630–10638 (2012)
8. Boyd, J.P.: A comparison of numerical algorithms for Fourier extension of the first, second, and third kinds. J. Comput. Phys. **178**, 118–160 (2002)
9. Bruno, O.P., Han, Y., Pohlman, M.M.: Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis. J. Comput. Phys. **227**, 1094–1125 (2007)
10. Curto, R.E., Fialkow, L.A.: A duality proof of Tchakaloff's theorem. J. Math. Anal. Appl. **269**, 519–532 (2002)
11. Da Fies, G.: Some results on subperiodic trigonometric approximation and quadrature. Master Thesis in Mathematics (advisor: Vianello, M.), University of Padova (2012)
12. Da Fies, G., Vianello, M.: Trigonometric Gaussian quadrature on subintervals of the period. Electron. Trans. Numer. Anal. **39**, 102–112 (2012)
13. Da Fies, G., Vianello, M.: On the Lebesgue constant of subperiodic trigonometric interpolation. J. Approx. Theory **167**, 59–64 (2013)
14. Da Fies, G., Vianello, M.: Product Gaussian quadrature on circular lunes. Numer. Math. Theory Methods Appl. **7**, 251–264 (2014)
15. Da Fies, G., Sommariva, A., Vianello, M.: Algebraic cubature by linear blending of elliptical arcs. Appl. Numer. Math. **74**, 49–61 (2013)
16. De Marchi, S. Vianello, M., Xu, Y.: New cubature formulae and hyperinterpolation in three variables. BIT Numer. Math. **49**, 55–73 (2009)
17. De Marchi, S., Sommariva, A., Vianello, M.: Multivariate Christoffel functions and hyperinterpolation. Dolomites Res. Notes Approx. DRNA **7**, 26–33 (2014)
18. Dominguez, V., Ganesh, M.: Interpolation and cubature approximations and analysis for a class of wideband integrals on the sphere. Adv. Comput. Math. **39**, 547–584 (2013)
19. Ganesh, M., Mhaskar, H.N.: Matrix-free interpolation on the sphere. SIAM J. Numer. Anal. **44**, 1314–1331 (2006)
20. Gautschi, W.: Computational aspects of three-term recurrence relations. SIAM Rev. **9**, 24–82 (1967)
21. Gautschi, W.: Orthogonal Polynomials: Computation and Approximation. Oxford University Press, New York (2004)

22. Gautschi, W.: Orthogonal polynomials (in Matlab). J. Comput. Appl. Math. **178**, 215–234 (2005)
23. Gautschi, W.: Sub-range Jacobi polynomials. Numer. Algorithms **61**, 649–657 (2012)
24. Gentile, M., Sommariva, A., Vianello, M.: Polynomial approximation and quadrature on geographic rectangles. Appl. Math. Comput. **297**, 159–179 (2017)
25. Hansen, O., Atkinson, K., Chien, D.: On the norm of the hyperinterpolation operator on the unit disc and its use for the solution of the nonlinear Poisson equation. IMA J. Numer. Anal. **29**, 257–283 (2009)
26. Hesse, K., Sloan, I.H.: Hyperinterpolation on the sphere. In: Frontiers in Interpolation and Approximation. Pure and Applied Mathematics, vol. 282, pp. 213–248. Chapman and Hall/CRC, Boca Raton, FL (2007)
27. Huybrechs, D.: On the Fourier extension of nonperiodic functions. SIAM J. Numer. Anal. **47**, 4326–4355 (2014)
28. Kosloff, D., Tal-Ezer, H.: A modified Chebyshev pseudospectral method with an $O(N^{-1})$ time step restriction. J. Comput. Phys. **104**, 457–469 (1993)
29. Leviatan, D., Sidon, J.: Monotone trigonometric approximation, Mediterr. J. Math. **12**, 877–887 (2015)
30. Matthysen, R., Huybrechs, D.: Fast algorithms for the computation of Fourier extensions of arbitrary length. SIAM J. Sci. Comput. **38**, A899–A922 (2016)
31. Nevai, P.G.: Orthogonal polynomials. Mem. Am. Math. Soc. **18**(213), 185 (1979)
32. Piciocchi, V.: Subperiodic trigonometric hyperinterpolation in tensor-product spaces, Master Thesis in Mathematics (advisor: Vianello, M.), University of Padova (2014)
33. Piessens, R.: Modified Clenshaw-Curtis integration and applications to numerical computation of integral transforms. In: Numerical Integration (Halifax, N.S., 1986). NATO Advanced Science Institutes Series C: Mathematical and Physical Sciences, vol. 203, pp. 35–51. Reidel, Dordrecht (1987)
34. Rudin, W.: Functional Analysis. McGraw-Hill, New York (1973)
35. Sloan, I.H.: Interpolation and hyperinterpolation over general regions. J. Approx. Theory **83**, 238–254 (1995)
36. Sommariva, A., Vianello, M.: Polynomial fitting and interpolation on circular sections. Appl. Math. Comput. **258**, 410–424 (2015)
37. Sommariva, A., Vianello, M.: Numerical hyperinterpolation over nonstandard planar regions. Math. Comput. Simul. **141**, 110–120 (2017)
38. Sommariva, A., Vianello, M.: `HYPERTRIG`: Matlab package for subperiodic trigonometric hyperinterpolation. Available online at: www.math.unipd.it/~marcov/subp.html
39. Tal-Ezer, H.: Nonperiodic trigonometric polynomial approximation. J. Sci. Comput. **60**, 345–362 (2014)
40. Vianello, M.: Norming meshes by Bernstein-like inequalities. Math. Inequal. Appl. **17**, 929–936 (2014)
41. Wade, J.: On hyperinterpolation on the unit ball. J. Math. Anal. Appl. **401**, 140–145 (2013)
42. Wang, H., Wang, K., Wang, X.: On the norm of the hyperinterpolation operator on the d-dimensional cube. Comput. Math. Appl. **68**, 632–638 (2014)

Discrete Data Fourier Deconvolution

Frank de Hoog, Russell Davies, Richard Loy, and Robert Anderssen

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In many practical situations, the recovery of information about some phenomenon of interest f reduces to performing Fourier deconvolution on indirect measurements $g = p*f$, corresponding to the Fourier convolution of f with a known kernel (point spread function) p. An iterative procedure is proposed for performing the deconvolution of $g = p * f$, which generates the partial sums of a Neumann series. However, the standard convergence analysis for the Neumann series is not applicable for such deconvolutions so a proof is given which is based on using Fourier properties in L^2.

In practice, only discrete measurements $\{g_m\}$ of g will be available. Consequently, the construction of a discrete approximation $\{f_m\}$ to f reduces to performing a deconvolution using a discrete version $\{g_m\} = \{p_m\} * \{f_m\}$ of $g = p * f$. For $p(x) = \operatorname{sech}(x)/\pi$, it is shown computationally, using the discrete version of the proposed iteration, that the resulting accuracy of $\{f_m\}$ will depend on the form and smoothness of f, the size of the interval truncation, and the level of discretization of the measurements $\{g_m\}$. Excellent accuracy for $\{f_m\}$ is obtained when $\{g_m\}$ and $\{p_m\}$ accurately approximate the essential structure in g and p, respectively, the support of p is much smaller than that for g, and the discrete measurements of $\{g_m\}$ are on a suitably fine grid.

F. de Hoog · R. Anderssen (✉)
CSIRO Data 61, Canberra, ACT, Australia
e-mail: Frank.deHoog@csiro.au; Bob.Anderssen@csiro.au

R. Loy
Mathematical Sciences Institute, Australian National University, Canberra, ACT, Australia
e-mail: Rick.Loy@anu.edu.au

R. Davies
School of Mathematics, Cardiff University, Cardiff, Wales, UK
e-mail: DaviesR@cardiff.ac.uk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_14

1 Introduction

In the recovery of information f from indirect measurements g, the relationship that connects g to f is often a Fourier convolution equation [8]

$$g(x) = p * f(x) = \int_{-\infty}^{\infty} p(x-y)f(y)dy$$
$$= \int_{-\infty}^{\infty} p(y)f(x-y)dy, \quad -\infty < x < \infty, \tag{1}$$

with known kernel p. Examples include phenomena where the theoretical range of the independent variable is infinite, such as the frequency response of the stress and strain in rheological oscillatory shear measurements and the frequency response of electronic amplifiers, of microphones and loudspeakers, and of brain waves [4, 7].

In this paper, it is assumed that the kernel p is a positive, even, and peaked at zero function such that [3]

$$\int_{-\infty}^{\infty} p(y)dy = 1, \tag{2}$$

and that the Fourier transform of p, $\hat{p}$, satisfies $\hat{p} > 0$. Though weaker conditions would suffice, the ones chosen are sufficient to clarify the matters under consideration. Of particular interest is $p(x) = \operatorname{sech}(x)/\pi$, because it corresponds to the recovery of information from oscillatory shear measurements in rheology [3, 9]. Various algorithms, including iterative procedures, have been proposed and implemented for this choice of p as well as for related rheological equations [1, 3].

Although, in practice, $g(x)$ is only measured at discrete values of x, it is nevertheless useful to initially focus on the continuous deconvolution (1). Therefore, we examine an iterative correction scheme for Eq. (1) and demonstrate its convergence. Furthermore, we observe that, in practice, useful approximations are obtained after a small number of iterations. This is the key to the subsequent developments, as it is well known that while deconvolution is an ill posed problem in the sense of Hadamard, the number of iterations applied in iterative methods, if small, provide effective regularisation. Here, the iterative correction procedure is implemented by approximating the iterates of the continuous problem using the discrete data or, equivalently, by applying the analogue of the iterative correction to the discretised convolution equation. In doing this, the fact that the actual discrete measurements $\{g_m\}$ of g will often be limited to a sub-interval $-L \leq x \leq L$ needs to be taken into account.

It is concluded that the iterative deconvolution of discretized versions of Eq. (1) converges rapidly and globally when the discretizations and the kernel p satisfy the following conditions

- the support of p is much smaller than that for g,
- the discretizations $\{g_m\}$ and $\{p_m\}$ accurately approximate the essential structure in g and p,
- any truncation of p recovers essentially all of the structure in p,
- for limited data, the discretization grid is sufficiently fine.

2 Preliminaries

2.1 The Discretization of $g = p * f$

Here, it is assumed that the discrete measurements correspond to a set of discrete values $\{g_m\}$ of g on an even grid with step-length h; namely, $\{g(mh)\}$. For the numerical solution of the convolution equation (1), some appropriate discretization of it must be performed. One possibility is given by

$$h \sum_{\ell=-\infty}^{\infty} p_{(m-\ell)} f_\ell = g_m, \quad p_m = p(mh), \quad f_\ell = f(\ell h), \quad m \in \mathbb{Z}. \tag{3}$$

In reality, since the $\{g_m\}$ values will have only been measured on some sub-grid $-L \leq m \leq L$, the actual equation that is solved for the discrete approximation$\{f_\ell\}$ is given by

$$h \sum_{\ell=-L}^{L} p_{(m-\ell)} f_\ell = g_m, \quad p_m = p(mh), \quad -L \leq m \leq L. \tag{4}$$

Consequently, the accuracy of the recovery of f by $\{f_\ell\}$ reduces to assessing, as a function of $\{p_m\}$, $\{g_m\}$, h and L, the accuracy with which the values of $\{f_\ell\}$ approximate the values of $\{f(\ell h)\}$.

Equation (4) is a discretization of the following truncation of Eq. (1)

$$g(x) = p * f_{[-a,a]}(x) = \int_{-a}^{a} p(x-y) f_{[-a,a]}(y)dy, \quad x \in (-a, a) \quad a = Lh. \tag{5}$$

2.2 Properties of Fourier Convolutions

Let the Fourier transform of an integrable function q be denoted by

$$\widehat{q}(\omega) = \int_{-\infty}^{\infty} q(x)\exp(-2\pi i\omega x)dx. \tag{6}$$

It follows that

(a) **Fourier transform of convolutions.** For the convolution of Eq. (1),

$$\widehat{g} = \widehat{p * f} = \widehat{p} \times \widehat{f}. \tag{7}$$

(b) **The spectrum of convolution operators.** Taking the Fourier transform of the convolution kernel p of Eq. (1) yields

$$\begin{aligned}\int_{-\infty}^{\infty} p(x-y)\exp(2\pi i\omega y)dy &= \int_{-\infty}^{\infty} p(x-y)\exp(-2\pi i\omega(x-y) + 2\pi i\omega x)dy \\ &= \left\{\int_{-\infty}^{\infty} p(x-y)\exp(-2\pi i\omega(x-y))\right\} \\ &\quad dy\exp(2\pi i\omega x) \\ &= \widehat{p}(\omega)\exp(2\pi i\omega x).\end{aligned} \tag{8}$$

It therefore follows that a convolution operator $p(x-y)$ has a continuous spectrum which is the Fourier transform $\widehat{p}$ of p.

(c) **The spectrum of discretized convolution operators.** For the chosen value h of the level of discretization, the Nyquist frequency constraint $-1/(2h) \leq \omega \leq 1/(2h)$ determines the range of frequencies ω that the sampling of the signal determines without involving oversampling, which would compromise the Fourier recovery of the signal. Consequently, the Fourier transform of the discretized kernel of Eq. (4) thereby yields, for that range of frequencies,

$$\begin{aligned}h\sum_{\ell=-\infty}^{\infty} p_{(n-\ell)}\exp(2\pi i\ell h\omega) &= h\sum_{\ell=-\infty}^{\infty} p_{(n-\ell)}\exp(-2\pi i(n-\ell)h\omega + 2\pi inh\omega) \\ &= \left\{h\sum_{\ell=-\infty}^{\infty} p_{(n-\ell)}\exp(-2\pi i(n-\ell)h\omega)\right\} \times \\ &\quad \exp(2\pi inh\omega) \\ &= \left\{h\sum_{\ell=-\infty}^{\infty} p_{\ell}\exp(-2\pi i\ell h\omega)\right\}\exp(2\pi inh\omega) \\ &= \widehat{p}(\omega; h)\exp(2\pi inh\omega),\end{aligned} \tag{9}$$

where

$$\widehat{p}(\omega;h) = h \sum_{\ell=-\infty}^{\infty} p_\ell \exp(-2\pi i\ell h\omega). \tag{10}$$

3 The Continuous Iterative Implementation and Convergence

In many practical situations, where the convolution kernel p is peaked at zero, such as for point spread functions and the rheological functions $\mathrm{sech}(x)/\pi$ and $\mathrm{sech}^2(x)/2$ [1, 2], the indirect measurements g correspond to a smoothed version of f.

For an approximation $\tilde{f}_1$ to f, we have the error $e_1 = f - \tilde{f}_1$, and residual $r_1 = g - p * \tilde{f}_1$. Since $r_1 = p * e_1$, r_1 is a smoothed version of e_1, this suggests the following iterative scheme

$$f_{n+1} = f_n + r_n = f_n + g - p * f_n\,, \quad n = 1, 2, \ldots, \tag{11}$$

with

$$f_1 = g, \qquad e_n = f - f_n, \qquad r_n = g - p * f_n = p * e_n.$$

In fact, f_n corresponds to the n^{th} partial sum of the Neumann series solution

$$f = \sum_{n=0}^{\infty} (I - A)^n g,$$

where $Af = p * f$ with $\|I - A\| = \|I - \widehat{p}\|_\infty = 1$. The subtraction of Eq. (11) from f yields the iterative error equation

$$e_{n+1} = f - f_{n+1} = e_n - r_n = e_n - p * e_n\,, \quad n = 1, 2, \ldots, e_1 = f - g\,. \tag{12}$$

Taking the Fourier transform of Eq. (12), invoking the Fourier convolution theorem and rearranging yields

$$\widehat{e_{n+1}} = (1 - \widehat{p})^n \widehat{e_1} \qquad (n \geq 1)\,. \tag{13}$$

Clearly, $\widehat{p}(0) = 1 = \|p\|_1$. Since $\hat{p} > 0$, which holds for sech and Gaussians, it follows that $0 \leq 1 - \widehat{p}(\omega) < 1$ on $[0, \infty)$. Consequently,

$$\|e_n\|_2^2 = \|\widehat{e_n}\|_2^2 = \int_{-\infty}^{\infty} (1 - \widehat{p}(\omega))^{2(n-1)} |\widehat{e_1}(\omega)|^2 \mathrm{d}\omega \searrow 0$$

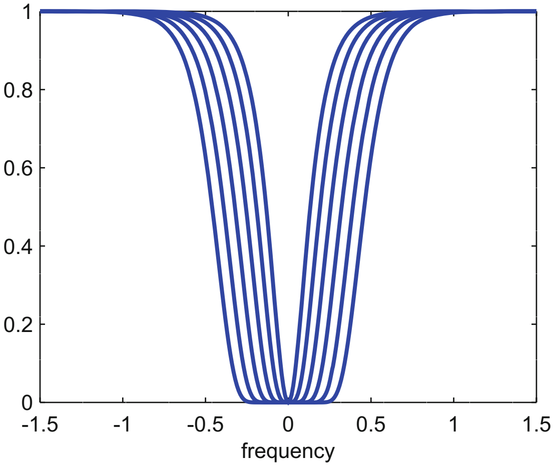


Fig. 1 Plots of $(1-\widehat{p})^n$ for $n = 1,\ 2,\ 4,\ 8,\ 16,\ 32$ (from the outer to the inner curve) for $p(x) = \operatorname{sech}(x)/\pi$, with the frequencies given on the horizontal axis

by monotone convergence. In fact, we have geometric $\|\cdot\|_2$ convergence on each bounded interval. So apart from some "tail" of $\widehat{e_1}$, we have geometric convergence. In addition,

$$e_n(x) = \int_{-\infty}^{\infty} (1-\widehat{p}(\omega))^{n-1}\widehat{e_1}(\omega)\exp(2\pi ix\omega)\mathrm{d}\omega\,,$$

so that

$$\|e_n\|_\infty = \sup_x |e_n(x)| \le \int_{-\infty}^{\infty} (1-\widehat{p}(\omega))^{n-1}|\widehat{e_1}(\omega)|\mathrm{d}\omega \searrow 0$$

provided that $\widehat{e_1} \in L^1(-\infty,\infty)$. Once again, geometric convergence holds on each bounded interval.

We note that, because of the factor $(1-\widehat{p}(\omega))$, the more peaked $\widehat{p}$ happens to be, such as for sech and gaussians, the slower will be the rate of convergence. This relates to the fact that the successive applications of $(1-\widehat{p}(\omega))$ correspond to a frequency filtering of $\widehat{e_1}(\omega)$ with a decrease in the removal of higher frequencies correlating with the stronger peakedness.

For $p(x) = \operatorname{sech}(x)/\pi$, for which $\widehat{p}(\omega) = \operatorname{sech}(\pi^2\omega) > 0$, and $n = 1,\ 2,\ 4,\ 8,\ 16,\ 32$, the convergence is plotted in Fig. 1.

4 Assessing the Effects of Working with the Discretized Data $\{g_m\}$

Because the deconvolution of Eq. (1) is performed numerically by solving the discrete equations (3) using the continuous iteration, the first matter to assess is the accuracy with which the Fourier transform $\widehat{p}(\omega; h)$ of the discretized kernel

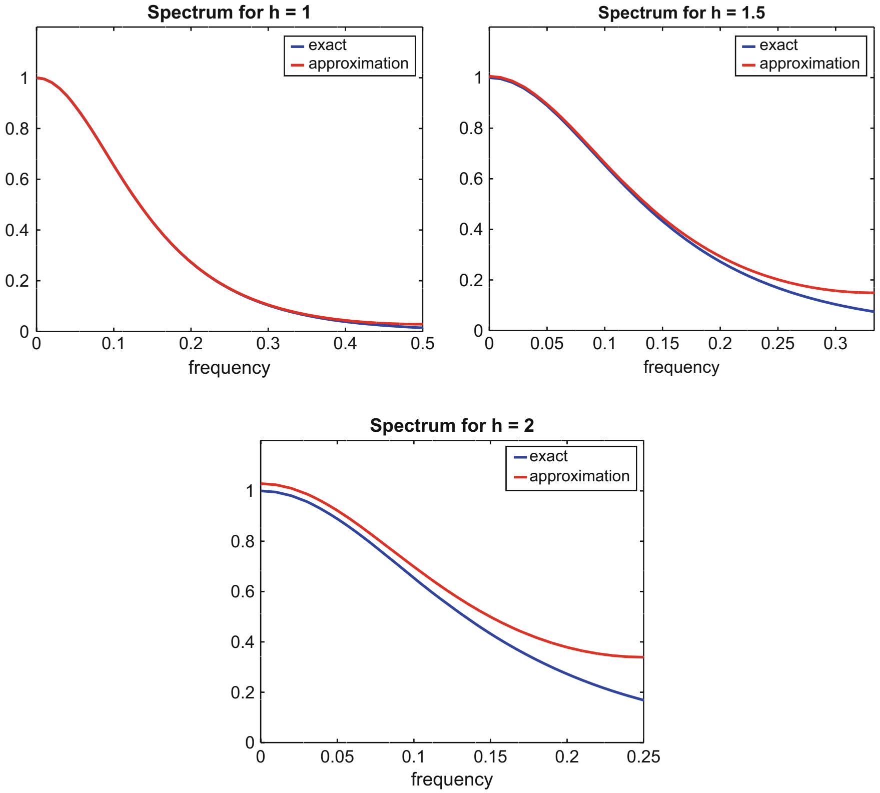


Fig. 2 Comparison of the accuracy with which $\widehat{p}(\omega; h)$ approximates $\widehat{p}(\omega)$, for different values of h

approximates the Fourier transform of the continuous kernel $\widehat{p}(\omega)$. This is illustrated in Fig. 2 for the rheology point spread function $p(x) = \text{sech}(x)/\pi$ [1] for the values $h = 2$, $h = 1.5$ and $h = 1$.

This highlights the fact that, for a sufficiently fine discretization (i.e. suitably small h), $\tilde{p}(\omega; h)$ yields an accurate recovery of the structure of $\widehat{p}(\omega)$.

The second matter to assess is the accuracy with which the matrix of the truncated equations (4) represents an accurate approximation of the structure encapsulated in Eq. (5). It is not a matter of comparing a selection of individual solutions of Eqs. (4) and (5), but of checking that the "low frequency" eigenvalues and eigenvectors of (4) yield accurate approximations for those of (5). To this end, for $p(x) = \text{sech}(x)/\pi$, $h = 0.5$ and $a = Lh = 6, 12, 24$, the eigenvalues of the Toeplitz matrices associated with the equations of (4) are compared in Fig. 3 with the values of $\widehat{p}(\omega)$. It highlights that the accuracy of the approximations improves as the sizes increase of the truncations applied to Eq. (1) to obtain the equations (4).

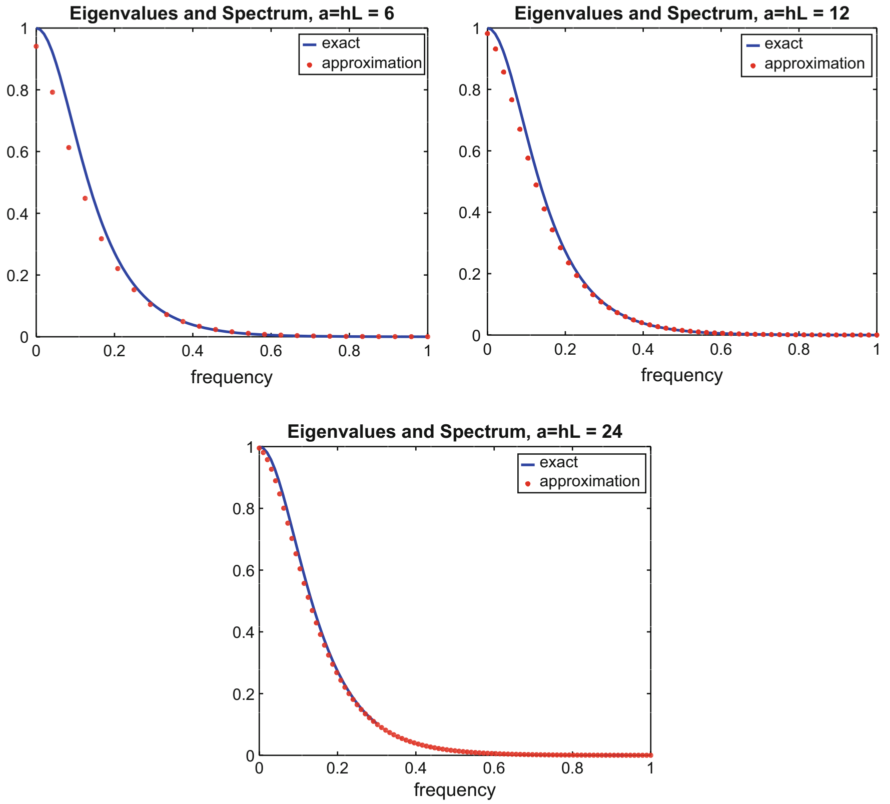


Fig. 3 Comparison, for $p(x) = \text{sech}(x)/\pi$, $h = 0.5$ and $a = Lh = 6,\ 12,\ 24$, of the eigenvalues of the corresponding matrices associated with equations of (4) with $\widehat{p}(\omega)$

4.1 Assessing the Numerical Performance of the Iteration

In order to implement the iteration numerically, it is necessary to evaluate $p * f_n$ numerically at each step of the iteration (11).

However, Fig. 1, though it indicates that convergence is faster at low frequencies, does not tell the full story about how the error propagates as highlighted in Eq. (13). Specifically, if there is a error η_n in the calculation of the convolution $p * e_n$, then it follows from Eqs. (11) and (12) that

$$\widehat{e}_{n+1} = (1-\widehat{p})^n\widehat{e}_1 - \sum_{k=0}^{n-1}\widehat{\eta}_{n-k}(1-\widehat{p})^k\,, \tag{14}$$

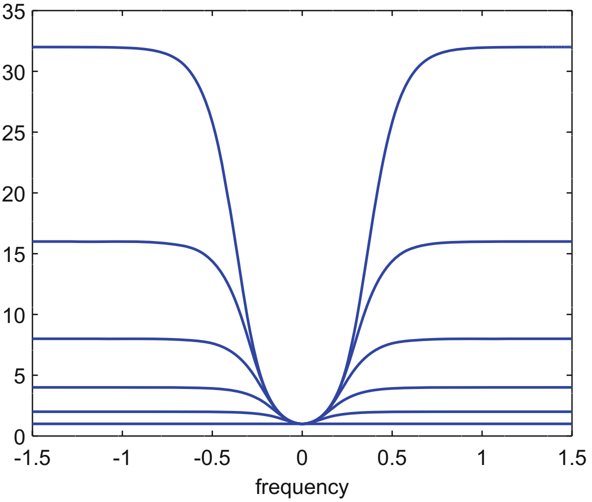


Fig. 4 Plots, for $p(x) = \mathrm{sech}(x)/\pi$ and $n = 1,\ 2,\ 4,\ 8,\ 16,\ 32$ (from the bottom to the top curve), of $[1-(1-\hat{p})^n]/\hat{p}$, with the frequencies given on the horizontal axis

and hence

$$
\begin{aligned}
|\widehat{e}_{n+1}| \ &\le |1-\widehat{p}|^n|\widehat{e}_1| + \left|\textstyle\sum_{k=0}^{n-1} \hat{\eta}_{n-k}(1-\widehat{p})^k\right| \\
&\le |1-\widehat{p}|^n|\widehat{e}_1| + \tfrac{1-(1-\widehat{p})^n}{\widehat{p}} \cdot \textstyle\sum_{k=0}^{n-1} |\widehat{\eta}_{n-k}|
\end{aligned}
\qquad (15)
$$

The final inequality follows because $\|\cdot\|_1$ is submultiplicative under convolution. Alternatively, one could also use Cauchy-Schwarz to obtain a bound involving $\|\cdot\|_2$ using $\|\widehat{\eta}_{n-k}\|_2 = \|\eta_{n-k}\|_2$.

The behaviour of the bound (15), as a function of n, is illustrated in Fig. 4, where $\frac{1-(1-\widehat{p})^n}{\widehat{p}}$ is plotted as a function of frequency.

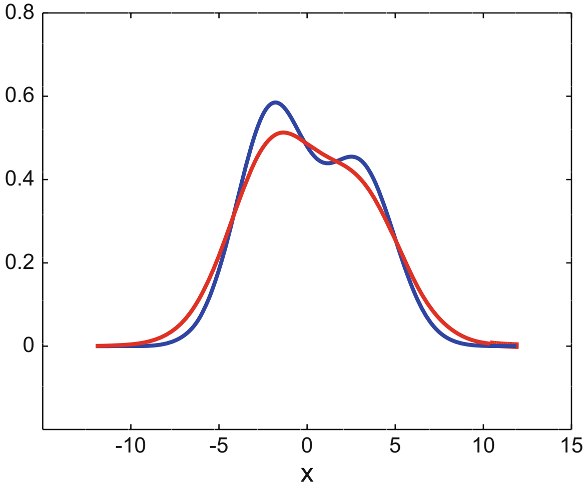
The plots in Fig. 4 illustrate that the low frequency components of the perturbation, where the convergence is most rapid, as highlighted in Fig. 1, are not as strongly amplified with each iteration as the high frequency components where the convergence is slow. Such situations are regularized by limiting the number of iterations using an appropriate stopping criterion.

5 Numerical Validation

With $p(x) = \mathrm{sech}(x)/\pi$, the validation was performed using the following synthetic data for the solution f

$$f(x) = \frac{1}{\sqrt{\pi}} \exp\left(-\frac{(x+2)^2}{8}\right) + \frac{3}{4\sqrt{\pi}} \exp\left(-\frac{(x-3)^2}{8}\right). \qquad (16)$$

Fig. 5 For the kernel $p(x) = \text{sech(x)}/\pi$, this shows that, visually, the bimodal nature of the solution f is not apparent in the data g. The blue curve is f, the red g



The importance of this choice is illustrated in Fig. 5, which shows that, in deconvolution situations, the available data g can often hide a multiple hump structure in f. As discussed elsewhere [2], such synthetic data is representative of practical deconvolution situations which arise in the study of the rheology of viscoelastic materials such as polystyrene and polybutadiene samples [5, 6]. In fact, the synthetic data is quite challenging in that, unlike the measurement data for polystyrene and polybutadiene samples, there is little evidence in the synthetic g that there are peaks in f.

In [2], the focus was joint inversion deconvolution algorithms for limited data where both $\text{sech}(x)/\pi$ and $\text{sech}^2(x)/2$ were utilized for p jointly. It acted as motivation of the more general limited data considerations developed in this paper.

5.1 The Effect of Truncation and Discretization

The accuracy with which the matrix of the truncated equations (4) represents an accurate approximation of the structure encapsulated in Eq. (5) has already been discussed in Sect. 4 and illustrated in Fig. 3.

Here, using the synthetic data of Eq. (16), the effect of different sizes of truncation in the algebraic equations (4) is examined in terms of its effect on the convergence of the iteration. For a grid spacing of $h = 0.5$, this is illustrated in Figs. 6 and 7, where g, corresponding to the synthetic f of Eq. (16), has been sampled on a uniform grid on the intervals $[-6, 6]$ and $[-12, 12]$, respectively.

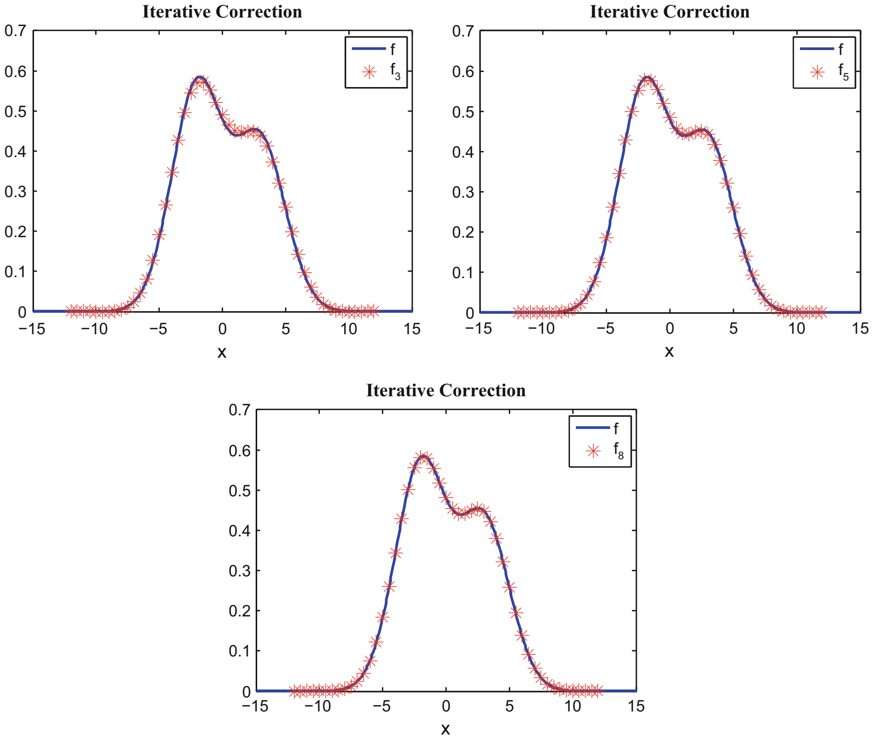


Fig. 6 For $p(x) = \text{sech}(x)/\pi$ and $h = 0.5$, the convergence of the 3rd, 5th and 8th iterates of (11) are compared when the measurements of g are truncated onto the interval [−12,12]

Figure 6 confirms that the successive approximations converge rapidly and globally to the correct solution if the truncation [−12,12] includes the bulk of the structure in g. This is clear from Fig. 7 that, if the size of the truncation [−6,6], on which the measurements of g are made, does not cover the full range of the structure in g, then the approximations generated by the iteration deteriorate away from the central region where they converges.

6 Conclusions

The various figures illustrate that excellent accuracy for the discrete approximation $\{f_m\}$ is obtained when the discrete measurements $\{g_m\}$ and the discrete kernel $\{p_m\}$ accurately approximate the essential structure in g and p, respectively, (Figs. 2 and 3) when the support of p is much smaller than that for g (Figs. 6 and 7) and the discrete measurements of $\{g_m\}$ are on a suitably fine grid (Fig. 3).

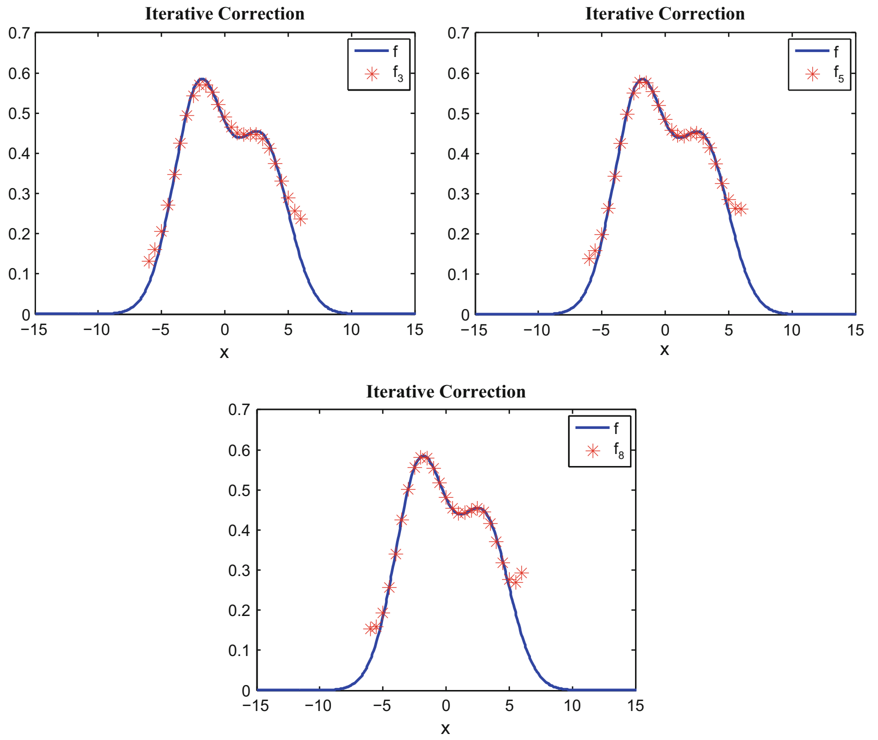


Fig. 7 For $p(x) = \text{sech}(x)/\pi$ and $h = 0.5$, the convergence of the 3rd, 5th and 8th iterates of (11) are compared when the measurements of g are truncated onto the interval [−6,6]

References

1. Anderssen, R.S., Davies, A.R., de Hoog, F.R., Loy, R.J.: Derivative based algorithms for continuous relaxation spectrum recovery. JNNFM **222**, 132–140 (2015)
2. Anderssen, R.S., Davies, A.R., de Hoog, F.R., Loy, R.J.: Simple joint inversion localized formulas for relaxation spectrum recovery. ANZIAM J. **58**, 1–9 (2016)
3. Davies, A.R., Goulding, N.J.: Wavelet regularization and the continuous relaxation spectrum. J. Non-Newtonian Fluid Mech. **189**, 19–30 (2012)
4. Gureyev, T.E., Nesterets, Y.I., Stevenson, A.W., Wilkins, S.W.: A method for local deconvolution. Appl. Opt. **42**, 6488–6494 (2003)
5. Honerkamp, J., Weese, J.: Determination of the relaxation spectrum by a regularization method. Macromolecules **22**, 4372–4377 (1989)
6. Honerkamp, J., Weese, J.: A nonlinear regularization method for the calculation of relaxation spectra. Rheol. Acta **32**, 65–73 (1993)
7. Starck, J.-L., Murtagh, F.D., Bijaoui, A.: Image Processing and Data Analysis: The Multiscale Approach. Cambridge University Press, Cambridge (1998)
8. Vogel, C.R.: Computational Methods for Inverse Problems, vol. 23. SIAM, Philadelphia (2002)
9. Walters, K.: Rheometry. Chapman and Hall, London (1975)

Kernels of a Class of Toeplitz Plus Hankel Operators with Piecewise Continuous Generating Functions

Victor D. Didenko and Bernd Silbermann

Dedicated to Ian H. Sloan on the occasion of his 80-th anniversary.

Abstract Toeplitz $T(a)$ and Toeplitz plus Hankel operators $T(a) + H(b)$ acting on sequence space l^p, $1 < p < \infty$, are considered. If $a \in PC_p$ is a piecewise continuous l^p-multiplier, a complete description of the kernel of the Fredholm operator $T(a)$ is derived. Moreover, the kernels of Fredholm Toeplitz plus Hankel operators $T(a) + H(b)$ the generating functions a and b of which belong to PC_p and satisfy the condition $a(t)a(1/t) = b(t)b(1/t)$, $t \in \mathbb{T}$, are also determined.

1 Introduction

Fredholm properties of Toeplitz plus Hankel operators with piecewise continuous generating functions acting on classical Hardy spaces and on spaces of p-power summable sequences are well studied. There are efficient necessary and sufficient conditions for their Fredholmness and index formulas (see Sections 4.95–4.102 in [1], Sections 4.5 and 5.7 in [12], and [13, 14]). Moreover, the Fredholmness of the operators with quasi piecewise continuous generating functions acting on Hardy spaces, have been investigated in [15] and index formulas have been derived in [2]. However, the invertibility, the kernels and cokernels of such operators have been little studied. In particular, the investigation of Toeplitz plus Hankel operators on sequence spaces faces considerable difficulties. Even for "pure" Toeplitz operators

V. D. Didenko (✉)
Department of Mathematics, Southern University of Science and Technology, Nanshan District, Shenzhen, Guangdong, China
e-mail: diviol@gmail.com

B. Silbermann
Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany
e-mail: silbermn@mathematik.tu-chemnitz.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_15

on l^p-spaces not much is known about the kernels of such operators, and in most cases there are no efficient representations for their inverses, either. On the other hand, for a wide class of Toeplitz plus Hankel operators acting on classical Hardy spaces H^p some results have been obtained recently [3–6]. This became possible due to the existence of a factorization of auxiliary functions in H^p. Such an approach generally does not work for operators acting on the spaces of sequences. The only exceptions are operators with generating functions from those multiplier algebras which allow factorizations inside the same algebra. One example of such a situation is the Wiener algebra of functions with absolutely convergent Fourier series.

Thus the aim of this paper is to describe the kernels of Toeplitz operators and Toeplitz plus Hankel operators with piecewise continuous generating functions acting on the spaces of p-summable sequences. Note that for Toeplitz plus Hankel operators we additionally assume that their generating functions satisfy an auxiliary algebraic condition.

This paper is organized as follows. In Sect. 2 we introduce Banach spaces and operators which are the main object of our study and recall some of their properties. In Sect. 3, connections between the kernels of Toeplitz plus Hankel operators acting in classical Hardy spaces and in spaces of summable sequences are described. In particular, we derive an efficient description for the kernels of Toeplitz operators $T(g)$ acting on l^p-spaces. Moreover, we establish a formula for the inverse of the operator $T(g)$ if the function g generates an invertible Toeplitz operator on the related Hardy space H^q, $1/p + 1/q = 1$. Section 4 deals with the kernels of Toeplitz plus Hankel operators the generating functions of which are l^p-multipliers and satisfy an algebraic condition. In conclusion, the corresponding results are specified for Toeplitz plus Hankel operators with piecewise continuous generating functions.

2 Spaces and Operators

In this section we introduce some operators and spaces we are interested in. Let $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle in the complex plane $\mathbb{C}$ equipped with the counterclockwise orientation, and let $L^p = L^p(\mathbb{T})$, $1 \leq p \leq \infty$ denote the complex Banach space of all Lebesgue measurable functions f on $\mathbb{T}$ such that

$$||f||_p := \left(\frac{1}{2\pi}\int_{\mathbb{T}} |f(t)|^p\,|dt|\right)^{1/p} < \infty, \quad 1 \leq p < \infty,$$

$$||f||_\infty := \operatorname*{ess\,sup}_{t\in\mathbb{T}} |f(t)| < \infty.$$

The Fourier transform $\mathscr{F}$ on the space L^1 and on its subspaces L^p, $p > 1$ is defined by

$$\mathscr{F} : L^1 \to c_0, \quad f \mapsto (\widehat{f}_n)_{n\in\mathbb{Z}},$$

where $\widehat{f}_n := (1/2\pi)\int_0^{2\pi} f(e^{in\theta})e^{-in\theta}\, d\theta$, $n \in \mathbb{Z}$, are the Fourier coefficients of f, and c_0 is the space of all sequences of complex numbers that tend to zero as $|n| \to \infty$. Note that in what follows, we often write $\widehat{f}$ instead of $\mathscr{F}f$.

For a non-empty subset $\mathbb{I}$ of the set of all integers $\mathbb{Z}$, let $l^p(\mathbb{I})$ denote the complex Banach space of all sequences $\xi = (\xi_n)_{n\in\mathbb{I}}$ of complex numbers with the norm

$$||\xi||_p = \left(\sum_{n\in\mathbb{I}} |\xi_n|^p\right)^{1/p}, \quad 1 \le p < \infty.$$

In this paper, the space $l^p(\mathbb{I})$ is considered as a natural subspace of $l^p(\mathbb{Z})$, and by $P_{\mathbb{I}}$ we denote the canonical projection from $l^p(\mathbb{Z})$ onto $l^p(\mathbb{I})$. Further, if $\mathbb{I}$ is the set of non-negative integers $\mathbb{Z}_+$, then instead of $l^p(\mathbb{Z}_+)$ and $P_{\mathbb{Z}_+}$ we will write l^p and P, respectively.

It is well known that the operator $\mathscr{F}$ maps L^2 isometrically onto $l^2(\mathbb{Z})$. For $p \neq 2$ a more general result, namely the celebrated Hausdorff-Young Theorem, describing relations between the spaces L^p and l^p will be recalled later on.

2.1 Toeplitz and Hankel Operators on l^p-Spaces

On the space $l^p(\mathbb{Z})$ we consider an operator J defined by

$$J\xi = J((\xi_n)_{n\in\mathbb{Z}}) = (\xi_{-n-1})_{n\in\mathbb{Z}},$$

and if I denotes the identity operator then we set $Q := I - P$. It is easily seen that the operators I, J, P and Q are connected by the following relations $J^2 = I$ and $JPJ = Q$.

Let $a \in L^\infty$. On the space $l_0(\mathbb{Z})$ of all finitely supported sequences on $\mathbb{Z}$ consider the Laurent operator $L(a)$ generated by a, i.e.

$$(L(a)\xi)_k := \sum_{m\in\mathbb{Z}} \widehat{a}_{k-m}\xi_m.$$

Note that for every $k \in \mathbb{Z}$ there is only a finite number of non-zero terms in this sum. We say that a is a multiplier on $l^p(\mathbb{Z})$ if $L(a)\xi \in l^p(\mathbb{Z})$ for any $\xi \in l_0(\mathbb{Z})$ and if

$$||L(a)|| := \sup\{||L(a)\xi||_p : \xi \in l_0(\mathbb{Z}),\ ||\xi||_p = 1\}$$

is finite. In this case, $L(a)$ extends to a bounded linear operator on $l^p(\mathbb{Z})$ and we denote it by $L(a)$ again. The set M^p of all multipliers on $l^p(\mathbb{Z})$ is a commutative Banach algebra under the norm $||a||_{M_p} := ||L(a)||$ (see, for example, [1]). Recall that $M^2 = L^\infty(\mathbb{T})$ and $M^1 = W(\mathbb{T})$ where $W(\mathbb{T})$ stands for the Wiener algebra of functions with absolutely convergent Fourier series. Moreover, every function

$a \in L^\infty(\mathbb{T})$ with bounded total variation Var (a) belongs to the algebra M^p for any $p \in (1, \infty)$, and the Stechkin inequality

$$||a||_{M^p} \leq c_p(||a||_\infty + \mathrm{Var}\,(a)),$$

with a constant c_p independent of a, holds. In particular, every trigonometric polynomial and every piecewise constant function on $\mathbb{T}$ are multipliers on any space $l^p(\mathbb{Z})$, $p \in (1, \infty)$. Moreover, trigonometric polynomials belong to the set $M^1 = W(\mathbb{T})$ as well. However, in this work we consider operators with piecewise continuous generating functions and we restrict ourselves to the case $p \in (1, \infty)$. Let $\mathscr{E}$ and $P\mathbb{C}$ be, respectively, the algebra of all trigonometric polynomials and the algebra of all piecewise constant functions on $\mathbb{T}$. By C_p and PC_p we, respectively, denote the closures of the sets $\mathscr{E}$ and $P\mathbb{C}$ in the algebra M^p. Note that C_2 is just the algebra $C(\mathbb{T})$ of all continuous functions on $\mathbb{T}$, and PC_2 is the algebra $PC(\mathbb{T})$ of all piecewise continuous functions on $\mathbb{T}$. It is well known that $C_p \subset C(\mathbb{T})$ and $C_p \subset PC_p \subset PC(\mathbb{T})$ (see [1] for this and other properties of multipliers). Let us recall that $\mathbb{T}$ is oriented counter-clockwise. Accordingly, hereinafter $a(t^-)$ and $a(t^+)$ stand for the one-sided limits of a function $a \in PC_p$ at the point $t \in \mathbb{T}$ from below and from above.

Remark 1 If $a \in M^p$ then $JL(a)J = L(\widetilde{a})$, where $\widetilde{a} = a(1/t)$.

Let $a \in M^p$. The operators $T(a) : l^p \to l^p$ and $H(a) : l^p \to l^p$ defined, respectively, by $f \mapsto PL(a)f$ and $f \mapsto PL(a)QJf$ are called Toeplitz and Hankel operators with generating function a. It is clear that for $a \in M^p$ the operators $T(a)$ and $H(a)$ are bounded on l^p. Moreover, the action of the operators $T(a)$ and $H(a)$ on the elements from l^p can be written as follows

$$T(a) : (\xi_j)_{j\in\mathbb{Z}_+} \to \left(\sum_{k\in\mathbb{Z}_+} \widehat{a}_{j-k}\xi_k\right)_{j\in\mathbb{Z}_+},$$

$$H(a) : (\xi_j)_{j\in\mathbb{Z}_+} \to \left(\sum_{k\in\mathbb{Z}_+} \widehat{a}_{j+k+1}\xi_k\right)_{j\in\mathbb{Z}_+}.$$

We remark that below we often use the notation $T_p(a)$ or $H_p(a)$ in order to underline that the corresponding Toeplitz or Hankel operator is considered on the space l^p for a fixed $p \in (1, \infty)$. Moreover, in what follows, we will operate in spaces defined by various indices p and q. In this connection, let us agree that whenever p and q appears in the text, they are related as: $1/p + 1/q = 1$.

By GM^p we denote the group of invertible elements in M^p.

Lemma 1 (Sections 2.30, 2.38, 6.5–6.6 in [1]) *Let $p \in (1, \infty)$.*

1. *If $T_p(a)$ is Fredholm, then $a \in GM^p$.*
2. *If $a \in M^p$, then one of the kernels of the operators $T_p(a)$ or $T_q^*(a)$ is trivial.*

3. *If $a \in GM^p$, then the operator $T_p(a)$ is Fredholm, and if* $\operatorname{ind} T_p(a) = 0$*, then $T(a)$ is invertible on l^p.*

Let us now recall the necessary and sufficient conditions for the Fredholmness of Toeplitz operators $T(a) : l^p \to l^p$, $1 < p < \infty$ with generating functions $a \in PC_p$. Consider the two-point compactification $\overline{\mathbb{R}}$ of the real line $\mathbb{R}$ and the function $\mu_p : \overline{\mathbb{R}} \to \mathbb{C}$ defined by

$$\mu_p(\lambda) := \begin{cases} \dfrac{1 + \coth(\pi(\lambda + i/p))}{2}, & \lambda \in \mathbb{R}, \\ 0, & \lambda = -\infty, \\ 1, & \lambda = +\infty. \end{cases}$$

Note that if λ runs from $-\infty$ to $+\infty$, then $\mu_p(\lambda)$ runs along a circular arc in $\mathbb{C}$ which connects the points 0 and 1 and passes through the point $(1 - i\cot(\pi/p))/2$. The next theorem is due to Roland Duduchava [7, 8]. For alternative proofs see [1, 11].

Theorem 1 *If $a \in PC_p$, then the operator $T_p(a)$ is Fredholm if and only if the function*

$$(\operatorname{smb} T_p(a))(t, \lambda) := a(t^-)(1 - \mu_q(\lambda)) + a(t^+)\mu_q(\lambda),$$

does not vanish on $\mathbb{T} \times \overline{\mathbb{R}}$.

For $a \in PC_p$, the index of the Fredholm operator T_p can be determined by means of the function $\operatorname{smb} T_p(a)$. First, let us suppose that $a \in PC_p$ is a piecewise smooth function with only finitely many jumps. Then the range of the function $\operatorname{smb} T_p(a)$ is a closed curve Γ with a natural orientation obtained from the essential range of a by filling in the circular arcs

$$\zeta_q(a(t^-), a(t^+)) = \{a(t^-)(1 - \mu_q(\lambda) + a(t^+))\mu_q(\lambda) : \lambda \in \overline{\mathbb{R}}\}),$$

at every point $t \in \mathbb{T}$ where a has a jump. If $T_p(a)$ is Fredholm, the curve Γ does not pass the origin, and we let $\operatorname{wind} \operatorname{smb} T_p(a)$ denote the winding number of Γ with respect to the origin, i.e. the integer $1/(2\pi)$ times the growth of the argument of $(\operatorname{smb} T_p(a))(t, \lambda)$ when t moves along $\mathbb{T}$ in the counter-clockwise direction and the arcs $\zeta_q(a(t^-), a(t^+))$ if a has a jump at the point t. The orientation of the arcs $\zeta_q(a(t^-), a(t^+))$ is chosen in such a way that the point $\zeta_q(a(t^-), a(t^+))(\lambda)$ runs from $a(t^-)$ to $a(t^+)$ if λ runs from $-\infty$ to $+\infty$. Then the index of the operator $T_p(a)$ is

$$\operatorname{ind} T_p(a) = -\operatorname{wind} \operatorname{smb} T_p(a), \tag{1}$$

(see [1, Section 2.73 and Section 6.32]). Moreover, analogously to Section 5.49 of [1] one can extend both the definition of the winding number and the index formula (1) to Fredholm operators $T_p(a)$ with arbitrary $a \in PC_p$.

2.2 Toeplitz and Hankel Operators on Hardy Spaces

For $1 \le p \le \infty$, let $H^p = H^p(\mathbb{T})$ and $\overline{H^p}$ stand for the Hardy spaces,

$$H^p := \{f \in L^p : \widehat{f}_n = 0 \quad \text{for all} \quad n < 0\},$$
$$\overline{H^p} := \{f \in L^p : \widehat{f}_n = 0 \quad \text{for all} \quad n > 0\}.$$

On the spaces L^p, $1 < p < \infty$ consider the operators $\mathbf{J}, \mathbf{P}$ and $\mathbf{Q}$ defined by

$$\begin{aligned} \mathbf{J} &: f(t) \mapsto t^{-1} f(t^{-1}), \\ \mathbf{P} &: \sum_{n\in\mathbb{Z}} \widehat{f}_n t^n \mapsto \sum_{n\in\mathbb{Z}_+} \widehat{f}_n t^n \\ \mathbf{Q} &:= \mathbf{I} - \mathbf{P}, \end{aligned}$$

where $\mathbf{I}$ is the identity operator on the space L^p. The operator of multiplication by a function $a \in L^\infty$ ia denoted by $a\mathbf{I}$. These operators satisfy the relations

$$\mathbf{J}^2 = \mathbf{I}, \quad \mathbf{Q} = \mathbf{JPJ}, \quad \mathbf{J}a\mathbf{J} = \widetilde{a}\mathbf{I},$$

where $\widetilde{a}(t) := a(1/t)$.

Any function $a \in L^\infty$ defines two bounded linear operators acting on the Hardy space H^p, $1 < p < \infty$, namely,

$$\begin{aligned} \mathbf{T}(a) &: \varphi \mapsto \mathbf{P}a\varphi, \\ \mathbf{H}(a) &: \varphi \mapsto \mathbf{P}a\mathbf{QJ}\varphi. \end{aligned}$$

Similarly to the case of l^p-spaces, the operators $\mathbf{T}(a)$ and $\mathbf{H}(a)$ are, respectively, called Toeplitz and Hankel operators, and we will write $\mathbf{T}_p$ or $\mathbf{H}_p$ if we want to emphasize that the corresponding operator is considered on a specific Hardy space H^p.

Let χ_n denote the function $\chi_n(t) = t^n$, $n \in \mathbb{Z}$. For $1 < p < \infty$ the function system $\mathscr{X}_0 = \{\chi_n : n \in \mathbb{Z}\}$ forms a Schauder basis in L^p, whereas the system $\mathscr{X} = \{\chi_n : n \in \mathbb{Z}_+\}$ forms a Schauder basis in H^p. It is easily seen that the matrix representations $[\mathbf{T}_p(a)]_{\mathscr{X}}$ and $[\mathbf{H}_p(a)]_{\mathscr{X}}$ of the operators $\mathbf{T}_p(a)$ and $\mathbf{H}_p(a)$ with respect to the above basis H^p are given by $(\widehat{a}_{i-j})_{i.j=0}^{\infty}$ and $(\widehat{a}_{i+j+1})_{i.j=0}^{\infty}$, respectively. For instance, the action of the operator $\mathbf{T}_p(a)$ on H^p can be described using the matrix representation of $\mathbf{T}_p(a)$ as follows. If $f = \sum_{n\in\mathbb{Z}_+} \widehat{f}_n \chi_n \in H^p$, then

$$\mathbf{T}_p(a) f = g, \quad g = \sum_{n\in\mathbb{Z}_+} \widehat{g}_n \chi_n,$$

where $\widehat{g}_n = \sum_{k\in\mathbb{Z}_+} \widehat{a}_{n-k}\widehat{f}_k$. Clearly, the action of $\mathbf{H}_p(a)$ on H^p can be described in the same manner.

It turns out that Theorem 1 is also valid for Toeplitz operators acting on the Hardy space H^p. More precisely, let us recall the following result of I. Gohberg and N. Krupnik (see [1]).

Theorem 2 *If $a \in PC$, then the Toeplitz operator $\mathbf{T}_p(a)$ is Fredholm if and only if the function*

$$(\operatorname{smb}\mathbf{T}_p(a))(t,\lambda) := a(t^-)(1-\mu_p(\lambda)) + a(t^+)\mu_p(\lambda),$$

does not vanish on $\mathbb{T}\times\overline{\mathbb{R}}$. If this condition is satisfied, then

$$\operatorname{ind}\mathbf{T}_p(a) = -\operatorname{wind}\operatorname{smb}\mathbf{T}_p(a).$$

Comparing Theorem 1 with Theorem 2, we observe a close relation between Toeplitz operators on l^p and H^q.

Corollary 1 *If $a \in PC_p$, $1 < p < \infty$, then the operators $T_p(a)$ and $\mathbf{T}_q(a)$ are simultaneously Fredholm or not and their indices as well as the kernel dimensions coincide.*

Indeed, the simultaneous Fredholmness of the operators $T_p(a)$ and $\mathbf{T}_q(a)$ and coincidence of their indices follow from Theorems 1 and 2, whereas the equality of kernel dimensions is the consequence of the Coburn-Simonenko Theorem (see assertion 2 in Lemma 1 or Section 6.6 in [1]).

Thus if $a \in PC_p$, then the Fredholmness of the operator $T_p(a)$ implies that of $\mathbf{T}_q(a)$ with $\operatorname{ind} T_p(a) = \operatorname{ind}\mathbf{T}_q(a)$. However, the authors do not know whether this statement is true without the assumption $a \in PC_p$. In other words the following problem appears.

Problem 1 Let $p \in (1,\infty)$. For which classes of functions a, the Fredholmness of the operator $T_p(a)$ implies the Fredholmness of $\mathbf{T}_q(a)$ and the parity of their indices?

3 Kernels of Toeplitz Operators Acting on Spaces l^p

As was already mentioned, the theory of Toeplitz operators on l^p-spaces is more complicated than the corresponding theory on H^p-spaces. The first obstacle in l^p-case is the multiplier problem which is not an issue for Toeplitz operators considered on the Hardy spaces H^p. Another complication is that in l^p-case there is no developed Wiener-Hopf factorization theory, whereas various factorizations are heavily used in description of the kernels of Toeplitz operators acting on H^p-spaces. In the present work, we propose an approach to overcome these difficulties. For we need the celebrated Hausdorff-Young theorem. Let us recall this important result.

Theorem 3 (Section 13.5 in [9]) *Let $1/p + 1/q = 1$ if $q \in (1, \infty)$, and let $p = \infty$ if $q = 1$.*

1. *If $g \in H^q$, $1 \le q \le 2$, then $\mathscr{F} g \in l^p$ and $||\mathscr{F} g||_{l^p} \le ||g||_{H^q}$.*
2. *If $\varphi \in l^q$, $1 \le q \le 2$, then there is an element $g \in H^p$ such that $\varphi = \mathscr{F} g$ and $||g||_{H^p} \le ||\varphi||_{l^q}$.*

In what follows we will also use the Wiener-Hopf factorization of functions $a \in L^\infty$, so let us remind some relevant definitions and facts.

Definition 1 A function $g \in L^\infty$ admits a generalized Wiener–Hopf factorization in H^p, if it can be represented in the form

$$g = g_- \chi_n g_+, \quad g_-(\infty) = 1, \tag{2}$$

where $n \in \mathbb{Z}$, $g_+ \in H^q$, $g_+^{-1} \in H^p$, $g_- \in \overline{H^p}$, $g_-^{-1} \in \overline{H^q}$, and the linear operator $g_+^{-1} \mathbf{P} g_-^{-1} \mathbf{I}$ defined on the set span $\{\chi_k : k \in \mathbb{Z}_+\}$ can be boundedly extended to the whole space H^p.

Let us emphasize that the generalized Wiener-Hopf factorization (2) strongly depends on the space H^p. However, if p is fixed, then it is unique. In what follows, the representation (2) is often called just Wiener-Hopf factorization. Let us also recall that the number n occurring in (2) is called the factorization index.

Theorem 4 (Section 5.5 in [1]) *If $g \in L^\infty$, then the Toeplitz operator $\mathbf{T}_p(g)$, $1 < p < \infty$ is Fredholm if and only if the generating function g admits the generalized Wiener-Hopf factorization (2). If $\mathbf{T}_p(g)$ is Fredholm, then*

$$\operatorname{ind} \mathbf{T}_p(g) = -n.$$

Now we are going to present a description of the kernels of Toeplitz $T(a)$ and Toeplitz plus Hankel operators $T(a) + H(b)$ acting on spaces l^p. Let us start with auxiliary results.

Proposition 1 *Let X_1 be a Banach space continuously and densely embedded into a Banach space X_2, and let $A_1 : X_1 \to X_1$ and $A_2 : X_2 \to X_2$ be bounded Fredholm operators such that $\operatorname{ind} A_1 = \operatorname{ind} A_2$. If A_2 is an extension of the operator A_1, then*

$$\ker A_1 = \ker A_2. \tag{3}$$

Proof The proof of this result is simple. Since $X_1 \subset X_2$, then

$$\dim \ker A_1 \le \dim \ker A_2. \tag{4}$$

On the other hand, the inclusion $X_2^* \subset X_1^*$ implies that

$$\dim \ker A_2^* \le \dim \ker A_1^*. \tag{5}$$

Using (4) and (5) and the relation $\operatorname{ind} A_1 = \operatorname{ind} A_2$, one obtains

$$\dim \ker A_1 = \dim \ker A_2,$$

and the identity (3) follows. □

Let us now describe general relations between the kernels of Toeplitz plus Hankel operators acting on the spaces l^p and H^q.

Lemma 2 *Let* $p \in (1,\infty)$*. Assume that* $a, b \in M^p$ *and that* $T_p(a) + H_p(b)$ *and* $\mathbf{T}_q(a) + \mathbf{H}_q(b)$ *are Fredholm operators. If*

$$\operatorname{ind}(T_p(a) + H_p(b)) = \operatorname{ind}(\mathbf{T}_q(a) + \mathbf{H}_q(b)),$$

then the Fourier transform $\mathscr{F}$ *is an isomorphism between the spaces* $\ker(\mathbf{T}_q(a) + \mathbf{H}_q(b))$ *and* $\ker(T_p(a) + H_p(b))$*. In particular, an element* $h \in l^p$ *is in* $\ker(T_p(a) + H_p(b))$ *if and only if* $h = \mathscr{F}\varphi$ *for a function* $\varphi \in \ker(\mathbf{T}_q(a) + \mathbf{H}_q(b))$*.*

Proof Let $\widehat{H}^p, p \in (1,\infty)$ be the space of all sequences $(g_k)_{k\in\mathbb{Z}_+}$ for which there is a function $g \in H^p$ such that $\mathscr{F}g = (g_k)_{k\in\mathbb{Z}_+}$. Let us equip $\widehat{H}^p$ with the norm

$$||(g_k)_{k\in\mathbb{Z}_+}|| := ||g||_{H^p}.$$

Apparently the spaces $\widehat{H}^p$ and H^p are isometrically isomorphic, and the operator $\mathbf{T}_q(a) + \mathbf{H}_q(b)$ induces a linear bounded operator on $\widehat{H}^p$ given by

$$[\mathbf{T}_p(a) + \mathbf{H}_p(a)]_{\mathscr{X}} = (\widehat{a}_{i-j})_{i,j=0}^{\infty} + (\widehat{a}_{i+j+1})_{i,j=0}^{\infty}.$$

Assume that $2 \leq p < \infty$. The first part of Hausdorff-Young Theorem shows that $\widehat{H}^q$ is continuously embedded in the space l^p. Moreover, $\widehat{H}^q$ is dense in l^p. The operators $\mathbf{T}_q(a)+\mathbf{H}_q(b)$ and $[\mathbf{T}_q(a)+\mathbf{H}_q(b)]_{\mathscr{X}} : \widehat{H}^q \to \widehat{H}^q$ have the same Fredholm properties as $T_p(a) + H_p(b)$, and the operator $T_p(a) + H_p(b)$ is an extension of $[\mathbf{T}_q(a) + \mathbf{H}_q(b)]_{\mathscr{X}} : \widehat{H}^q \to \widehat{H}^q$ on the whole space l^p. By Proposition 1, one has

$$\ker[\mathbf{T}_p(a) + \mathbf{H}_p(a)]_{\mathscr{X}} = \ker(T_p(a) + H_p(b)),$$

whence the assertion of Lemma 2 follows for $p \geq 2$.

Now let $1 < p < 2$. The second part of Hausdorff-Young Theorem assures that l^p is continuously embedded into $\widehat{H}^q$. Clearly, $[\mathbf{T}_q(a) + \mathbf{H}_q(b)]_{\mathscr{X}}$ is an extension of $T_p(a) + H_p(b)$. Using Proposition 1 once more, we obtain the result for $1 < p < 2$. □

The main problem in using Lemma 2 is the availability of information about Fredholm properties of the operators involved, including the equality of their Fredholm indices. For example, if $a \in PC_p$, then by Theorems 1 and 2 the operators $T_p(a)$ and $\mathbf{T}_q(a)$ are simultaneously Fredholm and their indices coincide. Consequently, we again arrive at Corollary 1.

Theorem 5 *Assume that $g \in PC_p$,* $\operatorname{ind} T_p(g) = k > 0$, *and let*

$$g = g_+ \chi_{-k} g_-$$

be the Wiener-Hopf factorization in H^q of the generating function a. Then

$$\ker T_p(g) = \operatorname{lin\,span}\{(\widehat{g_{+,n-l}^{-1}})_{n\in\mathbb{Z}_+} : l = 0, 1, \ldots, k-1\}, \tag{6}$$

where $\widehat{g_{+,j}^{-1}}$, $j \in \mathbb{Z}$ are the Fourier coefficients of the element g_+^{-1}.

Proof Recall that Fredholm Toeplitz operators $\mathbf{T}_q(g)$ are one-sided invertible. By Corollary 1, one has $\operatorname{ind} \mathbf{T}_q(g) = \operatorname{ind} T_p(g) = k > 0$, so the operator $\mathbf{T}_q(g)$ is right invertible. Let g_+ be the plus-factor in the Wiener-Hopf factorization of the function g in H^q. Then, it is easily seen that the functions $\{g_+^{-1}\chi_l\}_{l=0}^{k-1}$ form a basis in $\ker \mathbf{T}_q(g)$. Indeed, for $0 \le l \le k-1$ one has $-k + l < 0$, so that

$$\mathbf{T}_q(g) g_+^{-1} \chi_l = \mathbf{P} g_- \chi_{-k} g_+ g_+^{-1} \chi_l = \mathbf{P} g_- \chi_{-k+l} = 0.$$

It remains to employ Lemma 2 and the result follows. □

Remark 2 Note that the Fourier coefficients $\widehat{g_{+,-k+1}^{-1}}, \widehat{g_{+,-k+2}^{-1}}, \ldots, \widehat{g_{+,-1}^{-1}}$ in (6) are all equal to zero.

Remark 3 Theorem 5 shows that the kernel of $T_p(g)$ possesses a basis which is formed by a so-called V-chain, where V is the forward shift on l^p, that is $V = T_p(\chi_1)$. Indeed, relations (6) can be rewritten as

$$\ker T_p(g) = \operatorname{lin\,span}\{V^l \mathscr{F} g_+^{-l} : l = 0, 1, \ldots, k-1\}.$$

Note that for $g \in W(\mathbb{T})$ this result is well-known [10].

Corollary 2 *Let $g \in M^p$ and $T_p(g)$ and $\mathbf{T}_q(g)$ be Fredholm operators with coinciding indices. If g_+ is the plus factor in the Wiener-Hopf factorization of the function g in H^q, then $\mathscr{F} g_+^{-1} \in l^p$.*

Proof Without loss of generality, we may assume that

$$\operatorname{ind} \mathbf{T}_q(g) = 1.$$

Then $g_+^{-1} \in \ker \mathbf{T}_q(g)$ and $\mathscr{F} g_+^{-1} \in l^p$ by Lemma 2. □

Remark 4 Let $a, b \in M^p$. If the operator $T_p(a) + H_p(b)$ is subject to the Coburn-Simonenko Theorem, i.e. the kernel or cokernel of this operator is trivial, and if it is a Fredholm operator with the index zero, then $T_p(a) + H_p(b)$ is invertible.

For instance, let $a, b \in C_p$ and a be invertible in $C(\mathbb{T})$ with $\operatorname{wind} a = 0$. Then $T_p(a)$ is invertible on l^p and the operator $H_p(b)$ is compact. Hence, $T_p(a) + H_p(b)$ is Fredholm with the index zero. If (a, b) is a matching pair with the subordinated pair (c, d), then in the following cases

1. $\operatorname{ind} T(c) = 0$,
2. $\operatorname{ind} T(c) = 1$ and $\boldsymbol{\sigma}(c) = 1$,
3. $\operatorname{ind} T(c) = -1$ and $\boldsymbol{\sigma}(c) = -1$,

the Coburn-Simonenko Theorem is in force. This result can be proven similarly to Corollary 6.4 in [4].

Thus the kernel of a Toeplitz operator $T_p(a)$ admits the representation (6). Our next goal is to find efficient representations for the kernels of Toeplitz plus Hankel operators $T_p(a) + H_p(b)$. As we will see later in Sect. 4, such representations may contain inverses of certain auxiliary Toeplitz operators. However, to the best of authors' knowledge, exact formulas for the inverses of Toeplitz operators acting on spaces l^p, are not well developed. Therefore, we are going to discuss this issue here.

Theorem 6 *Let $g \in M^p$ be a function such that $\mathbf{T}_q(g)$ and $T_p(g)$ are invertible operators. Then the function g admits a Wiener-Hopf factorization*

$$g = g_- g_+, \tag{7}$$

in H^q and

$$T_p^{-1}(g) := \mathscr{T}(g_+^{-1})\mathscr{T}(g_-^{-1}),$$

where

$$\mathscr{T}(g_+^{-1}) := (\widehat{g}_{+,j-k}^{-1})_{j,k=0}^{\infty}, \quad \mathscr{T}(g_-^{-1}) := (\widehat{g}_{-,j-k}^{-1})_{j,k=0}^{\infty},$$

and $(\widehat{g}_{+,n}^{-1})_{n\in\mathbb{Z}}$ and $(\widehat{g}_{-,n}^{-1})_{n\in\mathbb{Z}}$ are the sequences of the Fourier coefficients of the functions g_+^{-1} and g_-^{-1}.

Proof Let us first recall that if $a \in L^q(\mathbb{T})$, $b \in L^p(\mathbb{T})$ then the n-th Fourier coefficient $\widehat{(ab)}_n$ of the function ab is

$$\widehat{(ab)}_n = \sum_{l\in\mathbb{Z}} \widehat{a}_{n-l}\widehat{b}_l, \quad n \in \mathbb{Z}, \tag{8}$$

where $\widehat{a}_l$ and $\widehat{b}_l$ are the Fourier coefficients of the functions a and b, correspondingly. Note that $ab \in L^1(\mathbb{T})$ and the series in the right-hand side of (8) converges for any $n \in \mathbb{Z}$.

Further, if the operator $\mathbf{T}_q(g)$ is invertible, then g admits a Wiener-Hopf factorization (7) such that $g_+ \in H^p$, $g_+^{-1} \in H^q$, $g_- \in \overline{H^q}$, $g_-^{-1} \in \overline{H^p}$, and the linear operator $g_+^{-1}\mathbf{P}g_-^{-1}\mathbf{I} : H^q \to H^q$ is bounded [1]. Consider the sequences $(\widehat{g}_{+,n})_{n\in\mathbb{Z}}$ and $(\widehat{g}_{-,n})_{n\in\mathbb{Z}}$ of the Fourier coefficients of the functions g_+ and g_- and the corresponding Toeplitz matrices

$$\mathscr{T}(g_+) := (\widehat{g}_{+,j-k})_{j,k=0}^{\infty}, \quad \mathscr{T}(g_-) := (\widehat{g}_{-,j-k})_{j,k=0}^{\infty}.$$

Computing the entry $g_{n,k}$ of the product $\mathscr{T}(g_-)\mathscr{T}(g_+)$ and taking into account that $\mathscr{T}(g_-)$ is a Toeplitz matrix, one obtains

$$g_{n,k} = \sum_{l=0}^{\infty} \widehat{g}_{-,n-l}\widehat{g}_{+,l-k} = \sum_{l=0}^{\infty} \widehat{g}_{-,n-k-(l-k)}\widehat{g}_{+,l-k}.$$

This leads to the relation

$$g_{n,k} = \sum_{j=0}^{\infty} \widehat{g}_{-,n-k-j}\widehat{g}_{+,j} = \widehat{g}_{n-k}, \tag{9}$$

since the function g allows representation (7), $\widehat{g}_{+,l-k} = 0$ for $l < k$ and the series in (9) converges by (8). Thus the matrix $\mathscr{T}(g) = (\widehat{g}_{j-k})_{j,k=0}^{\infty}$ can be represented as $\mathscr{T}(g) = \mathscr{T}(g_-)\mathscr{T}(g_+)$. We already know that the matrix $\mathscr{T}(g)$ generates a bounded linear operator, namely, $T_p(g)$. It is easily seen that

$$T_p(g)e_n = \mathscr{T}(g_-)\mathscr{T}(g_+)e_n, \tag{10}$$

where $e_n = (\delta_{i,n})_{i\in\mathbb{Z}_+}$ and $\delta_{i,n}$ is the Kronecker symbol. Clearly, the set $(e_n)_{n\in\mathbb{Z}_+}$ forms a Schauder basis in l^p. Relation (10) indicates that $\mathscr{T}(g_-)\mathscr{T}(g_+)$ is a bounded linear operator on the dense subset of l^p consisting of all linear combinations of a finite number of the basis elements from $(e_n)_{n\in\mathbb{Z}_+}$, i.e. on $\cup_{n=0}^{\infty} l_n^p$, where

$$l_n^p := \left\{ \sum_{i=0}^{n} c_i e_i : c_i \in \mathbb{C}, i = 0, 1, \cdots, n \right\}.$$

The equality (10) also suggests that $T_p^{-1}(g)$ can be expressed as $\mathscr{T}(g_+^{-1})\mathscr{T}(g_-^{-1})$. We will make sure that this is the case, indeed. It is easily seen that the compressions of $\mathscr{T}(g_-)$ and $\mathscr{T}(g_-^{-1})$ on l_n^p are well defined and map l_n^p into l_n^p, $n \in \mathbb{Z}_+$. Moreover, these compressions are the inverses to each other. Let us denote $\mathscr{T}(g_-)e_i$ by m_i. Then $(m_i)_{i=0}^{n}$ actually forms a basis in l_n^p. Now we get that $\mathscr{T}(g_+^{-1})\mathscr{T}(g_-^{-1})m_i = \mathscr{T}(g_+^{-1})e_i = V^i(\widehat{g_{+,k}^{-1}})_{k=0}^{\infty} \in l^p$ by Remark 3 (recall that V is the forward shift given by $T_p(\chi_1)$). Further, consider the element $T_p(g)V^i(\widehat{g_{+,k}^{-1}})_{k=0}^{\infty}$. The n_0-th entry in the last sequence is

$$\sum_{k=0}^{\infty} \widehat{g}_{n_0-k}\widehat{g_{+,k-i}^{-1}} = \sum_{l=0}^{\infty} \widehat{g}_{n_0+i-l}\widehat{g_{+,l}^{-1}}.$$

Now it is easily seen that the last sum is

$$\widehat{(gg_{+}^{-1}\chi_i)}_{n_0} = \widehat{(g_{-}^{-1}\chi_i)}_{n_0},$$

so that

$$T_p(g)V^i(\widehat{g}_{+,k}^{-1})_{k=0}^{\infty} = (\widehat{g}_{-,i}^{-1}, \widehat{g}_{-,i-1}^{-1}, \cdots, \widehat{g}_{-,0}^{-1}, 0, \cdots) = m_i.$$

The invertibility of $T_p(g)$ immediately leads to the identity

$$T_p^{-1}(g)m_i = \mathscr{T}(g_{+}^{-1})\mathscr{T}(g_{-}^{-1})m_i,$$

and

$$T_p(g)^{-1}m = \mathscr{T}(g_{+}^{-1})\mathscr{T}(g_{-}^{-1})m \tag{11}$$

for all $m \in \cup_{n=0}^{\infty} l_n^p$. Moreover, the relation (11) allows us to determine the matrix of the operator $T_p^{-1}(g)$ with respect to the basis $(e_i)_{i=0}^{\infty}$ and the proof of Theorem 6 is completed. □

4 Kernels of Toeplitz Plus Hankel Operators Acting on Spaces l^p

The considerations of the previous section suggest the idea to use the kernels of $\mathbf{T}_q(a)+\mathbf{H}_q(b)$ in order to obtain the description of ker $(T_p(a)+H_p(b))$. However, the situation is not so simple. One can prove that these operators are subject to Lemma 2 but the proof is involved and we are not going to follow along this line of thinking. One of the main reasons is the absence of the description of ker$(\mathbf{T}_q(a) + \mathbf{H}_q(b))$ in general situation. Nevertheless, for particular classes of Toeplitz plus Hankel operators on H^p some results have been obtained recently [4]. Thus if the generating functions a and b of the operators $\mathbf{T}_q(a)$ and $\mathbf{H}_q(b)$ satisfy the relations

$$a \in GL^{\infty},\ b \in L^{\infty} \quad \text{and} \quad a(t)a(1/t) = b(t)b(1/t), \quad t \in \mathbb{T}, \tag{12}$$

then under suitable conditions the kernel of the operator $\mathbf{T}_q(a) + \mathbf{H}_q(b)$ can be completely described via the kernels of two Toeplitz operators with generating functions from a special class. It turns out that this idea also works for Toeplitz plus Hankel operators acting on l^p-spaces.

As was mentioned in Section 3 of [4], the approach developed there can also be used to study Toeplitz plus Hankel operators acting on l^p-spaces. The proof of the corresponding results are similar to those for Toeplitz plus Hankel operators acting on Hardy spaces H^p. Note that the condition (12) reads now as follows:

$$a \in GM^p,\ b \in M^p \quad \text{and} \quad a(t)a(1/t) = b(t)b(1/t), \quad t \in \mathbb{T}. \tag{13}$$

If $a, b \in L^\infty$ ($a, b \in M^p$) satisfy the condition (12) ((13)), then the duo (a, b) is called matching pair. For a matching pair (a, b), the duo (c, d),

$$c := ab^{-1}(= \widetilde{b}\widetilde{a}^{-1}), \quad d := a\widetilde{b}^{-1}(= b\widetilde{a}^{-1})$$

is also a matching pair with the additional property

$$c\tilde{c} = 1, \quad d\tilde{d} = 1.$$

Recall that $\widetilde{a}(t) = a(1/t)$ for any $a \in L^\infty$.

The duo (c, d) is called the subordinated pair for the pair (a, b). Further, a matching pair (a, b) is called Fredholm if the Toeplitz operators with generating functions c and d are Fredholm on the spaces under consideration. It follows from relations (3.2) and (3.7) in [4] that $T_p(a) \pm H_p(b)$ are Fredholm if and only if so are both operators $T_p(c)$ and $T_p(d)$.

In what follows, any function $g \in L^\infty$ ($g \in M^p$) satisfying the relation $g\widetilde{g} = 1$ is called matching function. We recall some results from [4], formulating them for Toeplitz plus Hankel operators acting on the space l^p. To shorten the notation, we drop the subscript p when confusion is unlikely.

For $g \in M^p$ let us define the operators $P_g^\pm : l^p \to l^p$ by

$$P_g^\pm := \frac{1}{2}(I \pm JPL(g)Q).$$

Lemma 3 (Proposition 3.4 in [4]) *If $g \in M^p$ is a matching function, then the operators $P_g^\pm$ are complementary projections on the space* $\ker T(g)$.

Lemma 4 (Corollary 3.5 in [4]) *Let (c, d) be the subordinated pair for a matching pair $(a, b) \in M^p \times M^p$. Then the following relations*

$$\begin{aligned} \ker T(c) &= \operatorname{im} P_c^- \dotplus \operatorname{im} P_c^+, \\ \operatorname{im} P_c^- &\subset \ker(T(a) + H(b)), \\ \operatorname{im} P_c^+ &\subset \ker(T(a) - H(b)), \end{aligned}$$

hold.

Suppose that the operator $T(c)$ is right-invertible. An example of a right invertible Fredholm operator is $T(c)$ with $k = \operatorname{ind} T(c) \geq 0$, one of the right-inverses of which is the operator $T^{-1}(c\chi_k)T(\chi_k)$. Notice that $\operatorname{ind} T(c\chi_k) = 0$ and $T(c\chi_k)$ is invertible. If $T(c)$ is right-invertible, then $T_r^{-1}(c)$ stands for one of the right inverses of $T(c)$.

Now we can define the operators $\varphi_{\pm} : l^p \to l^p$ by

$$\varphi_{\pm}(s) := \frac{1}{2}(T_r^{-1}(c)T(\widetilde{a}^{-1}) \mp JQL(c)PT_r^{-1}(c)T(\widetilde{a}^{-1}) \pm JQL(\widetilde{a}^{-1}))s. \tag{14}$$

Lemma 5 (Proposition 3.7 in [4]) *Let (c, d) be the subordinated pair for a matching pair $(a, b) \in M^p \times M^p$. If the operator $T(c) : l^p \to l^p$ is right-invertible, then*

$$\begin{aligned} \ker(T(a) + H(b)) &= \varphi_+(\operatorname{im} P_d^+) \dotplus \operatorname{im} P_c^-, \\ \ker(T(a) - H(b)) &= \varphi_-(\operatorname{im} P_d^-) \dotplus \operatorname{im} P_c^+. \end{aligned} \tag{15}$$

If $(a, b) \in M^p \times M^p$ is a Fredholm matching pair, and this will be assumed in what follows, then

$$\begin{aligned} \dim\ker(T(a) + H(b)) &= \dim\operatorname{im} P_d^+ \dotplus \dim\operatorname{im} P_c^-, \\ \dim\ker(T(a) - H(b)) &= \dim\operatorname{im} P_d^- \dotplus \dim\operatorname{im} P_c^+, \end{aligned}$$

provided that $\operatorname{ind} T(c) \geq 0$. These relations are also valid if $\operatorname{ind} T(c) < 0$ and $\operatorname{ind} T(d) \leq 0$. However, in the situation where $\operatorname{ind} T(c) < 0$ and $\operatorname{ind} T(d) > 0$ they may fail. Nevertheless, the last case still can be studied by using a special representation of the Toeplitz plus Hankel operator under consideration. More precisely, if n is a natural number, then the operator $T(a) + H(b)$ can be represented as the product of two operators, namely,

$$T(a) + H(b) = (T(a\chi_{-n}) + H(b\chi_n))T(\chi_n), \tag{16}$$

(see relation (2.4) in [4]). Let us choose $n \in \mathbb{N}$ such that

$$0 \leq 2n + \operatorname{ind} T(c) \leq 1. \tag{17}$$

Such a number n is uniquely defined and

$$2n + \operatorname{ind} T(c) = \begin{cases} 0, \text{ if } \operatorname{ind} T(c) \text{ is even,} \\ 1, \text{ if } \operatorname{ind} T(c) \text{ is odd.} \end{cases}$$

In addition, we observe that $(a\chi_{-n}, b\chi_n)$ is also a matching pair with the subordinated pair $(c\chi_{-2n}, d)$. However, $\operatorname{ind} T(c\chi_{-2n}) \geq 0$, so that the operator $T(a\chi_{-n}) + H(b\chi_n)$ is subject to Lemma 5. This leads to the following result.

Lemma 6 *Assume that $(a, b) \in M^p \times M^p$ be a Fredholm matching pair,* $\operatorname{ind} T(c) < 0$ *and* $\operatorname{ind} T(d) > 0$*, and let n be the natural number defined by the relation* (17).

1. *If n is even, then $T(c\chi_{-2n})$ is invertible and*

$$\ker(T(a\chi_{-n})+H(b\chi_n))=\varphi_+(\operatorname{im}P_d^+),$$
$$\ker(T(a\chi_{-n})-H(b\chi_n))=\varphi_-(\operatorname{im}P_d^-).$$

2. *If n is odd, then $\operatorname{ind}T(c\chi_{-2n}) = 1$ and $\ker T(c\chi_{-2n})$ is a one-dimensional subspace of l^p. Moreover,*

$$\ker(T(a\chi_{-n})+H(b\chi_n))=\varphi_+(\operatorname{im}P_d^+)\dot{+}\operatorname{im}P^-_{c\chi_{-2n}},$$
$$\ker(T(a\chi_{-n})-H(b\chi_n))=\varphi_-(\operatorname{im}P_d^-)\dot{+}\operatorname{im}P^+_{c\chi_{-2n}}.$$

We would like to draw the reader's attention to the facts that the operators $\varphi_\pm$ in the last two identities are generated by the matching pair $(a\chi_{-n}, b\chi_n)$ and that one of the subspaces $\operatorname{im}P^\pm_{c\chi_{-2n}}$ has dimension zero whereas the dimension of the other is one.

If $\kappa_1 := \operatorname{ind}T(c)$, $\kappa_2 := \operatorname{ind}T(d)$, then relation (16) leads to the following result.

Lemma 7 *Assume that $(\kappa_1,\kappa_2) \in (-\mathbb{N})\times\mathbb{N}$ and let n be the integer defined by* (17).

1. *If κ_1 is even, then*

$$\ker(T(a)\pm H(b)) = T(\chi_{-n})\left(\{\varphi_\pm(\operatorname{im}P_d^\pm)\}\cap\operatorname{im}T(\chi_n)\right)$$
$$=\left\{\psi\in\{T(\chi_{-n})u\} : u\in\varphi_\pm(\operatorname{im}P_d^\pm) \text{ and } \widehat{u}_0=\cdots=\widehat{u}_{n-1}=0\right\}.$$

2. *If κ_1 is odd, then*

$$\ker(T(a)\pm H(b)) = T(\chi_{-n})\left(\left\{\operatorname{im}P^\mp_{c\chi_{-2n}}\dot{+}\varphi_\pm(\operatorname{im}P_d^\pm)\right\}\cap\operatorname{im}T(\chi_n)\right)=$$
$$\left\{\psi\in\{T(\chi_{-n})u\} : u\in\left\{\operatorname{im}P^\mp_{c\chi_{-2n}}\dot{+}\varphi_\pm(\operatorname{im}P_d^\pm)\right\} \text{ and } \widehat{u}_0=\cdots=\widehat{u}_{n-1}=0\right\}.$$

Thus what remains now is to describe the subspaces $\operatorname{im}P_c^\pm$, $\operatorname{im}P_d^\pm$ and $\operatorname{im}P^\pm_{c\chi_{-2n}}$. For this we again need a results from [4].

Proposition 2 (See Proposition 5.1, Corollary 5.3 and Theorem 5.4 of [4]) *Assume that $g\in L^\infty$ satisfy the relation $g\widetilde{g}=1$. Then*

1. *If the operator $\mathbf{T}_q(g)$ is Fredholm with the index n, then the function g admits Wiener-Hopf factorization in H^q,*

$$g(t)=g_+(t)\,\chi_{-n}\,g_-(t),\quad g_-(\infty)=1, \tag{18}$$

where $g_+\in H^p$, $g_+^{-1}\in H^q$, $g_-(t)=\boldsymbol{\sigma}(g)\widetilde{g}_+^{-1}(t)$, and $\boldsymbol{\sigma}(g)=g_+(0)=\pm1$ (see Remark 5 below).

2. *If* $k = \operatorname{ind} \mathbf{T}_q(g) > 0$, *then the operators*

$$\mathbf{P}_g^{\pm} := (1/2)(\mathbf{I} \pm \mathbf{J}\mathbf{Q}gP) : \ker \mathbf{T}_q(g) \to \ker \mathbf{T}_q(g),$$

are complementary projections.

3. *If* (18) *is the Wiener–Hopf factorization of g in* H^q *and* $n > 0$, *then the following systems of functions* $\mathbf{B}_{\pm}(g)$ *form bases in the spaces* $\operatorname{im} \mathbf{P}_g^{\pm}$*:*

   a. *If* $n = 2m$, $m \in \mathbb{N}$, *then*

$$\mathbf{B}_{\pm}(g) := \{g_+^{-1}(t^{m-k-1} \pm \boldsymbol{\sigma}(g)t^{m+k}) : k = 0, 1, \cdots, m-1\}.$$

   b. *If* $n = 2m + 1$, $m \in \mathbb{Z}_+$, *then*

$$\mathbf{B}_{\pm}(g) := \{g_+^{-1}(t^{m+k} \pm \boldsymbol{\sigma}(g)t^{m-k}) : k = 0, 1, \cdots, m\} \setminus \{0\}.$$

Remark 5 If g is a matching function and the operator $\mathbf{T}_q(g)$ is Fredholm, then $g_+(0)$ takes only two distinct values 1 and -1. The corresponding value, denoted by $\boldsymbol{\sigma}(g)$, is called the factorization signature of the matching function g. Its role can be seen from the relation $\boldsymbol{\sigma}(g)\widetilde{g}_+^{-1}(\infty) = g_-(\infty) = 1$.

Theorem 7 *Let* $g \in M^p$ *be a matching function such that both operators* $T_p(g)$ *and* $\mathbf{T}_q(g)$ *are Fredholm and* $\operatorname{ind} \mathbf{T}_q(g) = \operatorname{ind} T_p(g)$. *If* (18) *is the Wiener-Hopf factorization of g in* H^q *and* $n > 0$, *then the following systems of sequences* $\mathscr{B}_{\pm}$ *form bases in the spaces* $\operatorname{im} P_g^{\pm}$*:*

1. *If* $n = 2m, m \in \mathbb{N}$, *then*

$$\mathscr{B}_{\pm} := \Big\{(\widehat{g}_{+,j-(m-k-1)}^{-1})_{j \in \mathbb{Z}_+} \\ \pm \boldsymbol{\sigma}(g)(\widehat{g}_{+,j-(m+k)}^{-1})_{j \in \mathbb{Z}_+} : k = 0, 1, \cdots, m-1\Big\}, \tag{19}$$

   where here and in what follows $(\widehat{g}_{+,j}^{-1})_{j \in \mathbb{Z}_+} = \mathscr{F} g_+^{-1}$, *i.e.* $\widehat{g}_{+,j}^{-1}$ *are the Fourier coefficients of the function* g_+^{-1}.

2. *If* $n = 2m + 1, n \in \mathbb{N}$, *then*

$$\mathscr{B}_{\pm} := \Big\{(\widehat{g}_{+,j-(m+k)}^{-1})_{j \in \mathbb{Z}_+} \\ \pm \boldsymbol{\sigma}(g)(\widehat{g}_{+,j-(m-k)}^{-1})_{j \in \mathbb{Z}_+} : k = 0, 1, \cdots, m\Big\} \setminus \{0\}. \tag{20}$$

Proof Following the proof of Theorem 5, we get

$$\ker T_p(g) = \operatorname{lin\,span}\{(\widehat{g}_{+,n-l}^{-1})_{n \in \mathbb{Z}_+} : l = 0, 1, \ldots, k-1\}.$$

It is clear that the sequences $(\widehat{g}_{+,n-l}^{-1})_{n \in \mathbb{Z}_+}$ form a basis in $\ker T_p(g)$ and so does the system $\mathscr{B}_+ \cup \mathscr{B}_-$.Thus we only have to prove that $P_g^{\pm}\mathscr{B}_{\pm} = \mathscr{B}_{\pm}$. Observe that the matrix representation $[(1/2)(\mathbf{I} \pm \mathbf{JQ}g\mathbf{P})]_{\mathscr{X}}$ of the operator $(1/2)(\mathbf{I} \pm \mathbf{JQ}g\mathbf{P})$ is

$$\left[\frac{1}{2}(\mathbf{I} \pm \mathbf{JQ}g\mathbf{P})\right]_{\mathscr{X}} = \frac{1}{2}\left(I \pm (g_{-j-k-1})_{j,k=0}^{\infty}\right),$$

because $\mathbf{JQ}g\mathbf{P} = \mathbf{P}\widetilde{g}\mathbf{QJ}$. Note that the matrix generating the operator $JQL(g)P = PL(\widetilde{g})QJ$ on l^p coincides with the matrix representation of $\mathbf{JQ}g\mathbf{P}$. Taking into account the relations $\mathbf{P}_g^{\pm}\mathbf{B}_{\pm} = \mathbf{B}_{\pm}$ and $\mathscr{B}_{\pm} = \mathscr{F}\mathbf{B}_{\pm}$, we obtain

$$\mathscr{B}_{\pm} = \left[\frac{1}{2}(\mathbf{I} \pm \mathbf{JQ}g\mathbf{P})\right]_{\mathscr{X}} \mathscr{F}\mathbf{B}_{\pm} = \frac{1}{2}(I \pm JQL(g)P)\mathscr{B}_{\pm},$$

which completes the proof. □

Remark 6 Let (a, b) be a Fredholm matching pair. It is easily seen that the adjoints to the operators $\mathbf{T}_q(a) + \mathbf{H}_q(b)$ and $T_p(a) + H_p(b)$ are $\mathbf{T}_p(\overline{a}) + \mathbf{H}_p(\widetilde{\overline{b}})$ and $T_q(\overline{a}) + H_q(\widetilde{\overline{b}})$, respectively. Recall that $(\overline{a}, \widetilde{\overline{b}})$ forms again a matching pair with the subordinated pair $(\overline{d}, \overline{c})$. Therefore, the Fredholm indices of the operators $\mathbf{T}_p(\overline{c})$ and $T_q(\overline{c})$ coincide if so are those of $\mathbf{T}_q(c)$ and $T_p(c)$. Similar assertion is also true for the operators $\mathbf{T}_p(\overline{d})$ and $T_q(\overline{d})$. Now it can be shown that the Fredholm indices of the operators $\mathbf{T}_p(\overline{a}) + \mathbf{H}_p(\widetilde{\overline{b}})$ and $T_q(\overline{a}) + H_q(\widetilde{\overline{b}})$ coincide. Moreover, it is also possible to study the invertibility of the operators under consideration.

Consider examples where a complete description of the kernels of the operators $T_p(a) + H_p(b)$ can be derived.

Theorem 8 *Let $(a, b) \in PC_p \times PC_p$ be a Fredholm matching pair with the subordinated pair (c, d), and let $\widehat{c_{+,j}^{-1}}$, $j \in \mathbb{Z}_+$ be the Fourier coefficients of the function c_+^{-1}, where c_+ is the plus factor in the Wiener-Hopf factorization (18) of the function c in H^q. If $\kappa_1 := \operatorname{ind} T_p(c) > 0$, $\kappa_2 := \operatorname{ind} T_p(d) \leq 0$, then the kernel of the operator $T_p(a) + H_p(b)$ admits the following representation:*

1. *If $\kappa_1 = 1$ and $\sigma(c) = 1$, then*

$$\ker(T_p(a) + H_p(b)) = \{0\}.$$

2. *If $\kappa_1 = 1$ and $\sigma(c) = -1$, then*

$$\ker(T_p(a) + H_p(b)) = \operatorname{lin\,span}\{(\widehat{c_{+,j}^{-1}})_{j\in\mathbb{Z}_+}\}.$$

3. *If $\kappa_1 > 1$ is odd, then*

$$\begin{aligned} &\ker(T_p(a) + H_p(b)) = \\ &= \operatorname{lin\,span}\{(\widehat{c_{+,j-(\kappa_1-1)/2-l}^{-1}} - \boldsymbol{\sigma}(c)\widehat{c_{+,j-(\kappa_1-1)/2+l}^{-1}})_{j\in\mathbb{Z}_+} : l = 0, 1, \cdots, (\kappa_1 - 1)/2\}. \end{aligned}$$

4. *If κ_1 is even, then*

$$\begin{aligned}&\ker(T_p(a)+H_p(b)) =\\ &\quad = \operatorname{lin\,span}\{(\widehat{c}^{-1}_{+,j-\kappa_1/2+l+1} - \boldsymbol{\sigma}(c)\widehat{c}^{-1}_{+,j-\kappa_1/2-l})_{j\in\mathbb{Z}_+} : l = 0, 1, \cdots, \kappa_1/2-1\}.\end{aligned}$$

Proof Let the operators $T_p(c)$ and $T_p(d)$ be as above. According to Lemma 5 the kernel of the operator $T_p(a) + H_p(b)$ can be expressed in the form

$$\ker(T_p(a) + H_p(b)) = \operatorname{im} P_c^-.$$

Further, since the matching function $c \in PC_p$, by Corollary 1 the operator $\mathbf{T}_q(c)$ is also Fredholm and it has the same index as the operator $T_p(c)$. It remains to use Theorem 7 and representations (19)–(20). □

Thus in the case where $\operatorname{ind} T_p(c) > 0$ and $\operatorname{ind} T_p(d) \le 0$ we have a complete description of the $\ker T_p(a) + H_p(b)$. The kernel spaces arising in Theorem 8 we denote by $\mathfrak{N}^p(c)$.

Next we are going to derive the kernel description of the operator $T_p(a) + H_p(b)$ in another situation. Let c and d be as above. Assume now that $\kappa_1 = \operatorname{ind} T_p(c) > 0$ and $\kappa_2 = \operatorname{ind} T_p(d) > 0$ and let the sequences $(\psi_j^{(l)}(d))_{j\in\mathbb{Z}_+}$ be defined by

$$\begin{aligned}&\psi_j^{(l)}(d) :=\\ &\begin{cases} \widehat{d}^{-1}_{+,j-\kappa_2/2+l+1} + \boldsymbol{\sigma}(d)\widehat{d}^{-1}_{+,j-\kappa_2/2-l}, & l = 0, 1, \ldots, \kappa_2/2-1 \quad \text{if } \kappa_2 \text{ is even,}\\ \widehat{d}^{-1}_{+,j-\kappa_2+1)/2-l} + \boldsymbol{\sigma}(d)\widehat{d}^{-1}_{+,j-(\kappa_2-1)/2+l}, & l = 0, 1, \ldots, (\kappa_2-1)/2 \text{ if } \kappa_2 \text{ is odd,}\end{cases}\end{aligned}$$

where $\widehat{d}^{-1}_{+,j}, j \in \mathbb{Z}_+$ are the Fourier coefficients of the function d_+^{-1}, and d_+ is the plus factor in the Wiener-Hopf factorization (18) of the matching function d in H^q.

Let $\varphi_+ : l^p \to l^P$ be the operator defined by (14). We specify the set $\mathfrak{N}^p_{\varphi_+}(d) \subset l^p$ as follows

1. If $\kappa_2 = 1$ and $\sigma(d) = -1$, then

$$\mathfrak{N}^p_{\varphi_+}(d) := \{0\}.$$

2. If $\kappa_2 = 1$ and $\sigma(d) = 1$, then

$$\mathfrak{N}^p_{\varphi_+}(d) := \operatorname{lin\,span}\{\varphi_+(\psi_j^{(0)}(d))\}.$$

3. If $\kappa_2 > 1$, then

$$\mathfrak{N}^p_{\varphi_+}(d) =: \operatorname{lin\,span}\{\varphi_+(\psi_j^{(l)}(d))\},$$

and $l = 0, 1, \cdots, (\kappa_2-1)/2$ if κ_2 is odd or $l = 0, 1, \cdots, \kappa_2/2-1$ if κ_2 is even.

Theorem 9 *Let* $(a, b) \in PC_p \times PC_p$ *be a Fredholm matching pair with the subordinated pair* (c, d) *and let* $\kappa_1 := \operatorname{ind} T_p(c) > 0$, $\kappa_2 := \operatorname{ind} T_p(d) > 0$. *Then the operator* $T_p(a) + H_p(b)$ *is also Fredholm and its kernel admits the following representation:*

$$\ker(T_p(a) + H_p(b)) = \mathfrak{N}^p(c) \dotplus \mathfrak{N}^p_{\varphi_+}(d).$$

Proof Once again one can employ Corollary 1 and obtain that the operators $\mathbf{T}_q(c)$ and $\mathbf{T}_q(d)$ are Fredholm and have the indices κ_1 and κ_2, correspondingly. Therefore, by Theorem 7, $\operatorname{im} P_c^- = \mathfrak{N}^p$ and $\varphi_+(\operatorname{im} P_d^+) = \mathfrak{N}^p_{\varphi_+}(d)$. Taking in the account the first relation in (15), one obtains the result. □

It remains to study the situation where $\kappa_1 := \operatorname{ind} T_p(c) < 0, \kappa_2 := \operatorname{ind} T_p(d) > 0$. This can be done analogously to previous considerations in Theorem 8 and Theorem 9, if one uses Lemma 7.

5 Concluding Remarks

The description of the kernels of the operators $T_p(a) \pm H_p(b)$ has been obtained under the condition that the related operators $T_p(c)$ and $T_p(d)$ are Fredholm. It turns out that even if this condition is violated, then one of the operators $T_p(a) - H_p(b)$ or $T_p(a) + H_p(b)$ can still be Fredholm. In the case where $a, b \in PC_p$, the kernel of the corresponding Fredholm operator can be also described. For, one has to use the methods of [4] where this situation is considered in the H^p-setting.

Moreover, the approach developed in the present work can also be used to study other objects where the multiplier problem is involved. For example, it can be employed to Wiener-Hopf plus Hankel operators. However, such an extension is highly non-trivial and will be reported elsewhere.

References

1. Böttcher, A., Silbermann, B.: Analysis of Toeplitz operators. Springer Monographs in Mathematics. Springer, Berlin (2006)
2. Didenko, V.D., Silbermann, B.: Index calculation for Toeplitz plus Hankel operators with piecewise quasi-continuous generating functions. Bull. Lond. Math. Soc. **45**(3), 633–650 (2013)
3. Didenko, V.D., Silbermann, B.: Some results on the invertibility of Toeplitz plus Hankel operators. Ann. Acad. Sci. Fenn. Math. **39**(1), 443–461 (2014)
4. Didenko, V.D., Silbermann, B.: Structure of kernels and cokernels of Toeplitz plus Hankel operators. Integr. Equ. Oper. Theory **80**(1), 1–31 (2014)
5. Didenko, V.D., Silbermann, B.: Generalized inverses and solution of equations with Toeplitz plus Hankel operators. Bol. Soc. Mat. Mex. **22**(2), 645–667 (2016)

6. Didenko, V.D., Silbermann, B.: Invertibility and inverses of Toeplitz plus Hankel operators. J. Operator Theory **72**(2), 293–307 (2017)
7. Duduchava, R.: On discrete Wiener-Hopf equations in l^p spaces with weight. Soobsh. Akad. Nauk Gruz. SSR **67**(1), 17–20 (1972)
8. Duduchava, R.: The discrete Wiener-Hopf equations. Proc. A. Razmadze Math. Inst. **50**, 42–59 (1975)
9. Edwards, R.: Fourier Series. A Modern Introduction. Vol. 2. Graduate Texts in Mathematics, vol. 85. Springer, New York (1982)
10. Gohberg, I.C., Feldman, I.A.: Convolution Equations and Projection Methods for Their Solution. American Mathematical Society, Providence (1974)
11. Hagen, R., Roch, S., Silbermann, B.: Spectral theory of approximation methods for convolution equations, Operator Theory: Advances and Applications, vol. 74. Birkhäuser Verlag, Basel (1995)
12. Roch, S., Santos, P.A., Silbermann, B.: Non-commutative Gelfand Theories. A Tool-kit for Operator Theorists and Numerical Analysts. Universitext. Springer-Verlag London Ltd., London (2011)
13. Roch, S., Silbermann, B.: Algebras of convolution operators and their image in the Calkin algebra. Report MATH, vol. 90. Akademie der Wissenschaften der DDR Karl-Weierstrass-Institut für Mathematik, Berlin (1990)
14. Roch, S., Silbermann, B.: A handy formula for the Fredholm index of Toeplitz plus Hankel operators. Indag. Math. **23**(4), 663–689 (2012)
15. Silbermann, B.: The C^*-algebra generated by Toeplitz and Hankel operators with piecewise quasicontinuous symbols. Integr. Equ. Oper. Theory **10**(5), 730–738 (1987)

Probabilistic Lower Bounds for the Discrepancy of Latin Hypercube Samples

Benjamin Doerr, Carola Doerr, and Michael Gnewuch

Dedicated to Ian H. Sloan on the occasion of his 80*th birthday.*

Abstract We provide probabilistic lower bounds for the star discrepancy of Latin hypercube samples. These bounds are sharp in the sense that they match the recent probabilistic upper bounds for the star discrepancy of Latin hypercube samples proved in Gnewuch and Hebbinghaus (Discrepancy bounds for a class of negatively dependent random points including Latin hypercube samples. Preprint 2016). Together, this result and our work implies that the discrepancy of Latin hypercube samples differs at most by constant factors from the discrepancy of uniformly sampled point sets.

1 Introduction

Discrepancy measures are well established and play an important role in fields like computer graphics, experimental design, pseudo-random number generation, stochastic programming, numerical integration or, more general, stochastic simulation.

The prerelevant and most intriguing discrepancy measure is arguably the *star discrepancy*, which is defined in the following way:

Let $P \subset [0,1)^d$ be an N-point set. (We always understand an "N-point set" as a "multi-set", i.e., it consists of N points, but these points do not have to be

B. Doerr
École Polytechnique, LIX - UMR 7161, Palaiseau, France
e-mail: doerr@lix.polytechnique.fr

C. Doerr
Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, Paris, France
e-mail: Carola.Doerr@mpi-inf.mpg.de

M. Gnewuch (✉)
Christian-Albrechts-Universität Kiel, Mathematisches Seminar, Kiel, Germany
e-mail: gnewuch@math.uni-kiel.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_16

pairwise different.) We define the *local discrepancy* of P with respect to a Lebesgue-measurable test set $T \subseteq [0,1)^d$ by

$$D_N(P,T) := \left| \frac{1}{N} |P \cap T| - \lambda^d(T) \right|,$$

where $|P \cap T|$ denotes the size of the finite set $P \cap T$ (again understood as a multi-set) and λ^d the d-dimensional Lebesgue measure on $\mathbb{R}^d$. For vectors $x = (x_1, x_2, \dots, x_d)$, $y = (y_1, y_2, \dots, y_d) \in \mathbb{R}^d$ we write

$$[x,y) := \prod_{j=1}^{d} [x_j, y_j) = \{z \in \mathbb{R}^d \mid x_j \le z_j < y_j \text{ for } j = 1, \dots, d\}.$$

The *star discrepancy* of P is then given by

$$D_N^*(P) := \sup_{y \in [0,1]^d} D_N(P, [0,y)).$$

We will refer to the sets $[0,y)$, $y \in [0,1]^d$, as *anchored test boxes*.

The star discrepancy is intimately related to quasi-Monte Carlo integration via the Koksma-Hlawka inequality: For every N-point set $P \subset [0,1)^d$ we have

$$\left| \int_{[0,1)^d} f(x)\, d\lambda^d(x) - \frac{1}{N} \sum_{p \in P} f(p) \right| \le D_N^*(P) \mathrm{Var}_{\mathrm{HK}}(f),$$

where $\mathrm{Var}_{\mathrm{HK}}(f)$ denotes the variation in the sense of Hardy and Krause see, e.g., [13]. The Koksma-Hlawka inequality is sharp, see again [13]. (An alternative version of the Koksma-Hlawka inequality can be found in [8]; it says that the worst-case error of equal-weight cubatures based on a set of integration points P over the norm unit ball of some Sobolev space is exactly the star discrepancy of P.) The Koksma-Hlawka inequality shows that equal-weight cubatures based on integration points with small star discrepancy yield small integration errors. (Deterministic equal-weight cubatures are commonly called *quasi-Monte Carlo algorithms*; for a recent survey we refer to [2].) For the very important task of high-dimensional integration, which occurs, e.g., in mathematical finance, physics or quantum chemistry, it is therefore of interest to know sharp bounds for the smallest achievable star discrepancy and to be able to construct integration points that satisfy those bounds. To avoid the "curse of dimensionality" it is crucial that such bounds scale well with respect to the dimension.

The best known upper and lower bounds for the smallest achievable star discrepancy with explicitly given dependence on the number of sample points N as well as on the dimension d are of the following form: For all $d, N \in \mathbb{N}$ there exists an N-point set $P \subset [0, 1)^d$ satisfying

$$D_N^*(P) \le C\sqrt{\frac{d}{N}} \tag{1}$$

for some universal constant $C > 0$, while for all N-point sets $Q \subset [0, 1)^d$ it holds that

$$D_N^*(Q) \ge \min\left\{c_0, c\frac{d}{N}\right\}, \tag{2}$$

where $c_0, c \in (0, 1]$ are suitable constants. The upper bound (1) was proved by Heinrich et al. [7] without providing an estimate for the universal constant C. The first estimate for this constant was given by Aistleitner [1]; he showed that $C \le 9.65$. This estimate has recently been improved to $C \le 2.5287$ in [5]. All these results are based on probabilistic arguments and do not provide an explicit point construction that satisfies (1). The lower bound (2) was established by Hinrichs [9]. Observe that there is a gap between the upper bound (1) and the lower bound (2). In [6, Problem 1 & 2] Heinrich asked the following two questions:

(a) Does any of the various known constructions of low discrepancy point sets satisfy an estimate like (1) or at least some slightly weaker estimates?
(b) What are the correct sharp bounds for the smallest achievable star discrepancy?

It turned out that these two questions are very difficult to answer.

To draw near an answer, it was proposed in [5] to study the following related questions:

(c) What kind of randomized point constructions satisfy (1) in expectation and/or with high probability?
(d) Can it even be shown, by probabilistic constructions, that the upper bound (1) is too pessimistic?

As mentioned, the upper bound (1) was proved via probabilistic arguments. Indeed, Monte Carlo points, i.e., independent random points uniformly distributed in $[0, 1)^d$, satisfy this bound with high probability. In [3] it was rigorously shown that the star discrepancy of Monte Carlo point sets X behaves like the right hand side in (1). More precisely, there exists a constant $K > 0$ such that the expected star discrepancy of X is bounded from below by

$$\mathrm{E}[D_N^*(X)] \ge K\sqrt{\frac{d}{N}} \tag{3}$$

and additionally we have the probabilistic discrepancy bound

$$\mathrm{P}\left(D_N^*(X) < K\sqrt{\frac{d}{N}} \right) \leq \exp(-\Omega(d)). \tag{4}$$

The upper bound (1) is thus sharp for Monte Carlo points, showing that they cannot be employed to improve it.

What about other randomized point constructions? In [5] it is shown that so-called Latin hypercube samples satisfy the upper bound (1) with high probability, see Theorem 1 below. In this note we show that this estimate is tight. More precisely, we prove that the bounds (3) and (4) for Monte Carlo point sets also apply to Latin hypercube samples.

2 Probabilistic Discrepancy Bounds for Latin Hypercube Sampling

For $N \in \mathbb{N}$ we denote the set $\{1, 2, \ldots, N\}$ by $[N]$. The definition of Latin hypercube sampling presented below was introduced by McKay et al. [12] for the design of computer experiments.

Definition 1 A *Latin hypercube sample* (LHS) $(X_n)_{n\in[N]}$ in $[0, 1)^d$ is of the form

$$X_{n,j} = \frac{\pi_j(n) - u_{n,j}}{N},$$

where $X_{n,j}$ denotes the jth coordinate of X_n, π_j is a permutation of $[N]$ that is chosen uniformly at random, and $u_{n,j}$ obeys the uniform distribution on $[0, 1)$. The d permutations π_j and the dN random variables $u_{n,j}$ are mutually independent.

The following result was proved in [5].

Theorem 1 *Let $d, N \in \mathbb{N}$, and let $X = (X_n)_{n\in[N]}$ be a Latin hypercube sample in $[0, 1)^d$. Then for every $c > 0$*

$$\mathrm{P}\left(D_N^*(X) \leq c\sqrt{\frac{d}{N}} \right) \geq 1 - \exp\left(-(1.6741\, c^2 - 11.7042)\, d\right).$$

In particular, there exists a realization $P \subset [0, 1)^d$ of X such that

$$D_N^*(P) \leq 2.6442 \cdot \sqrt{\frac{d}{N}}$$

and the probability that X satisfies

$$D_N^*(X) \le 3 \cdot \sqrt{\frac{d}{N}} \quad \text{and} \quad D_N^*(X) \le 4 \cdot \sqrt{\frac{d}{N}}$$

is at least 0.965358 *and* 0.999999*, respectively.*

The result of our note complements the previous theorem and shows that it is sharp from a probabilistic point of view.

Theorem 2 *There exists a constant $K > 0$ such that for all $d, N \in \mathbb{N}$ with $d \ge 2$ and $N \ge 1600d$, the discrepancy of a Latin hypercube sample $X = (X_n)_{n \in [N]}$ in $[0,1)^d$ satisfies*

$$\mathrm{E}[D_N^*(X)] \ge K\sqrt{\frac{d}{N}}$$

and

$$\mathrm{P}\left(D_N^*(X) < K\sqrt{\frac{d}{N}}\right) \le \exp(-\Omega(d)).$$

We note that Theorem 2 does not hold for $d = 1$. Indeed, it is easily verified that in dimension $d = 1$ we have $D_N^*(X) \le 1/N$ almost surely.

3 Proof of Theorem 2

We now list the results that we need to prove Theorem 2. We will employ the fact that (under suitable conditions) the hypergeometric distribution resembles the binomial distribution. Let us make this statement more precise.

Consider an urn that contains N balls among which W are white and $N - W$ are black. Now we draw a random sample of size n. The number of white balls in the sample has the *hypergeometric distribution* $H(N, W, n)$ if we sample without replacement and the *binomial distribution* $B(n, p)$ with

$$p := W/N$$

if we sample with replacement. The deviation of both distributions can be measured by the *total variation distance*

$$\delta\big(H(N,W,n), B(n,p)\big) := \max_{A \subseteq \{0,1,\dots,n\}} |H(N,W,n)(A) - B(n,p)(A)|.$$

The following theorem can be found in [11, p. 1]; here we only need the upper bound, which is due to Ehm, see [4].

Theorem 3 *Let* $n, N, W \in \mathbb{N}$ *with* $W, n \le N$ *and let* $p \in (0, 1)$ *such that* $np(1-p) \ge 1$. *Then*

$$\frac{1}{28}\frac{n-1}{N-1} \le \delta\big(H(N, W, n), B(n, p)\big) \le \frac{n-1}{N-1}.$$

Furthermore, we will make use of the following lemma from [3].

Lemma 1 *Let* $n \ge 16$ *and* $1/n \le p \le 1/4$. *Then*

$$B(n, p)\left(\left[0, np - \frac{1}{2}\sqrt{np}\right]\right) \ge \frac{3}{160}.$$

Finally, we need the following Chernoff-Hoeffding bound for sums of independent Bernoulli random variables, see [10]. Recall that a Bernoulli random variable is simply a random variable that takes only values in $\{0, 1\}$.

Theorem 4 *Let* $k \in \mathbb{N}$, *and let* $\xi_1, \ldots, \xi_k$ *be independent (not necessarily identically distributed) Bernoulli random variables. Put* $S := \sum_{i=1}^{k}(\xi_i - \mathrm{E}[\xi_i])$. *Then we have for all* $t > 0$ *that*

$$\mathrm{P}\left(S < -tk\right) \le \exp\left(-2t^2k\right). \tag{5}$$

The Bernoulli random variables η_i, $i = 1, \ldots, d$, that appear in our proof of Theorem 2 are actually not independent; to cope with that we need the following lemma.

Lemma 2 *Let* $k \in \mathbb{N}$ *and* $q \in (0, 1)$. *Let* $\xi_1, \ldots, \xi_k$ *be independent Bernoulli random variables with* $\mathrm{P}(\xi_j = 1) = q$ *for all* $j \in [k]$, *and let* $\eta_1, \ldots, \eta_k$ *be (not necessarily independent) Bernoulli random variables satisfying*

$$\mathrm{P}(\eta_j = 1 \mid \eta_1 = v_1, \ldots, \eta_{j-1} = v_{j-1}) \ge q \quad \textit{for all } j \in [k] \textit{ and all } v \in \{0, 1\}^{j-1}.$$

Then we have

$$\mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t\right) \le \mathrm{P}\left(\sum_{i=1}^{j} \xi_i < t\right) \quad \textit{for all } j \in [k] \textit{ and all } t > 0. \tag{6}$$

Since we do not know a proper reference for this lemma, we provide a proof. For a finite bit string $v \in \{0, 1\}^j$ we put $|v|_1 := v_1 + \cdots + v_j$.

Proof We verify statement (6) by induction on j. For $j = 1$ statement (6) is true, since for $t \in (0, 1]$ we have $\mathrm{P}(\eta_1 < t) \le 1 - q = \mathrm{P}(\xi_1 < t)$ and for the trivial case $t > 1$ we have $\mathrm{P}(\eta_1 < t) = 1 = \mathrm{P}(\xi_1 < t)$.

Now assume that statement (6) is true for $j \in [k-1]$. This gives for $t > 0$

$$\mathrm{P}\left(\sum_{i=1}^{j+1} \eta_i < t\right) = \mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t-1\right) + \mathrm{P}\left(\eta_{j+1} = 0\,,\ \sum_{i=1}^{j} \eta_i \in [t-1,t)\right)$$

$$= \mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t-1\right) + \sum_{\substack{v\in\{0,1\}^j \\ |v|_1 \in [t-1,t)}} \mathrm{P}\left(\eta_{j+1} = 0 \,|\, \eta_1 = v_1, \ldots, \eta_j = v_j\right) \times$$

$$\times\, \mathrm{P}\left(\eta_1 = v_1, \ldots, \eta_j = v_j\right)$$

$$\leq \mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t-1\right) + (1-q)\mathrm{P}\left(\sum_{i=1}^{j} \eta_i \in [t-1,t)\right)$$

$$= q\,\mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t-1\right) + (1-q)\,\mathrm{P}\left(\sum_{i=1}^{j} \eta_i < t\right)$$

$$\leq q\,\mathrm{P}\left(\sum_{i=1}^{j} \xi_i < t-1\right) + (1-q)\,\mathrm{P}\left(\sum_{i=1}^{j} \xi_i < t\right)$$

$$= \mathrm{P}\left(\sum_{i=1}^{j} \xi_i < t-1\right) + (1-q)\mathrm{P}\left(\sum_{i=1}^{j} \xi_i \in [t-1,t)\right)$$

$$= \mathrm{P}\left(\sum_{i=1}^{j} \xi_i < t-1\right) + \mathrm{P}\left(\xi_{j+1} = 0\,,\ \sum_{i=1}^{j} \xi_i \in [t-1,t)\right)$$

$$= \mathrm{P}\left(\sum_{i=1}^{j+1} \xi_i < t\right).$$

□

For a given N-point set $P \subset [0,1)^d$ and a measurable set $B \subseteq [0,1)^d$ let us define the *excess of points from P in B* by

$$\mathrm{exc}(P,B) := |P \cap B| - N\lambda^d(B).$$

For an arbitrary anchored test box B we always have

$$D_N^*(P) \geq D_N(P,B) \geq \frac{1}{N}\mathrm{exc}(P,B). \tag{7}$$

Proof of Theorem 2 We adapt the proof approach of [3, Theorem 1] and construct recursively a random test box $B_d = B_d(X)$ that exhibits with high probability a (relatively) large excess of points $\mathrm{exc}(X, B_d)$. Due to (7) this leads to a (relatively) large local discrepancy $D_N(X, B_d)$. Put $I := [0, \lfloor N/4 \rfloor /N)$, where $\lfloor N/4 \rfloor := \max\{z \in \mathbb{Z} \mid z \leq N/4\}$. We start with $B_1 := I \times [0,1)^{d-1}$. Notice that there are exactly $\lfloor N/4 \rfloor$ points of X inside the box B_1, implying $\mathrm{exc}(X, B_1) = 0$. The recursion step is as follows: Let $j \geq 2$ and assume we already have a test box B_{j-1} that satisfies $\mathrm{exc}(X, B_{j-1}) \geq 0$ and is of the form

$$B_{j-1} := I \times \prod_{i=2}^{j-1} [0, x_i) \times [0,1)^{d-j+1},$$

where $x_i \in \{1 - c/d, 1\}$ for $i = 2, \ldots, j-1$ and c is the largest value in $(1/84, 1/80]$ that ensures $Nc/d \in \mathbb{N}$. Observe that due to $N \geq 1600\,d$ we have $Nc/d \geq 20$ and $\lambda^d(B_1) = \lambda^1(I) \in (1/5, 1/4]$. Let

$$S_j := [0,1)^{j-1} \times [1 - c/d, 1) \times [0,1)^{d-j} \quad \text{and} \quad C_j := B_{j-1} \cap S_j,$$

and put

$$Y_j := |X \cap C_j|.$$

Looking at Definition 1 one sees easily that Y_j has the hypergeometric distribution $H(N, W, n)$ with

$$W := |X \cap B_{j-1}| \quad \text{and} \quad n := |X \cap S_j| = N\frac{c}{d}.$$

Observe that

$$\frac{1}{4} \geq \lambda^d(B_{j-1}) \geq \frac{1}{5}(1 - c/d)^{d-2} \geq \frac{1}{5}(1 - c/d)^d \geq \frac{1}{5}(1 - c/2)^2 =: v \geq \frac{1}{6}, \tag{8}$$

and, due to $\mathrm{exc}(X, B_{j-1}) \geq 0$,

$$W = |X \cap B_{j-1}| \geq N\lambda^d(B_{j-1}) \geq Nv. \tag{9}$$

Put

$$p := W/N.$$

We now want to check that the conditions on p and n in Theorem 3 and Lemma 1 hold. Due to $B_{j-1} \subseteq B_1$ and $\mathrm{exc}(X, B_1) = 0$ we have $p \leq 1/4$. Furthermore, we have $n = Nc/d \geq 20$ and, due to (9) and (8), $p \geq v \geq 1/6 \geq 1/n$. This leads to

$$np(1-p) \geq 20 \cdot \frac{1}{6}\left(1 - \frac{1}{4}\right) = \frac{5}{2} > 1.$$

Hence we may apply Theorem 3 and Lemma 1 to obtain

$$\begin{aligned} \mathrm{P}\left(Y_j \le np - \frac{1}{2}\sqrt{np} \right) &\ge B(n,p)\left(\left[0, np - \frac{1}{2}\sqrt{np} \right] \right) - \delta\big(H(N,W,n), B(n,p)\big) \\ &\ge \frac{3}{160} - \frac{c}{d} \\ &\ge \frac{1}{80}. \end{aligned} \tag{10}$$

If

$$Y_j = |X \cap C_j| \le np - \frac{1}{2}\sqrt{np},$$

then put $x_j := 1 - c/d$, and otherwise put $x_j := 1$. We define

$$B_j := I \times \prod_{i=2}^{j} [0, x_i) \times [0,1)^{d-j}.$$

Before we go on, let us make a helpful observation: Put

$$\eta_i := 1_{[x_i = 1 - c/d]}(X) \qquad \text{for } i = 2, \dots, j.$$

Then η_i is a Bernoulli random variable and (10) says that $\mathrm{P}(\eta_j = 1) \ge 1/80$. Actually, due to our construction we proved a slightly stronger result, namely:

$$\mathrm{P}\left(\eta_j = 1 \mid \eta_2 = v_1, \dots, \eta_{j-1} = v_{j-2} \right) \ge 1/80 \quad \text{for all } v \in \{0,1\}^{j-2} \tag{11}$$

(since (10) holds for all values of $\eta_2, \dots, \eta_{j-1}$ that have been determined previously in the course of the construction of B_j).

We now want to estimate the excess of points of X in B_j. In the case $x_j = 1 - c/d$ we have $\lambda^d(B_j) = (1 - c/d)\lambda^d(B_{j-1})$ and thus

$$\begin{aligned} \mathrm{exc}(X, B_j) &= |X \cap B_{j-1}| - |X \cap C_j| - N(1 - c/d)\lambda^d(B_{j-1}) \\ &\ge |X \cap B_{j-1}| - np + \frac{1}{2}\sqrt{np} - N(1 - c/d)\lambda^d(B_{j-1}) \\ &= (1 - c/d)\left(|X \cap B_{j-1}| - N\lambda^d(B_{j-1}) \right) + \frac{1}{2}\sqrt{np} \\ &= (1 - c/d)\mathrm{exc}(X, B_{j-1}) + \frac{1}{2}\sqrt{W\frac{c}{d}} \\ &\ge (1 - c/d)\mathrm{exc}(X, B_{j-1}) + \frac{\sqrt{cv}}{2}\sqrt{\frac{N}{d}}, \end{aligned}$$

where in the last step we used (9).

In the case $x_j = 1$ we obviously have $B_j = B_{j-1}$ and consequently $\operatorname{exc}(X, B_j) = \operatorname{exc}(X, B_{j-1})$.

Put

$$k = k(X) := |\{\, i \in \{2, \ldots, d\} \,|\, x_i = 1 - c/d\}|.$$

Due to $(1 - c/d)^k \geq 5v$ (cf. (8)) we obtain

$$\operatorname{exc}(X, B_d) \geq k(1 - c/d)^k \frac{\sqrt{cv}}{2} \sqrt{N/d} \geq \frac{5}{2} \sqrt{cv^3}\, k \sqrt{N/d}. \tag{12}$$

Thus we get on the one hand from (7)

$$\begin{aligned} \mathrm{E}[D_N^*(X)] &\geq \frac{1}{N} \mathrm{E}[\operatorname{exc}(X, B_d)] \\ &\geq \sum_{\kappa=0}^{d-1} \frac{5}{2} \sqrt{cv^3}\, \kappa \sqrt{1/Nd}\, \mathrm{P}(k(X) = \kappa) \\ &= \frac{5}{2} \sqrt{cv^3}\, \sqrt{1/Nd} \sum_{\kappa=0}^{d-1} \kappa\, \mathrm{P}(k(X) = \kappa) \\ &= \frac{5}{2} \sqrt{cv^3}\, \sqrt{1/Nd}\, \mathrm{E}[k(X)] \\ &\geq (\sqrt{cv^3}/32\sqrt{2}) \sqrt{(d-1)/N}, \end{aligned}$$

where in the last step we used (10) to obtain

$$\mathrm{E}[k(X)] = \sum_{i=2}^{d} \mathrm{E}[\eta_i] = \sum_{i=2}^{d} \mathrm{P}\left(Y_i \leq np - \frac{1}{2}\sqrt{np} \right) \geq (d-1)/80.$$

On the other hand we get from (12) for $K := \sqrt{cv^3}/80$

$$\begin{aligned} \mathrm{P}\left(D_N^*(X) < K\sqrt{d/N}\right) &\leq \mathrm{P}\left(\operatorname{exc}(X, B_d) < K\sqrt{dN}\right) \\ &\leq \mathrm{P}\left(\frac{5}{2}\sqrt{cv^3}\, k(X) \sqrt{N/d} < K\sqrt{dN}\right) \\ &= \mathrm{P}\left(k(X) < d/200\right) \\ &= \mathrm{P}\left(\sum_{i=2}^{d} \eta_i < d/200\right). \end{aligned}$$

Let ξ_i, $i = 2, \dots, d$, be independent Bernoulli random variables with

$$\mathrm{P}(\xi_i = 1) = 1/80 \quad \text{and} \quad \mathrm{P}(\xi_i = 0) = 79/80.$$

Clearly, $\mathrm{E}[\xi_i] = 1/80$. Since estimate (11) holds for each $j \in \{2, \dots, d\}$, we have due to Lemma 2

$$\mathrm{P}\left(\sum_{i=2}^{d} \eta_i < d/200\right) \leq \mathrm{P}\left(\sum_{i=2}^{d} \xi_i < d/200\right).$$

Hence we get from Theorem 4

$$\begin{aligned}\mathrm{P}\left(D_N^*(X) < K\sqrt{d/N}\right) &\leq \mathrm{P}\left(\sum_{i=2}^{d}(\xi_i - \mathrm{E}[\xi_i]) < \left(\frac{1}{200} - \frac{1}{80}\frac{d-1}{d}\right)d\right)\\ &\leq \mathrm{P}\left(\sum_{i=2}^{d}(\xi_i - \mathrm{E}[\xi_i]) < -\frac{d}{800}\right)\\ &\leq \exp\left(-\frac{2d^2}{(800)^2(d-1)}\right)\\ &= \exp\left(-\Omega(d)\right).\end{aligned}$$

This concludes the proof of the theorem. □

Acknowledgements The authors thank two anonymous referees for their comments which helped to improve the presentation of the paper.

Part of this work was done while Michael Gnewuch was visiting the Laboratoire d'Informatique (LIX) as Chercheur Invité of École Polytechnique. He likes to thank the colleagues from LIX for their hospitality.

References

1. Aistleitner, C.: Covering numbers, dyadic chaining and discrepancy. J. Complex. **27**, 531–540 (2011)
2. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
3. Doerr, B.: A lower bound for the discrepancy of a random point set. J. Complex. **30**, 16–20 (2014)
4. Ehm, W.: Binomial approximation to the Poisson binomial distribution. Stat. Probab. Lett. **11**, 7–16 (1991)
5. Gnewuch, M., Hebbinghaus, N.: Discrepancy bounds for a class of negatively dependent random points including Latin hypercube samples. Preprint (2016)
6. Heinrich, S.: Some open problems concerning the star-discrepancy. J. Complex. **19** (Oberwolfach Special Issue), 416–419 (2003)

7. Heinrich, S., Novak, E., Wasilkowski, G.W., Woźniakowski, H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arith. **96**, 279–302 (2001)
8. Hickernell, F. J., Sloan, I.H., Wasilkowski, G.W.: On tractability of weighted integration over bounded and unbounded regions in $\mathbb{R}^s$. Math. Comput. **73**, 1885–1902 (2004)
9. Hinrichs, A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complex. **20**, 477–483 (2004)
10. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Statist. Assoc. **58**, 13–30 (1963)
11. Künsch, H.R.: The difference between the hypergeometric and the binomial distribution. Note, ETH Zürich (1998)
12. McKay, M., Beckman, R., Conover, W.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**, 239–245 (1979)
13. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 63. SIAM, Philadelphia (1992)

Hyperinterpolation for Spectral Wave Propagation Models in Three Dimensions

Mahadevan Ganesh and Stuart C. Hawkins

Dedicated to Professor Ian H. Sloan on the occasion of his 80th birthday.

Abstract In this review article, we describe some advances in applications of the hyperinterpolation operator introduced by Sloan about two decades ago (J Approx Theory 83:238–254, 1995). In particular, our focus is on reviewing the application of the scalar and vector-valued hyperinterpolation approximations for developing, analyzing and implementing fully-discrete high-order algorithms. Such approximations facilitate efficient simulation of scattering of acoustic, electromagnetic and elastic waves, exterior to connected and disconnected bounded three dimensional domains. The main contributions of this article are: (1) a unified (acoustic, electromagnetic, and elastic) approach for the three important classes of waves; (2) theoretical and numerical comparisons of the hyperinterpolation approximations in these three applications; and (3) new results for a class of unbounded heterogeneous media.

1 Introduction

In this survey article on a class of fully-discrete spectral Galerkin wave propagation models based on spherical hyperinterpolation approximations, we mark more than two decades since Professor Ian H. Sloan introduced polynomial hyperinterpolation in his seminal 1995 paper "Polynomial Interpolation and Hyperinterpolation over General Regions" [47].

M. Ganesh (✉)
Department of Applied Mathematics & Statistics, Colorado School of Mines, Golden, CO, USA
e-mail: mganesh@mines.edu

S. C. Hawkins
Department of Mathematics, Macquarie University, Sydney, NSW, Australia
e-mail: stuart.hawkins@mq.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_17

In the two decades since the publication of the paper [47], the results in the paper, and the hyperinterpolation technique it introduced, have been applied extensively in fields including approximation of functions, cubature, and methods for solution of PDEs. Particular applications include approximation of functions on various shaped regions using interpolation [51] and hyperinterpolation [5, 22, 30, 36], and generalisations of hyperinterpolation [6, 46, 49]. The results in [47] have been applied to inform development of cubature techniques [1, 24, 29, 44], and inspire development of new cubature rules suitable for hyperinterpolation [35, 37]. The hyperinterpolation technique in [47] has been a powerful tool for development and analysis of mesh-free numerical methods for solving partial differential equations (PDEs) [8, 38, 39] and integral equations [7, 11–14, 23, 25, 27, 32, 34, 40, 41].

In particular, global or local polynomial approximations are fundamental to understanding physical processes governed by partial differential equations in a bounded domain $\Omega \subset \mathbb{R}^d$ (or in its complement $\mathbb{R}^d \setminus \overline{\Omega}$) and by appropriate conditions on the boundary $\partial\Omega$. Development of an associated spectral computational PDE model requires projection of some unknown quantity onto finite dimensional spaces V_n, which are spanned by global polynomials of degree of at most n that are eigenfunctions of a second-order elliptic differential operator. A key tool to analyze the supremum norm error in the associated computer model is the growth of the Lebesgue constant of the projection operator with respect to n.

If the PDE is described in the unbounded region $\mathbb{R}^d \setminus \overline{\Omega}$ (together with a radiation condition) and if a fundamental solution of the PDE is known (as in the homogeneous constant coefficient PDE case), it is efficient to start with a surface integral ansatz. The integrand in the ansatz usually depends on the fundamental solution and an unknown density u defined on $\partial\Omega$. This approach reformulates the PDE model into a surface integral equation (SIE) for the unknown density u.

For two dimensional PDE models, spectrally accurate approximation of the SIE on $\partial\Omega$ was investigated thoroughly in the last century and is detailed in [4, 31] and references therein. Spectrally accurate approximation for the two dimensional acoustic scattering problem is comprehensively presented in [4]. The starting point for such approximations is to assume that $\partial\Omega \subset \mathbb{R}^2$ is a simply connected closed (almost everywhere) differentiable manifold and use a diffeomorphic map $\mathbf{q}^1$ from the smooth unit circle $\mathbb{S}^1$ to the boundary $\partial\Omega \subset \mathbb{R}^2$ (or equivalently, to use a 2π-periodic parameterization from $[0, 2\pi]$ to $\partial\Omega$).

The corresponding fully discrete Fourier projection $\mathscr{L}_N^1$ is defined from $\mathscr{C}(\mathbb{S}^1)$ onto V_N^1, where $V_N^1 \subset \mathscr{C}(\mathbb{S}^1)$ is spanned by orthonormal polynomials $\phi_j(\widehat{\boldsymbol{x}}) = \frac{1}{\sqrt{2\pi}} \exp(\mathrm{i}j\theta)$ for $j = -N, \cdots, N$ and $\widehat{\boldsymbol{x}} = (\cos\theta, \sin\theta)$. This projection and its properties are well known.

For a given function f in $\mathscr{C}(\mathbb{S}^1)$, the fully-discrete approximation $\mathscr{L}_N^1(f)$ to f is obtained by approximating the Fourier coefficients $\langle f, \phi_j \rangle$ in the truncated semi-discrete Fourier expansion

$$\mathscr{P}_N^1 f = \sum_{j=-N}^{N} \langle f, \phi_j \rangle \, \phi_j, \tag{1}$$

using a quadrature rule with positive weights and a certain exactness property. We denote the quadrature approximation to the Fourier coefficients as $\langle f, \phi_j \rangle_N$ and the required exactness property is

$$\langle v, w \rangle_N = \langle v, w \rangle, \quad \text{for all} \quad v, w \in V_N^1. \tag{2}$$

The rectangle rule, for example, has positive weights and satisfies the exactness property. The fully-discrete approximation, corresponding to $\mathscr{P}_N^1 f$, is

$$\mathscr{L}_N^1 f = \sum_{j=-N}^{N} \langle f, \phi_j \rangle_N \, \phi_j. \tag{3}$$

Under the assumption of (2), the approximation (3) satisfies $\mathscr{L}_N^1 v = \mathscr{P}_N^1 v$, for all $v \in V_N^1$. It is well known that the Lebesgue constant of $\mathscr{P}_N^1$ is $\mathscr{O}(\log N)$ [54, p. 67]. If the quadrature rule can be chosen so that it satisfies the exactness property (2) and that the total number of quadrature points is $Q_N^1 = 2N + 1$, then $\mathscr{L}_N^1$ in (3) is also an interpolatory operator. Indeed, $Q_N^1 = \mathrm{Dim}(V_N)$ is a necessary condition for $\mathscr{L}_N^1$ to be an interpolatory operator [47, Lemma 2]. In case $Q_N^1 > \mathrm{Dim}(V_N^1)$, then $\mathscr{L}_N^1$ is called a hyperinterpolation operator [47].

Such quadrature rules with $Q_N^1 = \mathrm{Dim}(V_N^1)$ to approximate integrals on $\mathbb{S}^1$ exist. An example is the rectangle rule with equally spaced quadrature points, and the Lebesgue constant of $\mathscr{L}_N^1$ is then of the same order as that of $\mathscr{P}_N^1$. Due to this fact, the interpolatory property plays an important role in a large number of applications, and the operator is efficiently applied using the fast Fourier transform (FFT).

The fundamental problem addressed by Sloan in [47] is the generalization of the above fully discrete Fourier projection to the higher dimensional unit sphere $\mathbb{S}^{d-1}$ for $d \geq 3$, by replacing the trigonometric basis functions ϕ_j in (3) with spherical polynomial basis functions.

In this article we focus on the SIE reformulations of three dimensional homogeneous PDEs associated with acoustic, electromagnetic, and elastic waves propagating in a heterogeneous medium exterior to a bounded three dimensional obstacle. The key is therefore efficient approximation of an unknown surface density (the surface current) on a simply connected closed almost everywhere differentiable boundary surface $\partial\Omega \subset \mathbb{R}^3$.

We assume that the boundary $\partial\Omega \subset \mathbb{R}^3$ can be mapped onto the unit sphere $\mathbb{S}^2$ by an isomorphic map $\mathbf{q}^2$. Such maps are available in many applications of interest, such as light scattering by red blood cells [28, 53]; scattering from model atmospheric particles, such as ice crystals [42] and dust particles [50], and all of the non-smooth benchmark radar targets in [52].

For approximating spherical functions in $\mathbb{S}^2$, the natural polynomial projection space is

$$V_n = \mathrm{span}\left\{ Y_{l,j} \ : \ |j| \leq l, \ 0 \leq l \leq n \right\}. \tag{4}$$

Here the orthonormal spherical harmonics $Y_{l,j}$ are eigenfunctions of the second-order spherical Laplace-Beltrami differential operator, and we use the following representation for any $\widehat{\boldsymbol{x}} \in \mathbb{S}^2$,

$$Y_{l,j}(\widehat{\boldsymbol{x}}) = (-1)^{(j+|j|)/2}\sqrt{\frac{2l+1}{4\pi}\frac{(l-|j|)!}{(l+|j|)!}}P_l^{|j|}(\cos\theta)e^{ij\phi}, \qquad l = 0, 1, \ldots; \ |j| \le l, \tag{5}$$

defined using the associated Legendre functions $P_l^{|j|}$ and the spherical polar coordinates map

$$\widehat{\boldsymbol{x}} = \boldsymbol{p}(\theta,\phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)^T, \qquad \widehat{\boldsymbol{x}} \in \mathbb{S}^2, \tag{6}$$

where θ and ϕ are the polar and azimuthal angles respectively. (The Condon-Shortly phase factor $(-1)^{(j+|j|)/2}$ in (5) is convenient for our algorithms but is not used in [47].)

The dimension of $V_n = (n+1)^2$. Here orthonormality is with respect to the $L^2(\mathbb{S}^2)$ inner product $(\cdot,\cdot)$. One of the key reasons for using V_n is the following spectrally accurate Jackson-type approximation property: For any $F \in \mathscr{C}^{r,\alpha}(\mathbb{S}^2)$, the space of all r-times differentiable Hölder continuous functions with Hölder constant $0 < \alpha < 1$, there exists $\Psi_n \in V_n$ such that

$$\| F - \Psi_n\|_\infty \le Cn^{-(r+\alpha)}\| F\|_{r,\alpha}, \tag{7}$$

where, throughout the article, C is a generic constant independent of n and F.

Similar to the $\mathbb{S}^1$ case, we consider the corresponding semi-discrete and fully-discrete projection operators:

$$\mathscr{P}_n F = \sum_{l=0}^{n}\sum_{|j|\le l}(F, Y_{l,j})Y_{l,j}, \quad \mathscr{L}_n F = \sum_{l=0}^{n}\sum_{|j|\le l}(F, Y_{l,j})_n Y_{l,j}, \quad F \in \mathscr{C}(\mathbb{S}^2), \tag{8}$$

where the discrete approximation $(\cdot,\cdot)_n$ to the inner product, obtained using a cubature rule with positive weights, is required to satisfy the exactness property

$$(v, w) = (v, w)_n, \qquad \text{for all } v, w \in V_n, \tag{9}$$

so that $\mathscr{P}_n v = \mathscr{L}_n v$ for all $v \in V_n$. It has been known for more than a century that, in the $\mathbb{S}^2$ case, the Lebesgue constant of $\mathscr{P}_n$ is $\mathscr{O}(\sqrt{n})$ [26]. (Throughout this article, we use the standard definition that the Lebesgue constant of an operator is the norm of the operator from $\mathscr{C}(\mathbb{S}^2)$ to itself.) The slowest possible growth with respect to n of the Lebesgue constant for the operator $\mathscr{L}_n$ in (8) is that of $\mathscr{P}_n$. However, establishing this optimal order Lebesgue constant for $\mathscr{L}_n$ was not achieved until the turn of this century.

The quest to solve this challenging problem, which was open for several decades, began in 1995 with the seminal work of Sloan [47]. The key result [47, Theorem 1] is proving that the norm of $\mathscr{L}_n$ as an operator from $\mathscr{C}(\mathbb{S}^2)$ to $L^2(\mathbb{S}^2)$ is independent of n, generalizing a theorem of Erdős and Turán. Another important observation in [47, Theorem 2] is that, unlike in the $\mathbb{S}^1$ case, for $n \geq 3$ the fully-discrete operator $\mathscr{L}_n$ in (8) *cannot* be an interpolation operator, because the total number of points Q_n in any cubature on $\mathbb{S}^2$ with positive weights and exactness property (9) must satisfy $Q_N > (n+1)^2$, leading to the term "hyperinterpolation" in [47].

The two operators in (8) were first studied and analyzed for approximating a nonlinear elasticity model and associated SIE operators in the 1994 article by Ganesh et al. [20]. The sub-optimal Lebesgue constant proved in [20] was recalled in the 2000 article by Sloan and Womersley [48] and they proved (see [48, Theorem 5.5.4]) that the hyperinterpolation operator $\mathscr{L}_n$ achieves the optimal $\mathscr{O}(\sqrt{n})$ Lebesgue constant, under certain mild conditions on the cubature rule. This mild condition was subsequently removed by Reimer in the 2000 article [45].

In the two decades since hyperinterpolation was introduced by Sloan in 1995 [47], it has played a pivotal role in the development and analysis of a family of spectral algorithms for solving a class of partial differential equations (PDEs) posed exterior to a connected body $D \subseteq \mathbb{R}^3$ that can be reformulated as equivalent SIEs posed on the body's surface $\partial\Omega$ (with the restriction that $\partial\Omega$ is isomorphic to $\mathbb{S}^2$). In this article, we focus on PDEs associated with propagation of time harmonic acoustic, electromagnetic and elastic waves using SIE reformulations on $\partial\Omega$.

From a computational perspective, such SIE formulations have three key advantages compared with the original PDE formulations. Firstly, the surface $\partial\Omega$ is two-dimensional and hence the domain of the SIE is of lower dimension than the domain of the PDE. Secondly, the surface $\partial\Omega$ is bounded, whereas the domain $\mathbb{R}^3 \setminus \overline{\Omega}$ of the PDE is unbounded and so standard discretisation techniques such as finite differences and the finite element method are not applicable without modification. Thirdly, any radiation condition associated with the PDE is incorporated exactly into the SIE formulation.

However, a disadvantage of the SIE reformulation is the need to solve dense complex linear systems. This disadvantage can be avoided by developing algorithms that require substantially fewer degrees of freedom (DoF) compared to that required for local polynomial based approximations of the PDE and SIE models. The hyperinterpolation-based global polynomial approximations, which are spectrally accurate, facilitate such a reduction in the DoF. As we demonstrate in the last section of this article, to solve scalar acoustic, and vector electromagnetic and elastic scattering models exterior to a sphere of 8-wavelengths diameter, we can achieve high-order accuracy using only about one hundred DoFs.

A similar reduction of the number of DoFs, compared to local polynomial-based low-order finite- and boundary-element approximations, which are used in many industrial standard software packages, has been demonstrated [12–14] for a large class of smooth and non-smooth test obstacles used in the literature. The non-smooth obstacles include the benchmark radar targets in [52].

Of course the restriction imposed on $\partial\Omega$ in this article does not allow general Lipschitz geometries such as aircraft. For such geometries, and for heterogeneous models, it is important to include local polynomial based finite element approximations. However, heterogeneous wave propagation models that use only finite element approximations demand truncation of the domain and approximation of the radiation condition, and require a very large bounded domain to justify the truncation. However, this major restriction on local polynomial approximations can be avoided by using a hybrid of finite element and hyperinterpolation approximations [2, 19].

A practical high-order algorithm for a general Lipschitz domain was developed in [19] by closely circumscribing the domain with an artificial closed smooth interface (that is diffeomorphic to $\mathbb{S}^{d-1}$) and reformulating the problem as an interface model comprising interior and exterior problems. The algorithm in [19] uses high-order local polynomial finite element approximations for the PDE in the bounded domain with the Lipschitz and smooth boundary, and spectrally accurate approximations for the exterior part of the model. The method in [19] was demonstrated in two dimensions for a Lipschitz (acoustic-horn) geometry using approximation based on the hyperinterpolation operator $\mathscr{L}_N^1$ for the exterior part of the model. In a future work, we plan to develop a similar FEM-SIE software for a general class of 3D models using the hyperinterpolation approximations described in this article.

A fundamental result that is needed in the analysis of all such high-order SIE algorithms for unbounded, exterior regions and homogeneous media, is that the Lebesgue constant of the hyperinterpolation operator satisfies $\mathscr{O}(n^s)$ with $s < 1$. Below we will consider two vector-valued counterparts of the scalar hyperinterpolation operator $\mathscr{L}_n$ for electromagnetic scattering, namely $\underline{\mathscr{O}}_n$ and $\underline{\mathscr{L}}_n$. It is still an open problem to prove such a bound for the hyperinterpolation operator $\underline{\mathscr{O}}_n$, which is based on only tangential basis functions.

For scalar acoustic problems using $\mathscr{L}_n$, the apex practical algorithm is given by Ganesh and Graham [11]. This algorithm incorporates the excellent work of Sloan and collaborators [25, 47, 48] who contributed to a full convergence analysis for the fully discrete scheme. In their landmark paper [25], Graham and Sloan provided a full analysis for the fully discrete scheme and established superalgebraic convergence for the solution of the SIE. Ganesh and Graham [11] subsequently proved superalgebraic convergence of derived quantities such as the far field and exterior field, and demonstrated the algorithm with extensive numerical examples. The algorithm in [11] includes several important details required for efficient implementation, including high order evaluation of the inner integrals using a numerically stable rotation of the coordinate system and hyperinterpolation (see Sect. 4).

The vector hyperinterpolation operator $\underline{\mathscr{L}}_n$ based on componentwise application of $\mathscr{L}_n$ was introduced and analysed in [12]. A fully discrete high order method for electromagnetic scattering, based on the magnetic dipole equation, was then developed utilising this hyperinterpolation operator. The analysis in [12] establishes the optimal order $\mathscr{O}(\sqrt{n})$ Lebesgue constant for $\underline{\mathscr{L}}_n$, which facilitates a full convergence proof for the algorithm.

The method in [12] requires modification of the magnetic dipole operator to enforce the tangential property of the solution of the SIE. A different hyperinterpolation operator $\underline{\mathscr{O}}_n$ based on vector spherical harmonics was introduced in [13]. For spherical scatterers the operator $\underline{\mathscr{O}}_n$ preserves the tangential property of the solution of the SIE and so the number of unknowns is reduced by about one third. The restriction to spherical scatterers was subsequently removed in [14] by incorporating a rotation of the tangential vector spherical harmonics. The convergence of the methods in [13, 14] is based on the conjecture that $\underline{\mathscr{O}}_n$ has $\mathscr{O}(\sqrt{n})$ Lebesgue constant. Pieper [43] established an order $\mathscr{O}(n)$ estimate for the Lebesgue constant. Proving the conjecture in [13, 14] remains an open problem.

A key element of the analysis in [25] is proving an error bound for evaluating the weakly singular inner surface integral. A corresponding bound for the vector case was established for $\underline{\mathscr{L}}_n$ in [12]. Such an error bound is yet to be proven when $\underline{\mathscr{O}}_n$ is used for evaluating the inner integral (see [12, Remark 2]). The electromagnetic scattering algorithms [12–14] are restricted to SIEs with weakly-singular kernels. Le Louër recently developed similar spectral algorithms for SIEs with hypersingular operators arising in elastic and electromagnetic wave PDEs [40, 41] using the hyperinterpolation operators in [12–14]. In Le Louër's method, the rotation of the tangential vector spherical harmonics is effected using the Piola transformation. Integration by parts is then used to derive, for example, the relation [41, Section 4.2]

$$\underline{\mathscr{O}}_n \,\mathrm{curl}_{\mathbb{S}^2} = \mathrm{curl}_{\mathbb{S}^2} \,\mathscr{L}_n, \tag{10}$$

which reduces the degree of singularity of the kernel.

The rest of this article assumes (as in the literature cited above) that Ω is a closed and simply-connected domain in $\mathbb{R}^3$. However, in addition to our earlier discussion on a class of practical smooth and non-smooth geometries and general Lipschitz domains, we note that the restriction to simply-connected domains is not necessary for the application of the hyperinterpolation operators; techniques for domains Ω comprising up to several thousands of disjoint sub-domains are presented in [15–17], and in [18] for the case when there is uncertainty in the description of the domains.

In Sect. 2 we recall the three dimensional homogeneous PDEs associated with acoustic, electromagnetic and elastic waves and their corresponding SIE reformulations. In Sect. 3 we describe the spectral algorithm in a unified framework that includes all of the PDE models. In Sect. 4 we describe in detail how the weakly singular inner integrals are evaluated using hyperinterpolation. Finally, in Sect. 5 we present numerical results.

2 Acoustic, Electromagnetic, and Elastic Wave PDEs

In this section, we recall the exterior PDE models for propagation of the three important classes of waves and reformulation of the PDEs as SIEs.

2.1 Acoustic Wave Propagation

The time-harmonic acoustic wave scattered by Ω is described by the complex acoustic velocity potential $u(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}$, which satisfies the Helmholtz equation

$$\triangle u(\boldsymbol{x}) + k^2 u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}, \tag{11}$$

where $k = 2\pi/\lambda$ is the wavenumber and λ is the wavelength, and satisfies the Sommerfeld radiation condition

$$\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{x}| \left(\frac{\partial u}{\partial \boldsymbol{x}} - iku \right) = 0. \tag{12}$$

The scattered field is induced by a known incident field u^{inc} through a boundary condition applied on $\partial\Omega$. In the case that Ω represents a sound soft obstacle, the boundary condition is of the form

$$u(\boldsymbol{x}) = -u^{\mathrm{inc}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{13}$$

Analogous Neumann and Robin boundary conditions arise when the obstacle is sound hard or absorbing [11].

Following [4], we use the surface integral ansatz for the scattered field

$$u(\boldsymbol{x}) = \int_{\partial\Omega} \left(\frac{\partial \Phi(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{n}(\boldsymbol{y})} - i\gamma\Phi(\boldsymbol{x},\boldsymbol{y}) \right) v(\boldsymbol{y})\, ds(\boldsymbol{y}), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega},$$

where $\gamma \neq 0$ is a coupling parameter, and

$$\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{e^{ik|\boldsymbol{x}-\boldsymbol{y}|}}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}, \qquad \boldsymbol{x} \neq \boldsymbol{y},$$

is the fundamental solution for the Helmholtz equation and v is the unknown surface current. The surface current $v \in \mathscr{C}(\mathbb{S}^2)$ satisfies the second kind SIE

$$v(\boldsymbol{x}) + \mathfrak{K}v(\boldsymbol{x}) - i\gamma\mathfrak{S}v(\boldsymbol{x}) = -2u^{\mathrm{inc}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega,$$

where $\mathfrak{K}$ is the weakly singular double layer potential,

$$\mathfrak{K}v(\boldsymbol{x}) = \int_{\partial\Omega} \frac{\partial\Phi}{\partial \boldsymbol{n}_{\boldsymbol{y}}}(\boldsymbol{x},\boldsymbol{y})v(\boldsymbol{y})\, ds(\boldsymbol{y}),$$

with $\boldsymbol{n}_{\boldsymbol{y}}$ denoting the unit outward normal at $\boldsymbol{y} \in \partial\Omega$, and $\mathfrak{S}$ is the weakly singular single layer potential,

$$\mathfrak{S}v(\boldsymbol{x}) = \int_{\partial\Omega} \Phi(\boldsymbol{x},\boldsymbol{y})v(\boldsymbol{y})\, ds(\boldsymbol{y}).$$

We refer to [4, 11] for full details, for other surface integral ansatzes and related SIEs.

2.2 Electromagnetic Wave Propagation

The vector time-harmonic electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$ scattered by Ω satisfy the reduced Maxwell equations

$$\operatorname{curl} \boldsymbol{E}(\boldsymbol{x}) - ik\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{0}, \qquad \operatorname{curl} \boldsymbol{H}(\boldsymbol{x}) + ik\boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}, \tag{14}$$

and the Silver-Müller radiation condition

$$\lim_{|\boldsymbol{x}|\to\infty} [\boldsymbol{H}(\boldsymbol{x}) \times \boldsymbol{x} - |\boldsymbol{x}|\boldsymbol{E}(\boldsymbol{x})] = \boldsymbol{0}.$$

The scattered field is induced by a known incident field $\boldsymbol{E}^{\mathrm{inc}}, \boldsymbol{H}^{\mathrm{inc}}$ through a boundary condition applied on $\partial\Omega$. In the case that Ω represents a perfectly conducting obstacle, the boundary condition is

$$\boldsymbol{n}(\boldsymbol{x}) \times \boldsymbol{E}(\boldsymbol{x}) = -\boldsymbol{n}(\boldsymbol{x}) \times \boldsymbol{E}^{\mathrm{inc}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{15}$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the unit outward normal at $\boldsymbol{x} \in \partial\Omega$.

Following [4, 12], we use the surface integral ansatz for the electric field

$$\boldsymbol{E}(\boldsymbol{x}) = \operatorname{curl} \int_{\partial\Omega} \Phi(\boldsymbol{x},\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y})\, ds(\boldsymbol{y}), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}.$$

The tangential surface field $\boldsymbol{w} \in \underline{\mathscr{C}}(\partial\Omega)$ satisfies the magnetic dipole equation

$$\boldsymbol{w}(\boldsymbol{x}) + \mathfrak{M}\boldsymbol{w}(\boldsymbol{x}) = -2\, \boldsymbol{n}(\boldsymbol{x}) \times \boldsymbol{E}^{\mathrm{inc}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega},$$

where $\mathfrak{M}$ is the magnetic dipole operator

$$\mathfrak{M}\boldsymbol{w}(\boldsymbol{x}) = 2 \int_{\partial\Omega} \operatorname{curl}_{\boldsymbol{x}} \{\Phi(\boldsymbol{x},\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y})\}\, ds(\boldsymbol{y}), \qquad \boldsymbol{x} \in \partial\Omega.$$

The magnetic dipole operator is weakly singular for operands that are tangential to $\partial\Omega$.

In the case that Ω is a dielectric obstacle, the PDEs above are augmented with similar equations posed in Ω (with the wavenumber replaced by the corresponding *interior* parameters) and the perfect conductor boundary condition is replaced by a transmission boundary condition.

Unlike the acoustic case, a major issue for the electromagnetic SIE reformulation is the breakdown at low-frequencies. This breakdown is described in detail in [9], which surveys over a century of research and provides a large number of references. In particular, the SIE reformulation for the Maxwell system dates back to the work of Lorentz (1890), Mie (1907), and Debye (1908) for scattering by spheres and leads to the 1949 work of Maue, who proposed the electric field SIE (EFIE) and the magnetic field SIE (MFIE).

The recent works [9, 21] were motivated by the lack of an SIE that is stable at all-frequencies (with robust mathematical analysis and properties). Such an SIE is highly desired and the lack thereof is both odd and unsatisfactory, given more than a century of work in this area. The most desirable class of SIE is the weakly-singular Fredholm SIE of the second-kind, governed by the identity plus a weakly-singular operator of negative-order that does not suffer from low-frequency breakdown. Recently a new class of second kind weakly singular SIE reformulations for the dielectric case, with rigorous mathematical analysis, was developed [21]. A computational implementation of the all-frequency model based on hyperinterpolation approximations will be a future work. Preliminary results are presented in Sect. 5.

2.3 Elastic Wave Propagation

The time-harmonic elastic wave scattered by Ω is described by the complex vector field $\boldsymbol{u}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}$, which satisfies the Navier equation

$$\mu \triangle u(\boldsymbol{x}) + (\lambda + \mu) \operatorname{grad} \operatorname{div} \boldsymbol{u}(\boldsymbol{x}) + \rho\omega^2 \boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}, \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}, \tag{16}$$

where μ and λ are the Lamé parameters, ρ is the density, and ω is the frequency, as well as the Kupradze radiation conditions

$$\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{x}| \left(\frac{\partial \boldsymbol{u}_p}{\partial \boldsymbol{x}} - ik\boldsymbol{u}_p \right), \quad \lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{x}| \left(\frac{\partial \boldsymbol{u}_s}{\partial \boldsymbol{x}} - ik\boldsymbol{u}_s \right). \tag{17}$$

Here $\boldsymbol{u}_p$ and $\boldsymbol{u}_s$ are the longitudinal and transverse components of $\boldsymbol{u}$ and k_p and k_s are the corresponding wavenumbers, with

$$\boldsymbol{u}_p = -k_p^2 \operatorname{grad} \operatorname{div} \boldsymbol{u}, \qquad \boldsymbol{u}_s = \boldsymbol{u} - \boldsymbol{u}_p,$$

and

$$k_p = \omega\sqrt{\frac{\rho}{\lambda + 2\mu}}, \qquad k_s = \omega\sqrt{\frac{\rho}{\mu}}.$$

The scattered field is induced by a known incident field $\boldsymbol{u}^{\mathrm{inc}}$ through a boundary condition applied on $\partial\Omega$. In the case that Ω represents a rigid body, the boundary condition is

$$\boldsymbol{u}(\boldsymbol{x}) = -\boldsymbol{u}^{\mathrm{inc}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{18}$$

An analogous Neumann boundary condition arises when Ω corresponds to a cavity [33, 40].

Following [33, 40], we use the surface integral ansatz for the elastic field

$$\boldsymbol{u}(\boldsymbol{x}) = -\int_{\partial\Omega} G(\boldsymbol{x},\boldsymbol{y})\boldsymbol{P}\boldsymbol{v}(\boldsymbol{y})\, ds(\boldsymbol{y}), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega},$$

where $G(\boldsymbol{x},\boldsymbol{y})$ is the fundamental solution of the Navier equation and the traction derivative $\boldsymbol{P}$ is given by

$$\boldsymbol{P}\boldsymbol{u} = 2\mu\frac{\partial\boldsymbol{u}}{\partial\boldsymbol{n}} + \lambda\boldsymbol{n}\operatorname{div}\boldsymbol{u} + \mu\boldsymbol{n}\times\operatorname{curl}\boldsymbol{u}.$$

The surface field $\boldsymbol{v} \in \mathscr{C}(\partial\Omega)$ satisfies the combined field integral equation

$$\boldsymbol{v}(\boldsymbol{x}) + \mathfrak{D}'\boldsymbol{v}(\boldsymbol{x}) + i\eta\mathfrak{S}\boldsymbol{v}(\boldsymbol{x}) = 2\boldsymbol{P}\boldsymbol{u}^{\mathrm{inc}}(\boldsymbol{x}) + 2i\eta\boldsymbol{u}^{\mathrm{inc}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega,$$

where $\mathfrak{S}$ is the single layer potential for the Navier equation,

$$\mathfrak{S}\boldsymbol{v}(\boldsymbol{x}) = 2\int_{\partial\Omega} \Phi(\boldsymbol{x},\boldsymbol{y})\boldsymbol{v}(\boldsymbol{y})\, ds(\boldsymbol{y}),$$

and $\mathfrak{D}'$ its traction derivative,

$$\mathfrak{D}'\boldsymbol{v}(\boldsymbol{x}) = 2\int_{\partial\Omega} [\boldsymbol{P}\Phi(\boldsymbol{x},\boldsymbol{y})\boldsymbol{v}(\boldsymbol{y})]\, ds(\boldsymbol{y}).$$

Here $\boldsymbol{P}$ is applied to each column of $\Phi(\boldsymbol{x},\boldsymbol{y})$ and the differentiation in $\boldsymbol{P}$ is with respect to the variable $\boldsymbol{x}$. We refer to [33, 40] for full details.

3 Unified Vector and Scalar SIE Spectral Algorithms

The second kind vector SIEs in the previous section can be written in the unified form

$$\boldsymbol{w}(\boldsymbol{x}) + (\mathfrak{K}\boldsymbol{w})(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{19}$$

where $\mathfrak{K}$ is an appropriate surface integral operator. The right hand side $\boldsymbol{f}$ is known and the equation is to be solved for the unknown vector valued surface potential $\boldsymbol{w}$. The SIE corresponding to the Helmholtz equation is identical, with suitable modification so that all functions are scalar valued.

The key to utilizing hyperinterpolation on the sphere for high order approximation of (19) is to derive an equivalent integral equation

$$\boldsymbol{W}(\widehat{\boldsymbol{x}}) + (\mathscr{K}\boldsymbol{W})(\widehat{\boldsymbol{x}}) = \boldsymbol{F}(\widehat{\boldsymbol{x}}), \qquad \widehat{\boldsymbol{x}} \in \mathbb{S}^2, \tag{20}$$

posed on the unit sphere $\mathbb{S}^2$. In [14] this derivation incorporates an orthogonal transformation that maps tangential fields on $\partial\Omega$ to tangential fields on $\mathbb{S}^2$. In [40, 41] a similar transformation is achieved using the Piola transformation. An appropriate vector variant of the space V_n in (4) is required to project the unknown $\boldsymbol{W}$ and the SIE (20) onto a finite dimensional space.

We use the notation $\underline{\mathbb{Q}}_n \subset \underline{\mathscr{C}}(\mathbb{S}^2)$ to denote such a finite dimensional space for the vector valued electromagnetic and elastic models. Similar to V_n with the property (7), the space $\underline{\mathbb{Q}}_n$ is chosen to satisfy a Jackson-type property: For any $\mathbf{F} \in \underline{\mathscr{C}}^{r,\alpha}(\mathbb{S}^2)$, the space of all r-times differentiable Hölder vector-valued continuous functions with Hölder constant $0 < \alpha < 1$, there exists $\boldsymbol{\Psi}_n \in \underline{\mathbb{Q}}_n$ such that

$$\| F - \boldsymbol{\Psi}_n\|_\infty \leq Cn^{-(r+\alpha)}\|\mathbf{F}\|_{r,\alpha}. \tag{21}$$

The semi-discrete Galerkin scheme for the SIE (20) is to solve

$$(\boldsymbol{W}_n + \mathscr{K}\boldsymbol{W}_n, \boldsymbol{\Psi}_n)_n = (\boldsymbol{F}, \boldsymbol{\Psi}_n)_n, \qquad \text{for all } \boldsymbol{\Psi}_n \in \underline{\mathbb{Q}}_n, \tag{22}$$

where $\boldsymbol{W}_n \in \underline{\mathbb{Q}}_n$. Despite using the fully-discrete quadrature approximation (22) for the Galerkin integrals, we refer to the scheme as semi-discrete. This is because, in general, it is not possible to analytically evaluate the surface integrals $\mathscr{K}\boldsymbol{W}_n$ and we require an additional, non-trivial, spectrally accurate approximation of such integrals. We discuss such details in the next section and our focus is on fully-discrete computer implementable wave propagation models, with mathematical analysis to quantify the error in the models, and efficient computer implementation to simulate scattering from non-trivial curved surfaces.

It is convenient to choose an ansatz space $\underline{\mathbb{Q}}_n$ spanned by the tangential vector spherical harmonics (associated with $\underline{\mathscr{O}}_n$) or by the componentwise vector spherical harmonics (associated with $\underline{\mathscr{L}}_n$), depending on whether the solution $\boldsymbol{W}$ of (20) is tangential on $\mathbb{S}^2$ or not. In the case of componentwise vector spherical harmonics, the semi-discrete equation corresponding to (20) is

$$\boldsymbol{W}_n + \underline{\mathscr{L}}_n \mathscr{K} \boldsymbol{W}_n = \underline{\mathscr{L}}_n \boldsymbol{F} \tag{23}$$

where $\boldsymbol{W}_n \in \underline{\mathbb{Q}}_n$.

To elucidate the importance of the hyperinterpolation theory contributions made by Sloan for wave propagation applications, it is convenient to assume that $\mathscr{K}$ is weakly singular and bounded from $\underline{\mathscr{C}}(\mathbb{S}^2)$ to $\underline{\mathscr{C}}^{0,\alpha}(\mathbb{S}^2)$ for any $0 < \alpha < 1$. It is helpful to recap the key elements of the existence and uniqueness proof for the semi-discrete equation (23). The full proof for the scalar case is given by Sloan et al. in [25]. (A different proof of existence and uniqueness for the vector case is given in [12].)

As in [25, Theorem 3.1], the proof requires the assumption that $\underline{\mathscr{L}}_n$ is bounded on $\underline{\mathscr{C}}(\mathbb{S}^2)$ and satisfies

$$\|\underline{\mathscr{L}}_n\|_{\underline{\mathscr{C}}(\mathbb{S}^2)\to\underline{\mathscr{C}}(\mathbb{S}^2)} \leq C n^s, \tag{24}$$

for some $0 < s < 1$. Then for $\boldsymbol{F} \in \underline{\mathscr{C}}^{0,\alpha}(\mathbb{S}^2)$ and $\boldsymbol{\Psi}_n$ as in (21), using the exactness property $\underline{\mathscr{L}}_n \boldsymbol{\Psi}_n = \boldsymbol{\Psi}_n$ we obtain

$$\begin{aligned}
\|\boldsymbol{F} - \underline{\mathscr{L}}_n \boldsymbol{F}\|_\infty &= \|(I - \underline{\mathscr{L}}_n)(\boldsymbol{F} - \boldsymbol{\Psi}_n)\|_\infty \\
&\leq \left[1 + \|\underline{\mathscr{L}}_n\|_{\underline{\mathscr{C}}(\mathbb{S}^2)\to\underline{\mathscr{C}}(\mathbb{S}^2)}\right] \|\boldsymbol{F} - \boldsymbol{\Psi}_n\|_\infty \\
&\leq (1 + C n^s)\|\boldsymbol{F} - \boldsymbol{\Psi}_n\|_\infty \\
&\leq \frac{C}{n^{\alpha-s}} \|\boldsymbol{F}\|_{0,\alpha}.
\end{aligned} \tag{25}$$

For $\boldsymbol{A} \in \underline{\mathscr{C}}(\mathbb{S}^2)$ we have $\boldsymbol{F} = \mathscr{K}\boldsymbol{A} \in \underline{\mathscr{C}}^{0,\alpha}(\mathbb{S}^2)$ and from (25),

$$\|\mathscr{K}\boldsymbol{A} - \underline{\mathscr{L}}_n \mathscr{K}\boldsymbol{A}\|_\infty \leq \frac{C}{n^{\alpha-s}} \|\mathscr{K}\boldsymbol{A}\|_{0,\alpha} \leq \frac{C}{n^{\alpha-s}} \|\boldsymbol{A}\|_\infty. \tag{26}$$

Thus

$$\|(I - \underline{\mathscr{L}}_n)\mathscr{K}\|_{\underline{\mathscr{C}}(\mathbb{S}^2)\to\underline{\mathscr{C}}(\mathbb{S}^2)} \leq \frac{C}{n^{\alpha-s}}.$$

Existence and uniqueness of the solution $\boldsymbol{W}_n$ of (23) then follows from the Banach Lemma *provided* $s < \alpha \in (0, 1)$.

Thus establishing a bound of the form (24) with $s < 1$ is crucial for the analysis of the spectral method. As mentioned earlier, this crucial result was established in the scalar case by Sloan and Womersley [48, Theorem 5.5.4]. The corresponding result was established for $\underline{\mathscr{L}}_n$ in [12, Equation A.10]. Establishing a similar result for $\underline{\mathscr{O}}_n$ remains an open problem.

In the discussion above we have demonstrated the importance of hyperinterpolation approximations for projecting the unknown surface current and the SIE. The hyperinterpolation operator also plays a crucial role in developing spectrally accurate approximations of the various surface integral operators in the wave propagation model. However, the weakly singular surface integral operators have discontinuous kernels, and so the hyperinterpolation operator and best-approximation results (7) and (21) are not appropriate for direct application to obtain fully discrete approximations to the integrands. Indeed, a very careful hybrid analytic-numeric approach is required to evaluate such integrals to very high accuracy and establish associated error bounds. We describe such details in the next section for the general weakly singular operator $\mathscr{K}$.

4 Hyperinterpolation for Weakly-Singular Integrals

In this section we review the role of hyperinterpolation approximations for the description, analysis, and implementation of a fully-discrete scheme for (20) in the case that the operator $\mathscr{K}$ is weakly singular. In order to obtain such a fully-discrete scheme, we require a spectrally accurate approximation of the surface integral $\mathscr{K}$ by a summation operator $\mathscr{K}_{n'}$ that can be evaluated efficiently.

For hypersingular operators that arise in elastic wave propagation, relations such as (10) transform hypersingular kernels into weakly singular kernels of the kind considered in this section. The key task considered in this section is to evaluate the spectrally accurate surface integral approximation $(\mathscr{K}_{n'}\boldsymbol{A})(\widehat{\boldsymbol{x}})$ for any given vector potential $\boldsymbol{A} \in \underline{\mathscr{C}}(\partial\Omega)$ and observation point $\widehat{\boldsymbol{x}} \in \mathbb{S}^2$.

We proceed by splitting $\mathscr{K}$ into singular and regular components

$$\mathscr{K}\boldsymbol{A} = \mathscr{K}_1\boldsymbol{A} + \mathscr{K}_2\boldsymbol{A} \tag{27}$$

where

$$(\mathscr{K}_1\boldsymbol{A})(\widehat{\boldsymbol{x}}) = \int_{\mathbb{S}^2} \frac{1}{|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}|} K_1(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \boldsymbol{A}(\widehat{\boldsymbol{y}})\, ds(\widehat{\boldsymbol{y}}), \tag{28}$$

$$(\mathscr{K}_2\boldsymbol{A})(\widehat{\boldsymbol{x}}) = \int_{\mathbb{S}^2} K_2(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \boldsymbol{A}(\widehat{\boldsymbol{y}})\, ds(\widehat{\boldsymbol{y}}), \tag{29}$$

and $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ are 3×3 matrix valued functions that are infinitely continuously differentiable on $\mathbb{R}^3 \times \mathbb{R}^3$. The process for scalar valued weakly singular operators is similar, with suitable modification so that all functions are scalar valued.

First we consider evaluation of the weakly singular integral (28) for fixed $\widehat{\boldsymbol{x}} \in \mathbb{S}^2$. The first step is to introduce a new coordinate system in which the weak singularity at $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{x}}$ is transformed to the north pole $\widehat{\boldsymbol{n}} = (0, 0, 1)^T$. We achieve this by rotating the coordinate system using the orthogonal matrix

$$T_{\hat{x}} := \mathscr{R}_z(\phi)\mathscr{R}_y(-\theta)\mathscr{R}_z(-\phi) \tag{30}$$

where $\widehat{\boldsymbol{x}} = \boldsymbol{p}(\theta, \phi)$ and

$$\mathscr{R}_z(\psi) := \begin{pmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathscr{R}_y(\psi) := \begin{pmatrix} \cos\psi & 0 & \sin\psi \\ 0 & 1 & 0 \\ -\sin\psi & 0 & \cos\psi \end{pmatrix}. \tag{31}$$

To describe the approximation procedure it is helpful to introduce a linear transformation

$$\mathscr{T}_{\hat{x}}\boldsymbol{A}(\widehat{z}) := \boldsymbol{A}(T_{\hat{x}}^{-1}\widehat{z}), \qquad \widehat{z} \in \mathbb{S}^2, \qquad \boldsymbol{A} \in \underline{\mathscr{C}}(\mathbb{S}^2), \tag{32}$$

and its bivariate analogue

$$\mathscr{T}_{\hat{x}}\boldsymbol{A}(\widehat{z}_1, \widehat{z}_2) := \boldsymbol{A}(T_{\hat{x}}^{-1}\widehat{z}_1, T_{\hat{x}}^{-1}\widehat{z}_2), \qquad \widehat{z}_1, \widehat{z}_2 \in \mathbb{S}^2, \qquad \boldsymbol{A} \in \underline{\mathscr{C}}(\mathbb{S}^2 \times \mathbb{S}^2). \tag{33}$$

For $\widehat{z} = T_{\hat{x}}\widehat{\boldsymbol{y}} \in \mathbb{S}^2$ we have, using the orthogonality of $T_{\hat{x}}$,

$$|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}| = |T_{\hat{x}}^{-1}(\widehat{\boldsymbol{n}} - \widehat{z})| = |\widehat{\boldsymbol{n}} - \widehat{z}|. \tag{34}$$

Using the fact that the surface measure on $\mathbb{S}^2$ is invariant, we get

$$\mathscr{K}_1\boldsymbol{A}(\widehat{\boldsymbol{x}}) = \int_{\mathbb{S}^2} \frac{1}{|\widehat{\boldsymbol{n}} - \widehat{z}|} \mathscr{T}_{\hat{x}}K_1(\widehat{\boldsymbol{n}}, \widehat{z})\ \mathscr{T}_{\hat{x}}\boldsymbol{A}(\widehat{z})\ ds(\widehat{z}). \tag{35}$$

In the transformed coordinate system, the function $(\theta', \phi') \mapsto \mathscr{T}_{\hat{x}}K_1(\widehat{\boldsymbol{n}}, \boldsymbol{p}(\theta', \phi'))$ is infinitely continuously differentiable, with all derivatives 2π-periodic in each variable, and each partial derivative is uniformly bounded with respect to $\widehat{\boldsymbol{x}} \in \mathbb{S}^2$ (see [12, Theorem 1], and references therein for the corresponding scalar case). Furthermore, the denominator $|\widehat{\boldsymbol{n}} - \widehat{z}| = 2\sin\theta'/2$ for $\widehat{z} = \boldsymbol{p}(\theta', \phi')$ and so when the surface integral in (28) is expressed in spherical polar coordinates, the singularity is cancelled out by the surface element $\sin\theta'\ d\theta'\ d\phi'$.

Next we choose $n' > n$ and approximate $\mathscr{K}_1\boldsymbol{A}(\widehat{\boldsymbol{x}})$ by

$$\begin{aligned}
\mathscr{K}_{1,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) &= \int_{\mathbb{S}^2} \frac{1}{|\widehat{\boldsymbol{n}} - \widehat{z}|} \underline{\mathscr{L}}_{n'} \left\{ \mathscr{T}_{\hat{x}}K_1(\widehat{\boldsymbol{n}}, \cdot)\ \mathscr{T}_{\hat{x}}\boldsymbol{A}(\cdot) \right\} (\widehat{z})\ ds(\widehat{z}) \\
&= \int_{\mathbb{S}^2} \frac{1}{|\widehat{\boldsymbol{n}} - \widehat{z}|} \sum_{l=0}^{n'} \sum_{|j| \le l} \sum_{k=1}^{3} \left(\mathscr{T}_{\hat{x}}K_1(\widehat{\boldsymbol{n}}, \cdot)\ \mathscr{T}_{\hat{x}}\boldsymbol{A}(\cdot), \boldsymbol{Y}_{l,j,k}(\cdot) \right)_{n'} \boldsymbol{Y}_{l,j,k}(\widehat{z})\ ds(\widehat{z}) \\
&= \sum_{l=0}^{n'} \sum_{|j| \le l} \sum_{k=1}^{3} \left(\mathscr{T}_{\hat{x}}K_1(\widehat{\boldsymbol{n}}, \cdot)\ \mathscr{T}_{\hat{x}}\boldsymbol{A}(\cdot), \boldsymbol{Y}_{l,j,k}(\cdot) \right)_{n'} \int_{\partial B} \frac{1}{|\widehat{\boldsymbol{n}} - \widehat{z}|} \boldsymbol{Y}_{l,j,k}(\widehat{z})\ ds(\widehat{z}).
\end{aligned}$$

Here the componentwise vector spherical harmonics are

$$\boldsymbol{Y}_{l,j,k} = Y_{l,j}\boldsymbol{e}_k, \qquad 0 \le l \le n,\ |j| \le l,\ 1 \le k \le 3, \tag{36}$$

where $\boldsymbol{e}_k$ denotes the kth Euclidean vector.

We evaluate the integral using the property of the spherical harmonics that they are eigenfunctions of the single layer potential [10], that is

$$\int_{\mathbb{S}^2} \frac{1}{|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}|} Y_{l,j}(\widehat{\boldsymbol{y}})\, ds(\widehat{\boldsymbol{y}}) = \frac{4\pi}{2l+1} Y_{l,j}(\widehat{\boldsymbol{x}}). \tag{37}$$

Thus

$$\mathscr{K}_{1,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) = \sum_{l=0}^{n'} \sum_{|j|\le l} \sum_{k=1}^{3} \left(\mathscr{T}_{\widehat{\boldsymbol{x}}} K_1(\widehat{\boldsymbol{n}}, \cdot)\ \mathscr{T}_{\widehat{\boldsymbol{x}}} \boldsymbol{A}(\cdot), \boldsymbol{Y}_{l,j,k}(\cdot)\right)_{n'} \frac{4\pi}{2l+1} \boldsymbol{Y}_{l,j,k}(\widehat{\boldsymbol{n}}).$$

To expand the discrete inner product we use a $(2n'+2) \times (n'+1)$ point Gauss-rectangle quadrature rule

$$\int_{\mathbb{S}^2} G(\widehat{\boldsymbol{x}})\, ds(\widehat{\boldsymbol{x}}) \approx Q_{n'}(G) = \sum_{r=0}^{2n'+1} \sum_{s=1}^{n'+1} \mu_r \nu_s G(\widehat{\boldsymbol{z}}_{rs}), \quad G \in \mathscr{C}(\mathbb{S}^2), \tag{38}$$

where $\widehat{\boldsymbol{z}}_{rs} = \boldsymbol{p}(\theta_s, \phi_r)$. Here $\theta_s = \cos^{-1} z_s$, where z_s are the zeros of the Legendre polynomial of degree $n'+1$, ν_s are the corresponding Gauss-Legendre weights and

$$\mu_r = \frac{\pi}{n'+1}, \qquad \phi_r = \frac{r\pi}{n'+1}, \qquad r = 0, \ldots, 2n'+1. \tag{39}$$

Expanding the discrete inner product as sums over s' and r', we obtain

$$\begin{aligned}
&\mathscr{K}_{1,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) \\
&= \sum_{l=0}^{n'} \sum_{|j|\le l} \sum_{k=1}^{3} \sum_{s'=1}^{n'+1} \sum_{r'=0}^{2n'+1} \mu_{r'} \nu_{s'} \frac{4\pi}{2l+1} \overline{\boldsymbol{Y}_{l,j,k}(\widehat{\boldsymbol{z}}_{r's'})}^T \mathscr{T}_{\widehat{\boldsymbol{x}}} K_1(\widehat{\boldsymbol{n}}, \widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\widehat{\boldsymbol{x}}} \boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'}) \boldsymbol{Y}_{l,j,k}(\widehat{\boldsymbol{n}}) \\
&= \sum_{l=0}^{n'} \sum_{k=1}^{3} \sum_{s'=1}^{n'+1} \sum_{r'=0}^{2n'+1} \mu_{r'} \nu_{s'} \frac{4\pi}{2l+1} \boldsymbol{e}_k \boldsymbol{e}_k^T \mathscr{T}_{\widehat{\boldsymbol{x}}} K_1(\widehat{\boldsymbol{n}}, \widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\widehat{\boldsymbol{x}}} \boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'}) \sum_{|j|\le l} \overline{Y_{l,j}(\widehat{\boldsymbol{z}}_{r's'})} Y_{l,j}(\widehat{\boldsymbol{n}}).
\end{aligned}$$

From the addition theorem for the spherical harmonics [4, Theorem 2.8],

$$\frac{4\pi}{2l+1} \sum_{|j|\le l} \overline{Y_{l,j}(\widehat{\boldsymbol{a}})} Y_{l,j}(\widehat{\boldsymbol{b}}) = P_l(\widehat{\boldsymbol{a}} \cdot \widehat{\boldsymbol{b}}),$$

the last sum simplifies and we obtain

$$\begin{aligned}\mathscr{K}_{1,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) &= \sum_{l=0}^{n'}\sum_{k=1}^{3}\sum_{s'=1}^{n'+1}\sum_{r'=0}^{2n'+1} \mu_{r'}\nu_{s'}\boldsymbol{e}_k\boldsymbol{e}_k^T\,\mathscr{T}_{\hat{\boldsymbol{x}}}K_1(\widehat{\boldsymbol{n}},\widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'})P_l(\widehat{\boldsymbol{n}}\cdot\widehat{\boldsymbol{z}}_{r's'})\\ &= \sum_{s'=1}^{n'+1}\sum_{r'=0}^{2n'+1} \mu_{r'}\nu_{s'}\alpha_{s'}\,\mathscr{T}_{\hat{\boldsymbol{x}}}K_1(\widehat{\boldsymbol{n}},\widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'}),\end{aligned} \tag{40}$$

where $\alpha_{s'} = \sum_{l=0}^{n'} P_l(\cos\theta_{s'})$, and we have used $\widehat{\boldsymbol{n}}\cdot\widehat{\boldsymbol{z}}_{r's'} = \cos\theta_{s'}$.

Next we consider evaluation of the weakly singular integral (29). Using again the transformed coordinate system, we have

$$\mathscr{K}_2\boldsymbol{A}(\widehat{\boldsymbol{x}}) = \int_{\mathbb{S}^2} \mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\widehat{\boldsymbol{z}})\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\widehat{\boldsymbol{z}})\ ds(\widehat{\boldsymbol{z}}). \tag{41}$$

Similar to above, we approximate $\mathscr{K}_2\boldsymbol{A}$ by

$$\begin{aligned}\mathscr{K}_{2,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) &= \int_{\mathbb{S}^2} \underline{\mathscr{L}}_{n'}\left\{\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\cdot)\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\cdot)\right\}(\widehat{\boldsymbol{z}})\ ds(\widehat{\boldsymbol{z}}) \qquad (42)\\ &= \int_{\mathbb{S}^2} \sum_{l=0}^{n'}\sum_{|j|\le l}\sum_{k=1}^{3} \left(\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\cdot)\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\cdot),\boldsymbol{Y}_{l,j,k}(\cdot)\right)_{n'}\boldsymbol{Y}_{l,j,k}(\widehat{\boldsymbol{z}})\ ds(\widehat{\boldsymbol{z}})\\ &= \sum_{l=0}^{n'}\sum_{|j|\le l}\sum_{k=1}^{3} \left(\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\cdot)\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\cdot),\boldsymbol{Y}_{l,j,k}(\cdot)\right)_{n'}\int_{\mathbb{S}^2}\boldsymbol{Y}_{l,j,k}(\widehat{\boldsymbol{z}})\ ds(\widehat{\boldsymbol{z}}).\end{aligned}$$

Now,

$$\int_{\mathbb{S}^2} Y_{l,j,k}(\widehat{\boldsymbol{z}})\ ds(\widehat{\boldsymbol{z}}) = \begin{cases}\boldsymbol{e}_k/\sqrt{4\pi}, & \text{for } l=0, j=0,\\ 0, & \text{otherwise.}\end{cases}$$

It follows that

$$\mathscr{K}_{2,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) = \sum_{k=1}^{3}\left(\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\cdot)\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\cdot),\boldsymbol{e}_k\right)_{n'}\boldsymbol{e}_k.$$

Expanding the discrete inner product using (38),

$$\begin{aligned}\mathscr{K}_{2,n'}\boldsymbol{A}(\widehat{\boldsymbol{x}}) &= \sum_{k=1}^{3}\sum_{s'=1}^{n'+1}\sum_{r'=0}^{2n'+1} \mu_{r'}\nu_{s'}\boldsymbol{e}_k\boldsymbol{e}_k^T\,\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'})\\ &= \sum_{s'=1}^{n'+1}\sum_{r'=0}^{2n'+1} \mu_{r'}\nu_{s'}\,\mathscr{T}_{\hat{\boldsymbol{x}}}K_2(\widehat{\boldsymbol{n}},\widehat{\boldsymbol{z}}_{r's'})\ \mathscr{T}_{\hat{\boldsymbol{x}}}\boldsymbol{A}(\widehat{\boldsymbol{z}}_{r's'}).\end{aligned} \tag{43}$$

The fully discrete Galerkin scheme is obtained by replacing the operators $\mathscr{K}_1$ and $\mathscr{K}_2$ in (27) with the discrete approximations above. Following the corresponding scalar case analysis by Graham and Sloan [25, Theorem 4.2], it was proved in [12, Theorem 1] that $\mathscr{K}_{n'}\boldsymbol{A} = \mathscr{K}_{1,n'}\boldsymbol{A} + \mathscr{K}_{2,n'}\boldsymbol{A}$ converges to $\mathscr{K}\boldsymbol{A}$ with spectral accuracy for $\boldsymbol{A} \in \underline{\mathbb{P}}_n$ when $n' = an$ with $a > 1$ and $n' - n > 2$. In particular, for any $r \in \mathbb{N}$, there exists $C_r > 0$ independent of n and n' such that

$$\|(\mathscr{K} - \mathscr{K}_{n'})\boldsymbol{A}\|_\infty \leq C_r \frac{1}{n^r}\|\boldsymbol{A}\|_\infty, \qquad \text{for all } \boldsymbol{A} \in \mathbb{P}_n. \tag{44}$$

The trial functions $\boldsymbol{A}$ depend on the projection used in the semi-discrete equation (23). In particular, as discussed in Sect. 3, if the projection $\underline{\mathscr{L}}_n$ is used then the associated trial functions are the componentwise vector spherical harmonics. If the projection $\underline{\mathscr{O}}_n$ is used then the associated trial functions are the tangential vector spherical harmonics

$$\boldsymbol{Y}^{(1)}_{l,j}(\widehat{\boldsymbol{x}}) = \frac{1}{\sqrt{l(l+1)}}\text{Grad } Y_{l,j}(\widehat{\boldsymbol{x}}), \qquad \boldsymbol{Y}^{(2)}_{l,j}(\widehat{\boldsymbol{x}}) = \widehat{\boldsymbol{x}} \times \boldsymbol{Y}^{(1)}_{l,j}(\widehat{\boldsymbol{x}}), \qquad \widehat{\boldsymbol{x}} \in \mathbb{S}^2, \tag{45}$$

and the normal vector spherical harmonics

$$\boldsymbol{Y}^{(3)}_{l,j}(\widehat{\boldsymbol{x}}) = \widehat{\boldsymbol{x}}Y_{l,j}(\widehat{\boldsymbol{x}}), \quad \widehat{\boldsymbol{x}} \in \mathbb{S}^2. \tag{46}$$

It is well known, for over a century [3], that the space $\mathbb{P}_n$ of spherical polynomials is invariant under rotations, and so for fixed $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$,

$$\mathscr{T}_{\boldsymbol{p}(\theta,\phi)}Y_{l,j}(\widehat{\boldsymbol{z}}) = \sum_{|\tilde{j}|\leq l} R_{l,j,\tilde{j}}(\theta,\phi)Y_{l,\tilde{j}}(\widehat{\boldsymbol{z}}), \qquad \widehat{\boldsymbol{z}} \in \mathbb{S}^2, \tag{47}$$

for coefficients $R_{l,j,\tilde{j}}(\theta,\phi)$ that are independent of $\widehat{\boldsymbol{z}}$. The vector spherical harmonics defined in (45)–(46) and (36) are obtained from the spherical harmonics by application of particular linear operators, and hence the rotated vector spherical harmonics enjoy an analogous relation to (47), with the *same* coefficients.

The key consequence of (47) is that the vector spherical harmonics are evaluated only at the tensor product quadrature points given by (38). From (5) we see that on a tensor product grid the θ and ϕ parts of (5) decouple, facilitating a reduction in the complexity of the high order scheme by a factor n. Using a similar decoupling for the test function leads to a fast order n^5 assembly scheme for the Galerkin matrix corresponding to $\mathscr{K}_{n'}$ (see for example [12, Section 3] for full details). For comparison, naive implementation would have complexity order n^8. In Sect. 5 we present numerical results for n between 25 and 40.

Expressions for the rotation coefficients $R_{l,j,\tilde{j}}(\theta,\phi)$ in (47) were given in [3, 11, 25]. In particular, with focus on practical evaluation, Ganesh and Graham [11] used the following representation, which has enhanced numerical stability for higher degree harmonics,

$$R_{l,j,\tilde{j}}(\theta,\phi) = e^{i(j-\tilde{j})(\phi+\pi/2)} \sum_{|\tilde{m}|\le l} d^{(l)}_{\tilde{j}\tilde{m}}(\pi/2) d^{(l)}_{j\tilde{m}}(\pi/2) e^{i\tilde{m}\theta}, \tag{48}$$

where

$$d^{(l)}_{\tilde{j}j}(\pi/2) = 2^{\tilde{j}} \left[\frac{(l+\tilde{j})!(l-\tilde{j})!}{(l+j)!(l-j)!}\right]^{1/2} P^{(j-\tilde{j},-j-\tilde{j})}_{l+\tilde{j}}(0).$$

For given non-negative integers a, b and $s \ge 0$, $P^{(a,b)}_s(0)$ is the normalized Jacobi polynomial evaluated at zero,

$$P^{(a,b)}_s(0) = 2^{-s} \sum_{t=0}^{s} (-1)^t \binom{s+a}{s-t} \binom{s+b}{t}.$$

When a or b are negative, $d^{(l)}_{\tilde{j}j}$ can be computed using the symmetry relations

$$d^{(l)}_{\tilde{j}j}(\alpha) = (-1)^{\tilde{j}-j} d^{(l)}_{j\tilde{j}}(\alpha) = d^{(l)}_{-j-\tilde{j}}(\alpha) = d^{(l)}_{j\tilde{j}}(-\alpha).$$

5 Numerical Results

We demonstrate the high order convergence of this family of high order fully discrete methods based on hyperinterpolation by tabulating the error in the far field induced by an incident plane wave for the acoustic, electromagnetic and elasticity applications described in Sect. 2. The far field $\boldsymbol{u}^\infty$ is a physical quantity of interest (QoI) for various applications, including the inverse problem. For example, in the inverse scattering models, the input data are the experimentally observed or simulated far field measurements and the computational task is to reconstruct the shape of the scatterer that approximately matches the far field data [4]. Further hyperinterpolation approximations are needed to approximate the far field from the surface current approximations described earlier. For complete details of hyperinterpolation based spectrally accurate far field approximations, we refer to [11–14, 40].

For the electromagnetic example we give results for scattering by perfect electrical conducting (PEC) and dielectric obstacles. The tabulated error is a discrete approximation to

$$\|\boldsymbol{u}^\infty - \boldsymbol{u}^\infty_n\|_\infty, \tag{49}$$

Table 1 Error in the computed far field $\boldsymbol{u}_n^\infty$ for scattering by spheres of diameter 8 wavelengths with $n = n_0,\ n_0 + 5,\ n_0 + 10$

	Acoustic	EM (PEC)	EM (dielectric)	Elasticity
$n - n_0$	$n_0 = 30$	$n_0 = 25$	$n_0 = 35$	$n_0 = 30$
0	2.31e−03	9.84e−02	1.61e−01	6.17e−03
5	7.86e−07	1.12e−03	7.67e−04	3.23e−06
10	4.02e−11	5.63e−07	2.76e−08	1.44e−08

Here the value chosen for n_0 is problem dependent

where $\boldsymbol{u}^\infty$ denotes the true far field and $\boldsymbol{u}_n^\infty$ denotes the approximation computed using the spectral method.

For a spherical scatterer the true far field is given in series form by the Mie series in the acoustic and electromagnetic cases. A similar analytic solution for the elasticity case is given by Le Louër [40, Appendix A].

In Table 1 we present results for spheres of diameter 8 times the incident wavelength. The results for acoustic scattering by a sound soft sphere and for electromagnetic scattering by a perfect conductor were published in [11] and [12] respectively. The results for scattering of an elastic wave by a rigid sphere of diameter 8 times the transverse incident wavelength were published in [40]. The results for electromagnetic scattering by a glass sphere (with refractive index 1.9) are new and were obtained using a computational implementation of the new weakly singular SIE formulation in [21] based on hyperinterpolation approximations.

The key observation from the numerical results is that we obtain high accuracy even for small values of n, which is the key parameter for hyperinterpolation approximations, and correspondingly small DoFs.

The mathematical justification of such high-order accurate algorithms began with the seminal work of Sloan [47], and his intuition, more than two decades ago, about the need to study such approximations.

Acknowledgements Part of this work was carried out while MG was a Visiting Professorial Fellow at University of New South Wales (UNSW) in 2016, funded by an Australian Research Council (ARC) grant. Support of the ARC and the Colorado Golden Energy Computing Organization (GECO) are gratefully acknowledged.

References

1. Ahrens, C., Beylkin, G.: Rotationally invariant quadratures for the sphere. Proc. R. Soc. A **465**, 3103–3125 (2009)
2. Bagheri, S., Hawkins, S.C.: A coupled FEM-BEM algorithm for the inverse acoustic medium problem. ANZIAM J. **56**, C163–C178 (2015)
3. Brink, D.M., Satchler, G.R.: Angular Momentum, 2nd edn. Clarendon Press, Oxford (1968)
4. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 3rd edn. Springer, New York (2013)

5. Cools, R., Poppe, K.: Chebyshev lattices, a unifying framework for cubature with Chebyshev weight function. BIT **51**, 275–288 (2011)
6. Dai, F.: On generalized hyperinterpolation on the sphere. Proc. Am. Math. Soc. **134**, 2931–2941 (2006)
7. Das, P., Nelakanti, G.: Convergence analysis of discrete Legendre spectral projection methods for Hammerstein integral equations of mixed type. Appl. Math. Comput. **265**, 574–601 (2015)
8. Delbary, F., Hansen, P.C., Knudsen, K.: Electrical impedance tomography: 3D reconstructions using scattering transforms. Appl. Anal. **91**, 737–755 (2012)
9. Epstein, C., Greengard, L.: Debye sources and the numerical solution of the time harmonic Maxwell equations. Commun. Pure Appl. Math. **63**, 413–463 (2010)
10. Freeden, W., Gervens, T., Schreiner, M.: Constructive Approximation on the Sphere. Oxford University Press, Oxford (1998)
11. Ganesh, M., Graham, I.G.: A high-order algorithm for obstacle scattering in three dimensions. J. Comput. Phys. **198**, 211–242 (2004)
12. Ganesh, M., Hawkins, S.C.: A spectrally accurate algorithm for electromagnetic scattering in three dimensions. Numer. Algorithms **43**, 25–60 (2006)
13. Ganesh, M., Hawkins, S.C.: A hybrid high-order algorithm for radar cross section computations. SIAM J. Sci. Comput. **29**, 1217–1243 (2007)
14. Ganesh, M., Hawkins, S.C.: A high-order tangential basis algorithm for electromagnetic scattering by curved surfaces. J. Comput. Phys. **227**, 4543–4562 (2008)
15. Ganesh, M., Hawkins, S.C.: Simulation of acoustic scattering by multiple obstacles in three dimensions. ANZIAM J. **50**, 31–45 (2008)
16. Ganesh, M., Hawkins, S.C.: A high-order algorithm for multiple electromagnetic scattering in three dimensions. Numer. Algorithms **50**, 469–510 (2009)
17. Ganesh, M., Hawkins, S.C.: An efficient $\mathscr{O}(N)$ algorithm for computing $\mathscr{O}(N^2)$ acoustic wave interactions in large N-obstacle three dimensional configurations. BIT Numer. Math. **55**, 117–139 (2015)
18. Ganesh, M., Hawkins, S.C.: A high performance computing and sensitivity analysis algorithm for stochastic many-particle wave scattering. SIAM J. Sci. Comput. **37**, A1475–A1503 (2015)
19. Ganesh, M., Morgenstern, C.: High-order FEM-BEM computer models for wave propagation in unbounded and heterogeneous media: application to time-harmonic acoustic horn problem. J. Comput. Appl. Math. **37**, 183–203 (2016)
20. Ganesh, M., Graham, I.G., Sivaloganathan, J.: A pseudospectral three-dimensional boundary integral method applied to a nonlinear model problem from finite elasticity. SIAM. J. Numer. Anal. **31**, 1378–1414 (1994)
21. Ganesh, M., Hawkins, S.C., Volkov, D.: An all-frequency weakly-singular surface integral equation for electromagnetism in dielectric media: reformulation and well-posedness analysis. J. Math. Anal. Appl. **412**, 277–300 (2014)
22. Gentile, M., Sommariva, A., Vianello, M.: Polynomial approximation and quadrature on geographic rectangles. Appl. Math. Comput. **297**, 159–179 (2017)
23. Golberg, M.A., Chen, C.S., Bowman, H.: Some recent results and proposals for the use of radial basis functions in the BEM. Eng. Anal. Bound. Elem. **23**, 285–296 (1999)
24. Gräf, M., Kunis, S., Potts, D.: On the computation of nonnegative quadrature weights on the sphere. Appl. Comput. Harmon. Anal. **27**, 124–132 (2009)
25. Graham, I.G., Sloan, I.H.: Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in $\mathbb{R}^3$. Numer. Math. **92**, 289–323 (2002)
26. Gronwall, T.H.: On the degree of convergence of the Laplace series. Trans. Am. Math. Soc. **22**, 1–30 (1914)
27. Hansen, O., Atkinson, K., Chien, D.: On the norm of the hyperinterpolation operator on the unit disc and its use for the solution of the nonlinear Poisson equation. IMA J. Numer. Anal. **29**, 257–283 (2009)
28. Hellmers, J., Eremina, E., Wriedt, T.: Simulation of light scattering by biconcave Cassini ovals using the nullfield method with discrete sources. J. Opt. A Pure Appl. Opt. **8**, 1–9 (2006)

29. Hesse, K., Sloan, I.H., Womersley, R.S.: Numerical integration on the sphere. Handbook of Geomathematics, pp. 2671–2710. Springer, Berlin (2015)
30. Kazashi, Y.: A fully discretised polynomial approximation on spherical shells. GEM Int. J. Geomath. **7**, 299–323 (2016)
31. Kress, R.: Linear Integral Equations, 2nd edn. Springer, New York (1999)
32. Kulkarni, R.P., Gnaneshwar, N.: Iterated discrete polynomially based Galerkin methods. Appl. Math. Comput. **146**, 153–165 (2003)
33. Kupradze, V.D.: Three-Dimensional Problems of Elasticity and Thermoelasticity, vol. 25. Elsevier, Amsterdam (1979)
34. Langdon, S., Graham, I.G.: Boundary integral methods for singularly perturbed boundary value problems. IMA J. Numer. Anal. **21**, 217–237 (2001)
35. Le Gia, Q.T., Mhaskar, H.N.: Localised linear polynomial operators and quadrature formulas on the sphere. SIAM J. Numer. Anal. **47**, 440–466 (2008)
36. Le Gia, Q.T., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Needlet approximation for isotropic random fields on the sphere. J. Approx. Theory **216**, 86–116 (2017)
37. Leopardi, P.C.: Positive weight quadrature on the sphere and monotonicities of Jacobi polynomials. Numer. Algorithms **45**, 75–87 (2007)
38. Li, X.: Rate of convergence of the method of fundamental solutions and hyperinterpolation for modified Helmholtz equations on the unit ball. Adv. Comput. Math. **29**, 393–413 (2008)
39. Li, X., Chen, C.S.: A mesh free method using hyperinterpolation and fast Fourier transform for solving differential equations. Eng. Anal. Bound. Elem. **28**, 1253–1260 (2004)
40. Le Louër, F.: A high order spectral algorithm for elastic obstacle scattering in three dimensions. J. Comput. Phys. **279**, 1–17 (2014)
41. Le Louër, F.: Spectrally accurate numerical solution of hypersingular boundary integral equations for three-dimensional electromagnetic wave scattering problems. J. Comput. Phys. **275**, 662–666 (2014)
42. Nousiainen, T., McFarquhar, G.M.: Light scattering by quasi-spherical ice crystals. J. Atmos. Sci. **61**, 2229–2248 (2004)
43. Pieper, M.: Vector hyperinterpolation on the sphere. J. Approx. Theory **156**, 173–186 (2009)
44. Poppe, K., Cools, R.: CHEBINT: A MATLAB/Octave toolbox for fast multivariate integration and interpolation based on Chebyshev approximations over hypercubes. ACM Trans. Math. Softw. **40**(1), Article 2 (2013)
45. Reimer, M.: Hyperinterpolation on the sphere at the minimal projection order. J. Approx. Theory **104**, 272–286 (2000)
46. Reimer, M.: Generalized hyperinterpolation on the sphere and the Newman-Shapiro operators. Constr. Approx. **18**, 183–203 (2002)
47. Sloan, I.H.: Polynomial interpolation and hyperinterpolation over general regions. J. Approx. Theory **83**, 238–254 (1995)
48. Sloan, I.H., Womersley, R.S.: Constructive polynomial approximation on the sphere. J. Approx. Theory **103**, 91–118 (2000)
49. Sloan, I.H., Womersley, R.S.: Filtered hyperinterpolation: a constructive polynomial approximation on the sphere. GEM Int. J. Geomath. **3**, 95–117 (2012)
50. Veihelmann, B., Nousiainen, T., Kahnert, M., van der Zande, W.J.: Light scattering by small feldspar particles simulated using the Gaussian random sphere geometry. J. Quant. Spectrosc. Radiat. Transf. **100**, 393–405 (2005)
51. Womersley, R.S., Sloan, I.H.: How good can polynomial interpolation on the sphere be? Adv. Comput. Math. **14**, 195–226 (2001)
52. Woo, A.C., Wang, H.T., Schuh, M.J., Sanders, M.L.: Benchmark radar targets for the validation of computational electromagnetics programs. IEEE Antennas Propag. Mag. **35**, 84–89 (1993)
53. Wriedt, T., Hellmers, J., Eremina, E., Schuh, R.: Light scattering by single erythrocyte: comparison of different methods. J. Quant. Spectrosc. Radiat. Transf. **100**, 444–456 (2006)
54. Zygmund, A.: Trigonometric Series, vol. I, 3rd edn. Cambridge University Press, Cambridge (2002)

Multilevel QMC with Product Weights for Affine-Parametric, Elliptic PDEs

Robert N. Gantner, Lukas Herrmann, and Christoph Schwab

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We present an error analysis of higher order Quasi-Monte Carlo (QMC) integration and of randomly shifted QMC lattice rules for parametric operator equations with uncertain input data taking values in Banach spaces. Parametric expansions of these input data in locally supported bases such as splines or wavelets was shown in Gantner et al. (SIAM J Numer Anal 56(1):111–135, 2018) to allow for dimension independent convergence rates of combined QMC-Galerkin approximations. In the present work, we review and refine the results in that reference to the multilevel setting, along the lines of Kuo et al. (Found Comput Math 15(2):441–449, 2015) where randomly shifted lattice rules and globally supported representations were considered, and also the results of Dick et al. (SIAM J Numer Anal 54(4):2541–2568, 2016) in the particular situation of locally supported bases in the parametrization of uncertain input data. In particular, we show that locally supported basis functions allow for multilevel QMC quadrature with product weights, and prove new error vs. work estimates superior to those in these references (albeit at stronger, mixed regularity assumptions on the parametric integrand functions than what was required in the single-level QMC error analysis in the first reference above). Numerical experiments on a model affine-parametric elliptic problem confirm the analysis.

R. N. Gantner · L. Herrmann · C. Schwab (✉)
Seminar for Applied Mathematics, ETH Zürich, Zurich, Switzerland
e-mail: robert.gantner@sam.math.ethz.ch; lukas.herrmann@sam.math.ethz.ch; christoph.schwab@sam.math.ethz.ch

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_18

1 Introduction

A core task in computational uncertainty quantification (UQ for short) is to approximate the statistics of (functionals of) solutions to partial differential equations (PDEs for short) which depend on parameters describing uncertain input data. Upon placing probability measures on admissible parameters, the computation of mathematical expectations in so-called *forward UQ* and in *Bayesian inverse UQ* of such PDEs on possibly large sets of data amounts to a problem of *high dimensional numerical integration*; we refer to the surveys [4, 5, 26] and the references there.

In the present note, we address the numerical analysis of high-dimensional numerical integration methods of *Quasi-Monte Carlo* (QMC for short) type for the efficient numerical approximation of expectations of solutions of parametric PDEs over high-dimensional parameter spaces. Pioneering contributions to the mathematical foundation of dimension-independent convergence rates for QMC quadrature methods for such problems that we will draw on also in the present note, are by Sloan and Woźniakowski in [28], after earlier, foundational work by Sloan and Joe in [27].

Specifically, we consider the linear, affine-parametric elliptic PDE

$$\begin{aligned} -\nabla \cdot (a(x,\boldsymbol{y})\nabla u(x,\boldsymbol{y})) &= f(x) \text{ in } D, \\ u(x,\boldsymbol{y})\Big|_{\Gamma_1} = 0, \quad a(x,\boldsymbol{y})\nabla u(x,\boldsymbol{y})\cdot n(x)\Big|_{\Gamma_2} &= 0, \end{aligned} \tag{1}$$

with input $a(x,\boldsymbol{y})$ parametrized by $\boldsymbol{y} = (y_j)_{j\geq 1}$, fixed, $\boldsymbol{y}$-independent right hand side $f(x)$, and mixed boundary conditions. The domain $D \subset \mathbb{R}^d$, $d = 1,2$, is assumed to be either a bounded polygon with straight sides if $d = 2$ or, if $d = 1$, a bounded interval. The set $\Gamma_1 \neq \emptyset$ is assumed to be the union of some of the closed edges of ∂D, $\Gamma_2 := \partial D \backslash \Gamma_1$, and $n(x)$ denotes the unit outward pointing normal vector of D. Specifically, QMC rules with *product weights* are considered which are known to have linear complexity in the integration dimension, cp. [24, 25]. The purpose of the present paper is to prove error versus work bounds of these algorithms, with explicit estimation of the dependence of the constants on the dimension s of the domain of integration, and of the form $\mathscr{O}(\varepsilon^{-\theta})$, $\theta > 0$, for a given accuracy $\varepsilon > 0$.

Convergence analysis of QMC methods with randomly shifted lattice rules applied to a parametric PDE of the type (1) was first established in [21] together with the survey [20]. Randomly shifted lattice rules were first proposed in [29]. A multilevel version for parametric PDEs was first analyzed in [22]. This theory was extended in [7, 8] with interlaced polynomial lattice rules, which achieve higher order convergence rates. These convergence rates are independent of the number of scalar variables that is, of the dimension of the domain of integration. Conditions for such dimension independent error bounds of QMC algorithms were first shown in the seminal work [28] for integrand functions belonging to certain weighted function spaces with so-called *product weights*. In [20, 21], analogous results were shown to hold for randomly shifted lattice rules, and for input parametrizations in terms of globally supported basis functions (as, e.g., Karhunen-Loève expansions) with so-called *product and order dependent (POD for short) weights*. General references for QMC integration are [6, 20]; see also the survey [13] for multilevel Monte Carlo methods and [9, 19] for available software implementations.

As in the mentioned references, we admit parameter vectors $\boldsymbol{y} = (y_j)_{j\geq 1}$ whose components take values in the closed interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$, i.e., we will consider

$$\boldsymbol{y} \in U := \left[-\frac{1}{2}, \frac{1}{2}\right]^{\mathbb{N}}.$$

We model uncertainty in diffusion coefficients $a(x, \boldsymbol{y})$ to the PDE (1) by assuming the parameter vectors to be independent, identically distributed (i.i.d. for short) with respect to the uniform product probability measure

$$\mu(\mathrm{d}\boldsymbol{y}) := \bigotimes_{j\geq 1} \mathrm{d}y_j.$$

The triplet $(U, \bigotimes_{j\geq 1} \mathscr{B}([-1/2, 1/2]), \mu)$ is a probability space. For any Banach space B, the mathematical expectation of F with respect to the probability measure μ is a Bochner integral of the strongly measurable, integrable map $F : U \to B$ which will be denoted by

$$\mathbb{E}(F) := \int_U F(\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{y}). \tag{2}$$

The parametric input $a(x, \boldsymbol{y})$ of (1) is assumed to be of the form

$$a(x, \boldsymbol{y}) = \bar{a}(x) + \sum_{j\geq 1} y_j \psi_j(x), \quad \text{a.e. } x \in D, \boldsymbol{y} \in U, \tag{3}$$

where $\{\bar{a}, \psi_j : j \geq 1\} \subset L^\infty(D)$ and $\bar{a}$ is such that $0 < \bar{a}_{\min} \leq \bar{a}_{\max}$ exist and satisfy

$$\bar{a}_{\min} \leq \text{ess inf}_{x\in D}\{\bar{a}(x)\} \leq \text{ess sup}_{x\in D}\{\bar{a}(x)\} \leq \bar{a}_{\max}.$$

Convergence analysis for QMC with product weights was recently carried out in [11] under the assumption that there exists $\kappa \in (0, 1)$ and a sequence $(b_j)_{j\geq 1} \in (0, 1]^{\mathbb{N}}$ such that

$$\left\| \frac{\sum_{j\geq 1} |\psi_j|/b_j}{2\bar{a}} \right\|_{L^\infty(D)} \leq \kappa < 1. \tag{\textbf{A1}}$$

A (dimension independent) convergence rate of $1/p$ in terms of the number of QMC points for the approximate evaluation of (2) was shown in [11, Section 6] if $(b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$ for the range $p \in (0, 2]$. These rates coincide, in the mentioned range of summability exponents, with the convergence rates of best N-term approximation rates of generalized polynomial chaos expansions obtained in [3, Theorem 1.2 and Equation (1.11)]. As in [3], the assumption in (**A1**) can accommodate possible localization in D of the supports of the function system $(\psi_j)_{j\geq 1}$.

For every parameter instance $\boldsymbol{y} \in U$, in the physical domain D a standard, first order accurate Galerkin Finite Element (FE for short) discretization of the parametric PDE (1) will be applied. In the polygonal domain D, first order FE based on sequences of uniformly refined, regular simplicial meshes are well-known to converge at suboptimal rates due to *corner singularities* in the solution, even if $a(x,\boldsymbol{y})$ and $f(x)$ in (1) are smooth. To establish full FE convergence rates on locally refined meshes coupled with the QMC error estimates, parametric regularity estimates in *Kondrat'ev* spaces will be demonstrated.

In Sect. 2 well-posedness of the parametric solution and approximation by dimension truncation and FE is discussed. Particular weighted Sobolev spaces of parametric regularity that are required for the error analysis of multilevel QMC and general error estimates are reviewed in Sect. 3. Parametric regularity estimates of the dimension truncation and FE error are proven in Sect. 4. These estimates yield error bounds of multilevel QMC algorithms that are demonstrated in Sect. 5. In Sect. 6, parameter choices are derived that minimize the needed work for a certain error threshold. In the numerical experiments, we analyze piecewise (bi)linear wavelet bases to expand the diffusion coefficient in one and two spatial dimensions. The experiments confirm the theory and also show that the multilevel QMC algorithm outperforms the single-level version in terms of work versus achieved accuracy, in the engineering range of accuracy, and for a moderate number of integration points.

2 Well-Posedness and Spatial Approximation

The parametric problem in (1) admits a symmetric variational formulation with trial and test space $V := \{v \in H^1(D) : v|_{\Gamma_1} = 0\}$, with dual space denoted by $V^* = H^{-1}(D)$, where $v|_{\Gamma_1} = 0$ is to be understood as a trace in $H^{1/2}(\Gamma_1)$. Let $f \in V^*$ and let the assumption in (**A1**) be satisfied. Then, the *parametric weak formulation of* (1) reads: for every $\boldsymbol{y} \in U$ find $u(\cdot,\boldsymbol{y}) \in V$ such that

$$\int_D a(\cdot,\boldsymbol{y})\nabla u(\cdot,\boldsymbol{y}) \cdot \nabla v \mathrm{d}x = \langle f, v\rangle_{V^*,V}, \quad \forall v \in V, \tag{4}$$

where $\langle \cdot, \cdot\rangle_{V^*,V}$ denotes the dual pairing between V and V^*. Since the assumption in (**A1**) implies that

$$0 < (1-\kappa)\bar{a}_{\min} \leq \text{ess }\inf_{x\in D}\{a(x,\boldsymbol{y})\}, \quad \boldsymbol{y} \in U,$$

and

$$\text{ess sup}_{x\in D}\{a(x,\boldsymbol{y})\} \le (1+\kappa)\bar{a}_{\max}, \quad \boldsymbol{y} \in U,$$

the parametric bilinear form $(w,v) \mapsto \int_D a(\cdot,\boldsymbol{y})\nabla w \cdot \nabla v \mathrm{d}x$ is continuous and coercive on $V \times V$, uniformly with respect to the parameter vector $\boldsymbol{y} \in U$. By the Lax–Milgram lemma, the unique solution $u(\cdot,\boldsymbol{y}) \in V$ to (4) exists, is a strongly measurable mapping from U to V (by the second Strang lemma), and satisfies the a priori estimate

$$((1-\kappa)\bar{a}_{\min})\|u(\cdot,\boldsymbol{y})\|_V \le \|f\|_{V^*} \qquad \boldsymbol{y} \in U.$$

A finite dimensional domain of integration which is required for the use of QMC is achieved by truncating the expansion of $a(x,\boldsymbol{y})$ to a finite number of $s \in \mathbb{N}$ terms. We introduce the notation that for every $\boldsymbol{y} \in U$, $\boldsymbol{y}_{\{1:s\}}$ is such that $(\boldsymbol{y}_{\{1:s\}})_j = y_j$ if $j \le s$ and 0 otherwise, where $\{1 : s\}$ denotes the set $\{1,\dots,s\}$. Specifically, for every $s \in \mathbb{N}$ define

$$u^s(\cdot,\boldsymbol{y}) := u(\cdot,\boldsymbol{y}_{\{1:s\}}), \quad \boldsymbol{y} \in U.$$

Proposition 1 ([11, Proposition 5.1]) *Let the assumption in (**A1**) be satisfied for some $\kappa \in (0,1)$ and recall the right hand side $f \in V^*$ in (4). If for some $s_0 \in \mathbb{N}$*

$$\frac{\kappa\bar{a}_{\max}}{(1-\kappa)\bar{a}_{\min}} \sup_{j\ge s+1}\{b_j\} < 1,$$

then there exists a constant $C > 0$ such that for every $s \ge s_0$ and every $G(\cdot) \in V^*$

$$|\mathbb{E}(G(u)) - \mathbb{E}(G(u^s))| \le C\|G(\cdot)\|_{V^*}\|f\|_{V^*}\left(\sup_{j\ge s+1}\{b_j\}\right)^2.$$

For the study of the spatial regularity of $u(\cdot,\boldsymbol{y})$, we consider weighted Sobolev spaces of *Kondrat'ev* type, which allow for full regularity shifts in polygonal domains $D \subset \mathbb{R}^2$, cp. [2]. In our setting the domain D is either a polygon in $\mathbb{R}^2$ with corners $\{c_1,\dots,c_J\}$ or an interval. To introduce weighted Sobolev spaces, we define the functions $r_i(x) := |x - c_i|$, $x \in D$, $i = 1,\dots,J$, where $|\cdot|$ denotes the Euclidean norm. For a J-tuple $\boldsymbol{\beta} = (\beta_1,\dots,\beta_J)$ with $\beta_i \in [0,1)$, $i = 1,\dots,J$, we define the weight function $\Phi_{\boldsymbol{\beta}}$ by

$$\Phi_{\boldsymbol{\beta}}(x) := \prod_{i=1}^{J} r_i^{\beta_i}(x), \quad x \in D.$$

For multi-indices $\boldsymbol{\alpha} \in \mathbb{N}_0^2$, define the notation $\partial_x^{\boldsymbol{\alpha}} := \partial^{|\boldsymbol{\alpha}|}/(\partial x_1^{\alpha_1}\partial x_2^{\alpha_2})$. We define the weighted spaces $L^2_{\boldsymbol{\beta}}(D)$ and $H^2_{\boldsymbol{\beta}}(D)$ as the completion of $C^\infty(\overline{D})$ with respect to the corresponding norms which are given by

$$\|v\|_{L^2_{\boldsymbol{\beta}}(D)} := \|v\Phi_{\boldsymbol{\beta}}\|_{L^2(D)}, \quad \|v\|^2_{H^2_{\boldsymbol{\beta}}(D)} := \|v\|^2_{H^1(D)} + \sum_{|\boldsymbol{\alpha}|=2} \||\partial_x^{\boldsymbol{\alpha}} v|\|^2_{L^2_{\boldsymbol{\beta}}(D)}. \tag{5}$$

In the corresponding weighted Sobolev spaces, there is a full regularity shift of the Laplacean, cp. [2, Theorem 3.2 and Equation (3.2)], i.e., there exists a constant $C > 0$ such that for every $w \in V$ satisfying $\Delta w \in L^2_{\boldsymbol{\beta}}(D)$ there holds

$$\|w\|_{H^2_{\boldsymbol{\beta}}(D)} \leq C\|\Delta w\|_{L^2_{\boldsymbol{\beta}}(D)}, \tag{6}$$

where $\beta_i > 1 - \pi/\omega_i$ and $\beta_i \geq 0$, (ω_i denotes the interior angle of the corner c_i) for $i = 1, \ldots, J$ such that both edges that have c_i as an endpoint are both in Γ_1 or Γ_2. Otherwise (change of the boundary conditions at c_i), we require $\beta_i > 1 - \pi/(2\omega_i)$ and $\beta_i \geq 0$. Note that we allow the case $\omega_i = \pi$, which facilitates the case that the boundary conditions change within one edge of ∂D. Hence, in the case that the domain D is convex and $\Gamma_1 = \partial D$, we may choose $\boldsymbol{\beta} = (0, \ldots, 0)$. There holds an approximation property in FE spaces on D with local mesh refinement towards the corners of D. To state it, let $\{\mathcal{T}_\ell\}_{\ell\geq 0}$ denote a sequence of regular, simplicial triangulations of the polygon D, which can be generated either by judicious mesh grading in a vicinity of each corner of D, cp. [2, Section 4], or by *newest vertex bisection*, cp. [12]. Let $V_\ell := \{v \in V : v|_K \in \mathbb{P}_1(K), K \in \mathcal{T}_\ell\}$, $\ell \geq 0$, where $\mathbb{P}_1(K)$ denotes the affine functions on K. The FE space V_ℓ is of finite dimension $M_\ell := \dim(V_\ell)$, $\ell \geq 0$. Then, there exists a constant $C > 0$ such that for every $w \in H^2_{\boldsymbol{\beta}}(D)$ and every $\ell \geq 0$ there exists $w_\ell \in V_\ell$ such that

$$\|w - w_\ell\|_V \leq CM_\ell^{-1/d}\|w\|_{H^2_{\boldsymbol{\beta}}(D)}, \tag{7}$$

where $d = 1, 2$ is the dimension of the domain D. For $d = 2$, and in the case of graded meshes, this follows, for example, from [2, Lemmas 4.1 and 4.5]. An approximation property of this kind for newest vertex bisection is shown in [12]. The regularity shift in (6) and the approximation property in (7) also hold if D is an interval (for $d = 1$).

Assume that the right hand side $f \in L^2_{\boldsymbol{\beta}}(D)$, and that $\{|\nabla\bar{a}|\Phi_{\boldsymbol{\beta}}, |\nabla\psi_j|\Phi_{\boldsymbol{\beta}} : j \geq 1\} \subset L^\infty(D)$, and that there exists a bounded, positive sequence $(\bar{b}_j)_{j\geq 1}$, which satisfies

$$K := \left\| \left(|\nabla\bar{a}| + \sum_{j\geq 1} \frac{|\nabla\psi_j|}{\bar{b}_j} \right) \Phi_{\boldsymbol{\beta}} \right\|_{L^\infty(D)} < \infty. \tag{\textbf{A2}}$$

This assumption readily implies that $\sup_{\boldsymbol{y}\in U}\{\||\nabla a(\cdot,\boldsymbol{y})|\Phi_{\boldsymbol{\beta}}\|_{L^\infty(D)}\} < \infty$. As a consequence, $|\nabla a(\cdot,\boldsymbol{y})| \in L^\infty(\widetilde{D})$ and $\Delta u(\cdot,\boldsymbol{y}) \in L^2(\widetilde{D})$ for every compactly included subdomain $\widetilde{D} \subset\subset D$ and for every $\boldsymbol{y} \in U$. Then, by the divergence theorem and by the product rule for every $v \in C_0^\infty(D) \subset V$

$$\int_D a(\cdot,\boldsymbol{y})\nabla u(\cdot,\boldsymbol{y}) \cdot \nabla v \,\mathrm{d}x = -\int_D [\nabla \cdot (a(\cdot,\boldsymbol{y})\nabla u(\cdot,\boldsymbol{y}))]v \,\mathrm{d}x$$
$$= -\int_D (a(\cdot,\boldsymbol{y})\Delta u(\cdot,\boldsymbol{y}) + \nabla a(\cdot,\boldsymbol{y}) \cdot \nabla u(\cdot,\boldsymbol{y}))v \,\mathrm{d}x.$$

We have to show that $\Delta u(\cdot,\boldsymbol{y}) \in L^2_{\boldsymbol{\beta}}(D)$. By duality of the space $L^2_{\boldsymbol{\beta}}(D)$, the previous identity, and by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|a(\cdot,\boldsymbol{y})\Delta u(\cdot,\boldsymbol{y})\|_{L^2_{\boldsymbol{\beta}}(D)} &= \sup_{v\in L^2_{\boldsymbol{\beta}}(D),\|v\|_{L^2_{\boldsymbol{\beta}}(D)}\leq 1} \int_D a(\cdot,\boldsymbol{y})\Delta u(\cdot,\boldsymbol{y})v\Phi_{\boldsymbol{\beta}}^2 \,\mathrm{d}x \\ &= \sup_{v\in C_0^\infty(D),\|v\|_{L^2(D)}\leq 1} \int_D a(\cdot,\boldsymbol{y})\Delta u(\cdot,\boldsymbol{y})v\Phi_{\boldsymbol{\beta}} \,\mathrm{d}x \\ &= \sup_{v\in C_0^\infty(D),\|v\|_{L^2(D)}\leq 1} \int_D (f + \nabla a(\cdot,\boldsymbol{y}) \cdot \nabla u(\cdot,\boldsymbol{y}))v\Phi_{\boldsymbol{\beta}} \,\mathrm{d}x \\ &\leq \|f\|_{L^2_{\boldsymbol{\beta}}(D)} + \||\nabla a(\cdot,\boldsymbol{y})|\Phi_{\boldsymbol{\beta}}\|_{L^\infty(D)}\|u(\cdot,\boldsymbol{y})\|_V. \end{aligned} \tag{8}$$

Since ess $\inf_{x\in D}\{a(x,\boldsymbol{y})\} \geq (1-\kappa)\bar{a}_{\min}$, $\Delta u(\cdot,\boldsymbol{y}) \in L^2_{\boldsymbol{\beta}}(D)$. In (8), we applied that $C_0^\infty(D)$ is dense in $L^2(D)$ and used that the operator of pointwise multiplication $w \mapsto w\Phi_{-\boldsymbol{\beta}}$ is an isometry from $L^2(D)$ to $L^2_{\boldsymbol{\beta}}(D)$.

The *parametric FE solution* is defined as the unique solution of the variational problem: for $\boldsymbol{y} \in U$ and $\ell \geq 0$, find $u^{\mathcal{T}_\ell}(\cdot,\boldsymbol{y}) \in V_\ell$ such that

$$\int_D a(\cdot,\boldsymbol{y})\nabla u^{\mathcal{T}_\ell}(\cdot,\boldsymbol{y}) \cdot \nabla v \,\mathrm{d}x = \langle f, v\rangle_{V^*,V}, \quad \forall v \in V_\ell. \tag{9}$$

Well-posedness of the parametric FE solution also follows by the Lax–Milgram lemma. As above, we define for every truncation dimension $s \in \mathbb{N}$ and level $\ell \in \mathbb{N}_0$

$$u^{s,\mathcal{T}_\ell}(\cdot,\boldsymbol{y}) := u^{\mathcal{T}_\ell}(\cdot,\boldsymbol{y}_{\{1:s\}}), \quad \boldsymbol{y} \in U.$$

By Céa's lemma, an Aubin–Nitsche argument, Proposition 1, (6)–(8), there exists a constant $C > 0$ such that for every $s \in \mathbb{N}$, $\ell \geq 0$, and every $G(\cdot) \in L^2_{\boldsymbol{\beta}}(D)$,

$$\begin{aligned}|\mathbb{E}(G(u)) - \mathbb{E}(G(u^{s,\mathcal{T}_\ell}))| \leq C\|G(\cdot)\|_{V^*}\|f\|_{V^*}\left(\sup_{j\geq s+1}\{b_j\}\right)^2 \\ + C\|G(\cdot)\|_{L^2_{\boldsymbol{\beta}}(D)}\|f\|_{L^2_{\boldsymbol{\beta}}(D)}M_\ell^{-2/d}.\end{aligned} \tag{10}$$

Remark 1 If f and $G(\cdot)$ have less regularity, say $f \in (V^*, L^2_{\boldsymbol{\beta}}(D))_{t,\infty}$ and $G(\cdot) \in (V^*, L^2_{\boldsymbol{\beta}}(D))_{t',\infty}$, $t, t' \in (0, 1)$, then the estimate in (10) holds with $M_\ell^{-(t+t')/d}$. This follows by interpolation. The interpolation spaces are in the sense of the *K-method*, cp. [31]. Since $L^2_{\boldsymbol{\beta}}(D) \subset V^*$ continuously which follows by [2, Equation (3.2)], V^* and $L^2_{\boldsymbol{\beta}}(D)$ are an interpolation couple. Naturally the embedding $H^{-1+t'}(D) = (V^*, L^2(D))_{t',2} \subset (V^*, L^2_{\boldsymbol{\beta}}(D))_{t',\infty}$ is continuous, since $L^2(D)$ is continuously embedded in $L^2_{\boldsymbol{\beta}}(D)$.

3 Multilevel QMC Integration

Randomly shifted lattice rules and interlaced polynomial lattice rules are QMC rules that have well-known *worst case error estimates* in particular weighted Sobolev spaces of regularity with respect to the dimensionally truncated parameter vectors $\boldsymbol{y}_{\{1:s\}}$, $s \in \mathbb{N}$. Generally, these QMC rules approximate dimensionally truncated integrals

$$I_s(F) := \int_{[-\frac{1}{2},\frac{1}{2}]^s} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}.$$

Denote by $Q^{\mathrm{RS}}_{s,N}(\cdot)$ and $Q^{\mathrm{IP}}_{s,N}(\cdot)$ randomly shifted lattice rules and interlaced polynomial lattice rules in dimension s with N points, respectively. Subsequently, if the superscript is omitted either of the QMC rules is meant. For a nondecreasing sequence $(s_\ell)_{\ell=0,\dots,L}$ of truncation dimensions and numbers of QMC points $(N_\ell)_{\ell=0,\dots,L}$, $L \in \mathbb{N}$, and for the meshes $\{\mathscr{T}_\ell\}_{\ell \geq 0}$ from Sect. 2, the multilevel QMC quadrature for $L \in \mathbb{N}$ levels is, for every $G(\cdot) \in V^*$, defined by

$$Q_L(G(u^L)) := \sum_{\ell=0}^{L} Q_{s_\ell,N_\ell}(G(u^\ell - u^{\ell-1})),$$

where we introduced the notation $u^\ell := u^{s_\ell,\mathscr{T}_\ell}$, $\ell \in \mathbb{N}_0$, and have set $u^{-1} := 0$. Throughout, we shall assume that sequences of numbers of QMC points $(N_\ell)_{\ell=0,\dots,L}$ are nonincreasing. For the error analysis, we introduce for a collection of QMC weights $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u} \subset \mathbb{N}}$ the weighted Sobolev spaces $\mathscr{W}_{s,\boldsymbol{\gamma}}$ and $\mathscr{W}_{s,\alpha,\boldsymbol{\gamma},q,r}$ as closures of $C^\infty([-1/2, 1/2]^s)$ with respect to the norms

$$\|F\|_{\mathscr{W}_{s,\boldsymbol{\gamma}}} := \left(\sum_{\mathfrak{u} \subset \{1:s\}} \gamma_{\mathfrak{u}}^{-1} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{u}|}} \left| \int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathfrak{u}|}} \partial^{\mathfrak{u}}_{\boldsymbol{y}} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}_{\{1:s\} \setminus \mathfrak{u}} \right|^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \right)^{1/2}$$

and for $2 \le \alpha \in \mathbb{N}$, $q, r \in [1, \infty]$

$$\|F\|_{\mathscr{W}_{s,\alpha,\boldsymbol{\gamma},q,r}} := \left(\sum_{\mathfrak{u}\subset\{1:s\}} \left(\gamma_{\mathfrak{u}}^{-q} \sum_{\mathfrak{v}\subset\mathfrak{u}} \sum_{\boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}}\in\{1:\alpha\}^{|\mathfrak{u}\setminus\mathfrak{v}|}} \right.\right.$$
$$\left.\left. \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{v}|}} \left| \int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathfrak{v}|}} \partial_{\boldsymbol{y}}^{(\boldsymbol{\alpha}_{\mathfrak{v}},\boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}},\boldsymbol{0})} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{v}} \right|^{q} \mathrm{d}\boldsymbol{y}_{\mathfrak{v}} \right)^{r/q} \right)^{1/r}$$

with the obvious modifications if q or r is infinite. Here, $(\boldsymbol{\alpha}_{\mathfrak{v}}, \boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}}, \boldsymbol{0}) \in \{0 : \alpha\}^{s}$ denotes a multi-index such that $(\boldsymbol{\alpha}_{\mathfrak{v}}, \boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}}, \boldsymbol{0})_j = \alpha$ for $j \in \mathfrak{v}$, $(\boldsymbol{\alpha}_{\mathfrak{v}}, \boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}}, \boldsymbol{0})_j = \boldsymbol{\tau}_j$ for $j \in \mathfrak{u}\setminus\mathfrak{v}$, and $(\boldsymbol{\alpha}_{\mathfrak{v}}, \boldsymbol{\tau}_{\mathfrak{u}\setminus\mathfrak{v}}, \boldsymbol{0})_j = 0$ for $j \notin \mathfrak{u}$, for every $\mathfrak{u} \subseteq \{1 : s\}$, $\mathfrak{v} \subseteq \mathfrak{u}$, $\boldsymbol{\tau} \in \{1 : \alpha\}^{|\mathfrak{u}\setminus\mathfrak{v}|}$. Note that the integer $\alpha \ge 2$ is the interlacing factor. For every $\mathfrak{u} \subset \{1 : s\}$, $\mathrm{d}\boldsymbol{y}_{\mathfrak{u}}$ denotes the product measure $\bigotimes_{j\in\mathfrak{u}} \mathrm{d}y_j$. The following two estimates follow essentially from the worst case error estimates in [21, Theorem 2.1] and [7, Theorem 3.10] for $Q_{s,N}^{\mathrm{RS}}(\cdot)$ and $Q_{s,N}^{\mathrm{IP}}(\cdot)$, respectively. For every $\lambda \in (1/2, 1]$,

$$\mathbb{E}^{\boldsymbol{\Delta}}(|I_{s_L}(G(u^L)) - Q_L^{\mathrm{RS}}(G(u^L))|^2)$$
$$\le \sum_{\ell=0}^{L} \left(\sum_{\emptyset\neq\mathfrak{u}\subset\{1:s_\ell\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/\lambda} (\varphi(N_\ell))^{-1/\lambda} \|G(u^\ell - u^{\ell-1})\|^2_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}}, \tag{11}$$

cp. [22, Equation (25)], where $\boldsymbol{\Delta}$ denotes the random shift and φ denotes the Euler totient function. For every $\lambda \in (1/\alpha, 1]$,

$$|I_{s_L}(G(u^L)) - Q_L^{\mathrm{IP}}(G(u^L))|$$
$$\le \sum_{\ell=0}^{L} \left(\frac{2}{N_\ell - 1} \sum_{\emptyset\neq\mathfrak{u}\subset\{1:s_\ell\}} \gamma_{\mathfrak{u}}^{\lambda} (\rho_\alpha(\lambda))^{|\mathfrak{u}|} \right)^{1/\lambda} \|G(u^\ell - u^{\ell-1})\|_{\mathscr{W}_{s_\ell,\alpha,\boldsymbol{\gamma},\infty,\infty}}, \tag{12}$$

cp. [8, Equation (42)], where the constant $\rho_\alpha(\lambda)$ is finite if $\lambda > 1/\alpha$ as stated in [7, Equation (3.37)]. We remark that the choice of λ, α, $\boldsymbol{\gamma}$ in (11) and (12) may also depend on the level $\ell = 0, \ldots, L$, which is not explicit in the notation.

4 Parametric Regularity

As in the single-level QMC analysis in [11], we introduce the auxiliary parameter set $\widetilde{U} = [-1, 1]^{\mathbb{N}}$ with elements $\boldsymbol{z} \in \widetilde{U}$. Fix $\eta \in (\kappa, 1)$. We split the sparsity of the sequence $(b_j)_{j\ge1}$ between spatial approximation and QMC approximation rates, which naturally couple in multilevel integration methods. For a sequence $(\widehat{b}_j)_{j\geq 1}$ (to be specified in the following) which satisfies the assumption (**A1**), and for every $\boldsymbol{y} \in U$ define

$$\bar{a}_{\boldsymbol{y}}(x) = \bar{a}(x) + \sum_{j\geq 1} y_j \psi_j(x) \text{ and } \psi_{\boldsymbol{y},j}(x) = \frac{\eta^{-1} - 2|y_j|}{2\widehat{b}_j}\psi_j(x), \text{ a.e. } x \in D, j \in \mathbb{N}, \tag{13}$$

which are used to construct

$$\widetilde{a}_{\boldsymbol{y}}(x, \boldsymbol{z}) := \bar{a}_{\boldsymbol{y}}(x) + \sum_{j\geq 1} z_j \psi_{\boldsymbol{y},j}(x), \quad \text{a.e. } x \in D, \boldsymbol{z} \in \widetilde{U}.$$

We recall that for every $\boldsymbol{y} \in U$

$$\left\| \frac{\sum_{j\geq 1} |\psi_{\boldsymbol{y},j}|}{\bar{a}_{\boldsymbol{y}}} \right\|_{L^\infty(D)} \leq \frac{\kappa}{\eta} < 1, \tag{14}$$

which implies that the problem for arbitrary $\boldsymbol{y} \in U$ and $\boldsymbol{z} \in \widetilde{U}$ to find $\widetilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) \in V$ such that

$$\int_D \widetilde{a}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) \nabla \widetilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) \cdot \nabla v \mathrm{d}x = \langle f, v \rangle_{V^*,V}, \quad \forall v \in V,$$

is well-posed, cp. [11, Section 4]. Then, the affine mapping $T_{\boldsymbol{y}} : \widetilde{U} \to T_{\boldsymbol{y}}(\widetilde{U}) \subset \mathbb{R}^{\mathbb{N}}$, which is given by

$$(T_{\boldsymbol{y}}(\boldsymbol{z}))_j := y_j + \frac{\eta^{-1} - 2|y_j|}{2\widehat{b}_j} z_j, \quad j \geq 1, \boldsymbol{z} \in \widetilde{U}, \tag{15}$$

yields by construction, cp. [11, Section 4], a connection of $u(\cdot, \boldsymbol{y})$ and $\widetilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$, i.e.,

$$\widetilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) = u(\cdot, T_{\boldsymbol{y}}(\boldsymbol{z})) \quad \text{in } V.$$

Finally, by the chain rule for every $\boldsymbol{\tau} \in \mathscr{F} := \{\boldsymbol{\tau}' \in \mathbb{N}_0^{\mathbb{N}} : |\boldsymbol{\tau}'| < \infty\}$ holds

$$\partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} \widetilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{0}} = \left(\prod_{j\geq 1} \left(\frac{\eta^{-1} - 2|y_j|}{2\widehat{b}_j} \right)^{\tau_j} \right) \partial_{\boldsymbol{y}}^{\boldsymbol{\tau}} u(\cdot, \boldsymbol{y}). \tag{16}$$

A transformation of this type has been introduced in [3]. The dilated coordinate is analogously applied to dimensionally truncated solutions and the FE approximations, which are denoted by $\widetilde{u}_{\boldsymbol{y}}^s(\cdot, \boldsymbol{z})$, $\widetilde{u}_{\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot, \boldsymbol{z})$, and $\widetilde{u}_{\boldsymbol{y}}^{s,\mathscr{T}_\ell}(\cdot, \boldsymbol{z})$. As observed in

[11, Theorems 6.1 and 6.3], this sequence $(\widehat{b}_j)_{j\geq 1}$ will be the input for the product weights of the considered QMC rules and its summability properties will be a sufficient condition to achieve a certain dimension-independent convergence rate of either type of QMC rule. Note that the parametric regularity results of [11] also hold for homogeneous mixed boundary conditions, since the proof of [11, Lemma 4.1] relied on the variational formulation, which is the same and $v \mapsto (\int_D |\nabla v|^2 \mathrm{d}v)^{1/2}$ is also a norm on V.

4.1 Dimensionally Truncated Differences

Let $s \in \mathbb{N}$ be the truncation dimension of the series expansion of $a(\cdot, \boldsymbol{y})$ and also of $\widetilde{a}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$. The difference of solutions with respect to the full respectively to the truncated expansion of the parametric coefficient satisfies

$$\int_D \widetilde{a}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})\nabla(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z})) \cdot \nabla v \mathrm{d}x = -\int_D \sum_{j>s} z_j \psi_{\boldsymbol{y},j} \nabla \widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) \cdot \nabla v \mathrm{d}x, \quad \forall v \in V.$$

In this section we split the sequence $(b_j)_{j\geq 1}$ into two sequences by $b_j = b_j^{1-\theta} b_j^{\theta}$, $j \in \mathbb{N}$, $\theta \in [0,1]$, and consider the dilated coordinate in (13) and (15) with respect to the sequence $(b_j^{1-\theta})_{j\geq 1}$, i.e., here $(\widehat{b}_j)_{j\geq 1} = (b_j^{1-\theta})_{j\geq 1}$, which satisfies (**A1**) by the condition $b_j \in (0,1]$, $j \in \mathbb{N}$. By the assumption in (**A1**) and (14), for every $\boldsymbol{y} \in U$,

$$\left\| \frac{\sum_{j\geq 1} |\psi_{\boldsymbol{y},j}|/b_j^{\theta}}{\bar{a}_{\boldsymbol{y}}} \right\|_{L^\infty(D)} \leq \frac{\kappa}{\eta} < 1. \tag{17}$$

Theorem 1 *Let the assumption in* (**A1**) *be satisfied. There exists a constant $C > 0$ such that for every $\boldsymbol{y} \in U$ and for every $s \in \mathbb{N}$ and every $\theta \in [0,1]$*

$$\sum_{\boldsymbol{\tau}\in\mathscr{F}} \frac{1}{(\boldsymbol{\tau}!)^2} \left\| \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} \left(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) \right) \Big|_{\boldsymbol{z}=\boldsymbol{0}} \right\|_V^2 \leq C \|f\|_{V^*}^2 \sup_{j>s} \left\{ b_j^{2\theta} \right\}.$$

Proof As in the proof of [11, Lemma 4.1], we will consider the Taylor coefficients

$$t_{\boldsymbol{y},\boldsymbol{\tau}} := \frac{1}{\boldsymbol{\tau}!} \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} \widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) \Big|_{\boldsymbol{z}=\boldsymbol{0}} \quad \text{and} \quad t^s_{\boldsymbol{y},\boldsymbol{\tau}} := \frac{1}{\boldsymbol{\tau}!} \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} \widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) \Big|_{\boldsymbol{z}=\boldsymbol{0}}, \quad \forall \boldsymbol{\tau} \in \mathscr{F}. \tag{18}$$

We introduce a parametric energy norm $\|\cdot\|_{\bar{a}_{\boldsymbol{y}}}$ for every $\boldsymbol{y} \in U$ by

$$\|v\|_{\bar{a}_{\boldsymbol{y}}}^2 := \int_D \bar{a}_{\boldsymbol{y}} |\nabla v|^2 \mathrm{d}x, \quad \forall v \in V.$$

Evidently, $t^s_{\boldsymbol{y},\boldsymbol{\tau}} = 0$ in case that $\tau_j > 0$ for some $j > s$. For every $\boldsymbol{\tau} \in \mathscr{F}$,

$$\begin{aligned}\int_D \bar{a}_{\boldsymbol{y}} \nabla(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}) \cdot \nabla v \mathrm{d}x = &- \sum_{j(\boldsymbol{\tau})} \int_D \psi_{\boldsymbol{y},j} \nabla(t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} - t^s_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{e}_j}) \cdot \nabla v \mathrm{d}x \\ &- \sum_{j(\boldsymbol{\tau}),j>s} \int_D \psi_{\boldsymbol{y},j} \nabla t^s_{\boldsymbol{y},\boldsymbol{\tau}-e_j} \cdot \nabla v \mathrm{d}x, \qquad \forall v \in V,\end{aligned}$$

where we used the notation $j(\boldsymbol{\tau}) := \{j \in \mathbb{N} : \tau_j > 0\}$. Testing with $v = t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}$, we find for $\mathbf{0} \neq \boldsymbol{\tau} \in \mathscr{F}$,

$$\begin{aligned}\|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} \le &\int_D \sum_{j(\boldsymbol{\tau})} |\psi_{\boldsymbol{y},j}||\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} - t^s_{\boldsymbol{y},\boldsymbol{\tau}-e_j})||\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}})|\mathrm{d}x \\ &+ \int_D \sum_{j(\boldsymbol{\tau}),j>s} |\psi_{\boldsymbol{y},j}||\nabla t^s_{\boldsymbol{y},\boldsymbol{\tau}-e_j}||\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}})|\mathrm{d}x,\end{aligned}$$

where $e_j \in \mathscr{F}$ is such that $(e_j)_i = 1$ if $j = i$ and zero otherwise. We obtain with a twofold application of the Cauchy–Schwarz inequality using (**A1**) and (17)

$$\begin{aligned}\|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} \le &\left(\frac{\kappa}{\eta} \int_D \sum_{j(\boldsymbol{\tau})} |\psi_{\boldsymbol{y},j}||\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} - t^s_{\boldsymbol{y},\boldsymbol{\tau}-e_j})|^2 \mathrm{d}x\right)^{1/2} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|_{\bar{a}_{\boldsymbol{y}}} \\ &+ \left(\frac{\kappa}{\eta} \sup_{j>s} \{b_j^\theta\} \int_D \sum_{j(\boldsymbol{\tau}),j>s} |\psi_{\boldsymbol{y},j}||\nabla t^s_{\boldsymbol{y},\boldsymbol{\tau}-e_j}|^2 \mathrm{d}x\right)^{1/2} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|_{\bar{a}_{\boldsymbol{y}}}.\end{aligned}$$

Hence, by the Young inequality with $\varepsilon > 0$ and by (**A1**)

$$\begin{aligned}\sum_{k\ge1} \sum_{|\boldsymbol{\tau}|=k} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} \le &(1+\varepsilon)\frac{\kappa}{\eta} \sum_{k\ge1} \int_D \sum_{|\boldsymbol{\tau}|=k-1} \sum_{j\ge1} |\psi_{\boldsymbol{y},j}||\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}})|^2 \mathrm{d}x \\ &+ \left(1 + \frac{1}{\varepsilon}\right) \frac{\kappa}{\eta} \sup_{j>s} \{b_j^\theta\} \sum_{k\ge1} \int_D \sum_{|\boldsymbol{\tau}|=k-1} \sum_{j>s} |\psi_{\boldsymbol{y},j}||\nabla t^s_{\boldsymbol{y},\boldsymbol{\tau}}|^2 \mathrm{d}x \\ \le &(1+\varepsilon)\left(\frac{\kappa}{\eta}\right)^2 \sum_{k\ge0} \sum_{|\boldsymbol{\tau}|=k} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} \\ &+ \left(1 + \frac{1}{\varepsilon}\right)\left(\frac{\kappa}{\eta}\right)^2 \sup_{j>s} \{b_j^{2\theta}\} \sum_{k\ge0} \sum_{|\boldsymbol{\tau}|=k} \|t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}}.\end{aligned}$$

Since $\kappa < \eta$, we can choose ε such that $(1+\varepsilon)(\kappa/\eta)^2 < 1$ and conclude that

$$\sum_{\mathbf{0}\neq\boldsymbol{\tau}\in\mathscr{F}} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} \\ \leq \frac{1+\varepsilon}{1-(1+\varepsilon)(\kappa/\eta)^2}\left(\|t_{\boldsymbol{y},\mathbf{0}} - t^s_{\boldsymbol{y},\mathbf{0}}\|^2_{\bar{a}_{\boldsymbol{y}}} + \frac{1}{\varepsilon}\sup_{j>s}\{b_j^{2\theta}\}\sum_{\boldsymbol{\tau}\in\mathscr{F}}\|t^s_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}}\right),$$

which implies the assertion with [11, Lemma 4.1 and Proposition 5.1] using that $b_j \in (0,1]$, $j \in \mathbb{N}$. Note that [11, Lemma 4.1] gives an upper bound, since $\partial_z^{\boldsymbol{\tau}}\widetilde{u}^s_{\boldsymbol{y}}(\cdot,z) = \partial_z^{\boldsymbol{\tau}}\widetilde{u}_{\boldsymbol{y}_{\{1:s\}}}(\cdot,z)$ if $\tau_j = 0$ for every $j > s$ and $\partial_z^{\boldsymbol{\tau}}\widetilde{u}^s_{\boldsymbol{y}}(\cdot,z) = 0$ otherwise. □

Remark 2 The estimate in Theorem 1 also holds when the differences $\partial_z^{\boldsymbol{\tau}}(\widetilde{u}^{\mathscr{T}_\ell}_{\boldsymbol{y}}(\cdot,z) - \widetilde{u}^{s,\mathscr{T}_\ell}_{\boldsymbol{y}}(\cdot,z))|_{z=\mathbf{0}}$, $\boldsymbol{\tau} \in \mathscr{F}$, $\ell \geq 0$, are considered and the constant is independent of $\{\mathscr{T}_\ell\}_{\ell\geq 0}$. Since only the variational formulation was used in the proof, the corresponding variational formulation with trial and test space V_ℓ can be used instead.

4.2 FE Differences

We assume now that the sequence $(\bar{b}_j)_{j\geq 1}$ satisfies the assumptions in (**A1**) and in (**A2**). We consider the dilated coordinate in (13) and (15) with respect to this sequence $(\bar{b}_j)_{j\geq 1}$, i.e., here $(\widehat{b}_j)_{j\geq 1} = (\bar{b}_j)_{j\geq 1}$.

Proposition 2 *Let the assumption in* (**A1**) *and* (**A2**) *be satisfied for* $(\bar{b}_j)_{j\geq 1}$. *Then, there exists a constant* $C > 0$ *(independent of* f*) such that for every* $\boldsymbol{y} \in U$

$$\sum_{\boldsymbol{\tau}\in\mathscr{F}} \frac{1}{(\boldsymbol{\tau}!)^2}\left\|\Delta\partial_z^{\boldsymbol{\tau}}\widetilde{u}_{\boldsymbol{y}}(\cdot,z)\Big|_{z=\mathbf{0}}\right\|^2_{L^2_\beta(D)} \leq C\|f\|^2_{L^2_\beta(D)}.$$

Proof Recall that the Taylor coefficients $\{t_{\boldsymbol{y},\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathscr{F}\}$ have been defined in (18). We also recall that for any $\mathbf{0} \neq \boldsymbol{\tau} \in \mathscr{F}$,

$$\int_D \bar{a}_{\boldsymbol{y}}\nabla t_{\boldsymbol{y},\boldsymbol{\tau}}\cdot\nabla v\,\mathrm{d}x = -\sum_{j(\boldsymbol{\tau})}\int_D \psi_{\boldsymbol{y},j}\nabla t_{\boldsymbol{y},\boldsymbol{\tau}-e_j}\cdot\nabla v\,\mathrm{d}x, \quad \forall v \in V.$$

Similarly as in Sect. 2, by the divergence theorem for every $v \in C_0^\infty(D)$

$$-\int_D \bar{a}_{\boldsymbol{y}}\Delta t_{\boldsymbol{y},\boldsymbol{\tau}} v\,\mathrm{d}x = \int_D\left(\nabla\bar{a}_{\boldsymbol{y}}\cdot\nabla t_{\boldsymbol{y},\boldsymbol{\tau}} + \sum_{j(\boldsymbol{\tau})}\left(\nabla\psi_{\boldsymbol{y},j}\cdot\nabla t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} + \psi_{\boldsymbol{y},j}\Delta t_{\boldsymbol{y},\boldsymbol{\tau}-e_j}\right)\right)v\,\mathrm{d}x.$$

Since $\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\ \Phi_{\boldsymbol{\beta}} \in L^2(D)$, cp. (8), we may use $-\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\ \Phi_{\boldsymbol{\beta}}^2$ as a test function and obtain with the Young inequality for any $\varepsilon > 0$

$$
\begin{aligned}
\int_D \bar{a}_{\boldsymbol{y}} |\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}|^2 \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x &= -\int_D \left(\nabla \bar{a}_{\boldsymbol{y}} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}} + \sum_{j(\boldsymbol{\tau})} \nabla \psi_{\boldsymbol{y},j} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} \right) \Delta t_{\boldsymbol{y},\boldsymbol{\tau}} \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x \\
&\quad - \int_D \sum_{j(\boldsymbol{\tau})} \psi_{\boldsymbol{y},j} \Delta t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} \Delta t_{\boldsymbol{y},\boldsymbol{\tau}} \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x \\
&\le \varepsilon \int_D \bar{a}_{\boldsymbol{y}} |\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}|^2 \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x \\
&\quad + \frac{1}{4\varepsilon} \int_D \frac{\Phi_{\boldsymbol{\beta}}^2}{\bar{a}_{\boldsymbol{y}}} \left(\nabla \bar{a}_{\boldsymbol{y}} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}} + \sum_{j(\boldsymbol{\tau})} \nabla \psi_{\boldsymbol{y},j} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} \right)^2 \mathrm{d}x \\
&\quad + \frac{1}{2} \int_D \sum_{j(\boldsymbol{\tau})} |\psi_{\boldsymbol{y},j}| (|\Delta t_{\boldsymbol{y},\boldsymbol{\tau}-e_j}|^2 + |\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}|^2) \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x.
\end{aligned}
$$

For $k \ge 1$, by a twofold application of the Cauchy–Schwarz inequality (applied to the sum) and (**A2**) and $\eta \ge 1/2$

$$
\begin{aligned}
&\sum_{|\boldsymbol{\tau}|=k} \int_D \frac{\Phi_{\boldsymbol{\beta}}^2}{\bar{a}_{\boldsymbol{y}}} \left(\nabla \bar{a}_{\boldsymbol{y}} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}} + \sum_{j(\boldsymbol{\tau})} \nabla \psi_{\boldsymbol{y},j} \cdot \nabla t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} \right)^2 \mathrm{d}x \\
&\qquad \le 2K \int_D \frac{\Phi_{\boldsymbol{\beta}}}{\bar{a}_{\boldsymbol{y}}} \left(|\nabla \bar{a}_{\boldsymbol{y}}| \sum_{|\boldsymbol{\tau}|=k} |\nabla t_{\boldsymbol{y},\boldsymbol{\tau}}|^2 + \sum_{|\boldsymbol{\tau}|=k-1} \sum_{j\ge 1} |\nabla \psi_{\boldsymbol{y},j}| |\nabla t_{\boldsymbol{y},\boldsymbol{\tau}}|^2 \right) \mathrm{d}x \\
&\qquad \le \frac{4K^2}{(\bar{a}_{\boldsymbol{y},\min})^2} \left(\sum_{|\boldsymbol{\tau}|=k} \|t_{\boldsymbol{y},\boldsymbol{\tau}}\|_{\bar{a}_{\boldsymbol{y}}}^2 + \sum_{|\boldsymbol{\tau}|=k-1} \|t_{\boldsymbol{y},\boldsymbol{\tau}}\|_{\bar{a}_{\boldsymbol{y}}}^2 \right).
\end{aligned}
$$

Note that also by (**A1**) for every $k \ge 1$,

$$
\begin{aligned}
&\sum_{|\boldsymbol{\tau}|=k} \int_D \sum_{j(\boldsymbol{\tau})} |\psi_{\boldsymbol{y},j}| (|\Delta t_{\boldsymbol{y},\boldsymbol{\tau}-e_j}|^2 + |\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}|^2) \Phi_{\boldsymbol{\beta}}^2 \mathrm{d}x \\
&\qquad \le \frac{\kappa}{\eta} \left(\sum_{|\boldsymbol{\tau}|=k-1} \|\sqrt{\bar{a}_{\boldsymbol{y}}} \Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\|_{L_{\boldsymbol{\beta}}^2(D)}^2 + \sum_{|\boldsymbol{\tau}|=k} \|\sqrt{\bar{a}_{\boldsymbol{y}}} \Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\|_{L_{\boldsymbol{\beta}}^2(D)}^2 \right).
\end{aligned}
$$

We now choose $\varepsilon > 0$ such that $\varepsilon < 1/2(1-\kappa/\eta)$, which implies $\varepsilon + \kappa/(2\eta) < 1/2$. Then, we sum over $k \geq 1$ to obtain

$$\sum_{k\geq 1}\sum_{|\boldsymbol{\tau}|=k} \|\sqrt{\bar{a}_{\boldsymbol{y}}}\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{L^2_{\boldsymbol{\beta}}(D)} \leq C\sum_{\boldsymbol{\tau}\in\mathscr{F}} \|t_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} + C_\varepsilon \sum_{\boldsymbol{\tau}\in\mathscr{F}} \|\sqrt{\bar{a}_{\boldsymbol{y}}}\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{L^2_{\boldsymbol{\beta}}(D)},$$

where $C_\varepsilon = 1/2(1-\kappa/(2\eta)-\varepsilon)^{-1} < 1$ and $C = 2K^2/(\varepsilon(\bar{a}_{\boldsymbol{y},\min})^2)(1-\kappa/(2\eta)-\varepsilon)^{-1}$. It follows

$$\sum_{\boldsymbol{0}\neq\boldsymbol{\tau}\in\mathscr{F}} \|\sqrt{\bar{a}_{\boldsymbol{y}}}\Delta t_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{L^2_{\boldsymbol{\beta}}(D)} \leq \frac{1}{1-C_\varepsilon}\left(C\sum_{\boldsymbol{\tau}\in\mathscr{F}} \|t_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}} + \|\sqrt{\bar{a}_{\boldsymbol{y}}}\Delta t_{\boldsymbol{y},\boldsymbol{0}}\|^2_{L^2_{\boldsymbol{\beta}}(D)}\right),$$

which implies the assertion with [11, Lemma 4.1] and (8). □

Remark 3 For every truncation dimension $s \in \mathbb{N}$, the estimate in Proposition 2 also holds when $\Delta\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{0}}$, $\boldsymbol{\tau}\in\mathscr{F}$, are considered and the constant is independent of s. This follows from the observation that $\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{0}} = \partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}_{\boldsymbol{y}_{\{1:s\}}}(\cdot,\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{0}}$ if $\tau_j = 0$ for every $j > s$ and $\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}^s_{\boldsymbol{y}}(\cdot,\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{0}} = 0$ otherwise. Then, the sum of the estimate in Proposition 2 only consists of more terms and is an upper bound.

Proposition 3 *Let the assumptions in* (**A1**) *and* (**A2**) *be satisfied and let* $\boldsymbol{\beta}$ *satisfy* $\beta_i > 1-\pi/\omega_i$ *and if the boundary conditions change at* c_i *also* $\beta_i > 1-\pi/(2\omega_i)$, $i = 1,\ldots,J$. *Then, there exists a constant* $C > 0$ *such that for every* $\boldsymbol{y}\in U$ *and for every integer* $\ell \geq 0$

$$\sum_{\boldsymbol{\tau}\in\mathscr{F}} \frac{1}{(\boldsymbol{\tau}!)^2}\left\|\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\left(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}^{\mathscr{T}_\ell}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})\right)\Big|_{\boldsymbol{z}=\boldsymbol{0}}\right\|^2_V \leq CM_\ell^{-2/d}\|f\|^2_{L^2_{\boldsymbol{\beta}}(D)}.$$

Proof We argue similarly as in the proof of [11, Lemma 4.1] and consider the Taylor coefficients for fixed $\boldsymbol{y}\in U$

$$t_{\boldsymbol{y},\boldsymbol{\tau}} := \frac{1}{\boldsymbol{\tau}!}\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{0}} \quad\text{and}\quad t^\ell_{\boldsymbol{y},\boldsymbol{\tau}} := \frac{1}{\boldsymbol{\tau}!}\partial^{\boldsymbol{\tau}}_{\boldsymbol{z}}\widetilde{u}^{\mathscr{T}_\ell}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{0}}, \quad \boldsymbol{\tau}\in\mathscr{F}.$$

We observe that

$$\int_D \bar{a}_{\boldsymbol{y}}\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^\ell_{\boldsymbol{y},\boldsymbol{\tau}})\cdot\nabla v\,\mathrm{d}x = -\sum_{j(\boldsymbol{\tau})}\psi_{\boldsymbol{y},j}\nabla(t_{\boldsymbol{y},\boldsymbol{\tau}-e_j} - t^\ell_{\boldsymbol{y},\boldsymbol{\tau}-e_j})\cdot\nabla v\,\mathrm{d}x, \quad \forall v\in V_\ell.$$

For every $\boldsymbol{y}\in U$ and for every $\ell\in\mathbb{N}_0$, let $\mathscr{P}_{\boldsymbol{y},\ell} : V \to V_\ell$ denote the "dilated Galerkin projection". For every $w\in V$, $\mathscr{P}_{\boldsymbol{y},\ell}w$, it is defined by

$$\int_D \bar{a}_{\boldsymbol{y}}\nabla(w - \mathscr{P}_{\boldsymbol{y},\ell}w)\cdot\nabla v\,\mathrm{d}x = 0, \quad \forall v\in V_\ell. \tag{19}$$

By the definition of $\mathscr{P}_{y,\ell}$ in (19) and by testing with $v = \mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau}) \in V_\ell$,

$$\begin{aligned}
&\sum_{|\tau|=k} \int_D \bar{a}_y |\nabla \mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau})|^2 \mathrm{d}x \\
&\quad\le \int_D \sum_{|\tau|=k} \sum_{j(\tau)} |\psi_{y,j}| \frac{1}{2}(|\nabla(t_{y,\tau-e_j} - t^{\ell}_{y,\tau-e_j})|^2 + |\nabla \mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau})|^2)\mathrm{d}x \\
&\quad\le \frac{1}{2} \int_D \sum_{|\tau|=k-1} \sum_{j\ge 1} |\psi_{y,j}||\nabla(t_{y,\tau} - t^{\ell}_{y,\tau})|^2 \mathrm{d}x \\
&\qquad + \frac{1}{2} \int_D \sum_{|\tau|=k} \sum_{j\ge 1} |\psi_{y,j}||\nabla \mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau})|^2 \mathrm{d}x,
\end{aligned}$$

which implies with (**A1**)

$$\begin{aligned}
\sum_{|\tau|=k} \|\mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau})\|^2_{\bar{a}_y} &\le \frac{1}{2-\kappa/\eta} \int_D \sum_{|\tau|=k} \sum_{j(\tau)} |\psi_{y,j}||\nabla(t_{y,\tau-e_j} - t^{\ell}_{y,\tau-e_j})|^2 \mathrm{d}x \\
&\le \frac{1}{2-\kappa/\eta} \frac{\kappa}{\eta} \sum_{|\tau|=k-1} \|t_{y,\tau-e_j} - t^{\ell}_{y,\tau-e_j}\|^2_{\bar{a}_y}. \qquad (20)
\end{aligned}$$

Note that by the triangle inequality

$$\|t_{y,\tau} - t^{\ell}_{y,\tau}\|_{\bar{a}_y} \le \|\mathscr{P}_{y,\ell}(t_{y,\tau} - t^{\ell}_{y,\tau})\|_{\bar{a}_y} + \|(\mathscr{I} - \mathscr{P}_{y,\ell})t_{y,\tau}\|_{\bar{a}_y},$$

where $\mathscr{I} : V \to V$ denotes the identity. With the Young inequality and the previous two inequalities we obtain for any $\varepsilon > 0$

$$\begin{aligned}
&\sum_{|\tau|=k} \|t_{y,\tau} - t^{\ell}_{y,\tau}\|^2_{\bar{a}_y} \\
&\quad\le \frac{(1+\varepsilon)\kappa}{2\eta - \kappa} \sum_{|\tau|=k-1} \|t_{y,\tau} - t^{\ell}_{y,\tau}\|^2_{\bar{a}_y} + \left(1 + \frac{1}{\varepsilon}\right) \sum_{|\tau|=k} \|(\mathscr{I} - \mathscr{P}_{y,\ell})t_{y,\tau}\|^2_{\bar{a}_y}.
\end{aligned}$$

Since $\kappa < \eta < 1$, $2\eta - \kappa > \eta$ and so we choose $\varepsilon > 0$ such that $(1+\varepsilon)\kappa/\eta < 1$ and conclude by subtracting the first sum in the previous inequality that

$$\begin{aligned}
&\sum_{k\ge 1} \sum_{|\tau|=k} \|t_{y,\tau} - t^{\ell}_{y,\tau}\|^2_{\bar{a}_y} \\
&\quad\le \frac{1+\varepsilon}{1-(1+\varepsilon)\kappa/\eta} \left(\|t_{y,\mathbf{0}} - t^{\ell}_{y,\mathbf{0}}\|^2_{\bar{a}_y} + \frac{1}{\varepsilon} \sum_{k\ge 1} \sum_{|\tau|=k} \|(\mathscr{I} - \mathscr{P}_{y,\ell})t_{y,\tau}\|^2_{\bar{a}_y} \right),
\end{aligned}$$

which implies the assertion with (7), (6), and Proposition 2. □

Remark 4 The estimate in Proposition 3 holds if $f \in (V^*, L^2_{\boldsymbol{\beta}}(D))_{t,\infty}$ with the error being controlled by $M_\ell^{-2t/d}$, $t \in (0,1)$. This can be seen by interpolating the error bounds in the last step of the proof of Proposition 3 with the real method of interpolation, where (7), (6), and Proposition 2 were used (see also Remark 1).

For any $G \in V^*$, we introduce $u_G(\cdot,\boldsymbol{y})$ and $u_G^{\mathscr{T}_\ell}(\cdot,\boldsymbol{y})$, $\ell \in \mathbb{N}_0$, as the parametric solution to the dual problem of (4) and the parametric FE solution to the dual problem of (9), respectively, with right hand side G. Consideration of the dilated coefficient resulting from (13) gives $\widetilde{u}_{G,\boldsymbol{y}}(\cdot,\boldsymbol{z})$ and $\widetilde{u}_{G,\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z})$, $\ell \in \mathbb{N}_0$. By an Aubin–Nitsche argument, for every $\boldsymbol{y} \in U$ and every $\boldsymbol{z} \in \widetilde{U}$,

$$G(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}_{\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z})) = \int_D \widetilde{a}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})\nabla(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}_{\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z})) \cdot \nabla(\widetilde{u}_{G,\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}_{G,\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z}))\mathrm{d}x. \tag{21}$$

Theorem 2 *Let the assumptions in* (**A1**) *and* (**A2**) *be satisfied. Then, there exists a constant* $C > 0$ *such that for every* $G(\cdot) \in L^2_{\boldsymbol{\beta}}(D)$ *and for every integer* $\ell \geq 0$

$$\sum_{\boldsymbol{\tau}\in\mathscr{F}} \frac{1}{(\boldsymbol{\tau}+\mathbf{1})!\boldsymbol{\tau}!} \left| \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} G\left(\widetilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \widetilde{u}_{\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z})\right)\Big|_{\boldsymbol{z}=\mathbf{0}} \right|^2$$

$$\leq C M_\ell^{-4/d} \|f\|^2_{L^2_{\boldsymbol{\beta}}(D)} \|G(\cdot)\|^2_{L^2_{\boldsymbol{\beta}}(D)}.$$

Proof The Taylor coefficients of $\widetilde{u}_{G,\boldsymbol{y}}(\cdot,\boldsymbol{z})$ and $\widetilde{u}_{G,\boldsymbol{y}}^{\mathscr{T}_\ell}(\cdot,\boldsymbol{z})$ will be denoted by $\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}}$ and $\widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}}$, $\boldsymbol{\tau} \in \mathscr{F}$, respectively (see also (18)). By differentiating (21), for every $\mathbf{0} \neq \boldsymbol{\tau} \in \mathscr{F}$,

$$G(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}}) = \sum_{\boldsymbol{\nu}\leq\boldsymbol{\tau}} \int_D \left[\sum_{j(\boldsymbol{\nu})} \psi_{\boldsymbol{y},j} \nabla(t_{\boldsymbol{y},\boldsymbol{\nu}-e_j} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}-e_j}) \right] \cdot \nabla(\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}} - \widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}})\mathrm{d}x.$$

Squaring the previous equality and applying the Cauchy–Schwarz inequality yields

$$|G(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}})|^2 \leq \prod_{j(\boldsymbol{\tau})}(\tau_j + 1) \sum_{\boldsymbol{\nu}\leq\boldsymbol{\tau}} \left\| \sqrt{1/\bar{a}_{\boldsymbol{y}}}[\ldots] \right\|^2_{L^2(D)} \|\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}} - \widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}},$$

where we used that $\sum_{\boldsymbol{\nu}\le\boldsymbol{\tau}} = \prod_{j(\boldsymbol{\tau})}(\tau_j+1)$. The hidden term is given by $[\ldots] = \sum_{j(\boldsymbol{\nu})} \psi_{\boldsymbol{y},j}\nabla(t_{\boldsymbol{y},\boldsymbol{\nu}-e_j} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}-e_j})$. By changing the order of summation

$$\begin{aligned}
&\sum_{\boldsymbol{\tau}\in\mathscr{F}}\prod_{j(\boldsymbol{\tau})}(\tau_j+1)^{-1}|G(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}})|^2 \\
&\qquad \le \sum_{\boldsymbol{\nu}\in\mathscr{F}} \left\| \sqrt{1/\bar{a}_{\boldsymbol{y}}}[\ldots] \right\|^2_{L^2(D)} \sum_{\boldsymbol{\tau}\in\mathscr{F},\boldsymbol{\tau}\ge\boldsymbol{\nu}} \|\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}} - \widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}-\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} \qquad (22)\\
&\qquad = \sum_{\boldsymbol{\nu}\in\mathscr{F}} \left\| \sqrt{1/\bar{a}_{\boldsymbol{y}}}[\ldots] \right\|^2_{L^2(D)} \sum_{\boldsymbol{\tau}\in\mathscr{F}} \|\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}} - \widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}}.
\end{aligned}$$

By the Cauchy–Schwarz inequality we obtain with (**A1**)

$$\left\| \sqrt{1/\bar{a}_{\boldsymbol{y}}}[\ldots] \right\|^2_{L^2(D)} \le \frac{\kappa}{\eta}\int_D \sum_{j(\boldsymbol{\nu})} |\psi_{\boldsymbol{y},j}||\nabla(t_{\boldsymbol{y},\boldsymbol{\nu}-e_j} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}-e_j})|^2 \mathrm{d}x.$$

By another application of the Cauchy–Schwarz inequality and (**A1**)

$$\sum_{k\ge 1}\sum_{|\boldsymbol{\nu}|=k} \left\| \sqrt{1/\bar{a}_{\boldsymbol{y}}}[\ldots] \right\|^2_{L^2(D)} \le \left(\frac{\kappa}{\eta}\right)^2 \sum_{k\ge 1}\sum_{|\boldsymbol{\nu}|=k-1} \|t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}},$$

which implies with (22)

$$\sum_{\boldsymbol{\tau}\in\mathscr{F}}\prod_{j(\boldsymbol{\tau})}(\tau_j+1)^{-1}|G(t_{\boldsymbol{y},\boldsymbol{\tau}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}})|^2 \le \left(\sum_{\boldsymbol{\tau}\in\mathscr{F}} \|t_{\boldsymbol{y},\boldsymbol{\tau}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}}\right)\left(\sum_{\boldsymbol{\tau}\in\mathscr{F}} \|\widehat{t}_{\boldsymbol{y},\boldsymbol{\tau}} - \widehat{t}^{\ell}_{\boldsymbol{y},\boldsymbol{\tau}}\|^2_{\bar{a}_{\boldsymbol{y}}}\right).$$

The assertion now follows with Proposition 3. □

Remark 5 The estimate in Theorem 2 also holds if $f \in (V^*, L^2_{\boldsymbol{\beta}})_{t,\infty}$ and $G(\cdot) \in (V^*, L^2_{\boldsymbol{\beta}})_{t',\infty}$, $t, t' \in (0,1)$, with error bound $\mathscr{O}(M_\ell^{-2(t+t')/d})$, which follows by Remark 4.

Remark 6 For every truncation dimension $s \in \mathbb{N}$, the estimates in Proposition 3 and Theorem 2 also hold when the differences $\partial_z^{\boldsymbol{\tau}}(\widetilde{u}^s_{\boldsymbol{y}}(\cdot,z) - \widetilde{u}^{s,\mathscr{T}_\ell}_{\boldsymbol{y}}(\cdot,z))|_{z=\mathbf{0}}$, $\boldsymbol{\tau} \in \mathscr{F}$, are considered and the constant is independent of s. This follows by the same argument which is used to verify Remark 3.

5 Convergence of Multilevel QMC

The parametric regularity estimates from Sect. 4 will result in explicit error estimates of multilevel QMC. Let the sequence $(\mathfrak{b}_j)_{j\geq 1}$ be a generic input for the QMC weights. For interlaced polynomial lattice rules with interlacing factor $\alpha \geq 2$ we will consider the product weights $\boldsymbol{\gamma}^{\mathrm{IP}} = (\gamma_{\mathfrak{u}}^{\mathrm{IP}})_{\mathfrak{u}\subset\mathbb{N}}$ given by $\gamma_{\emptyset}^{\mathrm{IP}} := 1$ and

$$\gamma_{\mathfrak{u}}^{\mathrm{IP}} := \prod_{j\in\mathfrak{u}} \left(\sum_{\nu=1}^{\alpha} \left(\frac{2\mathfrak{b}_j}{1-\eta} \right)^{\nu} \sqrt{2^{\delta(\nu,\alpha)}\nu!} \right), \quad \mathfrak{u} \subset \mathbb{N}, |\mathfrak{u}| < \infty, \tag{23}$$

and for randomly shifted lattice rules the product weights $\boldsymbol{\gamma}^{\mathrm{RS}} = (\gamma_{\mathfrak{u}}^{\mathrm{RS}})_{\mathfrak{u}\subset\mathbb{N}}$ given by $\gamma_{\emptyset}^{\mathrm{RS}} := 1$ and

$$\gamma_{\mathfrak{u}}^{\mathrm{RS}} := \prod_{j\in\mathfrak{u}} \left(\frac{2\mathfrak{b}_j}{1-\eta} \right)^2, \quad \mathfrak{u} \subset \mathbb{N}, |\mathfrak{u}| < \infty. \tag{24}$$

We will apply one common QMC rule on discretization levels $\ell = 1, \ldots, L$ and allow a different rule on discretization level $\ell = 0$. The parametric regularity estimates that were derived in Sect. 4 are based on a dilated coordinate, cp. (13) and (15), with respect to sequences $(b_j^{1-\theta})_{j\geq 1}$ for the truncation error and $(\bar{b}_j)_{j\geq 1}$ for the FE error. These sequences will be the input for the product weights. Their summability in terms of membership in $\ell^{\bar{p}}(\mathbb{N})$, $\bar{p} \in (0, 2]$, will result in explicit bounds of the combined discretization and quadrature errors between discretization levels. On discretization levels $\ell = 1, \ldots, L$, we use $(b_j^{1-\theta} \vee 2\bar{b}_j)_{j\geq 1} := (\max\{b_j^{1-\theta}, 2\bar{b}_j\})_{j\geq 1}$ as input for the product weights in (23) and (24), i.e., here $(\mathfrak{b}_j)_{j\geq 1} = (b_j^{1-\theta} \vee 2\bar{b}_j)_{j\geq 1}$. On the lowest discretization level $\ell = 0$ we use $(b_j)_{j\geq 1}$ as an input for (23) and (24), which has potentially stronger summability properties.

Theorem 3 *Let the assumption in* (**A1**) *be satisfied by* $(b_j)_{j\geq 1}$ *and by* $(\bar{b}_j)_{j\geq 1}$. *Let the assumption in* (**A2**) *be satisfied by* $(\bar{b}_j)_{j\geq 1}$. *Let* $(b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$ *for some* $p \in (0, 2]$ *and assume that* $(b_j^{1-\theta} \vee \bar{b}_j)_{j\geq 1} \in \ell^{\bar{p}}(\mathbb{N})$ *for some* $\bar{p} \in [p, 2]$ *and any* $\theta \in [0, 1)$ *admitting this summability. For* $p \in (0, 1]$ *and* $\bar{p} \in [p, 1]$, $Q_L^{\mathrm{IP}}(\cdot)$, $L \in \mathbb{N}$, *satisfies with product weights* (23) *and order* $\alpha = \lfloor 1/p + 1 \rfloor$ *on discretization level* $\ell = 0$, *and of order* $\bar{\alpha} = \lfloor 1/\bar{p} + 1 \rfloor$ *on discretization levels* $\ell = 1, \ldots, L$, *the error estimate*

$$|\mathbb{E}(G(u)) - Q_L^{\mathrm{IP}}(G(u^L))| \leq C \Bigg(\sup_{j>s_L}\{b_j^2\} + M_L^{-2/d} + N_0^{-1/p} + \sum_{\ell=1}^{L} N_\ell^{-1/\bar{p}} \left(\xi_{\ell,\ell-1} \sup_{j>s_{\ell-1}} \{b_j^{\theta}\} + M_{\ell-1}^{-2/d} \right) \Bigg),$$

where $\xi_{\ell,\ell-1} := 0$ *if* $s_\ell = s_{\ell-1}$ *and* $\xi_{\ell,\ell-1} := 1$ *otherwise. For* $p \in (1,2]$ *and* $\bar{p} \in [p,2]$*,* $Q_L^{\mathrm{RS}}(\cdot)$*,* $L \in \mathbb{N}$*, satisfies with product weights* (24) *the error estimate*

$$\begin{aligned}
&\sqrt{\mathbb{E}^{\boldsymbol{\Delta}}(|\mathbb{E}(G(u)) - Q_L^{\mathrm{RS}}(G(u^L))|^2)} \\
&\quad \le C\Bigg(\sup_{j>s_L}\{b_j^4\} + M_L^{-4/d} + (\varphi(N_0))^{-2/p} \\
&\qquad + \sum_{\ell=1}^{L}(\varphi(N_\ell))^{-2/\bar{p}}\left(\xi_{\ell,\ell-1}\sup_{j>s_{\ell-1}}\{b_j^{2\theta}\} + M_{\ell-1}^{-4/d}\right)\Bigg)^{1/2}.
\end{aligned}$$

The constant C *is in particular independent of* L*,* $(N_\ell)_{\ell=0,\dots,L}$*,* $(M_\ell)_{\ell\ge 0}$*,* $(s_\ell)_{\ell=0,\dots,L}$*.*

Proof By the error estimates in (12) and (11), we have to estimate the difference $G(u^\ell - u^{\ell-1}) = G(u^{s_\ell,\mathscr{T}_\ell} - u^{s_{\ell-1},\mathscr{T}_{\ell-1}})$ in the $\mathscr{W}_{s_\ell,\alpha,\boldsymbol{\gamma},\infty,\infty}$-norm and in the $\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}$-norm, $\ell = 1,\dots,L$. We decompose by the triangle inequality

$$\begin{aligned}
&\|G(u^{s_\ell,\mathscr{T}_\ell} - u^{s_{\ell-1},\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} \\
&\quad \le \|G(u^{s_\ell,\mathscr{T}_\ell} - u^{s_\ell,\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} + \|G(u^{s_\ell,\mathscr{T}_{\ell-1}} - u^{s_{\ell-1},\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}},
\end{aligned}$$

and

$$\|G(u^{s_\ell,\mathscr{T}_\ell} - u^{s_\ell,\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} \le \|G(u^{s_\ell} - u^{s_\ell,\mathscr{T}_\ell})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} + \|G(u^{s_\ell} - u^{s_\ell,\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}}.$$

The contributions from the dimension truncation and the FE error have been separated in the $\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}$-norm. For the dimension truncation error, we obtain by the Jensen inequality, the relation of higher order partial derivatives in terms of the dilated coordinate in (16), Theorem 1, and Remark 2

$$\begin{aligned}
&\|G(u^{s_\ell,\mathscr{T}_{\ell-1}} - u^{s_{\ell-1},\mathscr{T}_{\ell-1}})\|^2_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} \\
&\quad \le \|G(\cdot)\|^2_{V^*}\int_{[-\frac{1}{2},\frac{1}{2}]^s}\sum_{\mathrm{u}\subset\{1:s\}}(\gamma_{\mathrm{u}}^{\mathrm{RS}})^{-1}\|\partial_{\boldsymbol{y}}^{\mathrm{u}}(u^{s_\ell,\mathscr{T}_{\ell-1}}(\cdot,\boldsymbol{y}) - u^{s_{\ell-1},\mathscr{T}_{\ell-1}}(\cdot,\boldsymbol{y}))\|^2_V \mathrm{d}\boldsymbol{y} \\
&\quad \le C\|G(\cdot)\|^2_{V^*}\|f\|^2_{V^*}\sup_{\mathrm{u}\subset\{1:s\}}(\gamma_{\mathrm{u}}^{\mathrm{RS}})^{-1}\prod_{j\in\mathrm{u}}\left(\frac{2b_j^{1-\theta}}{1-\eta}\right)^2\sup_{j>s_{\ell-1}}\{b_j^{2\theta}\}.
\end{aligned}$$

Due to the choice of the weights, there exists a constant $C > 0$ independent of the sequences $(s_\ell)_{\ell=0,\dots,L}$ and $\{\mathscr{T}_\ell\}_{\ell\ge 0}$ such that

$$\|G(u^{s_\ell,\mathscr{T}_{\ell-1}} - u^{s_{\ell-1},\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} \le C\|G(\cdot)\|_{V^*}\|f\|_{V^*}\sup_{j>s_{\ell-1}}\{b_j^\theta\}.$$

Note that if $s_\ell = s_{\ell-1}$ this difference is zero. Similarly, we obtain with Theorem 2 and Remark 4 that there exists a constant $C > 0$ which is independent of $(s_\ell)_{\ell=0,\dots,L}$ and $(M_\ell)_{\ell\geq 0}$ such that for every $\ell \in \mathbb{N}$

$$\|G(u^{s_\ell} - u^{s_\ell,\mathscr{T}_{\ell-1}})\|_{\mathscr{W}_{s_\ell,\boldsymbol{\gamma}}} \leq C\|G(\cdot)\|_{L^2_{\boldsymbol{\beta}}(D)}\|f\|_{L^2_{\boldsymbol{\beta}}(D)} M_{\ell-1}^{-2/d}.$$

Here the constant factor 2 in the sequence $(b_j^{1-\theta} \vee 2\bar{b}_j)_{j\geq 1}$, which is an input for the weight sequence, is necessary to compensate the factor $\prod_{j(\boldsymbol{\tau})}(\tau_j + 1)^{-1}$ in the estimate of Theorem 2. The corresponding estimate on the level $\ell = 0$ of $\|G(u^{s_0,\mathscr{T}_0})\|_{\mathscr{W}_{s_0,\boldsymbol{\gamma}}}$ is due to [11, Corollary 4.3], which is also applicable in the case of a dimensionally truncated FE solution, cp. Remarks 2 and 4. The estimate for the randomly shifted lattice rules follows then with (10) and (11) with $\lambda = p/2$.

The proof for interlaced polynomial lattice rules follows along the same lines, where the estimate $\|F\|_{\mathscr{W}_{s_\ell,\alpha,\boldsymbol{\gamma},\infty,\infty}} \leq \|F\|_{\mathscr{W}_{s_\ell,\alpha,\boldsymbol{\gamma},2,2}}$, $F \in \mathscr{W}_{s_\ell,\alpha,\boldsymbol{\gamma},2,2}$, is used and [11, Corollary 4.5] is used for the level $\ell = 0$ (see also the proof of [11, Proposition 4.4]).

□

Remark 7 The estimate in Theorem 3 also holds if $f \in (V^*, L^2_{\boldsymbol{\beta}})_{t,\infty}$ and $G(\cdot) \in (V^*, L^2_{\boldsymbol{\beta}})_{t',\infty}$, $t, t' \in [0, 1]$, with an error bound $\mathscr{O}(M_\ell^{-(t+t')/d})$ and $\mathscr{O}(M_\ell^{-2(t+t')/d})$, $\ell = 0, \dots, L$, in the estimates for $Q_L^{\mathrm{IP}}(\cdot)$ and $Q_L^{\mathrm{RS}}(\cdot)$, respectively. This follows by Remark 5.

Remark 8 The factor $2/(1-\eta)$ in the weights in (23) and (24) as well as the constant factor 2 in the sequence $(b_j^{1-\theta} \vee 2\bar{b}_j)_{j\geq 1}$ can be omitted. Then, the error estimates in Theorem 3 hold under the same assumptions with QMC convergence rates $1/p - \varepsilon$ and $1/\bar{p} - \varepsilon$ in the multilevel error estimates for every $\varepsilon > 0$. This can be seen by the same argument that we used to show [11, Corollary 6.2] (see also [11, Corollary 6.4]).

6 Error vs. Work Analysis

The error estimates in Theorem 3 are the key ingredient to calibrate and choose the parameters $(s_\ell)_{\ell=0,\dots,L}$, $(M_\ell)_{\ell\geq 0}$, $\theta \in [0, 1)$, and $(N_\ell)_{\ell=0,\dots,L}$ of either considered type of the multilevel QMC algorithm with $L \in \mathbb{N}$ levels. We seek to derive choices that optimize the work for a given error threshold. The analysis will be demonstrated for a class of multiresolution analyses (MRA for short), which will serve as the function system $(\psi_\lambda)_{\lambda\in\nabla}$, here indexed by $\lambda \in \nabla$. We will use notation that is standard for wavelets and MRA. Assume that $(\psi_\lambda)_{\lambda\in\nabla}$ is a MRA that is obtained by scaling and translation from a finite number of mother wavelets, i.e.,

$$\psi_\lambda(x) = \psi(2^{|\lambda|}x - k), \quad k \in \nabla_{|\lambda|}, x \in D.$$

The index set $\bigtriangledown_{|\lambda|}$ has cardinality $|\bigtriangledown_{|\lambda|}| = \mathscr{O}(2^{|\lambda|d})$ and $|\mathrm{supp}(\psi_\lambda)| = \mathscr{O}(2^{-|\lambda|d})$. Let $j : \bigtriangledown \to \mathbb{N}$ be a suitable bijective enumeration. We also assume that on every level $|\lambda|$ there is a finite overlap, i.e., there exists a *support overlap constant* $K > 0$ such that for every $i \in \mathbb{N}_0$ and for every $x \in D$

$$|\{\lambda \in \bigtriangledown : |\lambda| = i, \psi_\lambda(x) \neq 0\}| \leq K\,.$$

The work needed to assemble the stiffness matrix for a generic parameter instance $\boldsymbol{y} \in [-1/2, 1/2]^s$ is therefore $\mathscr{O}(M_\ell |j^{-1}(s_\ell)|) = \mathscr{O}(M_\ell \log(s_\ell))$. Assuming at hand a linear complexity solver (i.e., a procedure which delivers a numerical solution of the discrete problem (9) to accuracy $\mathscr{O}(M_\ell^{-2/d})$ in the goal functional $G(\cdot)$ as in (10) uniformly w.r.t. to $\boldsymbol{y} \in U$ in work and memory $\mathscr{O}(M_\ell)$) the *overall work for either multilevel QMC algorithm* with the number of levels $L \in \mathbb{N}_0$ satisfies

$$\text{work} = \mathscr{O}\left(\sum_{\ell=0}^{L} N_\ell M_\ell \log(s_\ell)\right).$$

We remark that error vs. work estimates for general function systems $(\psi_j)_{j\geq 1}$ in the uncertainty parametrization (3) have been derived in [8, 22].

The parameter θ in the coupled estimates of Theorem 3 allows to discuss two possible strategies in the choices of the dimension truncation levels $(s_\ell)_{\ell=0,\dots,L}$. We recall from [11, Section 8] that if $\|\psi_{j(\lambda)}\|_{L^\infty(D)} \leq \sigma 2^{-\widehat{\alpha}|\lambda|}$, then the sequence

$$b_{j(\lambda)} = \left(1 + \frac{\bar{a}_{\min}(1-\kappa)(1-2^{\widehat{\beta}-\widehat{\alpha}})}{\sigma 2K} 2^{\widehat{\beta}|\lambda|}\right)^{-1}, \quad j \in \mathbb{N}, \tag{25}$$

satisfies (**A1**) for $\widehat{\alpha} > \widehat{\beta} > 1$ and $b_j \sim j^{-\widehat{\beta}/d}, j \geq 1$, holds. The sequence

$$\bar{b}_j = b_j^{(\widehat{\beta}-1)/\widehat{\beta}}, \quad j \in \mathbb{N},$$

satisfies (**A2**) and (**A1**) and $\bar{b}_j \sim j^{-(\widehat{\beta}-1)/d}, j \geq 1$, holds. Note that $\|\nabla\psi_{j(\lambda)}\|_{L^\infty(D)} \leq C\sigma 2^{-(\widehat{\alpha}-1)|\lambda|}$ assuming $\|\nabla\psi\|_{L^\infty(D)} \leq C\|\psi\|_{L^\infty(D)}$ for some $C > 0$. The truncation levels $(s_\ell)_{\ell=0,\dots,L}$ are chosen so as to cover entire levels of the MRA expansion of the uncertain PDE input, so that we choose $s_\ell \in \{\sum_{i=0}^{I} |\nabla_i| : I \in \mathbb{N}_0\}$, $\ell \geq 0$. We also assume that

$$M_\ell \sim 2^{d\ell}, \quad \ell \geq 0. \tag{A3}$$

In this section we assume for simplicity that only one version of the QMC rule is applied with convergence rate $1/\bar{p}$. We remark that in some cases the application of two different weight sequences with different sparsity (as expressed by the summability exponents $p, \bar{p}$ of the sequences $(b_j)_{j\geq 1}, (\bar{b}_j)_{j\geq 1})$ may be beneficial. Also we assume that

$$f \in (V^*, L^2_{\boldsymbol{\beta}}(D))_{t,\infty} \quad \text{and} \quad G(\cdot) \in (V^*, L^2_{\boldsymbol{\beta}}(D))_{t',\infty}, \quad t, t' \in [0,1], \tag{\textbf{A4}}$$

which yields a FE convergence rate of $\tau := t + t' \in [0, 2]$, cp. Remark 7.

Strategy 1 We equilibrate the decay of the sequences $(b_j^{1-\theta})_{j\geq 1}$ and $(\bar{b}_j)_{j\geq 1}$, which determines the bound on the QMC error in Theorem 3. The parameter $\theta \in [0, 1)$ is chosen to be $\theta = 1/\widehat{\beta}$, which implies $b_j^{1-\theta} = \bar{b}_j$, $j \in \mathbb{N}$, and $(\bar{b}_j)_{j\geq 0} \in \ell^{\bar{p}}(\mathbb{N})$ for every $\bar{p} > d/(\widehat{\beta} - 1)$. We equilibrate the error contributions on the highest discretization level L. Since $M_L \sim 2^{dL}$, we choose

$$s_L \sim 2^{d\lceil L\tau/(2\widehat{\beta})\rceil}.$$

On the different discretization levels of the coupled error terms, we either increase the dimension truncation levels or leave them constant, which is reflected in the choice

$$s_\ell \sim \min\left\{2^{d\lceil \ell\tau/(\theta\widehat{\beta})\rceil}, s_L\right\}, \quad \ell = 0, \ldots, L-1.$$

Strategy 2 For certain function systems $(\psi_\lambda)_{\lambda\in\nabla}$ and meshes $\{\mathscr{T}_\ell\}_{\ell\geq 0}$ it may be interesting (also for implementation purposes) to couple their discretizations, i.e., we choose

$$s_\ell \sim M_\ell, \quad \ell = 0, \ldots, L.$$

To equilibrate the truncation and FE error on the levels we choose $\theta = \tau/\widehat{\beta}$, which imposes the constraint $\widehat{\beta} > \tau$ and implies that $b_j^{1-\theta} \sim j^{-(\widehat{\beta}-\tau)/d}$. Hence, $(b_j^{1-\theta} \vee \bar{b}_j)_{j\geq 1} \in \ell^{\bar{p}}(\mathbb{N})$ for every $\bar{p} > d/(\min\{\widehat{\beta} - \tau, \widehat{\beta} - 1\})$.

We will discuss interlaced polynomial lattice rules first and follow [8, Section 3.3]. In either of our parameter choices, the error estimate

$$\text{error} = \mathscr{O}\left(M_L^{-\tau/d} + \sum_{\ell=0}^{L} N_\ell^{-1/\bar{p}} M_\ell^{-\tau/d}\right)$$

holds, where we used that $M_\ell = \mathscr{O}(2^{d\ell})$. The QMC sample numbers $(N_\ell)_{\ell=0,\ldots,L}$ are chosen to optimize the error versus the required work. Optimizing error (bound) vs. cost as in [8, 22], we seek the stationary point of the function

$$g(\xi) = M_L^{-\tau/d} + \sum_{\ell=0}^{L} N_\ell^{-1/\bar{p}} M_\ell^{-\tau/d} + \xi \sum_{\ell=0}^{L} N_\ell M_\ell \log(s_\ell)$$

with respect to N_ℓ, i.e., choose N_ℓ such that $\partial g(\xi)/\partial N_\ell = 0$. We thus obtain

$$N_\ell = \left\lceil N_0 \left(M_\ell^{-1-\tau/d} \log(s_\ell)^{-1} \right)^{\bar{p}/(1+\bar{p})} \right\rceil, \quad \ell = 1, \ldots, L, \tag{26}$$

and for $E_\ell := (M_\ell^{1-\bar{p}\tau/d} \log(s_\ell))^{1/(\bar{p}+1)}$,

$$\text{error} = \mathscr{O}\left(M_L^{-\tau/d} + N_0^{-1/\bar{p}} \sum_{\ell=0}^{L} E_\ell \right) \quad \text{and} \quad \text{work} = \mathscr{O}\left(N_0 \sum_{\ell=0}^{L} E_\ell \right).$$

Since for every $0 \neq r_1 \in \mathbb{R}$ and $r_2 > 0$,

$$\sum_{\ell=0}^{L} 2^{r_1 \ell} \ell^{r_2} \leq \frac{2^{r_1(L+1)} - 1}{2^{r_1} - 1} L^{r_2},$$

$\log(s_\ell) = \mathscr{O}(\ell)$, which holds in the considered cases, implies that

$$\sum_{\ell=0}^{L} E_\ell = \begin{cases} \mathscr{O}(1) & \text{if } d < \bar{p}\tau, \\ \mathscr{O}(L^{(\bar{p}+2)/(\bar{p}+1)}) & \text{if } d = \bar{p}\tau, \\ \mathscr{O}(2^{(d-\bar{p}\tau)L/(\bar{p}+1)} L^{1/(\bar{p}+1)}) & \text{if } d > \bar{p}\tau. \end{cases}$$

We choose N_0 to equilibrate the error, i.e.,

$$N_0^{-1/\bar{p}} \sum_{\ell=0}^{L} E_\ell = \mathscr{O}\left(M_L^{-\tau/d} \right),$$

which yields

$$N_0 := \begin{cases} \lceil 2^{\tau\bar{p}L} \rceil & \text{if } d < \bar{p}\tau, \\ \lceil 2^{\tau\bar{p}L} L^{\bar{p}(\bar{p}+2)/(\bar{p}+1)} \rceil & \text{if } d = \bar{p}\tau, \\ \lceil 2^{\bar{p}(d+\tau)L/(\bar{p}+1)} L^{\bar{p}/(\bar{p}+1)} \rceil & \text{if } d > \bar{p}\tau. \end{cases} \tag{27}$$

This implies that an error $= \mathscr{O}(M_L^{-\tau/d})$ requires

$$\text{work} = \begin{cases} \mathscr{O}(2^{\bar{p}\tau L}) & \text{if } d < \bar{p}\tau, \\ \mathscr{O}(2^{\tau\bar{p}L} L^{\bar{p}+2}) & \text{if } d = \bar{p}\tau, \\ \mathscr{O}(2^{dL} L) & \text{if } d > \bar{p}\tau. \end{cases}$$

In the other case of randomly shifted lattice rules, sample numbers $(N_\ell)_{\ell=0,\ldots,L}$ are derived in [22, Section 3.7]. There, also the work functional from a MRA is considered, cp. [22, Equations (74) and (77)] with $\lambda = \bar{p}/2$ and $K_\ell = M_\ell \log(M_\ell) \sim 2^{d\ell}\ell$. Specifically, for randomly shifted lattice rules we choose

$$N_\ell = \left\lceil N_0 \left(M_\ell^{-1-2\tau/d} \log(s_\ell)^{-1} \right)^{\bar{p}/(2+\bar{p})} \right\rceil, \quad \ell = 1, \dots, L, \tag{28}$$

and

$$N_0 := \begin{cases} \lceil 2^{\tau\bar{p}L} \rceil & \text{if } d < \bar{p}\tau, \\ \lceil 2^{\tau\bar{p}L} L^{\bar{p}(\bar{p}+4)/(2\bar{p}+4)} \rceil & \text{if } d = \bar{p}\tau, \\ \lceil 2^{\bar{p}(d+2\tau)L/(\bar{p}+2)} L^{\bar{p}/(\bar{p}+2)} \rceil & \text{if } d > \bar{p}\tau. \end{cases} \tag{29}$$

The work estimates for these choices in the case of randomly shifted lattice rules are stated on [22, p. 443]. We collect the foregoing estimates in the following theorem.

Theorem 4 *Let the assumption in (A3) be satisfied and let for $L \in \mathbb{N}$ and $Q_L^{\mathrm{IP}}(\cdot)$, the sample numbers $(N_\ell)_{\ell=0,\dots,L}$ be given by (26) and (27) and for $Q_L^{\mathrm{RS}}(\cdot)$, be given by (28) and (29). Let the right hand side f and $G(\cdot)$ satisfy (A4). For $\bar{p} \in (d/(\widehat{\beta} - 1), 1]$, assuming $d < \widehat{\beta} - 1$ and error threshold $\varepsilon > 0$, we obtain*

$$|\mathbb{E}(G(u)) - Q_L^{\mathrm{IP}}(G(u^L))| = \mathscr{O}(\varepsilon)$$

with

$$\text{work} = \begin{cases} \mathscr{O}(\varepsilon^{-\bar{p}}) & \text{if } d < \bar{p}\tau, \\ \mathscr{O}(\varepsilon^{-\bar{p}} \log(\varepsilon^{-1})^{\bar{p}+2}) & \text{if } d = \bar{p}\tau, \\ \mathscr{O}(\varepsilon^{-d/\tau} \log(\varepsilon^{-1})) & \text{if } d > \bar{p}\tau. \end{cases}$$

For $\bar{p} \in (\max\{1, d/(\widehat{\beta} - 1)\}, 2]$ assuming $d < 2(\widehat{\beta} - 1)$ and an error threshold $\varepsilon > 0$, we obtain

$$\sqrt{\mathbb{E}^{\Delta}(|\mathbb{E}(G(u)) - Q_L^{\mathrm{RS}}(G(u^L))|^2)} = \mathscr{O}(\varepsilon)$$

with

$$\text{work} = \begin{cases} \mathscr{O}(\varepsilon^{-\bar{p}}) & \text{if } d < \bar{p}\tau, \\ \mathscr{O}(\varepsilon^{-\bar{p}} \log(\varepsilon^{-1})^{\bar{p}/2+2}) & \text{if } d = \bar{p}\tau, \\ \mathscr{O}(\varepsilon^{-d/\tau} \log(\varepsilon^{-1})) & \text{if } d > \bar{p}\tau. \end{cases}$$

Remark 9 The parameter choices for θ and $(s_\ell)_{\ell=0,\dots,L}$ in Theorem 4 reflect *Strategy 1*. For *Strategy 2*, the assumptions $\bar{p} > d/(\min\{\widehat{\beta} - \tau, \widehat{\beta} - 1\})$, $\widehat{\beta} > \tau$ are required, which is more restrictive if $\tau > 1$. However, aligning MRA and

FE meshes might be useful in certain cases. Note that the truncation dimension in *Strategy 2* could also be capped as in *Strategy 1*, which may be beneficial in some cases. Adopting this strategy would affect the work measure only by a constant factor.

7 Numerical Experiments

To illustrate the foregoing asymptotic error bounds, we present numerical experiments in space dimension $d = 1, 2$ with affine-parametric diffusion coefficient

$$a(x, \boldsymbol{y}) = \bar{a}(x) + \sum_{j \geq 1} y_j \psi_j(x),$$

where $\bar{a}(x) \equiv 1$ and we assume $\psi_j = \psi_{j(\ell,k)}$ to be a system of continuous, piecewise (bi)linear spline wavelets for $\ell \geq 0$, $k \in \{0, \dots, 2^\ell - 1\}^d$ with support overlap constant $K = 2^d$, see e.g. [18, Chapter 12]. We assume in the following the scaling $\|\psi_{j(\ell,k)}\|_{L^\infty(D)} = \sigma 2^{-\widehat{\alpha}\ell}$. We pursue Strategy 2 from Sect. 6, which yields for $\widehat{\alpha} > \widehat{\beta} > \tau$ a QMC weight sequence of the form

$$\mathfrak{b}_{j(\ell,k)} = \left(1 + c_2 2^{\widehat{\beta}\ell}\right)^{-(\widehat{\beta}-\tau)/\widehat{\beta}},$$

for $0 < c_2 \in \mathbb{R}$ as specified in (25) (see also Remark 8). We use the implementation from [9] for applying the single-level and multilevel methods in parallel, and use the Walsh coefficient bound $C = 0.1$ in the component-by-component (CBC for short) construction, cp. [10] for details. For the multilevel method, we choose $N_\ell = 2^{m_\ell}$, where m_ℓ follows from (26). The resulting expression is given by

$$m_\ell = \left\lceil \bar{p}\tau L + \frac{\bar{p}(\bar{p}+2)}{\bar{p}+1} \log_2(L+1) + \frac{\bar{p}}{\bar{p}+1}\big(-\ell(d+\tau) - \log_2(\ell+1)\big) \right\rceil, \tag{30}$$

with $m_\ell = 1$ if the expression is not positive. In the following examples, we consider the limiting case $d = \bar{p}\tau$ also with the limiting value $\bar{p}^{-1} = (\widehat{\beta} - \tau)/d$. This choice is based on the cost model

$$W_L^{\mathrm{ML}} = \sum_{\ell=0}^{L} N_\ell M_\ell \log_2(s_\ell) ,$$

which we use for computing the cost in the multilevel experiments below. We compare the multilevel computations to a single-level approach, where we equilibrate the QMC and FE discretization errors, yielding on a fixed level L with

$N_L^{-1/\bar{p}} \sim M_L^{-\tau/d}$ the choice $N_L = 2^{\bar{p}\tau(L+1)}$, i.e.,

$$m_L = \log_2(N_L) = \lceil \bar{p}\tau \rceil (L+1) .$$

In the single-level case, the work is simply $W_L^{\mathrm{SL}} = N_L M_L \log_2(s_L)$.

7.1 Univariate Model Problem

We consider the domain $D = (0,1)$ and homogeneous Dirichlet boundary conditions, i.e., $\Gamma_1 = \partial D$, with right hand side $f(x) = 10x$, $x \in D$. As goal functional, we consider point evaluation of the solution at $\bar{x} = e^{-1}$ (which is not a node on any mesh used in our simulations), $G(u(\cdot,\boldsymbol{y})) = u(\bar{x},\boldsymbol{y})$, which implies the FE convergence rate $\tau = 1.5 - \varepsilon$ for arbitrary $\varepsilon > 0$. The parameter calibration will be done under the formal case $\tau = 1.5$. For a given discretization level ℓ, we solve the parametric PDE (9) with the finite element method using piecewise linear basis functions on an equidistant mesh with meshwidth $h_\ell = 2^{-\ell-1}$ to approximate the solution of (1). Considering the wavelet basis for the coefficients on the same mesh, we obtain $s_\ell = h_\ell^{-1} - 1 = 2^{\ell+1} - 1$ parametric dimensions on level ℓ. We choose $\widehat{\alpha} = 3$, $\widehat{\beta} = 2.99$, $\sigma = 0.15$, yielding the expected QMC convergence rate $\widehat{\beta} - \tau = 1.49$ (see Fig. 1). We use the same generating vectors as above

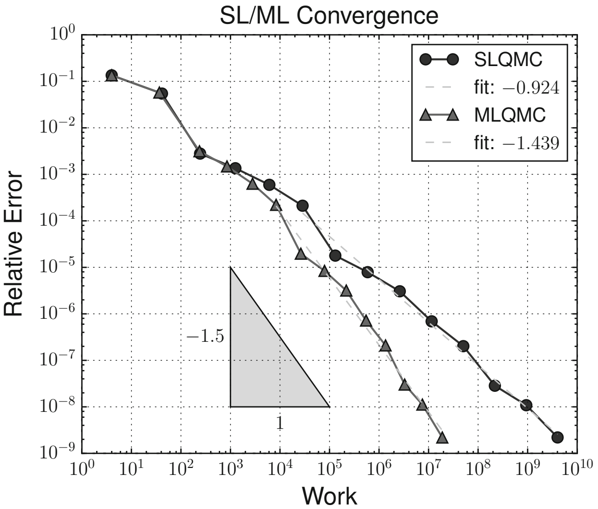


Fig. 1 Convergence of single-level and multilevel methods for a univariate diffusion coefficient given in wavelet representation. As a reference solution, the multilevel approximation on the level $L = 14$ with a total of $s_L = 32{,}767$ dimensions was used. The measured rates were obtained by a linear least squares fit on the last 9 points. The expected rates are 0.75 for SLQMC and 1.5 for MLQMC ignoring log factors. The work is $W_L^{\mathrm{ML}} = \sum_{\ell=0}^{L} N_\ell h_\ell^{-1}(1 + \log_2(s_\ell))$ for multilevel and $W_L^{\mathrm{SL}} = N_L h_L^{-1}(1 + \log_2(s_L))$ for single-level

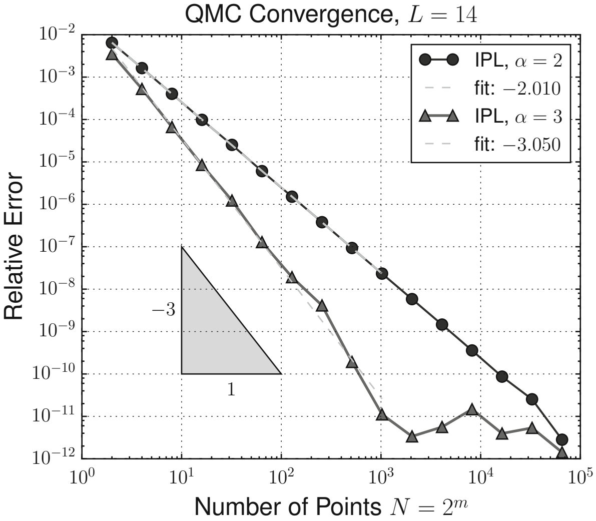


Fig. 2 Convergence of the QMC approximation for the univariate model problem using interlaced polynomial lattice (IPL) rules with $N = 2^m$ points, $m = 1, \dots, 17$ and for digit interlacing factors $\alpha = 2, 3$. We use the results with $m = 17$ as the reference value and keep the maximal discretization level $L = 14$ fixed, resulting in $s_L = 2^{15} - 1 = 32{,}767$ parameter dimensions and smallest FE meshwidth $h_L = 2^{-15}$

for the single-level method; this is justified since the weight sequence used in the CBC construction majorizes the weight sequence for the single-level quadrature, theoretically capping the rate at $N^{-1.5}$. With these generating vectors, as observed in Fig. 2, the measured QMC convergence rate is independent of the parameter dimension, and equals $N^{-\alpha}$ for $\alpha = 2, 3$ rather than the expected rate $N^{-1.5}$.

7.2 Two Spatial Dimensions

For $d = 2$, we consider the domain $D = (-1, 1) \times (0, 1)$ with mixed boundary conditions. Specifically, the Neumann boundary is given by $\Gamma_2 = (-1, 0) \times \{0\}$ and the Dirichlet boundary is $\Gamma_1 = \partial D \backslash \Gamma_2$. Although the domain is convex, the change in boundary conditions at the origin induces a point singularity in the parametric solutions corresponding to an interior angle equal to π. Due to isotropy of the parametric diffusion coefficient, this leads to a non-$H^2(D)$ singularity of ($\boldsymbol{y}$-independent) strength $O(\sqrt{r})$ of the parametric solution $u(\cdot, \boldsymbol{y})$ concentrated at the origin. The boundary conditions change also at the corner $(-1, 0)^\top \in \partial D$, inducing a weaker singularity there as well. The considered goal functional is here integration over the domain D, which is an element of $L^2(D)$. Since the parametric coefficients $a(x, \boldsymbol{y})$ are isotropic, i.e., scalar valued, the full regularity shift of the Laplacean in weighted Hilbert spaces is applicable as detailed in Sect. 2, we obtain $\tau = 2$. Analogous to the univariate problem considered in the previous subsection, we use continuous, bilinear FE on quadrilaterals on sequences of nested, locally refined meshes of the domain D which were obtained by a suitable bisection refinement, cp. [12].

Here, we have $J = 5$ singular points or corners and $\boldsymbol{\beta} \in [0,1)^J$ satisfies that $\beta_i > 1 - \pi/\omega_i$, $i = 1,2,3$, and $\beta_i > 1 - \pi/(2\omega_i)$, $i = 4,5$, cp. Sect. 2. Then, for the Laplacean with mixed boundary conditions in D there holds a full regularity shift in weighted Sobolev spaces, i.e. $(-\Delta)^{-1} : L^2_{\boldsymbol{\beta}}(D) \to H^1_0(D) \cap H^2_{\boldsymbol{\beta}}(D)$ is bounded with $\boldsymbol{\beta} = (0,0,0,\beta_4,\beta_5)$, $1 > \beta_4 > 0$, and $1 > \beta_5 > 1/2$, where singular points are enumerated counter clockwise, i.e., $c_1 = (1,0)^\top$, $c_2 = (1,1)^\top$, $c_3 = (-1,1)^\top$, $c_4 = (-1,0)^\top$, and $c_5 = (0,0)^\top$. We observe that solutions will in general have a weak non-$H^2(D)$ singularity at the corner c_4, i.e., $u(x,\boldsymbol{y}) \in H^{2-\varepsilon}(D_4)$ for every $\varepsilon > 0$, where $D_4 \subset D$ is a sufficiently small neighborhood of c_4. We use the values $\beta_1 = \beta_2 = \beta_3 = 0$, $\beta_4 = 0.05$, and $\beta_5 = 0.55$ as inputs for a bisection refinement algorithm, which results in 1-irregular quadrilateral meshes. In polar coordinates $(r,\phi) \in (0,\infty) \times (0,\pi)$, where $x = r(\cos(\phi),\sin(\phi))^\top$, the function $\bar{u}(r,\phi) = \sqrt{r}\sin(\phi/2)$ is harmonic, i.e., $\Delta\bar{u} = 0$, and satisfies the homogeneous Neumann boundary conditions. We solve the parametric boundary value problem

$$-\nabla \cdot (a(x,\boldsymbol{y})\nabla u(x,\boldsymbol{y})) = 0, \quad u(x,\boldsymbol{y})\Big|_{\Gamma_1} = \bar{u}(x)\Big|_{\Gamma_1}, \quad a(x,\boldsymbol{y})\nabla u(x,\boldsymbol{y}) \cdot n(x)\Big|_{\Gamma_2} = 0.$$

Clearly, $u(x,\mathbf{0}) = \bar{u}(x)$. The inhomogeneous Dirichlet boundary terms can be incorporated into the right hand side, for example by solving $-\nabla \cdot (a\nabla(u - \bar{u})) = \nabla \cdot (a\nabla\bar{u})$ and adding $\bar{u}$ to the solution afterwards. Instead of $\bar{u}$ one may use any other suitable extension of $\bar{u}|_{\partial D}$ to the domain D. The difference $u - \bar{u}$ satisfies the homogeneous mixed boundary conditions. The parametric right hand side is given by $f(x,\boldsymbol{y}) := \nabla \cdot (a(x,\boldsymbol{y})\nabla\bar{u}(x)) \in L^2_{\boldsymbol{\beta}}(D)$ for $\boldsymbol{\beta}$ stated above. This right hand side $f(x,\boldsymbol{y})$ depends affinely on the parameter vector $\boldsymbol{y}$. In previous sections, we assumed a fixed right hand side only for simplicity and conciseness of the presentation. A right hand side, which only depends linearly on the coefficient $a(x,\boldsymbol{y})$ under the made assumptions is admissible by a straightforward extension of our theory. The implementation of the spatial discretization in two space dimensions of bilinear FE uses `deal.II`, cp. [1].

For the uncertain diffusion coefficient, we consider the parametrization obtained by tensorizing the univariate continuous, piecewise linear biorthogonal spline wavelets. Specifically, we choose

$$\widehat{\psi}_{\ell,k_1,k_2}(x_1,x_2) = \sigma 2^{-\widehat{\alpha}\ell}\psi_{\ell,k_1}(x_1)\psi_{\ell,k_2}(x_2), \quad k_1,k_2 \in \{0,\dots,2^\ell - 1\}, \tag{31}$$

where $\psi_{\ell,k}(x)$ denotes the univariate continuous, piecewise linear wavelet function with scaling $\|\psi_{\ell,k}\|_{L^\infty(D)} = 1$ and $\sigma = 0.01$. Thus, $\|\widehat{\psi}_{\ell,k_1,k_2}\|_{L^\infty(D)} = \sigma 2^{-\widehat{\alpha}\ell}$ with $\widehat{\alpha} = 4$. This choice of parametrization results in $s_L = \sum_{\ell=0}^{L} 4^\ell = (4^{L+1} - 1)/3$ dimensions on the discretization level L. The generating vectors were constructed

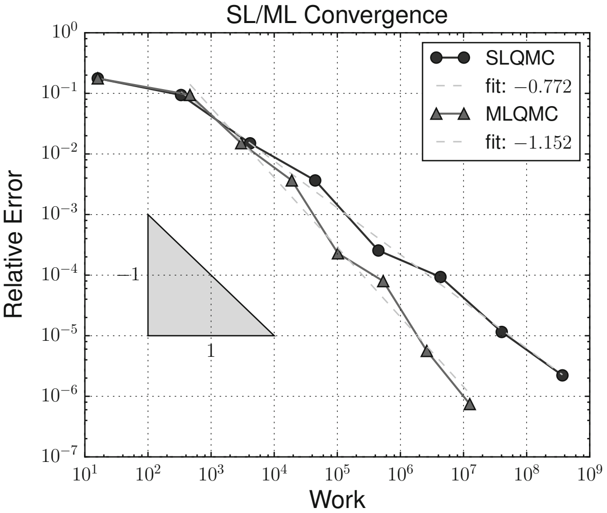


Fig. 3 Convergence of single-level and multilevel methods for a 2d diffusion equation with parametric coefficient given in wavelet representation. Continuous, piecewise bilinear biorthogonal spline wavelets (31) on uniform partitions of the domain D with meshwidth $O(2^{-\ell})$, $\ell = 0, \dots, L$, were used. As a reference solution, the multilevel approximation on the level $L = 8$ with a total of $s_L = 87{,}381$ dimensions was used. The measured rates were obtained by a linear least squares fit on all points but the first and the two last ones. The rates expected from the theory for this problem are 0.67 for SLQMC and 1 for MLQMC ignoring log factors. The work measure is $W_L^{\mathrm{ML}} = \sum_{\ell=0}^{L} N_\ell 2^{2\ell}(1 + \log_2(s_\ell))$ for multilevel and $W_L^{\mathrm{SL}} = N_L 2^{2L}(1 + \log_2(s_L))$ for single-level

by the CBC algorithm based on a QMC weight sequence analogous to the univariate case, given here by $\mathfrak{b}_{j(\ell,k_1,k_2)} = \left(1 + c_2 2^{\widehat{\beta}\ell}\right)^{-(\widehat{\beta}-\tau)/\widehat{\beta}}$ where $\widehat{\beta} = 3.99$ and $\tau = 2$.

For the multilevel method, the number of samples per level is given by $N_\ell = 2^{m_\ell}$ where the exponent m_ℓ is given as in (30) with $d = 2$. To compare to a single-level approach, we equilibrate the finite element and QMC sampling error to obtain $N_L = 2^{L\tau/r} \sim M_L^{\tau/(dr)}$, where r is the QMC convergence rate, here $r \approx 2$ for interlacing factor $\alpha = 2$ and we take $r = 2$ to obtain the value of N_L (see Fig. 3).

8 Conclusions

We provided the convergence rate analysis of randomly shifted and higher order, interlaced polynomial lattice rules for the numerical evaluation of linear functionals $G(\cdot)$ of solutions of countably affine-parametric, linear second order elliptic partial differential equations. The spatially inhomogeneous diffusion coefficient was assumed to be represented by a multiresolution analysis with local supports, rather than the globally supported Karhunen-Loève expansion considered, for example, in [7, 8, 14, 21, 22] and the references there. As in the corresponding single-level QMC Petrov–Galerkin approaches considered in [11], we proved that QMC with product weights, originally proposed by Sloan and Woźniakowski in [28], can provide optimal QMC convergence rates which are independent of the parameter dimension. Unlike the so-called *product and order dependent weights* which are mandated by globally supported representation systems of uncertain input data, the use of product weights results in linear w.r.t. to dimension scaling of fast CBC constructions from [24, 25], which originate in a dimension-wise, greedy strategy to minimize the worst case error, as proposed originally in [30]. The present analysis addressed linear, affine-parametric random input data where the supports of the parameters are bounded. The extension for log-Gaussian diffusion coefficients in the present setting, along the lines of [14, 23] (where the case of globally supported ψ_j were treated) and in the setting of the single-level analysis in [15], is given in [16]. Numerical experiments were given for a model, linear elliptic problem in one and in two space dimensions with local spatial mesh refinement. The present mathematical analysis holds, however, also for PDEs on polyhedra in three space dimensions, for proper choice of (corner- and edge-weighted) function spaces, and corresponding mesh refinements. We refer to [16]. Analogous error bounds for product weight QMC also hold for log-Gaussian representations of uncertain PDE inputs. Details are presented in [16, 17].

Acknowledgements This work was supported in part by the Swiss National Science Foundation (SNSF) under grant SNF 159940 and SNF 149819. The authors thank Fabian Keller for his contribution to the implementation in two space dimensions and Fabian Müller for his help with generating the bisection-tree meshes.

References

1. Arndt, D., Bangerth, W., Davydov, D., Heister, T., Heltai, L., Kronbichler, M., Maier, M., Pelteret, J.P., Turcksin, B., Wells, D.: The `deal.II` library, version 8.5. J. Numer. Math. (2017). https://doi.org/10.1515/jnma-2017-0058
2. Babuška, I., Kellogg, R.B., Pitkäranta, J.: Direct and inverse error estimates for finite elements with mesh refinements. Numer. Math. **33**(4), 447–471 (1979)
3. Bachmayr, M., Cohen, A., Migliorati, G.: Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients. ESAIM Math. Model. Numer. Anal. **51**(1), 321–339 (2017)
4. Chen, P., Schwab, Ch.: Model order reduction methods in computational uncertainty quantification. In: Handbook of Uncertainty Quantification, pp. 1–53. Springer International Publishing, Cham (2016)
5. Dashti, M., Stuart, A.: The Bayesian approach to inverse problems. In: Handbook of Uncertainty Quantification, pp. 1–118. Springer International Publishing, Cham (2016)
6. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
7. Dick, J., Kuo, F.Y., Le Gia, Q.T., Nuyens, D., Schwab, Ch.: Higher order QMC Petrov-Galerkin discretization for affine parametric operator equations with random field inputs. SIAM J. Numer. Anal. **52**(6), 2676–2702 (2014)

8. Dick, J., Kuo, F.Y., Le Gia, Q.T., Schwab, Ch.: Multilevel higher order QMC Petrov-Galerkin discretization for affine parametric operator equations. SIAM J. Numer. Anal. **54**(4), 2541–2568 (2016)
9. Gantner, R.N.: A generic C++ library for multilevel quasi-Monte Carlo. In: Proceedings of the Platform for Advanced Scientific Computing Conference, PASC'16, pp. 11:1–11:12. ACM, New York, NY (2016)
10. Gantner, R.N., Schwab, Ch.: Computational higher order quasi-Monte Carlo integration. In: Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, April 2014, vol. 163, pp. 271–288. Springer, Cham (2016)
11. Gantner, R.N., Herrmann, L., Schwab, Ch.: Quasi-Monte Carlo integration for affine-parametric, elliptic PDEs: local supports and product weights. SIAM J. Numer. Anal. **56**(1), 111–135 (2018)
12. Gaspoz, F.D., Morin, P.: Convergence rates for adaptive finite elements. IMA J. Numer. Anal. **29**(4), 917–936 (2009)
13. Giles, M.B.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)
14. Graham, I.G., Kuo, F.Y., Nichols, J.A., Scheichl, R., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. Numer. Math. **131**(2), 329–368 (2015)
15. Herrmann, L., Schwab, Ch.: QMC integration for lognormal-parametric, elliptic PDEs: local supports and product weights. Technical Report 2016-39 (revised), Seminar for Applied Mathematics, ETH Zürich (2016)
16. Herrmann, L., Schwab, Ch.: Multilevel quasi-Monte Carlo integration with product weights for elliptic PDEs with lognormal coefficients. Technical Report 2017-19, Seminar for Applied Mathematics, ETH Zürich, Zürich (2017)
17. Herrmann, L., Schwab, Ch.: QMC algorithms with product weights for lognormal-parametric, elliptic PDEs. Technical Report 2017-04 (revised), Seminar for Applied Mathematics, ETH Zürich, Zürich (2017)
18. Hilber, N., Reichmann, O., Schwab, Ch., Winter, Ch.: Computational methods for quantitative finance. In: Finite Element Methods for Derivative Pricing. Springer Finance. Springer, Heidelberg (2013)
19. Kuo, F.Y., Nuyens, D.: Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients: a survey of analysis and implementation. Found. Comput. Math. **16**(6), 1631–1696 (2016)
20. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo methods for high-dimensional integration: the standard (weighted Hilbert space) setting and beyond. ANZIAM J. **53**(1), 1–37 (2011)
21. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**(6), 3351–3374 (2012)
22. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. Found. Comput. Math. **15**(2), 411–449 (2015)
23. Kuo, F., Scheichl, R., Schwab, Ch., Sloan, I., Ullmann, E.: Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. Math. Comput. **86**(308), 2827–2860 (2017)
24. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comput. **75**(254), 903–920 (electronic) (2006)
25. Nuyens, D., Cools, R.: Fast component-by-component construction of rank-1 lattice rules with a non-prime number of points. J. Complex. **22**(1), 4–28 (2006)
26. Schwab, Ch., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numer. **20**, 291–467 (2011)
27. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford Science Publications. The Clarendon Press/Oxford University Press, Oxford/New York (1994)

28. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals? J. Complex. **14**(1), 1–33 (1998)
29. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. SIAM J. Numer. Anal. **40**(5), 1650–1665 (2002)
30. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comput. **71**(240), 1609–1640 (2002)
31. Triebel, H.: Interpolation Theory, Function Spaces, Differential Operators, 2nd edn. Johann Ambrosius Barth, Heidelberg (1995)

An Adaptive Filon Algorithm for Highly Oscillatory Integrals

Jing Gao and Arieh Iserles

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Based on the error analysis of Extended Filon Method (EFM), we present an adaptive Filon method to calculate highly oscillatory integrals. The main idea is to allow interpolation points depend upon underlying frequency in order to minimize the error. Typically, quadrature error need be examined in two regimes. Once frequency is large, asymptotic behaviour dominates and we need to choose interpolation points accordingly, while for small frequencies good choice of interpolation points is similar to classical, non-oscillatory quadrature. In this paper we choose frequency-dependent interpolation points according to a smooth homotopy function and the accuracy is superior to other EFMs. The basic algorithm is presented in the absence of stationary points but we extend it to cater for highly oscillatory integrals with stationary points. The presentation is accompanied by numerical experiments which demonstrate the power of our approach.

J. Gao
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an,
People's Republic of China
e-mail: jgao@xjtu.edu.cn

A. Iserles (✉)
DAMTP, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK
e-mail: ai10@cam.ac.uk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_19

1 Introduction

The focus of this paper is on the computation of the highly oscillatory integral

$$I_\omega[f] = \int_{-1}^{1} f(x)\mathrm{e}^{\mathrm{i}\omega g(x)}\mathrm{d}x, \tag{1}$$

where $f, g \in \mathbf{C}^\infty[-1,1]$ and $\omega \geq 0$ is the frequency. We assume that the phase function $g(x)$ is normalised so that $\max_{x\in[-1,1]}|g(x)| = 1$. Since this integral abounds in mathematics and computational engineering [2, 14, 15] and standard quadrature methods fail to calculate it well, it has been subjected to very active research effort in the last two decades. This has resulted in a significant number of efficient quadrature methods, such as the asymptotic expansion and Filon methods [5, 10, 11], Levin's method [12, 13], numerical steepest descent [9], complex Gaussian quadrature [1, 3] and other efficient algorithms [4, 6].

Each of these methods has its own advantages and disadvantages and it would be rash to proclaim one as the definite approach to the integration of (1). They require the availability of different information (e.g., Filon methods and complex Gaussian quadrature require the computation of moments, numerical steepest descent relies on practical computation of steepest-descent paths in the complex plane) and might have critical shortcomings in some situations (Levin's method cannot work in the presence of stationary points and explicit asymptotic expansions are exceedingly difficult once (1) is generalised to multivariate setting—a setting in which nothing is known of complex Gaussian quadrature).

Popularity of Filon-type methods owes much to their simplicity and flexibility. We just need to replace f by an interpolating polynomial and, assuming that moments $\int_{-1}^{1} x^m \mathrm{e}^{\mathrm{i}\omega g(x)}\mathrm{d}x$, $m \geq 0$, are explicitly available, the new integral can be computed easily. The make-or-break issue, however, is the location of suitable interpolation points. The basic imperative is to select interpolation points that ensure good behaviour for large ω, and this is entirely governed by asymptotic analysis. Let us recap some basic facts from [10]. Assume first that there are no stationary points, i.e. that $g' \neq 0$ in $[-1,1]$. Letting $\tilde{p}$ be the interpolating polynomial, the error can be expanded into asymptotic series,

$$\begin{aligned} I_\omega[\tilde{p}] - I_\omega[f] &= I_\omega[\tilde{p} - f] \\ &\sim -\sum_{m=0}^{s-1} \frac{1}{(-\mathrm{i}\omega)^{m+1}} \left[\frac{\sigma_m[\tilde{p}-f](1)}{g'(1)}\mathrm{e}^{\mathrm{i}\omega g(1)} - \frac{\sigma_m[\tilde{p}-f](-1)}{g'(-1)}\mathrm{e}^{\mathrm{i}\omega g(-1)} \right] \\ &\quad + \mathscr{O}\big(\omega^{-(s+1)}\big), \end{aligned} \tag{2}$$

where

$$\sigma_0[h](x) = h(x), \qquad \sigma_m[h](x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\sigma_{m-1}[h](x)}{g'(x)}, \quad m \geq 1.$$

Moreover, $\sigma_m[h](x)$ is a linear combination (with coefficients depending on derivatives of g) of $h^{(j)}(x), j = 0, \ldots, m$ [11]. It immediately follows that the Hermite-type interpolation conditions

$$\tilde{p}^{(j)}(1) = f^{(j)}(1), \quad \tilde{p}^{(j)}(-1) = f^{(j)}(-1), \qquad j = 0, 1, \cdots, s-1, \tag{3}$$

imply that the error is $\sim \mathscr{O}(\omega^{-s-1})$ for $\omega \gg 1$. The outcome is the (plain-vanilla) *Filon method,*

$$\mathscr{Q}_\omega^{\mathsf{F},s,0}[f] = \int_{-1}^{1} \tilde{p}(x) \mathrm{e}^{\mathrm{i}\omega g(x)} \mathrm{d}x.$$

Once g' vanishes somewhere in $[-1, 1]$, the oscillation of the integrand slows down in the vicinity of that point and the behaviour of (1) changes. In particular, the asymptotic expansion (2) is no longer valid. For example, if $g'(c) = 0$, $g''(c) \neq 0$, for $c \in (-1, 1)$ and $g'(x) \neq 0$ elsewhere in $[-1, 1]$, then

$$\begin{aligned} I_\omega[\tilde{p} - f] \sim \mu_0(\omega) \sum_{m=0}^{\infty} \frac{\tilde{\sigma}_m[\tilde{p} - f](c)}{(-\mathrm{i}\omega)^m} \\ - \sum_{m=0}^{\infty} \frac{1}{(-\mathrm{i}\omega)^{m+1}} \Bigg\{ \frac{\tilde{\sigma}_m[\tilde{p} - f](1) - \tilde{\sigma}_m[\tilde{p} - f](c)}{g'(1)} \mathrm{e}^{\mathrm{i}\omega g(1)} \\ - \frac{\tilde{\sigma}_m[\tilde{p} - f](-1) - \tilde{\sigma}_m[\tilde{p} - f](c)}{g'(-1)} \mathrm{e}^{\mathrm{i}\omega g(-1)} \Bigg\}, \end{aligned} \tag{4}$$

where

$$\mu_0(\omega) = \int_{-1}^{1} \mathrm{e}^{\mathrm{i}\omega g(x)} \mathrm{d}x = \mathscr{O}(\omega^{-1/2})$$

and

$$\tilde{\sigma}_m[h](x) = h(x), \qquad \tilde{\sigma}_m[h](x) = \frac{\mathrm{d}}{\mathrm{d}x} \frac{\tilde{\sigma}_{m-1}[h](x) - \tilde{\sigma}_{m-1}[h](c)}{g'(x)}, \quad m \geq 1$$

[11]. Note that the functions $\tilde{\sigma}_m$ are $\mathbf{C}^\infty[-1, 1]$, since the singularity at $x = c$ is removable. This removable singularity is the reason why, while $\sigma_m[h](x)$ is a linear combination of $h^{(j)}(x), j = 0, \ldots, m$, for $x \in [-1, 1] \setminus \{c\}$, at $x = c$ we have a linear combination of $h^{(j)}(c), j = 0, \ldots, 2m$. The clear implication is that once, *in addition to* (3), we also impose the interpolation conditions

$$\tilde{p}^{(j)}(c) = f^{(j)}(c), \qquad j = 0, 1, \ldots, 2s-2,$$

the plain-vanilla Filon method bears an error of $\tilde{O}(\omega^{-s-1/2})$ for $\omega \gg 1$.

For reasons that will become apparent in the sequel, it is important to consider also the case when c is at an endpoint: without loss of generality we let $c = -1$. In that case (4) need be replaced by

$$\begin{aligned} I_\omega[\tilde{p}-f] \sim \mu_0(\omega)\sum_{m=0}^{\infty}\frac{\tilde{\sigma}_m[\tilde{p}-f](-1)}{(-\mathrm{i}\omega)^m} \\ -\sum_{m=0}^{\infty}\frac{1}{(-\mathrm{i}\omega)^{m+1}}\left\{\frac{\tilde{\sigma}_m[\tilde{p}-f](1)-\tilde{\sigma}_m[\tilde{p}-f](-1)}{g'(1)}\mathrm{e}^{\mathrm{i}\omega g(1)}\right. \\ \left. -\frac{\tilde{\sigma}'_m[\tilde{p}-f](-1)}{g''(-1)}\mathrm{e}^{\mathrm{i}\omega g(-1)}\right\} \end{aligned}$$

and $\tilde{\sigma}'_m(-1)$ is a linear combination of $h^{(j)}(-1), j = 0, \dots, 2m+1$ [8].

A plain-vanilla Filon method can be also implemented in a derivative-free manner, e.g. when the derivatives of f are unknown or not easily available. In that case we need to replace derivatives by finite differences with an $\mathcal{O}(\omega^{-1})$ spacing and this procedure does not lead to loss of asymptotic accuracy [10]. In particular, in place of (3), we may interpolate at the points

$$c_k(\omega) = \begin{cases} -1+\frac{\theta k}{\omega+1}, & k = 0, \dots, s-1, \\ 1-\frac{\theta(2s-k-1)}{\omega+1}, & k = s, \dots, 2s-1, \end{cases} \tag{5}$$

where the denominator $\omega + 1$ ensures that the interpolation points do not blow up near $\omega = 0$, while $0 < \theta < (s-1)^{-1}$ implies that the interpolation points are all distinct and live in $[-1, 1]$.

As an example, consider $f(x) = (1 + x + x^2)^{-1}$, $g(x) = x$ in (1). In Fig. 1 we plot on the left the interpolation points (5) with $s = 5$. The errors committed by Filon methods for $s = 2$ (hence with an asymptotic error decay of $\mathcal{O}(\omega^{-3})$) based on (3) and (5) are displayed on the right in logarithmic scale. As can be seen, that the points (5) are equidistant at $\omega = 0$ and bunch at the endpoints when ω increases. The derivative-free Filon method (5) (black dotted line) has essentially the same good behaviour as (3) (green solid line) for large ω.

The addition of extra interpolation points to (3) (or, for that matter, (5)) can be highly beneficial in reducing an error committed by a Filon method. Specifically, in the $g' \neq 0$ case, we choose distinct *inner nodes* $c_1, \dots, c_\nu \in (-1, 1)$ and impose that $2s + \nu$ interpolation conditions

$$\begin{aligned} &p^{(j)}(1) = f^{(j)}(1), \quad p^{(j)}(-1) = f^{(j)}(-1), \qquad j = 0, 1, \cdots, s-1, \\ &p(c_k) = f(c_k), \qquad k = 1, \cdots, \nu. \end{aligned} \tag{6}$$

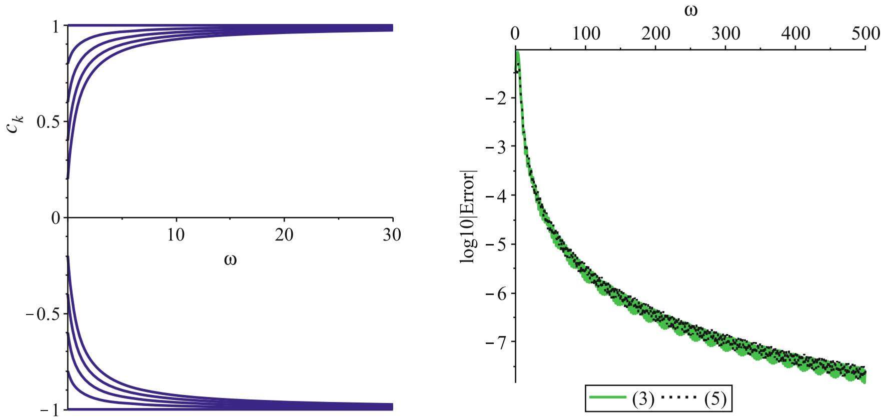


Fig. 1 The left: the interpolation points $c_k(\omega)$ of (5) with $k = 0, \cdots, 2s-1$ (from the bottom line to top) for $s = 5, \theta = \frac{1}{5}$ and $\omega \in [0, 30]$; The right: the logarithm (to base 10) of the error of both Filon methods, (3) (green solid line) and (5) (black dotted line), for $f(x) = (1 + x + x^2)^{-1}$, $g(x) = x$, $s = 2$ and $\omega \in [0, 500]$

This is the *Extended Filon Method (EFM),*

$$Q_\omega^{\mathsf{F},s,\nu}[f] = \int_{-1}^{1} p(x)\mathrm{e}^{\mathrm{i}\omega g(x)}\mathrm{d}x$$

that has been carefully analysed in [7, 8]. Different choices of internal nodes result in different behaviour for small $\omega \geq 0$ or in greater simplicity in implementation although, for large ω, the rate of asymptotic decay of the error is always $\mathscr{O}(\omega^{-s-1})$. In particular, [8] examined two choices of internal nodes: Zeros of the Jacobi polynomial $\mathscr{P}_\nu^{(s,s)}$ and Clenshaw–Curtis points. In the first instance we have the best-possible behaviour for $\omega = 0$ and in the second the coefficients are substantially simpler and, for large ν can be evaluated in just $\mathscr{O}(\nu \log \nu)$ operations.

Regardless of the choice of internal nodes, the leading term of the asymptotic error can be expressed as

$$\begin{aligned} &Q_\omega^{\mathsf{F},s,\nu}[f] - I_\omega[f] \\ \sim &-\frac{1}{(-\mathrm{i}\omega)^{s+1}}\left[\frac{f^{(s)}(1) - \tilde{p}^{(s)}(1)}{g'^{s+1}(1)}\mathrm{e}^{\mathrm{i}\omega g(1)} - \frac{f^{(s)}(-1) - \tilde{p}^{(s)}(-1)}{g'^{s+1}(-1)}\mathrm{e}^{\mathrm{i}\omega g(-1)}\right] \\ &+ \mathscr{O}\big(\omega^{-s-2}\big). \end{aligned} \tag{7}$$

Similar formula applies in the presence of stationary points: quadrature error is reduced to interpolation error at the endpoints and stationary points. This error, in turn, can be analysed very precisely using the Peano Kernel Theorem [8] and the decrease in asymptotic error (as distinct to the asymptotic *rate of decay* of the error) can be very substantial.

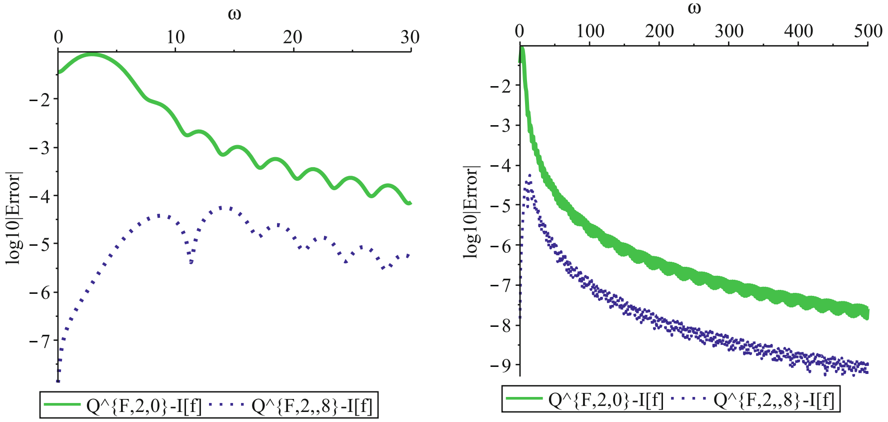


Fig. 2 The logarithmic error $\log_{10}|Q_\omega^{\mathsf{F},2,0}[f]-I[f]|$ (the lime green solid line, the top) and $\log_{10}|Q_\omega^{\mathsf{F},2,8}[f]-I[f]|$ (the dark blue dotted line, the bottom) for $f(x)=(1+x+x^2)^{-1}$, $g(x)=x$, $s=2$, $\omega\in[0,30]$ (the left) and $\omega\in[0,500]$ (the right)

To illustrate this we revisit the example from Fig. 1. Logarithmic errors of plain-vanilla Filon (the lime green solid line) and EFM with Jacobi points (the dark blue dotted line) are displayed in Fig. 2 with $s=2$ and $\nu=8$. It can be observed that the rates of decay between plain-vanilla Filon and EFM are very different. For small ω, EFM is definitely superior by design, while as ω increases both of them decay as the asymptotic order $\mathcal{O}\big(\omega^{-3}\big)$ but EFM has much smaller error.

Based on the above research, it is legitimate to ask *what is the optimal choice of internal nodes*. In reality, these are two questions. If we are concerned with choosing the same nodes for all ω then the two main choices in [8] are probably the best: if 'optimal' means the least uniform error then Jacobi wins but once we wish to optimize computation then Clenshaw–Curtis is the better choice. However, the situation is entirely different once the c_ks are allowed to depend on ω. Now the answer is clear at the 'extremities':

- For $\omega=0$ the optimal choice is Legendre points, lending themselves to classical Gaussian quadrature;
- For $\omega\gg 1$ the optimal choice maximizes the asymptotic rate of error decay, whereby (5) emerges as the natural preference.

The challenge, though, is to bridge $\omega=0$ with $\omega\gg 1$, and this forms the core of this paper.

This is the place to mention a recent paper of Zhao and Huang [16], which combines the plain Filon with Exponentially Fitted method (EF), to propose an alternative version of adaptive Filon method. For large ω, the nodes in [16] are reduced to $\mp 1\pm\frac{k}{\omega}$, which is similar to our method, inspired by plain Filon method in [10]. For small ω, since the EF method introduces complex points, the method of [16] employs complex nodes derived from the computation of asymptotic expansion. The outcome is considerably more complicated and restricted to integrals without stationary points. The algorithm in this paper employs altogether different strategy for small ω. To connect the optimal nodes between Gauss–Legendre points when $\omega = 0$ and $\mp\left(1 - \frac{\theta k}{1+\omega}\right)$ of large ω, the real nodes dependent ω are presented by constructing Filon homotopy. Moreover, our method is extended to the case of stationary points.

In Sect. 2 we discuss different choices of *homotopy functions,* connecting Gaussian weights for $\omega = 0$ and points (5) for $\omega \gg 1$ in the absence of stationary points. Numerical experiments are provided to illustrate the effectiveness of the adaptive method. The adaptive approach to the Filon method is extended in Sect. 3 to the case of stationary points. Finally, in Sect. 4 we discuss the advantages and limitations of this approach.

2 Adaptive Filon Method Without Stationary Points

2.1 The Construction of ω-Dependent Interpolation Points

Throughout this section we assume that (1) has no stationary points, i.e. that $g' \neq 0$ in $[-1, 1]$. We define the vector function $\mathbf{c}(\omega) = \{c_k(\omega)\}_{k=0}^{2s-1}$ as *Filon homotopy* once it obeys the following conditions:

1. Each c_k is a piecewise-smooth function of $\omega \geq 0$;
2. $c_k(0) = \xi_{k+1}^{(2s)}$, the $(k+1)$st zero of the Legendre polynomial P_{2s} (in other words, the $(k+1)$st Gauss–Legendre point), arranged in a monotone order;
3. $$c_k(\omega) = \begin{cases} -1 + \dfrac{\theta k}{\omega + 1}, & k = 0, \ldots, s-1, \\ 1 - \dfrac{\theta(2s-k-1)}{\omega+1}, & k = s, \ldots, 2s-1 \end{cases} + \mathscr{O}(\omega^{-2}), \qquad \omega \gg 1,$$
   where $0 < \theta < (s-1)^{-1}$;
4. For every $\omega \geq 0$
   $$-1 \leq c_0(\omega) < c_1(\omega) < \cdots < c_{2s-1}(\omega) \leq 1.$$

In other words, $\mathbf{c}$ is a vector of s trajectories connecting Gauss–Legendre points with (5), all distinct and living in $[-1, 1]$.

A convenient way to construct Filon homotopy is by choosing any piecewise-smooth weakly monotone function κ such that $\kappa(0) = 1$, $\kappa(\omega) = \mathscr{O}(\omega^{-2})$ (or smaller) for $\omega \gg 1$ (therefore $\lim_{\omega\to\infty} \kappa(\omega) = 0$), and setting

$$c_k(\omega) = \xi_{k+1}^{(2s)}\kappa(\omega) + \varphi_k(\omega)[1 - \kappa(\omega)], \qquad k = 0, \ldots, 2s-1, \tag{8}$$

where

$$\varphi_k(\omega) = \begin{cases} -1 + \dfrac{\theta k}{\omega + 1}, & k = 0, \ldots, s-1, \\ 1 - \dfrac{\theta(2s-k-1)}{\omega + 1}, & k = s, \ldots, 2s-1. \end{cases}$$

It is easy to prove that conditions 1–4 are satisfied and (8) is a Filon homotopy.

To illustrate our argument and in search for a 'good' Filon homotopy, we consider four functions κ,

a. $\kappa_1(\omega) = \text{Heaviside}(10 - \omega)$, where

$$\text{Heaviside}(y) = \begin{cases} 1, & y \geq 0, \\ 0, & y < 0 \end{cases}$$

is the Heaviside function;

b. $\kappa_2(\omega) = (1 + \omega^2)^{-1}$;

c. $\kappa_3(\omega) = 2/\left[1 + \exp\left(\log^4(1+\omega)\right)\right]$;

d. $\kappa_4(\omega) = \cos\left(\frac{\pi}{2}\frac{e^{\omega/2}-1}{256+e^{\omega/2}}\right)$.

Figure 3 displays the four functions κ but perhaps more interesting is Fig. 4, where we depict the homotopy curves $c_k(\omega)$ of (8) for the four choices of κ and $s = 4$. κ_1 essentially stays put at Gauss–Legendre points until $\omega = 10$ and then jumps to the points (5), while κ_4 represents a smooth approximation to κ_1. κ_2 and κ_3 abandon any memory of Gauss–Legendre points fairly rapidly, implicitly assuming very early onset of asymptotic behaviour in the integral (1).

To gain basic insight into the differences among the functions κ_j, we have applied them to the evaluation of the integral

$$\int_{-1}^{1} \frac{e^{i\omega x}dx}{1 + x + x^2} \tag{9}$$

using ten function evaluations and letting $\theta = 1/s$. To set the stage, in Fig. 5 we have calculated the integral using five different Extended Filon–Jacobi methods (6) with $\nu = 10-2s$ referenced from [8]: (1) $s = 1$, $\nu = 8$; (2) $s = 2$, $\nu = 6$; (3) $s = 3$, $\nu = 4$, (4) $s = 4$, $\nu = 2$ and (5) $s = 5$, $\nu = 0$. The errors (to logarithmic scale) are displayed separately for $\omega \in [0, 20]$ and $\omega \in [0, 200]$.

So far, the figure is not very surprising and we recall from the previous section that "large s, small ν" strategy is better for $\omega \gg 1$, while "small s, large ν" wins for small $\omega \gg 0$. However, let us instead solve (10) with adaptive Filon, using one of the four κ_j functions above. Again, we need to distinguish between small and

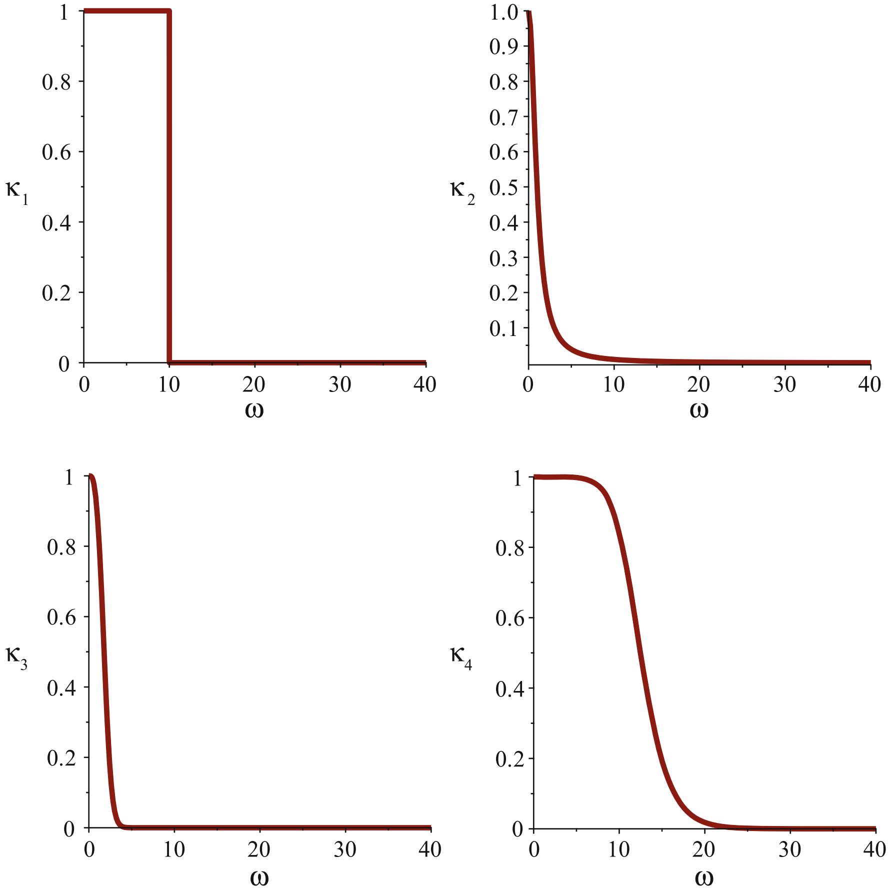


Fig. 3 The functions $\kappa_j, j = 1, 2, 3, 4$ (from the left to right)

large ω and the corresponding plots are Figs. 6 and 7 respectively. It is clear that for large ω there is little to distinguish adaptive Filon from EFJ with $s = 5$ (which is also plain Filon): everything in this regime is determined by asymptotic analysis and the only relevant observation is that nothing of essence is lost once we replace derivatives by suitable finite differences. The big difference is for small $\omega \geq 0$, before the onset of asymptotics. At $\omega = 0$ all four methods use Gauss–Legendre points and the error beats even EFJ with $\nu = 8$, which corresponds to Lobatto points. However, the errors for κ_2 and κ_3 deteriorate rapidly and this is explained by the homotopy curves in Fig. 4, because interpolation points very rapidly move to their 'asymptotic regime'. κ_1 and κ_4 are much better, except that κ_1 has an ungainly jump at $\omega = 10$, a consequence of its discontinuity, while κ_4 seems to be the winner. Similar outcome is characteristic to all other numerical experiments that we have undertook.

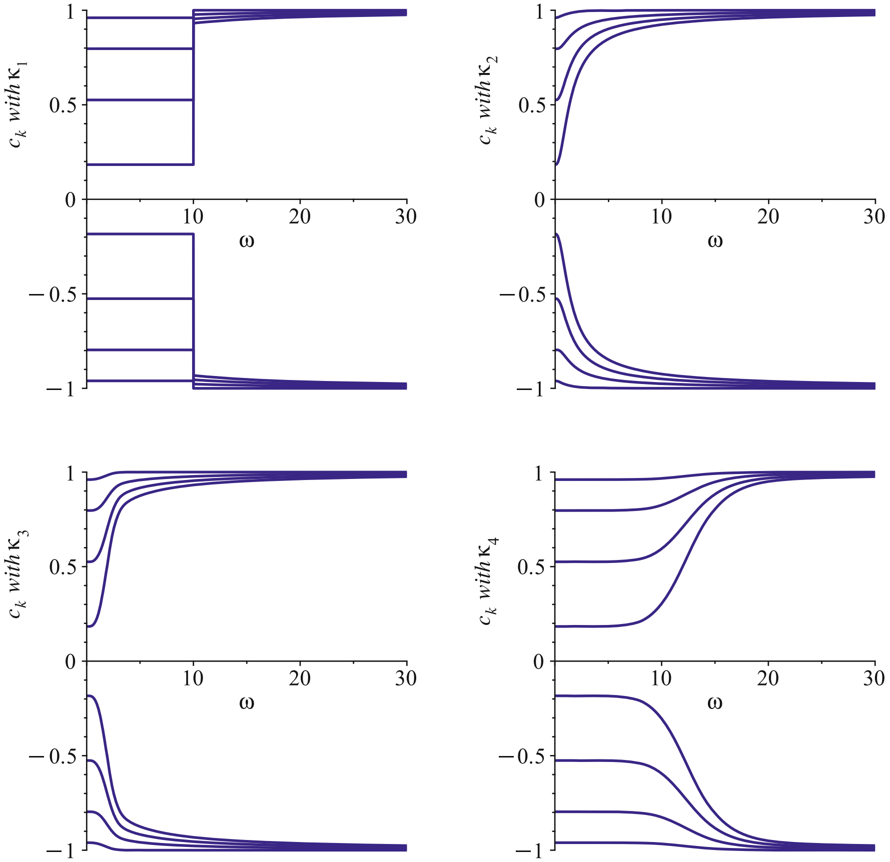


Fig. 4 Homotopy curves (8) $c_k(\omega)$, $k = 0, \cdots, 2s - 1$ (from the bottom to top line) with $s = 4$ for each functions $\kappa_j, j = 1, 2, 3, 4$ (from the left to right)

Another interpretation of κ_4 is that it tends to represent for every ω the best outcome for *any* EFJ with the same number of function evaluations. In other words, denoting the error of EFJ with $\nu = 10 - 2s$ by $e_\omega^{[s]}$ (the dark blue dotted line) and the error of adaptive Filon by $\tilde{e}_\omega$ (the orange red solid line) derived by κ_4, we plot in Fig. 8

$$\log_{10} \left|\min\{|e_\omega^{[j]}| : j = 1, \ldots, 5\}\right| \qquad \text{and} \qquad \log_{10} |\tilde{e}_\omega|.$$

For larger values of ω the two curves overlap to all intents and purposes. For small ω, though, adaptive Filon is better than the best among the different EFJ schemes—the difference is directly attributable to Gauss–Legendre points being superior to Lobatto points.

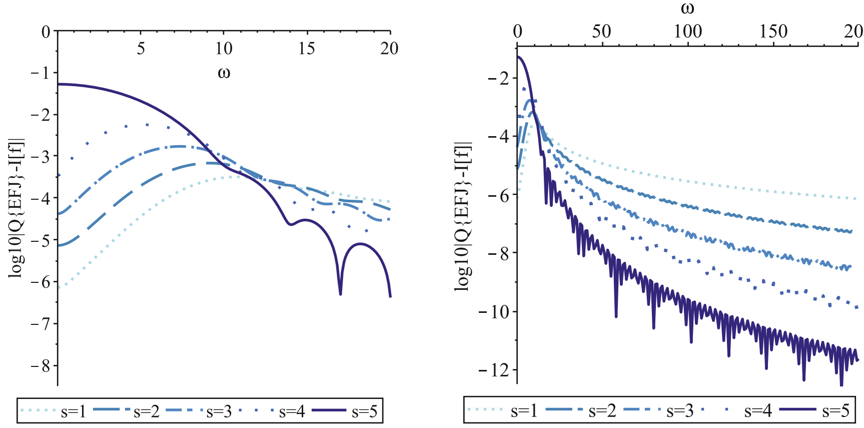


Fig. 5 Logarithmic errors for EFJ, applied to (9), with ten function evaluations: The lines corresponding to s vary in shades of blue between 1 (light) and 5 (dark), as well as in the line style, with $\nu = 10 - 2s$

The function κ_4 is a special case of

$$\kappa_{a,b}(\omega) = \cos\left(\frac{\pi}{2}\,\frac{\mathrm{e}^{a\omega}-1}{b+\mathrm{e}^{a\omega}}\right), \tag{10}$$

using $a = \frac{1}{2}$ and $b = 256$. In general, any $\kappa_{a,b}$ with small $a > 0$ and large $b > 0$ obeys the conditions for a Filon homotopy and, in addition, exhibits favourable behaviour—essentially, it is a smooth approximation to a Heaviside function, allowing for Gauss–Legendre points seamlessly segueing into (5), a finite-difference approximation of derivatives at the endpoints.

What is the optimal function κ? Clearly, this depends on the functions f and g, as does the pattern of transition from 'small ω' to asymptotic behaviour. Our choice, $\kappa_{\frac{1}{2},256}$, is in our experience a good and practical compromise.

2.2 *The Adaptive Filon Algorithm*

Let us commence by gathering all the threads into an algorithm. Given the integral (1) (without stationary points) and a value of ω,

1. Compute the interpolation points $c_0, \ldots, c_{2s-1}$ using $\theta = 1/s$, (8) and $\kappa = \kappa_{\frac{1}{2},256}$ given by (10).
2. Evaluate the polynomial $\tilde{p}$ of degree $2s-1$ which interpolates f at $c_0, \ldots, c_{2s-1}$.
3. Calculate

$$\mathscr{Q}_{\omega}^{\mathsf{AF},s}[f] = \int_{-1}^{1} \tilde{p}(x)\mathrm{e}^{\mathrm{i}\omega g(x)}\mathrm{d}x. \tag{11}$$

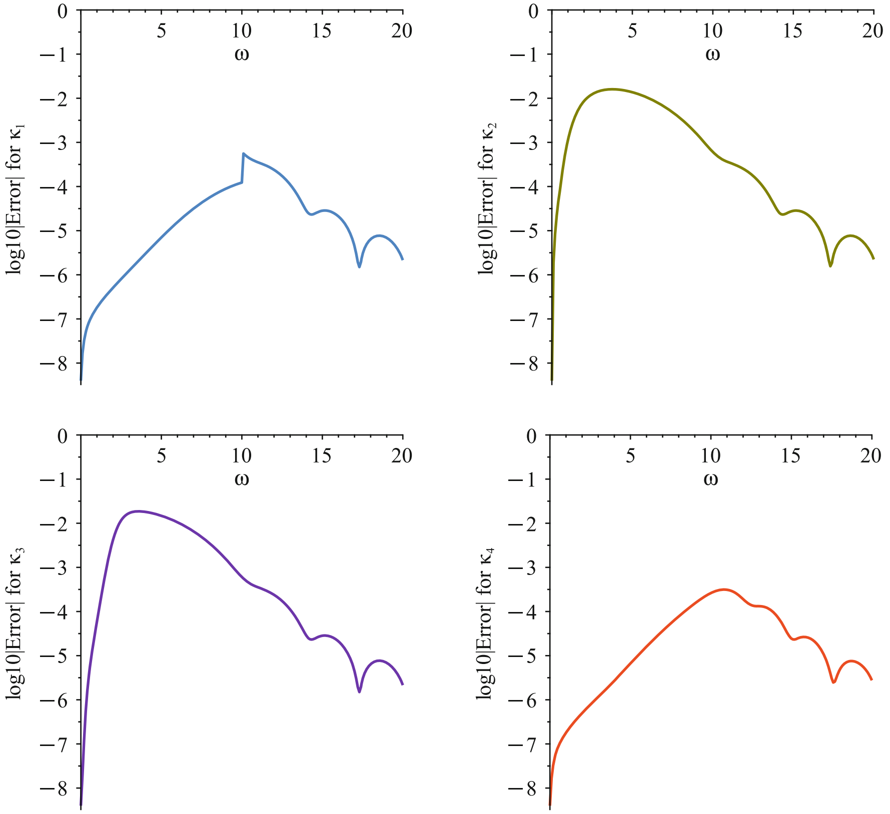


Fig. 6 Logarithmic errors for adaptive Filon, applied to (9), with ten function evaluations, $\theta = \frac{1}{5}$, $\omega \in [0, 20]$ and $\kappa_j, j = 1, 2, 3, 4$ (top left to bottom right)

Proposition 1 *The asymptotic error of the adaptive Filon method* $\mathscr{Q}_\omega^{\mathsf{AF},s}[f]$ *is* $\mathscr{O}\left(\omega^{-s-1}\right)$.

Proof For a fixed ω, adaptive Filon is a special case of EFM with derivatives at the endpoints replaced by suitable finite differences—we already know from [10] that this is consistent with the stipulated asymptotic behaviour. □

Alternatively, we can prove the proposition acting directly on the error term (7), this has the advantage of resulting in an explicit expression for the leading error term.

Needless to say, Proposition 1 represents just one welcome feature of adaptive Filon. The other is that it tends to deliver the best uniform behaviour for all $\omega \geq 0$.

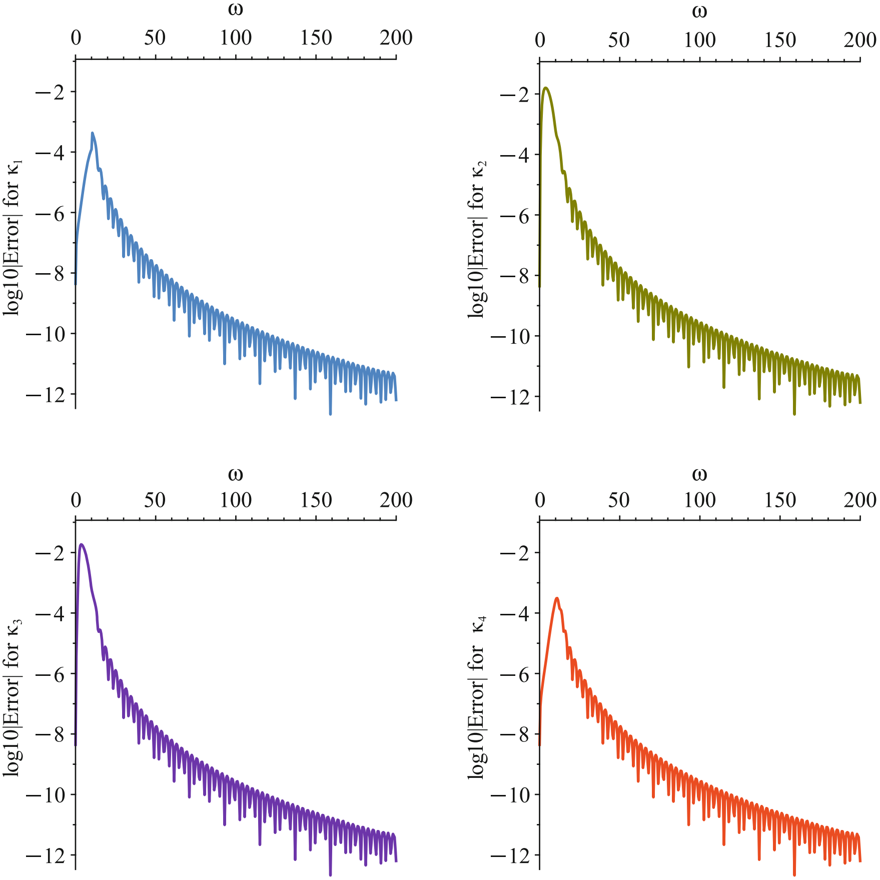


Fig. 7 Logarithmic errors for Adaptive Filon, applied to (9), with ten function evaluations, $\theta = \frac{1}{5}$, $\omega \in [0, 200]$ and $\kappa_j, j = 1, 2, 3, 4$ (top left to bottom right)

3 Stationary Points

Let us suppose that g' vanishes at $r \geq 1$ points in $[-1, 1]$. We split the interval into subintervals I_k such that in each $I_k = [\alpha_k, \beta_k]$ there is a single stationary point residing *at one of the endpoints*—it is trivial to observe that there are at least $\max\{1, 2r - 2\}$ and at most $2r$ such subintervals. We use a linear transformation to map each I_k to the interval $[-1, 1]$ so that the stationary point resides at -1:

$$\text{Stationary point at } \alpha_k : \; x \to \frac{2x - (\beta_k + \alpha_k)}{\beta_k - \alpha_k},$$

$$\text{Stationary point at } \beta_k : \; x \to -\frac{2x - (\beta_k + \alpha_k)}{\beta_k - \alpha_k}.$$

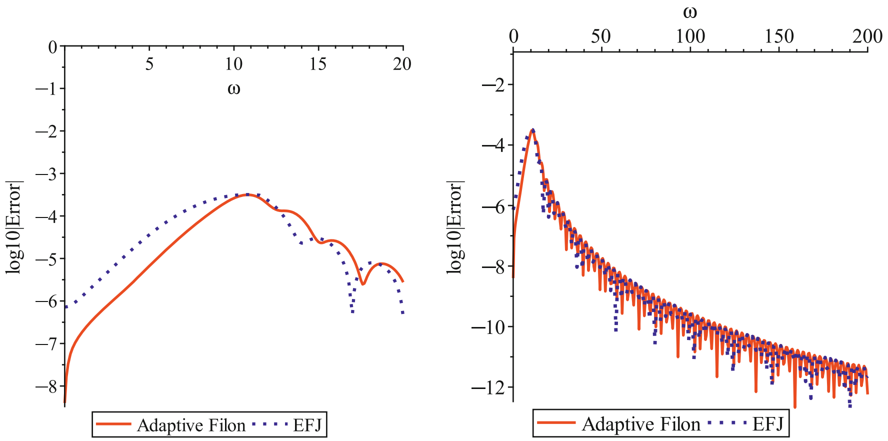


Fig. 8 A comparison between adaptive Filon (the orange red solid line) and the pointwise best scheme among different EFJ methods (the dark blue dotted line)

We thus reduce the task at hand into a number of computations of (1) with a single stationary point at $x = -1$.

In the sequel we assume that -1 is a simple stationary point, i.e. that $g'(-1) = 0$ and $g''(-1) \neq 0$. The extension of our narrative to higher-order stationary points is straightforward.

We commence with the EFM method and recall from [8] its asymptotic expansion,

$$I_\omega[f] \sim \mu_0(\omega) \sum_{m=0}^{\infty} \frac{\rho_m[f](-1)}{(-\mathrm{i}\omega)^m} - \sum_{m=0}^{\infty} \frac{1}{(-\mathrm{i}\omega)^{m+1}} \left[\frac{\rho_m[f](1) - \rho_m[f](-1)}{g'(1)} \mathrm{e}^{\mathrm{i}\omega g(1)} - \frac{\rho_m'[f](-1)}{g''(-1)} \mathrm{e}^{\mathrm{i}\omega g(-1)} \right], \tag{12}$$

where

$$\mu_0(\omega) = \int_{-1}^{1} \mathrm{e}^{\mathrm{i}\omega g(x)} \mathrm{d}x,$$

$$\rho_0[f](x) = f(x) \qquad \rho_m[f](x) = \frac{\mathrm{d}}{\mathrm{d}x} \frac{\rho_{m-1}[f](x) - \rho_{m-1}[f](-1)}{g'(x)}, \quad m \geq 0.$$

We recall that $\mu_0(\omega) = \int_{-1}^{1} \mathrm{e}^{\mathrm{i}\omega g(x)} \mathrm{d}x \sim \mathscr{O}(\omega^{-1/2})$ and that $\sigma_m[f](1)$ is a linear combination of $f^{(j)}(1)$, $j = 0, \ldots, m$, while $\sigma_m[f]'(-1)$ is a linear combination of $f^{(j)}(-1)$, $j = 0, \ldots, 2m+1$. Putting all this together, we need to impose the interpolation conditions

$$\begin{aligned} p^{(k)}(-1) &= f^{(k)}(-1), \qquad k = 0, \ldots, 2s, \\ p^{(k)}(1) &= f^{(k)}(1), \qquad k = 0, \ldots, s-1, \end{aligned} \tag{13}$$

to ensure that the error of (12) is $\mathscr{O}(\omega^{-s-1})$. (Alternatively, we can interpolate at -1 up to $j = 2s-1$, resulting in an asymptotic error of $\mathscr{O}(\omega^{-s-1/2})$—we do not pursue this route here.) Alternatively to (13) (and the proof is identical to the case when stationary points are absent), we can take a leaf off (5) and interpolate at

$$\varphi_k(\omega) = -1 + \frac{\theta k}{\omega+1}, \qquad k = 0, \ldots, 2s, \tag{14}$$

$$\varphi_k(\omega) = 1 - \frac{\theta(3s-k)}{\omega+1}, \qquad k = 2s+1, \ldots, 3s, \tag{15}$$

where $\theta < 2/(3s-1)$ ensures that all interpolation points are distinct, by a polynomial $\tilde{p}$ of degree $3s$. This gives a derivative-free Filon *á la* [10]. To extend this to adaptive Filon we need to use (8) again by replacing the superscript $2s$ by $3s+1$, blending the φ_ks with Gauss–Legendre points and employing $\kappa = \kappa_{\frac{1}{2},256}$. The outcome is no longer symmetric, as demonstrated in Fig. 9, but this should cause no alarm.

The construction of adaptive Filon proceeds exactly along the same lines as when stationary points are absent. All that remains is to present a numerical example:

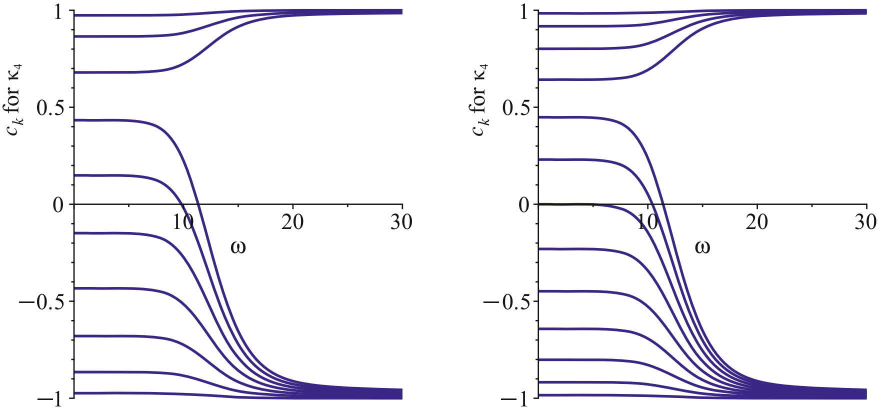


Fig. 9 Homotopy curves $c_k(\omega)$, $k = 0, \cdots, 3s$ (from the bottom to top line), for (1) $s = 3$, $\theta = 2/9$ (the left) and (2) $s = 4$, $\theta = \frac{1}{6}$ (the right)

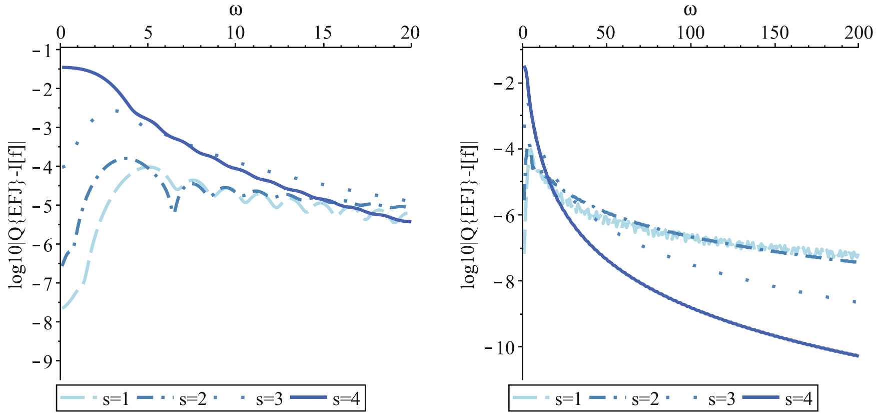


Fig. 10 Logarithmic errors for EFJ, applied to (16), with 13 function evaluations: s varies between 1 (light) and 4 (dark), with $\nu = 12 - 3s$. The colours correspond to different values of s: the larger s, the darker the colour

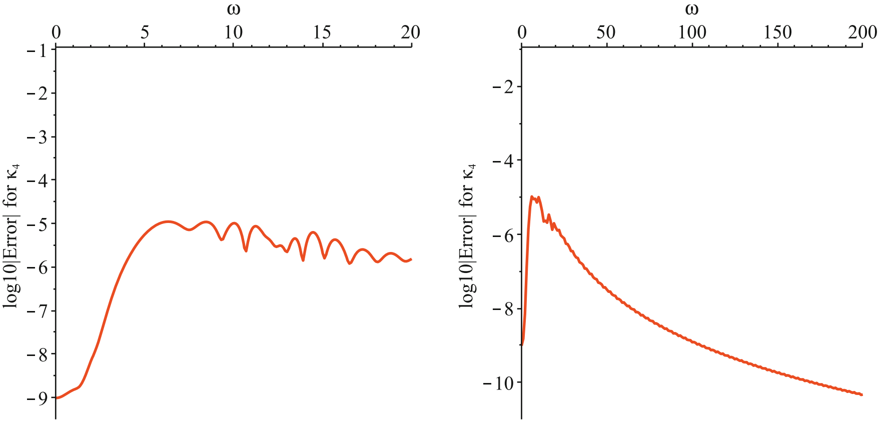


Fig. 11 Logarithmic errors for adaptive Filon, applied to (16), with 13 function evaluations

instead of (9), we consider

$$\int_{-1}^{1} \frac{e^{i\omega(x+1)^2}dx}{1+x+x^2} \tag{16}$$

and present the counterparts of Figs. 5, 6, 7, and 8, except that we plot only the results for $\kappa = \kappa_4 = \kappa_{\frac{1}{2},256}$. All figures compare an implementation with 13 function evaluations.

It is vividly clear from Figs. 10, 11, and 12 that, again, adaptive Filon represents the best of all worlds: for small ω is it as good as Gaussian quadrature, for large ω it matches plain Filon and in the intermediate interval it converts smoothly and seamlessly between these two regimes.

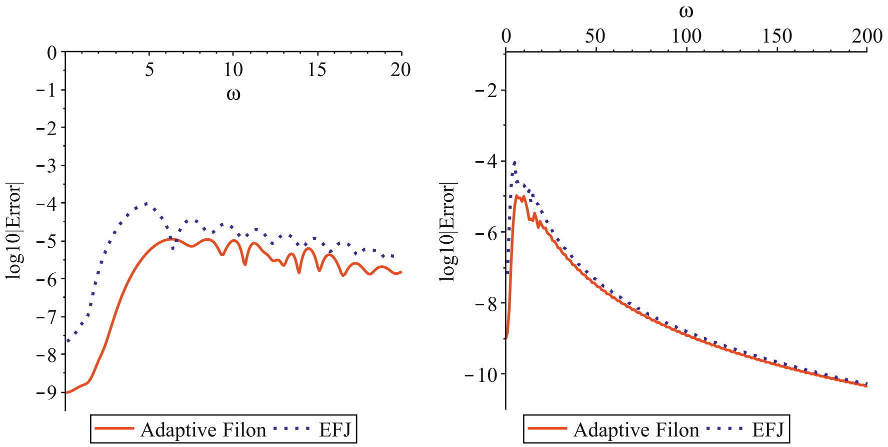


Fig. 12 A comparison between adaptive Filon (the orange red solid line) and the pointwise best scheme among different EFJ methods (the dark blue dotted line) for 13 function evaluations

4 Conclusions

In this paper, we have developed an adaptive Filon method for the computation of a highly oscillatory integral with or without stationary points. The main feature of this method is that it optimises the choice of interpolation points between different oscillatory regimes relative to those EFM based on the analysis in [8].

Is adaptive Filon the best-possible implementation of the 'Filon concept', a method for all seasons? Not necessarily! To define 'best' we must first define the purpose of the exercise. If the main idea is to compute (1) for a small number of values of ω and we cannot say in advance whether these values live in a highly oscillatory regime (or if we wish a method which is by design good uniformly for al $\omega \geq 0$) then adaptive Filon definitely holds the edge in comparison to other implementations of the Filon method, in particular to Extended Filon. However, the method is not competitive once we require the computation of a very large number of integrals for the same function f but many different values of ω. The reason is simple. Conventional Filon methods use interpolation points which are independent of ω, hence we need to compute the values of f (or its derivatives) and form an interpolating polynomial just once: it can be reused by any number of values of ω. Adaptive Filon, though, re-evaluates f afresh for every ω and subsequently forms a new interpolating polynomial. Thus, increased accuracy and better uniform behaviour are offset by higher cost.

Numerical methods must be always used with care and claims advanced on their behalf must be responsible. Adaptive Filon is probably optimal in the scenario when just few values of (1) need be computed but considerably more expensive once a multitude of computations with different values of ω is sought.

Acknowledgements The work of the first author has been supported by the Projects of International Cooperation and Exchanges NSFC-RS (Grant No. 11511130052), the Key Science and Technology Program of Shaanxi Province of China (Grant No. 2016GY-080) and the Fundamental Research Funds for the Central Universities.

References

1. Asheim, A., Huybrechs, D.: Complex Gaussian quadrature for oscillatory integral transforms. IMA J. Numer. Anal. **33**(4), 1322–1341 (2013)
2. Chandler-Wilde, S.N., Graham, I.G., Langdon S., Spence, E.A.: Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. Acta Numer. **21**, 89–305 (2012)
3. Deaño, A., Huybrechs, D., Iserles, A.: The kissing polynomials and their Hankel determinants. Technical Report, DAMTP, University of Cambridge (2015)
4. Domínguez, V., Ganesh, M.: Interpolation and cubature approximations and analysis for a class of wideband integrals on the sphere. Adv. Comput. Math. **39**(3–4), 547–584 (2013)
5. Domínguez, V., Graham, I.G., Smyshlyaev, V.P.: Stability and error estimates for Filon–Clenshaw–Curtis rules for highly oscillatory integrals. IMA J. Numer. Anal. **31**(4), 1253–1280 (2011)
6. Ganesh, M., Langdon, S., Sloan, I.H.: Efficient evaluation of highly oscillatory acoustic scattering surface integrals. J. Comput. Appl. Math. **204**(2), 363–374 (2007)
7. Gao, J., Iserles, A.: A generalization of Filon–Clenshaw–Curtis quadrature for highly oscillatory integrals. BIT Numer. Math. **57**, 943–961 (2017)
8. Gao, J., Iserles, A.: Error analysis of the extended Filon-type method for highly oscillatory integrals. Res. Math. Sci. **4**(21/24) (2017). https://doi.org/10.1186/s40687-017-0110-4
9. Huybrechs, D., Vandewalle, S.: On the evaluation of highly oscillatory integrals by analytic continuation. SIAM J. Numer. Anal. **44**(3), 1026–1048 (2006)
10. Iserles, A., Nørsett, S.P.: On quadrature methods for highly oscillatory integrals and their implementation. BIT Numer. Math. **44**(4), 755–772 (2004)
11. Iserles, A., Nørsett, S.P.: Efficient quadrature of highly oscillatory integrals using derivatives. Proc. R. Soc. London Ser. A Math. Phys. Eng. Sci. **461**(2057), 1383–1399 (2005)
12. Levin, D.: Fast integration of rapidly oscillatory functions. J. Comput. Appl. Math. **67**(1), 95–101 (1996)
13. Olver, S.: Moment-free numerical integration of highly oscillatory functions. IMA J. Numer. Anal. **26**(2), 213–227 (2006)
14. Trevelyan, J., Honnor, M.E.: A numerical coordinate transformation for efficient evaluation of oscillatory integrals over wave boundary elements. J. Integral Equ. Appl. **21**(3), 447–468 (2009)
15. Van't Wout, E., Gélat, P., Timo, B., Simon, A.: A fast boundary element method for the scattering analysis of high-intensity focused ultrasound. J. Acoust. Soc. Am. **138**(5), 2726–2737 (2015)
16. Zhao, L.B., Huang, C.M.: An adaptive Filon-type method for oscillatory integrals without stationary points. Numer. Algorithms **75**(3), 753–775 (2017)

MLMC for Nested Expectations

Michael B. Giles

Dedicated to Ian H. Sloan on the occasion of his 80th birthday, with warm appreciation for the way in which he welcomed me into the MC/QMC community and introduced me to QMC methods when I switched research fields in 2006.

Abstract This paper discusses progress and future research possibilities in applying MLMC ideas to nested expectations of the form $\mathbb{E}[\, g(\mathbb{E}[f(X,Y)|X])\,]$, with an outer expectation with respect to one random variable X, and an inner conditional expectation with respect to a second random variable Y. The difficulty in treating such applications is shown to depend on whether the function g is (1) smooth, (2) continuous and piecewise smooth, or (3) discontinuous.

1 Introduction

Considerable progress has been achieved over the past 10 years in the development, application and analysis of Multilevel Monte Carlo (MLMC) methods, applied to SDEs, SPDEs, continuous-time Markov processes, and a range of other stochastic models; see [10] and references therein.

This paper discusses an area of active research, the application of MLMC ideas to nested simulations, in which one is interested in estimating quantities of the form $\mathbb{E}[\, g(\mathbb{E}[f(X,Y)|X])\,]$ with an outer expectation with respect to one random variable X, and an inner conditional expectation with respect to a second random variable Y.

Such nested expectations arise in a number of applications; the two applications motivating the author's research are the evaluation of Expected Value of Partial Perfect Information (EVPPI) and Value-at-Risk (VaR).

M. B. Giles (✉)
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: mike.giles@maths.ox.ac.uk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_20

EVPPI arises in fields such as medicine [1, 4] and the exploration and exploitation of oil and gas reservoirs [3, 24], the common element being decision making under a large degree of uncertainty. In the medical case, models of the effectiveness of different medical treatments are based on a number of uncertain parameters which we group into two independent sets X and Y. Given no knowledge of X and Y other than that they come from prescribed probability distributions, then given a finite set of possible treatments D, the optimal choice d_{opt} is the one which maximises $\mathbb{E}\left[f_d(X,Y)\right]$ where $f_d(X,Y)$ represents some measure of the patient outcome, such as QALY's (quality-adjusted life-year, see Wikipedia), with a larger value being better. Thus, with no knowledge, the expected optimal outcome is $\max_d \mathbb{E}\left[f_d(X,Y)\right]$. On the other hand, given perfect information on X, Y due to additional medical research, the best treatment maximises $f_d(X,Y)$, giving the overall expected outcome $\mathbb{E}\left[\max_d f_d(X,Y)\right]$. In the intermediate situation, if X is determined but not Y, then the best treatment has expected outcome value $\mathbb{E}\left[\max_d \mathbb{E}\left[f_d(X,Y)\,|\,X\right]\right]$. EVPI, the expected value of perfect information, is the difference

$$\text{EVPI} = \mathbb{E}[\max_d f_d(X,Y)] - \max_d\; \mathbb{E}[f_d(X,Y)],$$

and EVPPI, the expected value of partial perfect information, is the difference

$$\text{EVPPI} = \mathbb{E}[\max_d\; \mathbb{E}\left[f_d(X,Y)\,|\,X\right]] - \max_d\; \mathbb{E}[f_d(X,Y)].$$

EVPPI represents the benefit, on average, of knowing the value of X. This can be compared to the cost of the research required to determine X, to judge whether or not the research is cost-effective.

Value-at-Risk (VaR) is a financial risk measure used by investment banks [16, 17, 22, 23]. In this application, X represent a set of risk factors affecting the value of the bank's portfolio over some short risk horizon. For a given X, the loss in value of the portfolio is $L(X) \equiv \mathbb{E}[f(X,Y)|X]$ where the expectation corresponds to risk-neutral pricing, with Y representing the stochastic drivers for the behaviour of the underlying assets beyond the risk horizon. The objective with VaR is to compute the loss threshold L_α such that $\mathbb{P}(L(X)\geq L_\alpha) = \alpha$, for some small value of α. This defines L_α implicitly, but in this paper we will consider the simpler situation of a given threshold L^* and then computing $\mathbb{P}(L(X)\geq L^*) \;\equiv\; \mathbb{E}\left[\mathbf{1}_{\mathbb{E}[f(X,Y)|X]\geq L^*}\right]$. Hence in this case the function g is a discontinuous indicator function.

The paper begins with a quick review of MLMC and two important variants, the randomised unbiased MLMC method due to Rhee and Glynn [25], and the Multi-Index Monte Carlo (MIMC) method of Haji-Ali et al. [21]. Based on material in [10], Sect. 3 addresses the case in which the function g is smooth, using an antithetic estimator to achieve a faster rate of multilevel variance convergence. Section 4 addresses the EVPPI problem; a similar antithetic estimator is used but the convergence is poorer due to the lack of smoothness when there is a switch in the optimal decision. Section 5 addresses the VaR problem, and the difficulty in dealing with the discontinuous indicator function, and the paper finishes with a few concluding comments.

2 MLMC and Two Important Variants

2.1 MLMC

The central idea behind MLMC is very simple: given a sequence $P_0, P_1, \ldots$ which approximates a random output variable P with increasing accuracy, but also increasing cost, we have the simple identity

$$\mathbb{E}[P_L] \;=\; \mathbb{E}[P_0] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell - P_{\ell-1}] \;=\; \sum_{\ell=0}^{L} \mathbb{E}[\Delta P_\ell], \tag{1}$$

if we define $\Delta P_\ell \equiv P_\ell – P_{\ell-1}$ and $P_{-1} \equiv 0$. Therefore, if Z_ℓ is an unbiased estimator for $\mathbb{E}[\Delta P_\ell]$ then $\sum_{\ell=0}^{L} Z_\ell$ is an estimator for $\mathbb{E}[P_L]$.

Combining this with a geometric sequence of levels, and choosing the finest level L to control the magnitude of the weak error $\mathbb{E}[P_L - P]$, leads to the usual MLMC theorem in which we assume that there exist independent estimators Z_ℓ based on N_ℓ Monte Carlo samples, each with expected cost C_ℓ and variance V_ℓ, such that there are positive constants $\alpha, \beta, \gamma, c_1, c_2, c_3$ with $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ and

1. $\left|\mathbb{E}[P_\ell - P]\right| \;\leq\; c_1\, 2^{-\alpha\, \ell}$
2. $\mathbb{E}[Z_\ell] \;=\; \mathbb{E}[\Delta P_\ell]$
3. $V_\ell \;\leq\; c_2\, 2^{-\beta\, \ell}$
4. $C_\ell \;\leq\; c_3\, 2^{\gamma\, \ell}$,

and then conclude that there exists a positive constant c_4 such that for any desired root-mean-square accuracy $\varepsilon < e^{-1}$ there are values L and N_ℓ for which the multilevel estimator

$$Z = \sum_{\ell=0}^{L} Z_\ell,$$

has a mean-square-error with bound

$$MSE \equiv \mathbb{E}\left[(Z - \mathbb{E}[P])^2\right] < \varepsilon^2$$

with a computational complexity C with bound

$$\mathbb{E}[C] \leq \begin{cases} c_4\, \varepsilon^{-2}, & \beta > \gamma, \\ c_4\, \varepsilon^{-2} (\log \varepsilon)^2, & \beta = \gamma, \\ c_4\, \varepsilon^{-2-(\gamma-\beta)/\alpha}, & \beta < \gamma. \end{cases}$$

In each new application, the objective is to design an estimator so that $\beta > \gamma$ to achieve the best order of complexity.

2.2 Randomised MLMC for Unbiased Estimation

An important extension has been introduced by Rhee and Glynn in [25]. Rather than choosing the finest level of simulation L, based on the desired accuracy, and then using the optimal number of samples on each level based on an estimate of the variance, their "single term" estimator instead uses N samples in total, and for each sample they perform a level ℓ simulation with probability $p_\ell > 0$, with $\sum_{\ell=0}^{\infty} p_\ell = 1$.

The estimator is

$$Z = \frac{1}{N} \sum_{n=1}^{N} \Delta P^{(n)}_{\ell^{(n)}} / p_{\ell^{(n)}}$$

with the level $\ell^{(n)}$ for each sample being selected randomly with the relevant probability, so that

$$\mathbb{E}[Z] = \sum_{\ell} \mathbb{E}[\Delta P_\ell] = \mathbb{E}[P].$$

Hence, it is an unbiased estimator.

The choice of the probabilities p_ℓ is crucial. For both the variance and the expected cost to be finite, it is necessary that

$$\sum_{\ell=0}^{\infty} V_\ell / p_\ell \; < \; \infty, \quad \sum_{\ell=0}^{\infty} p_\ell\, C_\ell \; < \; \infty.$$

Under the conditions of the usual MLMC theorem, this is possible when $\beta > \gamma$ by choosing $p_\ell \propto 2^{-(\gamma+\beta)\ell/2}$, so that

$$V_\ell / p_\ell \propto 2^{-(\beta-\gamma)\ell/2}, \quad p_\ell\, C_\ell \propto 2^{-(\beta-\gamma)\ell/2}.$$

It is not possible when $\beta \leq \gamma$, and for these cases the estimators in [25] have infinite expected cost.

2.3 Multi-Index Monte Carlo

In standard MLMC, there is a one-dimensional set of levels, with a scalar level index ℓ, although in some applications changing ℓ can change more than one aspect of the computation, such as both timestep and spatial discretisation in a parabolic SPDE application [13]. In [21], Haji-Ali, Nobile and Tempone generalised this, with the Multi-Index Monte Carlo (MIMC) method defining "levels" in multiple directions, so that the level index $\boldsymbol{\ell}$ is now a vector of integer indices. This is illustrated in Fig. 1 for a 2D MIMC application.

Generalising (1) to D dimensions in [21], Haji-Ali, Nobile and Tempone first define a backward difference operator in one particular dimension, $\Delta_d P_{\boldsymbol{\ell}} \equiv P_{\boldsymbol{\ell}} - P_{\boldsymbol{\ell}-\boldsymbol{e}_d}$ where $\boldsymbol{e}_d$ is the unit vector in direction d, and then define the cross-difference

$$\boldsymbol{\Delta} P_{\boldsymbol{\ell}} \equiv \left(\prod_{d=1}^{D} \Delta_d \right) P_{\boldsymbol{\ell}}$$

so that the telescoping sum becomes

$$\mathbb{E}[P] = \sum_{\boldsymbol{\ell} \geq \mathbf{0}} \mathbb{E}[\boldsymbol{\Delta} P_{\boldsymbol{\ell}}]. \tag{2}$$

As an example, Fig. 1 marks the four locations at which $P_{\boldsymbol{\ell}}$ must be computed to determine the value of $\boldsymbol{\Delta} P_{(5,4)}$ in the 2D application.

Following the presentation in [10], the MIMC theorem formulated in [21] can be expressed in a form which matches quite closely the formulation of the MLMC

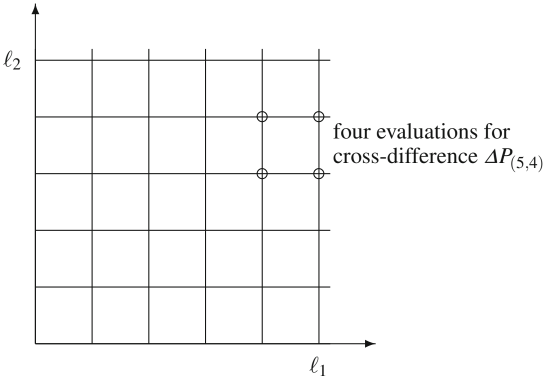


Fig. 1 "Levels" in 2D multi-index Monte Carlo application

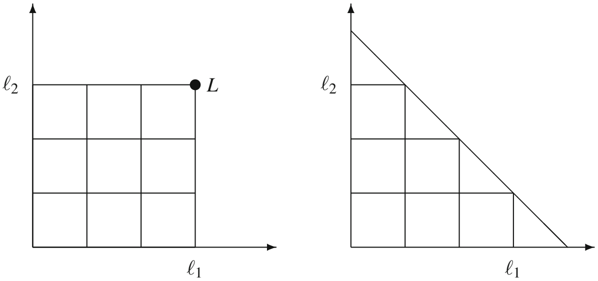


Fig. 2 Two choices of 2D MIMC summation region $\mathscr{L}$

theorem. If the level $\boldsymbol{\ell}$ MIMC estimator $Z_{\boldsymbol{\ell}}$, with variance $V_{\boldsymbol{\ell}}$ and cost $C_{\boldsymbol{\ell}}$, per sample, satisfies

1. $\left|\mathbb{E}[P_{\boldsymbol{\ell}} - P]\right| \longrightarrow 0 \;\text{ as }\; \min\limits_d \ell_d \longrightarrow \infty$
2. $\mathbb{E}[Z_{\boldsymbol{\ell}}] \;= \mathbb{E}[\boldsymbol{\Delta} P_{\boldsymbol{\ell}}]$
3. $\left|\mathbb{E}[Z_{\boldsymbol{\ell}}]\right| \;\le\; c_1\, 2^{-\boldsymbol{\alpha}\cdot\boldsymbol{\ell}}$
4. $V_{\boldsymbol{\ell}} \;\le\; c_2\, 2^{-\boldsymbol{\beta}\cdot\boldsymbol{\ell}}$
5. $C_{\boldsymbol{\ell}} \;\le\; c_3\, 2^{\boldsymbol{\gamma}\cdot\boldsymbol{\ell}}$,

then the complexity is $O(\varepsilon^{-2})$ provided $\beta_d > \gamma_d$ for all dimensions d, with additional $|\log \varepsilon|$ factors introduced if $\beta_d = \gamma_d$ for some d.

This complexity is achieved by truncating the set of increments in Eq. (2). It might seem natural that the summation region $\mathscr{L}$ should be rectangular, as illustrated on the left in Fig. 2, so that

$$\sum_{\boldsymbol{\ell}\in\mathscr{L}} \mathbb{E}[Z_{\boldsymbol{\ell}}] = \mathbb{E}[P_{\boldsymbol{L}}]$$

where $\boldsymbol{L}$ is the outermost point on the rectangle. However, [21] proves that in general this does not give the optimal order of complexity, and instead it is often best to use a region which in 2D is triangular, as illustrated on the right in Fig. 2. This is very similar to the use of sparse grid methods in high-dimensional PDE approximations [7], and indeed MIMC can be viewed as a combination of sparse grid methods and Monte Carlo sampling.

3 The General Smooth Case

In this first section, we consider the case in which g is a smooth function. A particular case of interest is the VaR application which was discussed in the Sect. 1. If one can estimate moments of the loss function $L(X)$, then an approximation of the loss CDF can be generated using Maximum Entropy reconstruction [2, 19]. The critical loss value L_α can then be determined from this CDF approximation.

3.1 MLMC Treatment

Following the presentation in [10], we are interested in estimating quantities of the form $\mathbb{E}\left[g\left(\mathbb{E}[f(X,Y)|X]\right)\right]$ where X is an outer random variable, and $\mathbb{E}[f(X,Y)|X]$ is a conditional expectation with respect to an independent inner random variable Y.

This can be simulated using nested Monte Carlo simulation with N outer samples $X^{(n)}$, M inner samples $Y^{(m,n)}$ and a standard Monte Carlo estimator:

$$Z = N^{-1}\sum_{n=1}^{N} g\left(M^{-1}\sum_{m=1}^{M} f(X^{(n)}, Y^{(m,n)})\right)$$

Note that to improve the accuracy of the estimate we need to increase both M and N, and this will significantly increase the cost. In fact, it can be proved [18] that the root-mean-square error is $O(M^{-1}+N^{-1/2})$, so to achieve r.m.s. accuracy of ε it is best to choose $M=O(\varepsilon^{-1})$, $N=O(\varepsilon^{-2})$, giving a complexity which is $O(\varepsilon^{-3})$.

An MLMC implementation is straightforward; on level ℓ we can use $M_\ell = 2^\ell$ inner samples. To construct a low variance estimate for $\mathbb{E}[P_\ell - P_{\ell-1}]$ where

$$\mathbb{E}[P_\ell] \equiv \mathbb{E}\left[g\left(M_\ell^{-1}\sum_{m=1}^{M_\ell} f(X, Y^{(m)})\right)\right],$$

we use an *antithetic* approach and split the M_ℓ samples for the "fine" value into two subsets of size $M_{\ell-1}$ for the "coarse" value:

$$\begin{aligned} Z_\ell = N_\ell^{-1}\sum_{n=1}^{N_\ell} \Bigg\{ & g\left(M_\ell^{-1}\sum_{m=1}^{M_\ell} f(X^{(n)}, Y^{(m,n)})\right) \\ & - \tfrac{1}{2} g\left(M_{\ell-1}^{-1}\sum_{m=1}^{M_{\ell-1}} f(X^{(n)}, Y^{(m,n)})\right) \\ & - \tfrac{1}{2} g\left(M_{\ell-1}^{-1}\sum_{m=M_{\ell-1}+1}^{M_\ell} f(X^{(n)}, Y^{(m,n)})\right)\Bigg\} \end{aligned}$$

Note that this has the correct expectation, i.e. $\mathbb{E}[Z_\ell] = \mathbb{E}[P_\ell - P_{\ell-1}]$.

If we define

$$M_{\ell-1}^{-1}\sum_{m=1}^{M_{\ell-1}} f(X^{(n)},Y^{(m,n)}) = \mathbb{E}[f(X^{(n)},Y)] + \Delta f_1^{(n)},$$

$$M_{\ell-1}^{-1}\sum_{m=M_{\ell-1}+1}^{M_\ell} f(X^{(n)},Y^{(m,n)}) = \mathbb{E}[f(X^{(n)},Y)] + \Delta f_2^{(n)},$$

then if g is twice differentiable a Taylor series expansion gives

$$Z_\ell \approx -\frac{1}{4\,N_\ell}\sum_{n=1}^{N_\ell} g''\left(\mathbb{E}[f(X^{(n)},Y)]\right)\left(\Delta f_1^{(n)} - \Delta f_2^{(n)}\right)^2.$$

By the Central Limit Theorem, $\Delta f_1^{(n)}, \Delta f_2^{(n)} = O(M_\ell^{-1/2})$ and therefore

$$g''\left(\mathbb{E}[f(X^{(n)},Y)]\right)\left(\Delta f_1^{(n)} - \Delta f_2^{(n)}\right)^2 = O(M_\ell^{-1}).$$

It follows that $\mathbb{E}[Z_\ell] = O(M_\ell^{-1})$ and $V_\ell = O(M_\ell^{-2})$. For the MLMC theorem, this corresponds to $\alpha = 1$, $\beta = 2$, $\gamma = 1$, so the complexity is $O(\varepsilon^{-2})$.

This antithetic approach to nested simulation has been developed independently by several authors [6, 8, 20], and is related to an earlier use of an antithetic MLMC estimator for SDEs [14].

Haji-Ali [20] used it in a mean field model for the motion of crowds, in which each person is modelled as a independent agent subject to random forcing and an additional force due to the collective influence of the crowd. This same approach is also relevant to mean field problems which arise in plasma physics [26].

Bujok et al. [6] used multilevel nested simulation for a financial credit derivative application. In their case, the function g was piecewise linear, not twice differentiable, and so the rate of variance convergence was slightly lower, with $\beta = 1.5$. This will be discussed in Sect. 4, but it is still sufficient to achieve an overall $O(\varepsilon^{-2})$ complexity.

3.2 MIMC Treatment

The previous analysis assumes we can compute $f(X,Y)$ with $O(1)$ cost, but suppose now that Y represents a complete Brownian path, and $f(X,Y)$ cannot be evaluated exactly; it can only be approximated using some finite number of timesteps. Using MLMC, on level ℓ we could use 2^ℓ timesteps and a Milstein discretisation (giving first order weak and strong convergence) which would still give $\alpha = 1$, $\beta = 2$. However, we would now have $\gamma = 2$, because on successive levels we would be using twice as many timesteps as well as twice as many inner samples. This then leads to an overall MLMC complexity which is $O(\varepsilon^{-2}(\log \varepsilon)^{-2})$.

Instead we can use MIMC to recover an optimal complexity of $O(\varepsilon^{-2})$. We now have a pair of level indices (l_1, l_2), with the number of inner samples equal to 2^{ℓ_1} and the number of timesteps proportional to 2^{ℓ_2}. If we use the natural extension of the MLMC estimator to the corresponding MIMC estimator, which means (for $l_1 > 0, l_2 > 0$) using

$$
\begin{aligned}
Z_\ell = N_\ell^{-1} \sum_{n=1}^{N_\ell} \Bigg\{ & g\left(2^{-\ell_1} \sum_{m=1}^{2^{\ell_1}} f_{\ell_2}(X^{(n)}, Y^{(m,n)})\right) - \tfrac{1}{2} g\left(2^{-\ell_1+1} \sum_{m=1}^{2^{\ell_1-1}} f_{\ell_2}(X^{(n)}, Y^{(m,n)})\right) \\
& - \tfrac{1}{2} g\left(2^{-\ell_1+1} \sum_{m=2^{\ell_1-1}+1}^{2^{\ell_1}} f_{\ell_2}(X^{(n)}, Y^{(m,n)})\right) \\
& - g\left(2^{-\ell_1} \sum_{m=1}^{2^{\ell_1}} f_{\ell_2-1}(X^{(n)}, Y^{(m,n)})\right) + \tfrac{1}{2} g\left(2^{-\ell_1+1} \sum_{m=1}^{2^{\ell_1-1}} f_{\ell_2-1}(X^{(n)}, Y^{(m,n)})\right) \\
& + \tfrac{1}{2} g\left(2^{-\ell_1+1} \sum_{m=2^{\ell_1-1}+1}^{2^{\ell_1}} f_{\ell_2-1}(X^{(n)}, Y^{(m,n)})\right) \Bigg\}
\end{aligned}
$$

The subscript on the$f$ terms denotes the level of timestep approximation.

Carrying out the same analysis as before, performing the Taylor series expansion around $\mathbb{E}[f(X^{(n)}, Y)]$, we obtain

$$
Z_\ell \approx -\frac{1}{4N_\ell} \sum_{n=1}^{N_\ell} g''\left(\mathbb{E}[f(X(n), Y)]\right) \left\{ \left(\Delta f_{1,\ell_2}^{(n)} - \Delta f_{2,\ell_2}^{(n)}\right)^2 - \left(\Delta f_{1,\ell_2-1}^{(n)} - \Delta f_{2,\ell_2-1}^{(n)}\right)^2 \right\} .
$$

The difference of squares can be re-arranged as

$$
\begin{aligned}
& \left(\Delta f_{1,\ell_2}^{(n)} - \Delta f_{2,\ell_2}^{(n)}\right)^2 - \left(\Delta f_{1,\ell_2-1}^{(n)} - \Delta f_{2,\ell_2-1}^{(n)}\right)^2 \\
& \qquad = \left((\Delta f_{1,\ell_2}^{(n)} + \Delta f_{1,\ell_2-1}^{(n)}) - (\Delta f_{2,\ell_2}^{(n)} + \Delta f_{2,\ell_2-1}^{(n)}) \right) \times \\
& \qquad\quad \left((\Delta f_{1,\ell_2}^{(n)} - \Delta f_{1,\ell_2-1}^{(n)}) - (\Delta f_{2,\ell_2}^{(n)} - \Delta f_{2,\ell_2-1}^{(n)}) \right)
\end{aligned}
$$

Due to the Central Limit Theorem, we have

$$
\Delta f_{1,\ell_2}^{(n)} + \Delta f_{1,\ell_2-1}^{(n)} = O(2^{-\ell_1/2}), \quad \Delta f_{2,\ell_2}^{(n)} + \Delta f_{2,\ell_2-1}^{(n)} = O(2^{-\ell_1/2}),
$$

and assuming first order strong convergence we also have

$$\Delta f^{(n)}_{1,\ell_2} - \Delta f^{(n)}_{1,\ell_2-1} = O(2^{-\ell_1/2-\ell_2}), \quad \Delta f^{(n)}_{2,\ell_2} - \Delta f^{(n)}_{2,\ell_2-1} = O(2^{-\ell_1/2-\ell_2}).$$

Combining these results we obtain

$$\left(\Delta f^{(n)}_{1,\ell_2} - \Delta f^{(n)}_{2,\ell_2}\right)^2 - \left(\Delta f^{(n)}_{1,\ell_2-1} - \Delta f^{(n)}_{2,\ell_2-1}\right)^2 = O(2^{-\ell_1-\ell_2})$$

and therefore $\mathbb{E}[Z_\ell] = O(2^{-\ell_1-\ell_2})$ and $V_\ell = O(2^{-2\ell_1-2\ell_2})$ with a cost per sample which is $O(2^{\ell_1+\ell_2})$. In the MIMC theorem this corresponds to $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 2$, and $\gamma_1 = \gamma_2 = 1$, so the overall complexity is $O(\varepsilon^{-2})$.

3.3 Nested MLMC

MIMC is not the only way in which to generalise MLMC to multiple dimensions. Another option, which can sometimes be equivalent, but is often not, is to use nested MLMC, with an inner MLMC being used to generate samples within an outer MLMC computation.

The application in the previous section gives rise to a natural example of this. Ideally, we would like to generate exact samples of $f(X, Y)$ with $O(1)$ cost per sample. However, it is just as good to produce samples which have the correct expected value $\mathbb{E}[f(X, Y)|X]$, with an expected cost which is $O(1)$. This can be achieved by using the randomised MLMC discussed in Sect. 2.2, so that $f(X^{(n)}, Y^{(m,n)})$ is replaced by

$$\left(f(X^{(n)}, Y^{(m,n)}_\ell) - f(X^{(n)}, Y^{(m,n)}_{\ell-1})\right) \,/\, p_\ell,$$

where the level ℓ which determines the number of timesteps is a random variable taking integer value $\ell' \geq 0$ with probability $p_{\ell'} > 0$. The only requirement is that the variance for this inner randomised MLMC must decay faster with the number of timesteps than the increase in the computational cost, so that $p_{\ell'}$ can be specified appropriately to achieve both finite variance and finite expected cost.

4 EVPPI

For the estimation of the difference EVPI − EVPPI defined in the Sect. 1, we define a level ℓ approximation as

$$P_\ell = \overline{\max_d f_d}^{\,\ell} - \max_d \overline{f_d}^{\,\ell}$$

where $\overline{\max_d f_d}^{\ell}$ and $\overline{f_d}^{\ell}$ represent averages over 2^ℓ independent values of $Y^{(i)}$ for one particular value of X, so that

$$\text{EVPI} - \text{EVPPI} = \lim_{\ell\to\infty} \mathbb{E}[P_\ell].$$

Following the ideas in [6, 10, 20] we use the antithetic MLMC estimator

$$Z_\ell \;=\; \tfrac{1}{2}\left(\max_d \overline{f_d}^{(a)} + \max_d \overline{f_d}^{(b)}\right) \;-\; \max_d \overline{f_d}$$

where

- $\overline{f_d}^{(a)}$ is an average of $f_d(X,Y)$ over $2^{\ell-1}$ independent samples for Y;
- $\overline{f_d}^{(b)}$ is an average over a second independent set of $2^{\ell-1}$ samples;
- $\overline{f_d}$ is an average over the combined set of 2^ℓ inner samples.

The MLMC variance can be analysed by following the approach used by Giles and Szpruch for Theorem 5.2 in [14], which is also similar to the analysis by Bujok et al. in [6]. Define

$$F_d(X) = \mathbb{E}_Y\left[f_d(X,Y)\right], \quad d_{opt}(X) = \arg\max_d F_d(X)$$

so the domain for X is divided into a number of regions in which the optimal decision $d_{opt}(X)$ is uniform, with a dividing lower-dimensional decision manifold K on which $d_{opt}(X)$ is not uniquely-defined.

Note that $\frac{1}{2}(\overline{f_d}^{(a)} + \overline{f_d}^{(b)}) - \overline{f_d} \;=\; 0$, and therefore $Z_\ell = 0$ if the same decision d maximises each of the terms in its definition. This is the key advantage of the antithetic estimator, compared to the alternative $\overline{f_d}^{(a)} - \overline{f_d}$. When ℓ is large and so there are many samples, $\overline{f_d}^{(a)}, \overline{f_d}^{(b)}, \overline{f_d}$ will all be close to $F_d(X)$, and therefore it is highly likely that $Z_\ell = 0$ unless X is very close to K at which there is more than one optimal decision. This idea leads to a theorem on the MLMC variance, but first we need to make three assumptions.

Assumption 1 $\mathbb{E}\left[\,|f_d(X,Y)|^p\right]$ *is finite for all* $p \geq 2$.
Comment: this enables us to bound the difference between $\overline{f_d}^{(a)}, \overline{f_d}^{(b)}, \overline{f_d}$ *and* $F_d(X)$.

Assumption 2 *There exists a constant* $c_0 > 0$ *such that for all* $0 < \epsilon < 1$

$$\mathbb{P}\left(\min_{x\in K} \|X - x\| \leq \epsilon\right) \leq c_0\, \epsilon.$$

Comment: this bounds the probability of X being close to the decision manifold K.

Assumption 3 *There exist constants $c_1, c_2 > 0$ such that if $X \notin K$, then*

$$\max_d F_d(X) - \max_{d \neq d_{opt}(X)} F_d(X) \;>\; \min\left(c_1,\; c_2 \min_{x\in K} \|X - x\| \right).$$

Comment: on K itself there are at least 2 decisions d_1, d_2 which yield the same optimal value $F_d(X)$; this assumption ensures at least a linear divergence between the values as X moves away from K.

Theorem 1 *If Assumptions 1–3 are satisfied, and Z_ℓ is as defined previously for level ℓ, then for any $\delta > 0$*

$$\mathbb{V}[Z_\ell] = o(2^{-(3/2-\delta)\ell}), \quad \mathbb{E}[Z_\ell] = o(2^{-(1-\delta)\ell}).$$

The proof of the theorem is given in [11], but a heuristic explanation is as follows:

- Because of Assumption 1, for any $X, \overline{f_d} - F_d(X) = O(2^{-\ell/2})$;
- Because of Assumption 2, there is an $O(2^{-\ell/2})$ probability of X being within $O(2^{-\ell/2})$ of the decision manifold K, in which case $Z_\ell = O(2^{-\ell/2})$;
- Because of Assumption 3, if it is further away from K then there is a clear separation between the different decision values, and hence $Z_\ell = 0$ with very high probability.
- This results in $\mathbb{E}[Z_\ell^2] = O(2^{-\ell/2}) \times \big(O(2^{-\ell/2})\big)^2 = O(2^{-3\ell/2})$.

The conclusion from the theorem is that the parameters for the MLMC theorem are $\beta \approx 3/2$, $\alpha \approx 1$, and $\gamma = 1$, giving the optimal complexity of $O(\varepsilon^{-2})$. Numerical results support this prediction.

A final comment is that sometimes the random variables in X or Y correspond to Bayesian posterior distributions, with samples generated by MCMC methods. In that case, it is possible to pre-generate a large set of MCMC samples, after the initial burn-in, and then MLMC can uniformly and randomly take samples from this dataset as required.

5 Value-at-Risk

The Value-at-Risk problem has been defined in Sect. 1. In this section, we begin by introducing the idea of portfolio sub-sampling, and then proceed to discuss the difficulties in constructing efficient MLMC estimators for VaR because of the discontinuous nature of the indicator function.

5.1 Portfolio Sub-sampling

In Sects. 3.2 and 3.3, we considered Y to represent the driving Brownian motion, and the inner conditional expectation was with respect to this. However, in the context of the Value-at-Risk application, where we are considering estimation of moments of the loss for the purpose of Maximum Entropy reconstruction, there is an important second aspect to this conditional expectation.

The loss function $L(X)$ has contributions from a large number of financial options within a portfolio, so that it may be written as

$$L(X) = \sum_{i=1}^{N_o} L_i(X), \quad L_i(X) \equiv \mathbb{E}\left[f_i(X,Y) \,|\, X\right],$$

where $L_i(X)$ is the loss from the ith option. In the existing literature, standard treatments evaluate each sample of the total loss by summing the contributions from all of the financial options, and the computational cost is inevitably proportional to N_o, the number of options. However, instead we can express the loss as

$$L(X) = \mathbb{E}\left[L_i(X)/p_i\right].$$

where the integer index i is randomly sampled from the set $\{1, 2, \ldots, N_o\}$ with probability p_i.

Adding back in the expectation with respect to the Brownian motion we obtain the conditional expectation

$$L(X) = \mathbb{E}\left[f_i(X,Y)/p_i \,|\, X\right],$$

in which the expectation is now over both the Brownian motion and the index of the option being sampled. When 2^ℓ samples are generated to approximate the conditional expectation, they each can have a different option index as well as a different Brownian path sample. The overall benefit is to achieve a complexity, for a given accuracy ε expressed as a fraction of the total portfolio value, which no longer depends on N_o, the number of financial options in the portfolio.

This idea of sub-sampling a portfolio has been investigated by Wenhui Gou [19] whose research combined it with Maximum Entropy reconstruction of the loss distribution, but used an analytic expression for the conditional expectation with respect to the driving Brownian motion, and also used a control variate which substantially reduced the variance of the estimator.

5.2 Previous Work on VaR

As explained in the Sect. 1, we are interested in determining

$$\mathbb{P}\left[L(X) \geq L^*\right] \equiv \mathbb{E}\left[\mathbf{1}\left(L(X) \geq L^*\right)\right]$$

This is again a nested simulation problem, but the indicator function makes it much harder than EVPPI, because small differences between the "coarse" and "fine" estimates for the conditional expectation in $L(X)$ can lead to a ± 1 change in the indicator value.

Gordy and Juneja [18] considered this problem, using a single level Monte Carlo method with M inner samples $Y^{(m,n)}$ for each of N outer samples $X^{(n)}$ to estimate

$$L(X) \equiv \mathbb{E}[f(X,Y)]$$

for each of N outer samples $X^{(n)}$, so that the overall estimate for the probability of exceeding the loss threshold is

$$\mathbb{P}\left[L(X) \geq L^*\right] \approx N^{-1} \sum_{n=1}^{N} \mathbf{1}\left(M^{-1} \sum_{m=1}^{M} f(X^{(n)}, Y^{(m,n)} \geq L^*)\right)$$

This problem setup assumes that it is possible to exactly simulate $f(X^{(n)}, Y)$ at unit cost. Given this, they proved that the resulting RMS error is

$$O(M^{-1} + N^{-1/2}),$$

and hence, to achieve an ε RMS accuracy requires $M = O(\varepsilon^{-1})$, $N = O(\varepsilon^{-2})$ and so the complexity is $O(\varepsilon^{-3})$.

Broadie et al. [5] improved on this, by noting that unless $L(X) - L^*$ is small, we usually don't need many samples to determine whether $L(X) \geq L^*$. Their paper presents a rigorously analysed adaptive algorithm based on the theory of sequential sampling but here we give a simplified heuristic analysis. When using M inner samples, if

$$\sigma^2(X) = \mathbb{V}[f(X,Y)|X], \quad d(X) = \left| \mathbb{E}[f(X,Y)|X] - L^* \right|$$

then the usual CLT confidence interval for the estimate of $\mathbb{E}[f(X,Y)|X] - L^*$ has size $\pm 3\,\sigma/\sqrt{M}$. Hence, we need roughly

$$M = 9\,\sigma^2(X)/d^2(X)$$

inner samples to be sure whether or not $\mathbb{E}[f(X,Y)|X] \geq L^*$. If we now use

$$M = \min\left(c\,\varepsilon^{-1}, 9\,\sigma^2(X)/d^2(X)\right)$$

then the cross-over point between the two terms in the minimum is at $d = O(\varepsilon^{1/2})$, and it follows that the average number of inner samples required is

$$\overline{M} = O(\varepsilon^{-1/2}),$$

reducing the overall complexity to $O(\varepsilon^{-5/2})$.

This is clearly a significant improvement on the complexity of the uniform sampling algorithm of Gordy and Juneja, but in both papers they are not using the sub-sampling introduced in Sect. 5.1 but are instead evaluating the full portfolio each time so the complexity is also proportional to the number of options in the portfolio. Furthermore, their analysis does not consider the additional cost which is incurred when one needs to approximate an SDE for the underlying assets.

5.3 Current Research

Current research by the author and Abdul-Lateef Haji-Ali builds on the adaptive approach of Broadie et al. [5] by incorporating MLMC ideas.

The first step is to extend Wenhui Gou's work to Monte Carlo estimation of the inner conditional expectations:

$$\sum_{i=1}^{N_o} \mathbb{E}[f_i(X,Y)|X] \approx M^{-1} \sum_{m=1}^{M} f_{i_m}(X, W_m) \;/\; p_{i_m}$$

where W_m represents the Brownian path and any additional random inputs needed for the conditional expectation. This essentially combines, or unifies, the Monte Carlo averaging over the portfolio samples with the averaging over the Brownian paths.

If we do this with the uniform inner sampling with $M_\ell = 4^\ell$ samples on level ℓ, assuming that $f_{p_m}(X, W_m)$ can be computed exactly at unit cost, then the error in the inner estimate is $O(M_\ell^{-1/2}) = O(2^{-\ell})$. There is an $O(2^{-\ell})$ probability of being within $O(2^{-\ell})$ of the indicator step, producing an $O(1)$ value for the MLMC estimator sample, so the MLMC variance is $V_\ell \sim 2^{-\ell}$. In addition we get bias $\sim M_\ell^{-1} \sim 4^{-\ell}$, $C_\ell \sim M_\ell \sim 4^\ell$, so $\alpha \approx 2$, $\beta \approx 1$, $\gamma \approx 2$ and therefore the complexity is $O(\varepsilon^{-5/2})$. The advantage over the previous method due to Broadie et al is that the complexity is independent of the value of N_o the number of options in the portfolio, but it still falls short of our target of $O(\varepsilon^{-2})$.

To further improve things, we add in the adaptive approach of Broadie et al, with the number of inner samples dependent on both X and the level ℓ, along the lines of

$$M_\ell(X) = \max\left(c_1\, 2^\ell, \min\left(c_2\, 4^\ell, 9\,\sigma^2(X)/d^2(X)\right)\right).$$

This gives approximately the same asymptotic behaviour in the variance and the bias, i.e. bias $\sim 4^{-\ell}$, $V_\ell \sim 2^{-\ell}$, but the cost is reduced to approximately $C_\ell \sim 2^\ell$. This leads to $\alpha \approx 2$, $\beta \approx 1$, $\gamma \approx 1$ and hence the complexity is approximately $O(\varepsilon^{-2})$, independent of N_o.

The final challenge comes from the approximation of the underlying SDE. At first sight this looks very difficult, but the algorithm does not require the exact sampling of $f_{i_m}(X, W_m)$; it is sufficient to have an unbiased estimate with a unit expected cost. Following the ideas in Sect. 3.3, this is precisely what can be supplied in many cases by Rhee and Glynn's unbiased single-term estimator based on randomised MLMC. This requires the use of the Milstein time discretisation, because of the improved strong order of convergence and hence rate of MLMC variance convergence compared to an Euler-Maruyama discretisation. The complexity analysis is largely unchanged, and again we achieve an overall complexity of approximately $O(\varepsilon^{-2})$, to within log terms.

In practice, it is also very important to use an effective control variate, similar to the one used by Wenhui Gou [19], but the details are omitted here.

5.4 Future Research

There are other aspects of the VaR problem to be investigated in the future.

One is associated with the fact that the different financial options within a portfolio vary greatly, both in their variance (in part due to differences in their financial magnitude) and in the computational cost involved in their simulation. Both of these factors need to be taken into account in optimising the probability p_i for sampling the option with index i. It might even be desirable to identify a few options which should always be sampled because of their large value, and apply the randomised sub-sampling to the remainder.

Secondly, the discussion so far has been about the simpler problem of determining

$$\mathbb{P}(L(X)\geq L^*) \;\equiv\; \mathbb{E}\left[\mathbf{1}_{L(X)\geq L^*}\right],$$

for some given loss value L^*. The research must be extended to the full VaR definition which requires some root-finding algorithm to determine L_α defined implicitly for some α by

$$\mathbb{P}(L(X)\geq L_\alpha) = \alpha.$$

We also need to consider other risk measures such as CVaR, or expected shortfall,

$$\text{CVaR} \;=\; \mathbb{E}\left[L(X) \mid L(X) \geq L_\alpha\right] \;=\; \alpha^{-1}\,\mathbb{E}\left[L(X)\,\mathbf{1}_{L(X)\geq L_\alpha}\right].$$

6 Conclusions

In this paper we have reviewed progress in applying MLMC ideas to problems with nested expectations. Such applications lead quite naturally to the use of the Multi-Index Monte Carlo method and other generalisations of MLMC such as nested MLMC. Randomised MLMC for the inner conditional expectation is particularly helpful as it is unbiased, which simplifies the treatment.

One important nested expectation application is the estimation of EVPPI, the Expected Value of Partial Perfect Information. Substantial progress has been made on this topic, in both the construction and the analysis of efficient algorithms.

In the context of the financial Value-at-Risk application, we have pointed out the benefits to be achieved from sub-sampling the portfolio. Combining this with an adaptive MLMC estimator addresses the challenge due to the discontinuous indicator function in the outer expectation. This use of adaptive algorithms within MLMC fits well with other current research [9, 12, 15].

Acknowledgements The author is very grateful to a number of people for discussions and collaborations: Ian Sloan and Frances Kuo on alternatives to the Multi-Index Monte Carlo method; Takashi Goda and Howard Thom on EVPPI estimation; Wenhui Gou, Abdul-Lateef Haji-Ali, Ralf Korn and Klaus Ritter on Value-at-Risk estimation.

References

1. Ades, A., Lu, G., Claxton, K.: Expected value of sample information calculations in medical decision modeling. Med. Decis. Mak. **24**(2), 207–227 (2004)
2. Bierig, C., Chernov, A.: Approximation of probability density functions by the multilevel Monte Carlo maximum entropy method. J. Comput. Phys. **314**, 661–681 (2016)
3. Bratvold, R., Bickel, J., Lohne, H.: Value of information in the oil and gas industry: past, present, and future. SPE Reserv. Eval. Eng. **12**, 630–638 (2009)
4. Brennan, A., Kharroubi, S., O'Hagan, A., Chilcott, J.: Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. Med. Decis. Mak. **27**, 448–470 (2007)
5. Broadie, M., Du, Y., Moallemi, C.: Efficient risk estimation via nested sequential simulation. Manag. Sci. **57**(6), 1172–1194 (2011)
6. Bujok, K., Hambly, B., Reisinger, C.: Multilevel simulation of functionals of Bernoulli random variables with application to basket credit derivatives. Methodol. Comput. Appl. Probab. **17**(3), 579–604 (2015)
7. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)

8. Chen, N., Liu, Y.: Estimating expectations of functionals of conditional expected via multilevel nested simulation. In: Presentation at conference on Monte Carlo and Quasi-Monte Carlo Methods, Sydney (2012)
9. Fang, W., Giles, M.: Adaptive Euler-Maruyama method for SDEs with non-globally Lipschitz drift: Part I, finite time interval (2016). ArXiv preprint: 1609.08101
10. Giles, M.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)
11. Giles, M., Goda, T.: Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI (2017). ArXiv preprint: 1708.05531
12. Giles, M., Ramanan, K.: MLMC with adaptive timestepping for reflected Brownian diffusions (2018, in preparation)
13. Giles, M., Reisinger, C.: Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance. SIAM J. Financ. Math. **3**(1), 572–592 (2012)
14. Giles, M., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. Ann. Appl. Probab. **24**(4), 1585–1620 (2014)
15. Giles, M., Lester, C., Whittle, J.: Non-nested adaptive timesteps in multilevel Monte Carlo computations. In: Cools, R., Nuyens, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2014. Springer, Basel (2016)
16. Glasserman, P., Heidelberger, P., Shahabuddin, P.: Variance reduction techniques for estimating value-at-risk. Manag. Sci. **46**, 1349–1364 (2000)
17. Glasserman, P., Heidelberger, P., Shahabuddin, P.: Portfolio value-at-risk with heavy-tailed risk factors. Math. Finance **12**, 239–269 (2002)
18. Gordy, M., Juneja, S.: Nested simulation in portfolio risk measurement. Manag. Sci. **56**(10), 1833–1848 (2010)
19. Gou, W.: Estimating value-at-risk using multilevel Monte Carlo maximum entropy method. MSc Thesis, University of Oxford (2016)
20. Haji-Ali, A.L.: Pedestrian flow in the mean-field limit. MSc Thesis, KAUST (2012)
21. Haji-Ali, A.L., Nobile, F., Tempone, R.: Multi index Monte Carlo: when sparsity meets sampling. Numer. Math. **132**(4), 767–806 (2016)
22. Korn, R., Korn, E., Kronstadt, G.: Monte Carlo Methods and Models in Finance and Insurance. Chapman and Hall/CRC Financial Mathematics. CRC Press, Boca Raton (2010)
23. Korn, R., Pupashenko, M.: A new variance reduction technique for estimating value-at-risk. Appl. Math. Finance **22**(1), 83–98 (2015)
24. Nakayasu, M., Goda, T., Tanaka, K., Sato, K.: Evaluating the value of single-point data in heterogeneous reservoirs with the expectation maximization. SPE Econ. Manag. **8**, 1–10 (2016)
25. Rhee, C.H., Glynn, P.: Unbiased estimation with square root convergence for SDE models. Oper. Res. **63**(5), 1026–1043 (2015)
26. Rosin, M., Ricketson, L., Dimits, A., Caflisch, R., Cohen, B.: Multilevel Monte Carlo simulation of Coulomb collisions. J. Comput. Phys. **247**, 140–157 (2014)

A Note on Some Approximation Kernels on the Sphere

Peter Grabner

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We produce precise estimates for the Kogbetliantz kernel for the approximation of functions on the sphere. Furthermore, we propose and study a new approximation kernel, which has slightly better properties.

1 Introduction

For $d \geq 1$, let $\mathbb{S}^d = \{\mathbf{z} \in \mathbb{R}^{d+1} : \langle \mathbf{z}, \mathbf{z} \rangle = 1\}$ denote the d-dimensional unit sphere embedded in the Euclidean space $\mathbb{R}^{d+1}$ and $\langle \cdot, \cdot \rangle$ be the usual inner product. We use $\mathrm{d}\sigma_d$ for the surface element and set $\omega_d = \int_{\mathbb{S}^d} \mathrm{d}\sigma_d$.

In [3] E. Kogbetliantz studied Cesàro means of the ultraspherical Dirichlet kernel. Let C_n^λ denote the n-th Gegenbauer polynomial of index λ. Then for $\lambda = \frac{d-1}{2}$

$$K_n^{\lambda,0}(\langle \mathbf{x}, \mathbf{y} \rangle) = \sum_{k=0}^{n} \frac{k+\lambda}{\lambda} C_k^\lambda(\langle \mathbf{x}, \mathbf{y} \rangle)$$

is the projection kernel on the space of harmonic polynomials of degree $\leq n$ on the sphere $\mathbb{S}^d$. The kernel could be studied for all $\lambda > 0$, but since we have the application to polynomial approximation on the sphere in mind, we restrict ourselves to half-integer and integer values of λ. Throughout this paper d will denote the dimension of the sphere and $\lambda = \frac{d-1}{2}$ will be the corresponding Gegenbauer parameter.

P. Grabner (✉)
Institut für Analysis und Zahlentheorie, Technische Universität Graz, Graz, Austria
e-mail: peter.grabner@tugraz.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_21

Kogbetliantz [3] studied how higher Cesàro-means improve the properties of the kernel $K_n^{\lambda,0}$: for $\alpha \geq 0$ set

$$K_n^{\lambda,\alpha}(t) = \frac{1}{\binom{n+\alpha}{n}} \sum_{k=0}^{n} \binom{n-k+\alpha}{n-k} \frac{k+\lambda}{\lambda} C_k^{\lambda}(t).$$

He proved that the kernels $(K_n^{\lambda,\alpha})_n$ have uniformly bounded L^1-norm, if $\alpha > \lambda$ and that they are non-negative, if $\alpha \geq 2\lambda + 1$. There is a very short and transparent proof of the second fact due to Reimer [4]. In this paper, we will restrict our interest to the kernel $K_n^{\lambda,2\lambda+1}$, which we will denote by K_n^{λ} for short.

The purpose of this note is to improve Kogbetliantz' upper bounds for the kernel K_n^{λ}. Especially, the estimates for $K_n^{\lambda}(t)$ given in [3] exhibit rather bad behaviour at $t = -1$. This is partly a consequence of the actual properties of the kernel at that point, but to some extent the estimate used loses more than necessary. Furthermore, the estimates given in [3] contain unspecified constants. We have used some effort to provide good explicit constants.

In the end of this paper we will propose a slight modification of the kernel function, which is better behaved at $t = -1$ and still shares all desirable properties of K_n^{λ}.

2 Estimating the Kernel Function

In the following we will use the notation

$$A_n^{\alpha} = \binom{n+\alpha}{n}.$$

Notice that

$$\sum_{n=0}^{\infty} A_n^{\alpha} z^n = \frac{1}{(1-z)^{\alpha+1}}. \tag{1}$$

Let C_n^{λ} denote the n-th Gegenbauer polynomial with index λ. The Gegenbauer polynomials satisfy two basic generating function relations (cf. [1, 3])

$$\sum_{n=0}^{\infty} C_n^{\lambda}(\cos\vartheta) z^n = \frac{1}{(1-2z\cos\vartheta + z^2)^{\lambda}} \tag{2}$$

$$\sum_{n=0}^{\infty} \frac{n+\lambda}{\lambda} C_n^{\lambda}(\cos\vartheta) z^n = \frac{1-z^2}{(1-2z\cos\vartheta + z^2)^{\lambda+1}}. \tag{3}$$

Several different kernel functions for approximation of functions on the sphere and their saturation behaviour have been studied in [2]. We will investigate the kernel

$$K_n^\lambda(\cos\vartheta) = \frac{1}{A_n^{2\lambda+1}} \sum_{k=0}^{n} A_{n-k}^{2\lambda+1} \frac{k+\lambda}{\lambda} C_k^\lambda(\cos\vartheta),$$

which has been shown to be positive by E. Kogbetliantz [3] for $\lambda > 0$.

By the generating functions (1) and (3) it follows

$$\sum_{n=0}^{\infty} A_n^{2\lambda+1} K_n^\lambda(\cos\vartheta) z^n = \frac{1+z}{(1-2z\cos\vartheta + z^2)^{\lambda+1}(1-z)^{2\lambda+1}}. \tag{4}$$

Thus we can derive integral representations for K_n^λ using Cauchy's integral formula. As pointed out in the introduction, we will restrict the values of λ to integers or half-integers. The main advantage of this is the fact that the exponent of $(1-z)$ in (4) is then an integer.

For $\lambda = k \in \mathbb{N}_0$ we split the generating function (4) into two factors

$$\frac{1+z}{(1-2z\cos\vartheta + z^2)(1-z)} \times \frac{1}{(1-2z\cos\vartheta + z^2)^k(1-z)^{2k}}.$$

The first factor is essentially the generating function of the Fejér kernel, namely

$$\frac{1}{2\pi i} \oint_{|z|=\frac{1}{2}} \frac{1+z}{(1-2z\cos\vartheta + z^2)(1-z)} \frac{dz}{z^{n+1}} = \left(\frac{\sin(n+1)\frac{\vartheta}{2}}{\sin\frac{\vartheta}{2}} \right)^2 \leq \frac{1}{(\sin\frac{\vartheta}{2})^2}. \tag{5}$$

Notice that this is just the kernel $(n+1)K_n^0$.

We compute the coefficients of the second factor using Cauchy's formula

$$Q_n^k(\cos\vartheta) = \frac{1}{2\pi i} \oint_{|z|=\frac{1}{2}} \frac{1}{(1-2z\cos\vartheta + z^2)^k(1-z)^{2k}} \frac{dz}{z^{n+1}}. \tag{6}$$

In order to produce an estimate for Q_n^k, we first compute Q_n^1. This is done by residue calculus and yields

$$Q_n^1(\cos\vartheta) = \frac{1}{4\left(\sin\frac{\vartheta}{2}\right)^2} \left(n+2 - \frac{\sin(n+2)\vartheta}{\sin\vartheta} \right). \tag{7}$$

This function is obviously non-negative and satisfies

$$Q_n^1(\cos\vartheta) \leq \frac{n+2}{2\left(\sin\frac{\vartheta}{2}\right)^2}. \tag{8}$$

Now the functions Q_n^k are formed from Q_n^1 by successive convolution:

$$Q_n^{k+1}(\cos\vartheta) = \sum_{m=0}^{n} Q_m^k(\cos\vartheta) Q_{n-m}^1(\cos\vartheta).$$

Inserting the estimate (8) and an easy induction yields

$$Q_n^k(\cos\vartheta) \leq \frac{1}{2^k\left(\sin\frac{\vartheta}{2}\right)^{2k}} \sum_{r=0}^{k} \binom{k}{r}\binom{n+r+k-1}{n}. \tag{9}$$

Remark 1 Asymptotically, this estimate is off by a factor of 2^λ, but as opposed to Kogbetliantz' estimate it does not contain a negative power of $\sin\vartheta$, which would blow up at $\vartheta = \pi$. The size of the constant is lost in the transition from (7) to (8), where the trigonometric term (actually a Chebyshev polynomial of the second kind) is estimated by its maximum. On the one hand this avoids a power of $\sin\vartheta$ in the denominator, on the other hand it spoils the constant.

Putting (5) and (9) together yields

$$A_n^{2k+1} K_n^k(\cos\vartheta) \leq \frac{1}{2^k(\sin\frac{\vartheta}{2})^{2k+2}} \sum_{\ell=0}^{k} \binom{k}{\ell}\binom{n+k+\ell}{n}, \tag{10}$$

where we have used the identity

$$\sum_{i=0}^{n} \binom{i+m}{i} = \binom{n+m+1}{n}.$$

Remark 2 Since the generating function of $A_n^{2k+1}K_n^k(\cos\vartheta)$ is a rational function in this case, an application of residue calculus would have of course been an option. The calculation of the residues at $e^{\pm i\vartheta}$ produces a denominator containing $(\sin\vartheta)^{2k-1}$. Computation of the numerators for small values of k show that this denominator actually cancels, but we did not succeed in proving this in general. Furthermore, keeping track of the estimates through this cancellation seems to be difficult. This denominator could also be eliminated by restricting $\frac{C}{n} \leq \vartheta \leq \pi - \frac{C}{n}$, but this usually spoils any gain in the constants obtained before. This was actually the technique used in [3].

For $\lambda = \frac{1}{2} + k$ we split the generating function (4) into the factors

$$\frac{1}{\sqrt{1-2z\cos\vartheta+z^2}(1-z)} \times \frac{1+z}{(1-2z\cos\vartheta+z^2)^{k+1}(1-z)^{2k+1}} \tag{11}$$

with $k \in \mathbb{N}_0$. The second factor is exactly the generating function related to the case of integer parameter λ studied above.

For the coefficients of the first factor in (11) we use Cauchy's formula again

$$R_n(\cos\vartheta) = \frac{1}{2\pi i} \oint\limits_{|z|=\frac{1}{2}} \frac{1}{\sqrt{1-2z\cos\vartheta+z^2}(1-z)} \frac{\mathrm{d}z}{z^{n+1}}.$$

We deform the contour of integration to encircle the branch cut of the square root, which is chosen to be the arc of the circle of radius one connecting the points $e^{\pm i\vartheta}$ passing through -1 (Fig. 1). This deformation of the contour passes through ∞ and the simple pole at $z = 1$, where we collect a residue. This gives

$$R_n(\cos\vartheta) = \frac{1}{2\sin\frac{\vartheta}{2}} - \frac{1}{2\sqrt{2}\pi} \int\limits_{\vartheta}^{2\pi-\vartheta} \frac{\cos(n+1)t}{\sqrt{\cos\vartheta - \cos t}\sin\frac{t}{2}} \,\mathrm{d}t.$$

We estimate this by

$$R_n(\cos\vartheta) \le \frac{1}{2\sin\frac{\vartheta}{2}} + \frac{1}{2\sqrt{2}\pi} \int\limits_{\vartheta}^{2\pi-\vartheta} \frac{1}{\sqrt{\cos\vartheta - \cos t}\sin\frac{t}{2}} \,\mathrm{d}t = \frac{1}{\sin\frac{\vartheta}{2}}. \tag{12}$$

This estimate is the best possible independent of n, because $R_{2n}(-1) = 1$.

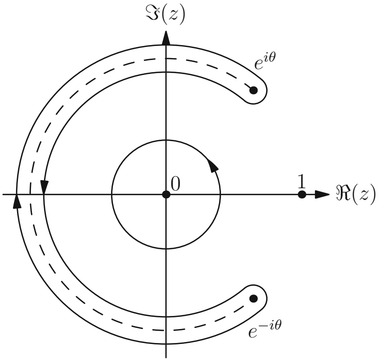


Fig. 1 The contour of integration used for deriving $R_n(\cos\vartheta)$

Putting the estimates (10) and (12) together we obtain

$$A_n^{2k+2}K_n^{k+\frac{1}{2}}(\cos\vartheta) \le \frac{1}{2^k\left(\sin\frac{\vartheta}{2}\right)^{2k+3}}\sum_{\ell=0}^{k}\binom{k}{\ell}\binom{n+k+\ell+1}{n}. \tag{13}$$

Summing up, we have proved the following.

Theorem 1 *Let $\lambda = \frac{d-1}{2}$ be a positive integer or half-integer. Then the kernel K_n^λ satisfies the following estimates*

$$K_n^\lambda(\cos\vartheta) \le \begin{cases} \dfrac{1}{2^{\lfloor\lambda\rfloor}\left(\sin\frac{\vartheta}{2}\right)^{2\lambda+2}}\displaystyle\sum_{\ell=0}^{\lfloor\lambda\rfloor}\binom{\lfloor\lambda\rfloor}{\ell}\frac{(2\lambda+1)_{\ell+1}}{(n+2\lambda+1)_{\ell+1}} & \textit{for}\quad 0<\vartheta\le\pi \\ \frac{(n+4\lambda+1)_n}{(n+2\lambda)_n} & \textit{for}\quad 0\le\vartheta\le\pi, \end{cases} \tag{14}$$

*where $(a)_n = a(a-1)\cdots(a-n+1)$ denotes the **falling** factorial (Pochhammer symbol).*

Remark 3 The estimate (14) is best possible with respect to the behaviour in n for a fixed $\vartheta \in (0,\pi)$, as well as for the power of $\sin\frac{\vartheta}{2}$. The constant in front of the main asymptotic term could still be improved, especially its dependence on the dimension. The second estimate is the trivial estimate by $K_n^\lambda(1)$.

3 A New Kernel

The kernel $K_n^\lambda(\cos\vartheta)$ exhibits a parity phenomenon at $\vartheta = \pi$, which occurs in the first asymptotic order term (see Fig. 2 for illustration). This comes from the fact that the two singularities at $e^{\pm i\vartheta}$ collapse to one singularity of twice the original order

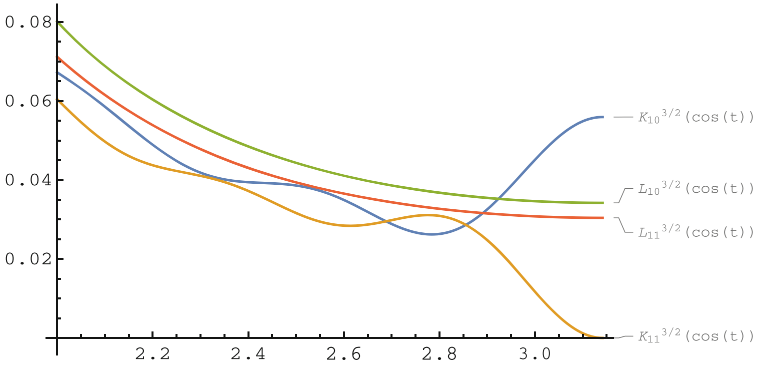


Fig. 2 Comparison between the kernels $K_{10}^{\frac{3}{2}}$, $K_{11}^{\frac{3}{2}}$, $L_{10}^{\frac{3}{2}}$, and $L_{11}^{\frac{3}{2}}$. The kernels K show oscillations and a parity phenomenon at $\vartheta = \pi$

for this value of ϑ. In order to avoid this, we propose to study the kernel given by the generating function

$$\frac{(1+z)^{2\lambda+2}}{(1-2z\cos\vartheta+z^2)^{\lambda+1}(1-z)^{2\lambda+1}} = \frac{1-z^2}{(1-2z\cos\vartheta+z^2)^{\lambda+1}} \times \frac{(1+z)^{2\lambda+1}}{(1-z)^{2\lambda+2}}. \tag{15}$$

Let B_n^λ be given by

$$\sum_{n=0}^{\infty} B_n^\lambda z^n = \frac{(1+z)^{2\lambda+1}}{(1-z)^{2\lambda+2}}, \tag{16}$$

then the kernel is given by

$$L_n^\lambda(\cos\vartheta) = \frac{1}{B_n^\lambda}\sum_{k=0}^{n} B_{n-k}^\lambda \frac{k+\lambda}{\lambda} C_k^\lambda(\cos\vartheta) \tag{17}$$

$$= \frac{1}{B_n^\lambda}\sum_{\ell=0}^{2\lambda+1}\binom{2\lambda+1}{\ell} A_{n-\ell}^{2\lambda+1} K_{n-\ell}^\lambda(\cos\vartheta). \tag{18}$$

The coefficients B_n^λ satisfy

$$B_n^\lambda = \sum_{\ell=0}^{2\lambda+1}\binom{2\lambda+1}{\ell}\binom{n-\ell+2\lambda+1}{n-\ell}$$
$$= \sum_{\ell=0}^{2\lambda+1}(-1)^\ell\binom{2\lambda+1}{\ell}2^{2\lambda+1-\ell}\binom{n-\ell+2\lambda+1}{n} \sim \frac{2^{2\lambda+1}n^{2\lambda+1}}{(2\lambda+1)!}.$$

The expression in the second line, which allows to read of the asymptotic behaviour immediately, is obtained by expanding the numerator in (16) into powers of $1-z$.

For $\lambda \in \mathbb{N}_0$ we write the generating function of $B_n^\lambda L_n^\lambda(\cos\vartheta)$ as

$$\left(\frac{(1+z)^2}{(1-2z\cos\vartheta+z^2)(1-z)^2}\right)^\lambda \times \frac{(1+z)^2}{(1-2z\cos\vartheta+z^2)(1-z)}. \tag{19}$$

The coefficients of the first factor are denoted by $S_n^\lambda(\cos\vartheta)$. They are obtained by successive convolution of

$$S_n^1(\cos\vartheta) = \frac{1}{2\pi i}\oint\limits_{|z|=\frac{1}{2}} \frac{(1+z)^2}{(1-2z\cos\vartheta+z^2)(1-z)^2}\frac{\mathrm{d}z}{z^{n+1}}$$
$$= \frac{n+1}{\left(\sin\frac{\vartheta}{2}\right)^2}\left(1-\frac{\cos\frac{\vartheta}{2}\sin(n+1)\vartheta}{2(n+1)\sin\frac{\vartheta}{2}}\right).$$

In order to estimate $S_n^1(\cos\vartheta)$, we estimate the sinc-function by its minimum

$$\operatorname{sinc}(t) = \frac{\sin t}{t} \geq -C' = -0.21723362821122165740827932556 2\ldots.$$

The value was obtained with the help of `Mathematica`. This gives

$$1-\cos\frac{\vartheta}{2}\frac{\sin(n+1)\vartheta}{2(n+1)\sin\frac{\vartheta}{2}} = 1-\operatorname{sinc}((n+1)\vartheta)\frac{\cos\frac{\vartheta}{2}}{\operatorname{sinc}(\frac{\vartheta}{2})}$$
$$\leq 1+C' =: C = 1.21723362821122165740827932556 2\ldots,$$

where we have used that $\cos\frac{\vartheta}{2} \leq \operatorname{sinc}(\frac{\vartheta}{2})$ for $0 \leq \vartheta \leq \pi$. From this we get the estimate

$$S_n^1(\cos\vartheta) \leq C\frac{n+1}{\left(\sin\frac{\vartheta}{2}\right)^2}$$

and consequently

$$S_n^\lambda(\cos\vartheta) \leq \frac{C^\lambda}{\left(\sin\frac{\vartheta}{2}\right)^{2\lambda}}\binom{n+2\lambda-1}{n} \tag{20}$$

by successive convolution as before.

Remark 4 This expression is bit simpler than the corresponding estimate for Q_n^λ, because the iterated convolution of the terms $n+1$ is a binomial coefficient, whereas the iterated convolution of terms $n+2$ can only be expressed as a linear combination of binomial coefficients. The growth order is the same.

In a similar way we estimate the coefficient of the second factor in (19)

$$\frac{1}{2\pi i}\oint\limits_{|z|=\frac{1}{2}} \frac{(1+z)^2}{(1-2z\cos\vartheta+z^2)(1-z)}\frac{\mathrm{d}z}{z^{n+1}}$$
$$= \frac{1}{2\left(\sin\frac{\vartheta}{2}\right)^2}\left(2-\cos n\vartheta-\cos(n+1)\vartheta\right) \leq \frac{2}{\left(\sin\frac{\vartheta}{2}\right)^2}.$$

As before, this is the kernel function for $\lambda = 0$.

Putting this estimate together with (20) we obtain

$$B_n^\lambda L_n^\lambda(\cos\vartheta) \le \frac{2C^\lambda}{\left(\sin\frac{\vartheta}{2}\right)^{2\lambda+2}}\binom{n+2\lambda}{n} \tag{21}$$

for $\lambda \in \mathbb{N}_0$.

For $\lambda = k + \frac{1}{2}$ ($k \in \mathbb{N}_0$) we factor the generating function as

$$\frac{(1+z)}{\sqrt{1-2z\cos\vartheta+z^2}(1-z)} \times \frac{(1+z)^{2k+2}}{(1-2z\cos\vartheta+z^2)^{k+1}(1-z)^{2k+1}}. \tag{22}$$

We still have to estimate the coefficient of the first factor, which is given by the integral

$$T_n(\cos\vartheta) = \frac{1}{2\pi i}\oint\limits_{|z|=\frac{1}{2}} \frac{(1+z)}{\sqrt{1-2z\cos\vartheta+z^2}(1-z)}\frac{\mathrm{d}z}{z^{n+1}}.$$

We transform this integral in the same way as we did before using the contour in Fig. 1 which yields

$$T_n(\cos\vartheta) = \frac{1}{\sin\frac{\vartheta}{2}} - \frac{1}{\pi\sqrt{2}}\int\limits_{\vartheta}^{2\pi-\vartheta} \frac{\cos\frac{t}{2}\cos(n+\frac{1}{2})t}{\sqrt{\cos\vartheta-\cos t}\sin\frac{t}{2}}\,\mathrm{d}t. \tag{23}$$

The modulus of the integral can be estimated by

$$\frac{\sqrt{2}}{\pi}\int\limits_{\vartheta}^{\pi} \frac{\cos\frac{t}{2}}{\sqrt{\cos\vartheta-\cos t}\sin\frac{t}{2}}\,\mathrm{d}t = \frac{\pi-\vartheta}{\pi\sin\frac{\vartheta}{2}} \le \frac{1}{\sin\frac{\vartheta}{2}}.$$

This gives the bound

$$T_n(\cos\vartheta) \le \frac{2}{\sin\frac{\vartheta}{2}}. \tag{24}$$

Putting this estimate together with (21) we obtain

$$B_n^\lambda L_n^\lambda(\cos\vartheta) \le \frac{4C^k}{\left(\sin\frac{\vartheta}{2}\right)^{2k+3}}\binom{n+2k+1}{n} \tag{25}$$

for $\lambda = k + \frac{1}{2}$.

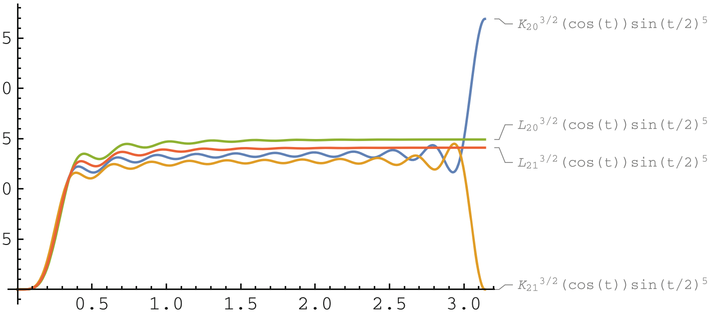


Fig. 3 Plots of the functions $K_{20}^{\frac{3}{2}}(\cos\vartheta)(\sin\frac{\vartheta}{2})^5$, $K_{21}^{\frac{3}{2}}(\cos\vartheta)(\sin\frac{\vartheta}{2})^5$, $L_{20}^{\frac{3}{2}}(\cos\vartheta)(\sin\frac{\vartheta}{2})^5$, and $L_{21}^{\frac{3}{2}}(\cos\vartheta)(\sin\frac{\vartheta}{2})^5$. Again the parity phenomenon for the kernel K is prominently visible

Summing up, we have proved the following. As before, the second estimate is just the trivial estimate by $L_n^\lambda(1)$.

Theorem 2 *Let $\lambda = \frac{d-1}{2}$ be a positive integer or half-integer. Then the kernel L_n^λ satisfies the following estimates*

$$L_n^\lambda(\cos\vartheta) \le \begin{cases} D_\lambda \dfrac{C^{\lfloor\lambda\rfloor}}{B_n^\lambda\left(\sin\frac{\vartheta}{2}\right)^{2\lambda+2}}\binom{n+2\lambda}{n} & \textit{for} \quad 0<\vartheta\le\pi \\ \frac{1}{B_n^\lambda}\sum_{\ell=0}^{2\lambda+2}\binom{2\lambda+2}{\ell}2^{2\lambda+2-\ell}(-1)^\ell\binom{n+4\lambda+2-\ell}{n} & \textit{for} \quad 0\le\vartheta\le\pi, \end{cases} \tag{26}$$

where $D_\lambda = 2$ for $\lambda \in \mathbb{N}$ and $D_\lambda = 4$, if $\lambda \in \frac{1}{2} + \mathbb{N}_0$.

Remark 5 Notice that the orders of magnitude in terms of n and the powers of $\sin\frac{\vartheta}{2}$ are the same for L_n^λ as for the kernel K_n^λ. This fact is illustrated by Fig. 3. The coefficient of the asymptotic leading term of the estimate decays like $(2\lambda+1)(C/4)^\lambda$ for L_n^λ, whereas this coefficient decays like $(2\lambda + 1)(1/2)^\lambda$ for K_n^λ.

Acknowledgements The author is supported by the Austrian Science Fund FWF projects F5503 (part of the Special Research Program (SFB) "Quasi-Monte Carlo Methods: Theory and Applications") and W1230 (Doctoral Program "Discrete Mathematics"). The author is grateful to two anonymous referees for their many helpful comments.

References

1. Andrews, G.E., Askey, R., Roy, R.: Special functions. In: Encyclopedia of Mathematics and its Applications, vol. 71. Cambridge University Press, Cambridge (1999)
2. Berens, H., Butzer, P.L., Pawelke, S.: Limitierungsverfahren von Reihen mehrdimensionaler Kugelfunktionen und deren Saturationsverhalten. Publ. Res. Inst. Math. Sci. Ser. A **4**, 201–268 (1968/1969)

3. Kogbetliantz, E.: Recherches sur la sommabilité; des séries ultra-sphériques par la méthode des moyennes arithmétiques. J. Math. Pures Appl. **3**, 107–188 (1924)
4. Reimer, M.: A short proof of a result of Kogbetliantz on the positivity of certain Cesàro means. Math. Z **221**(2), 189–192 (1996)

Modern Monte Carlo Variants for Uncertainty Quantification in Neutron Transport

Ivan G. Graham, Matthew J. Parkinson, and Robert Scheichl

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We describe modern variants of Monte Carlo methods for Uncertainty Quantification (UQ) of the Neutron Transport Equation, when it is approximated by the discrete ordinates method with diamond differencing. We focus on the mono-energetic 1D slab geometry problem, with isotropic scattering, where the cross-sections are log-normal correlated random fields of possibly low regularity. The paper includes an outline of novel theoretical results on the convergence of the discrete scheme, in the cases of both spatially variable and random cross-sections. We also describe the theory and practice of algorithms for quantifying the uncertainty of a functional of the scalar flux, using Monte Carlo and quasi-Monte Carlo methods, and their multilevel variants. A hybrid iterative/direct solver for computing each realisation of the functional is also presented. Numerical experiments show the effectiveness of the hybrid solver and the gains that are possible through quasi-Monte Carlo sampling and multilevel variance reduction. For the multilevel quasi-Monte Carlo method, we observe gains in the computational ε-cost of up to two orders of magnitude over the standard Monte Carlo method, and we explain this theoretically. Experiments on problems with up to several thousand stochastic dimensions are included.

1 Introduction

In this paper we will consider the Neutron Transport equation (NTE), sometimes referred to as the Boltzmann transport equation. This is an integro-differential equation which models the flux of neutrons in a reactor. It has particular applications

I. G. Graham (✉) · M. J. Parkinson · R. Scheichl
University of Bath, Bath, UK
e-mail: i.g.graham@bath.ac.uk; m.parkinson@bath.ac.uk; r.scheichl@bath.ac.uk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_22

for nuclear reactor design, radiation shielding and astrophysics [44]. There are many potential sources of uncertainty in a nuclear reactor, such as the geometry, material composition and reactor wear. Here, we will consider the problem of random spatial variation in the coefficients (the *cross-sections*) in the NTE, represented by correlated random fields with potentially low smoothness. Our aim is to understand how *uncertainty in the cross-sections propagates through to (functionals of) the neutron flux*. This is the forward problem of Uncertainty Quantification.

We will quantify the uncertainty using Monte Carlo (MC) type methods, that is, by simulating a finite number of pseudo-random instances of the NTE and by averaging the outcome of those simulations to obtain statistics of quantities of interest. Each statistic can be interpreted as an expected value of some (possibly nonlinear) functional of the neutron flux with respect to the random cross-sections. The input random fields typically need to be parametrised with a significant number of random parameters leading to a problem of high-dimensional integration. MC methods are known to be particularly well-suited to this type of problem due to their dimension independent convergence rates.

However, convergence of the MC algorithm is slow and determined by $\sqrt{\mathbb{V}(\cdot)/N}$, where $\mathbb{V}(\cdot)$ is the variance of the quantity of interest and N is the number of samples. For this reason, research is focussed on improving the convergence, whilst retaining dimensional independence. Advances in MC methods can broadly be split into two main categories: improved sampling and variance reduction. Improved sampling methods attempt to find samples that perform better than the pseudo-random choice. Effectively, they aim to improve the $\sqrt{1/N}$ term in the error estimate. A major advance in sampling methods has come through the development of quasi-Monte Carlo (QMC) methods. Variance reduction methods, on the other hand, attempt to reduce the $\mathbb{V}(\cdot)$ term in the error estimate and thus reduce the number of samples needed for a desired accuracy. Multilevel Monte Carlo (MLMC) methods (initiated in [18, 28] and further developed in, e.g., [7, 9, 10, 20, 27, 32, 34, 47]) fall into this category. A comprehensive review of MLMC can be found in [19].

The rigorous theory of all of the improvements outlined above requires regularity properties of the solution, the verification of which can be a substantial task. There are a significant number of published papers on the regularity of parametric elliptic PDEs, in physical and parameter space, as they arise, e.g., in flow in random models of porous media [9, 12, 13, 23, 32–34]. However, for the NTE, this regularity question is almost untouched. Our complementary paper [25] contains a full regularity and error analysis of the discrete scheme for the NTE with spatially variable and random coefficients. Here we restrict to a summary of those results.

The field of UQ has grown very quickly in recent years and its application to neutron transport theory is currently of considerable interest. There are a number of groups that already work on this problem, e.g. [4, 17, 21] and references therein. Up to now, research has focussed on using the polynomial chaos expansion (PCE), which comes in two forms; the intrusive and non-intrusive approaches. Both approaches expand the random flux in a weighted sum of orthogonal polynomials. The intrusive approach considers the expansion directly in the differential equation, which in turn requires a new solver ('intruding' on the original solver). In contrast, the non-intrusive approach attempts to estimate the coefficients of the PCE directly, by projecting onto the PCE basis cf. [4, Eq. (40)]. This means the original solver can be used as a 'black box' as in MC methods. Both of the approaches then use quadrature to estimate the coefficients in the PCE. The main disadvantage of standard PCE is that typically the number of terms grow exponentially in the number of stochastic dimensions and in the order of the PCE, the so-called *curse of dimensionality*.

Fichtl and Prinja [17] were some of the first to numerically tackle the 1D slab geometry problem with random cross-sections. Gilli et al. [21] improved upon this work by using (adaptive) sparse grid ideas in the collocation method, to tackle the curse of dimensionality. Moreover, [5] constructed a hybrid PCE using a combination of Hermite and Legendre polynomials, observing superior convergence in comparison to the PCE with just Hermite polynomials. More recently [4] tackled the (time-independent) full criticality problem in three spatial, two angular and one energy variable. They consider a second expansion, the high-dimensional model representation (HDMR), which allows them to expand the response (e.g. functionals of the flux) in terms of low-dimensional subspaces of the stochastic variable. The PCE is used on the HDMR terms, each with their own basis and coefficients. We note however, that none of these papers provide any rigorous error or cost analysis.

The structure of this paper is as follows. In Sect. 2, we describe the model problem, a 1D slab geometry simplification of the Neutron Transport Equation with spatially varying and random cross-sections. We set out the discretisation of this equation and discuss two methods for solving the resultant linear systems; a direct and an iterative solver. In Sect. 3, the basic elements of a fully-discrete error analysis of the discrete ordinates method with diamond differencing applied to the model problem are summarised. The full analysis will be given in [25]. In Sect. 4, we introduce a number of variations on the Monte Carlo method for quantifying uncertainty. This includes a summary of the theoretical computational costs for each method. Finally, Sect. 5 contains numerical results relating to the rest of the paper. We first present a hybrid solver that combines the benefits of both direct and iterative solvers. Its cost depends on the particular realisation of the cross-sections. Moreover, we present simulations for the UQ problem for the different variants of the Monte Carlo methods, and compare the rates with those given by the theory.

2 The Model Problem

The Neutron Transport Equation (NTE) is a physically derived balance equation, that models the angular flux $\psi(\mathbf{r}, \Theta, E)$ of neutrons in a domain, where $\mathbf{r}$ is position, Θ is angle and E is energy. Neutrons are modelled as non-interacting particles travelling along straight line paths with some energy E. They interact with the larger nuclei via absorption, scattering and fission. The rates σ_A, σ_S and σ_F at which these events occur are called the *absorption, scattering and fission cross-sections*, respectively. They can depend on the position **r** and the energy E of the neutron. The scattering cross-sections also depend on the energy E' after the scattering event, as well as on the angles Θ and Θ' before and after the event.

The two main scenarios of interest in neutron transport are the so-called *fixed source problem* and the *criticality problem*. We will focus on the former, which concerns the transport of neutrons emanating from some fixed source term f. It has particular applications in radiation shielding. We will further simplify our model to the *1D slab geometry case* by assuming

- no energy dependence;
- dependence only on one spatial dimension and infinite extent of the domain in the other two dimensions;
- no dependence of any cross-sections on angle;
- no fission.

The resulting simplified model is an integro-differential equation for the angular flux $\psi(x, \mu)$ such that

$$\mu \frac{\partial \bar{\psi}}{\partial x}(x, \mu) \ + \ \sigma(x)\psi(x, \mu) \ = \ \sigma_S(x)\phi(x) \ + \ f(x) \ , \tag{1}$$

$$\text{where} \qquad \phi(x) \ = \ \frac{1}{2}\int_{-1}^{1} \psi(x, \mu')\, \mathrm{d}\mu' \ , \tag{2}$$

for any $x \in (0, 1)$ and $\mu \in [-1, 1]$, subject to the no in-flow boundary conditions

$$\psi(0, \mu) \ = \ 0, \ \text{ for } \ \mu > 0 \quad \text{and} \quad \psi(1, \mu) = 0, \ \text{ for } \ \mu < 0 \, . \tag{3}$$

Here, the angular domain is reduced from $\mathbb{S}_2$ to the unit circle $\mathbb{S}_1$ and parametrised by the cosine $\mu \in [-1, 1]$ of the angle. The equation degenerates at $\mu = 0$, i.e. for neutrons moving perpendicular to the x-direction. The coefficient function $\sigma(x)$ is the total cross-section given by $\sigma = \sigma_S + \sigma_A$. For more discussion on the NTE see [11, 37].

2.1 Uncertainty Quantification

An important problem in industry is to quantify the uncertainty in the fluxes due to uncertainties in the cross-sections. Most materials, in particular shielding materials such as concrete, are naturally heterogeneous or change their properties over time through wear. Moreover, the values of the cross-sections are taken from nuclear data libraries across the world and they can differ significantly between libraries [36]. This means there are large amounts of uncertainty on the coefficients, and this could have significant consequences on the system itself.

To describe the random model, let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space with $\omega \in \Omega$ denoting a random event from this space. Consider a (finite) set of partitions of the spatial domain, where on each subinterval we assume that $\sigma_S = \sigma_S(x, \omega)$ and $\sigma = \sigma(x, \omega)$ are two (possibly dependent or correlated) random fields. Then the angular flux and the scalar flux become random fields and the model problem (1), (2) becomes

$$\mu \frac{\partial \bar{\psi}}{\partial x}(x, \mu, \omega) \; + \; \sigma(x, \omega) \psi(x, \mu, \omega) \; = \; \sigma_S(x, \omega) \phi(x, \omega) \; + \; f(x) \; , \tag{4}$$

$$\text{where} \qquad \phi(x, \omega) \; = \; \int_{-1}^{1} \psi(x, \mu', \omega) d\mu' \tag{5}$$

and $\psi(\cdot, \cdot, \omega)$ satisfies the boundary conditions (3). The set of Eqs. (4), (5), (3) have to hold for almost all realisations $\omega \in \Omega$.

For simplicity, we restrict ourselves to deterministic $\sigma_A = \sigma_A(x)$ with

$$0 \; < \; \sigma_{A,\min} \; \leq \; \sigma_A(x) \; \leq \; \sigma_{A,\max} \; < \; \infty \; , \quad \text{for all} \quad x \in [0, 1] \; , \tag{6}$$

and assume a log-normal distribution for $\sigma_S(x, \omega)$. The total cross-section $\sigma(x, \omega)$ is then simply the log-normal random field with values $\sigma(x, \omega) = \sigma_S(x, \omega) + \sigma_A(x)$. In particular, we assume that $\log \sigma_S$ is a correlated zero mean Gaussian random field, with covariance function defined by

$$C_\nu(x, y) \; = \; \sigma_{var}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(2\sqrt{\nu} \frac{|x - y|}{\lambda_C} \right)^\nu K_\nu \left(2\sqrt{\nu} \frac{|x - y|}{\lambda_C} \right) \; . \tag{7}$$

This class of covariances is called the Matérn class. It is parametrised by the smoothness parameter $\nu \geq 0.5$; λ_C is the correlation length, σ_{var}^2 is the variance, Γ is the gamma function and K_ν is the modified Bessel function of the second kind. The limiting case, i.e. $\nu \to \infty$, corresponds to the Gaussian covariance function $C_\infty(x, y) \; = \; \sigma_{var}^2 \exp(-|x - y|^2 / \lambda_C^2)$.

To sample from σ_S we use the Karhunen-Loève (KL) expansion of $\log \sigma_S$, i.e.,

$$\log \sigma_S(x, \omega) \; = \; \sum_{i=1}^{\infty} \sqrt{\xi_i} \; \eta_i(x) \; Z_i(\omega) \; , \tag{8}$$

where $Z_i \sim \mathscr{N}(0, 1)$ i.i.d. Here ξ_i and η_i are the eigenvalues and the $L^2(0, 1)$-orthogonal eigenfunctions of the covariance integral operator associated with kernel given by the covariance function in (7). In practice, the KL expansion needs to be truncated after a finite number of terms (here denoted d). The accuracy of this truncation depends on the decay of the eigenvalues [38]. For $\nu < \infty$, this decay is algebraic and depends on the smoothness parameter ν. In the Gaussian covariance case the decay is exponential. Note that for the Matérn covariance with $\nu = 0.5$, the eigenvalues and eigenfunctions can be computed analytically [38]. For other cases of ν, we numerically compute the eigensystem using the Nyström method—see, for example, [16].

The goal of stochastic uncertainty quantification is to understand how the randomness in σ_S and σ propagates to functionals of the scalar or angular flux. Such quantities of interest may be point values, integrals or norms of ϕ or ψ. They are random variables and the focus is on estimating their mean, variance or distribution.

2.2 Discretisation

For each realisation $\omega \in \Omega$, the stochastic 1D NTE (4), (5), (3) is an integro-differential equation in two variables, space and angle. For ease of presentation, we suppress the dependency on $\omega \in \Omega$ for the moment.

We use a $2N$-point quadrature rule $\int_{-1}^{1} f(\mu)d\mu \approx \sum_{|k|=1}^{N} w_k f(\mu_k)$ with nodes $\mu_k \in [-1,1]\backslash\{0\}$ and positive weights w_k to discretise in angle, assuming the (anti-) symmetry properties $\mu_{-k} = -\mu_k$ and $w_{-k} = w_k$. (In later sections, we construct such a rule by using N-point Gauss-Legendre rules on each of $[-1,0)$ and $(0,1]$.)

To discretise in space, we introduce a mesh $0 = x_0 < x_1 < \ldots < x_M = 1$ which is assumed to resolve any discontinuities in the cross-sections σ, σ_S and is also quasiuniform—i.e. the subinterval lengths $h_j := x_j - x_{j-1}$ satisfy $\gamma h \le h_j \le h := \max_{j=1,\ldots M} h_j$, for some constant $\gamma > 0$. Employing a simple Crank-Nicolson method for the transport part of (4), (5) and combining it with the angular quadrature rule above we obtain the classical *diamond-differencing* scheme:

$$\begin{aligned}\mu_k \frac{\Psi_{k,j} - \Psi_{k,j-1}}{h_j} &+ \sigma_{j-1/2}\frac{\Psi_{k,j} + \Psi_{k,j-1}}{2} \\ &= \sigma_{S,j-1/2}\Phi_{j-1/2} + F_{j-1/2}\,, \quad j = 1,\ldots,M,\ |k| = 1,\ldots,N, \end{aligned} \tag{9}$$

where

$$\Phi_{j-1/2} = \frac{1}{2}\sum_{|k|=1}^{N} w_k \frac{\Psi_{k,j} + \Psi_{k,j-1}}{2}\,, \ j = 1,\ldots,M\,. \tag{10}$$

Here $\sigma_{j-1/2}$ denotes the value of σ at the mid-point of the interval $I_j = (x_{j-1}, x_j)$, with the analogous meaning for $\sigma_{S,j-1/2}$ and $F_{j-1/2}$. The notation reflects the fact that (in the next section) we will associate the unknowns $\Psi_{k,j}$ in (9) with the nodal values $\psi_{k,h}(x_j)$ of continuous piecewise-linear functions $\psi_{k,h} \approx \psi(\cdot, \mu_k)$.

Finally, (9) and (10) have to be supplemented with the boundary conditions $\Psi_{k,0} = 0$, for $k > 0$ and $\Psi_{k,M} = 0$, for $k < 0$. If the right-hand side of (9) were known, then (9) could be solved simply by sweeping from left to right (when $k > 0$) and from right to left (when $k < 0$). The appearance of $\Phi_{j-1/2}$ on the right-hand side means that (9) and (10) consitute a coupled system with solution

$(\Psi, \Phi) \in \mathbb{R}^{2NM} \times \mathbb{R}^M$. It is helpful to think of Ψ as being composed of $2N$ subvectors Ψ_k, each with M entries $\Psi_{k,j}$, consisting of approximations to $\psi(x_j, \mu_k)$ with x_j ranging over all free nodes.

The coupled system (9) and (10) can be written in matrix form as

$$\begin{pmatrix} T & -\Sigma_S \\ -P & I \end{pmatrix} \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} = \begin{pmatrix} F \\ \mathbf{0} \end{pmatrix} . \tag{11}$$

Here, the vector $\Phi \in \mathbb{R}^M$ contains the approximations of the scalar flux at the M midpoints of the spatial mesh. The matrix T is a block diagonal $2NM \times 2NM$ matrix, representing the left hand side of (9). The $2N$ diagonal blocks of T, one per angle, are themselves bi-diagonal. The $2NM \times M$ matrix Σ_S simply consists of $2N$ identical diagonal blocks, one per angle, representing the multiplication of Φ by σ_S at the midpoints of the mesh. The $M \times 2NM$ matrix P represents the right hand side of (10), i.e. averaging at the midpoints and quadrature. The matrix I denotes the $M \times M$ identity matrix. The vector $F \in \mathbb{R}^{2NM}$ contains $2N$ copies of the source term evaluated at the M midpoints of the spatial mesh.

2.3 Direct and Iterative Solvers

We now wish to find the (approximate) fluxes in the linear system (11). We note that the matrix T is invertible and has a useful sparsity structure that allows its inverse to be calculated in $\mathscr{O}(MN)$ operations. However, the bordered system (11) is not as easy to invert, due to the presence of Σ_S and P.

To exploit the sparsity of T, we do block elimination on (11) obtaining the Schur complement system for the scalar flux, i.e.,

$$\big(I - PT^{-1}\Sigma_S\big)\,\Phi = PT^{-1}F , \tag{12}$$

which now requires the inversion of a smaller (dense) matrix. Note that (12) is a finite-dimensional version of the reduction of the integro-differential equation (4), (5) to the integral form of the NTE, see (20). In this case, the two dominant computations with $\mathscr{O}(M^2N)$ and $\mathscr{O}(M^3)$ operations respectively, are the triple matrix product $PT^{-1}\Sigma_S$ in the construction of the Schur complement and the LU factorisation of the $M \times M$ matrix $\big(I - PT^{-1}\Sigma_S\big)$. This leads to a total

$$\text{theoretical cost of the direct solver} \sim \mathscr{O}(M^2(M + N)) . \tag{13}$$

We note that for stability reasons (see Sect. 3, also [42] in a simpler context), the number of spatial and angular points should be related. A suitable choice is $M \sim N$, leading to a cost of the direct solver of $\mathscr{O}(M^3)$ in general.

The second approach for solving (11) is an iterative solver commonly referred to as *source iteration*, cf. [8]. The form of (12) naturally suggests the iteration

$$\Phi^{(k)} = PT^{-1}\left(\Sigma_S \Phi^{(k-1)} + F\right) , \tag{14}$$

where $\Phi^{(k)}$ is the approximation at the kth iteration, with $\Phi^{(0)} = PT^{-1}F$. This can be seen as a discrete version of an iterative method for the integral equation (20).

In practice, we truncate after K iterations. The dominant computations in the source iteration are the K multiplications with $PT^{-1}\Sigma_S$. Exploiting the sparsity of all the matrices involved, these multiplications cost $\mathscr{O}(MN)$ operations, leading to an overall

$$\text{theoretical cost of source iteration} \sim \mathscr{O}(MNK) . \tag{15}$$

Our numerical experiments in Sect. 5 show that for $N = 2M$ the hidden constants in the two estimates (13) and (15) are approximately the same. Hence, whether the iterative solver is faster than the direct solver depends on whether the number of iterations K to obtain an accurate enough solution is smaller or larger than M.

There are sharp theoretical results on the convergence of source iteration for piecewise smooth cross-sections [8, Thm 2.20]. In particular, if $\phi^{(K)}(\omega)$ denotes the approximation to $\phi(\omega)$ after K iterations, then

$$\left\| \sigma^{1/2}\left(\phi - \phi^{(K)}\right) \right\|_2 \le C' \left(\eta \left\| \frac{\sigma_S}{\sigma} \right\|_\infty \right)^K , \tag{16}$$

for some constant C' and $\eta \le 1$. That is, the error decays geometrically with rate no slower than the spatial maximum of σ_S/σ. This value depends on ω and will change pathwise. Using this result as a guide together with (6), we assume that the convergence of the L^2-error with respect to K can be bounded by

$$\|\phi - \phi^{(K)}\|_2 \le C \left\| \frac{\sigma_S}{\sigma} \right\|_\infty^K , \tag{17}$$

for some constant C that we will estimate numerically in Sect. 5.

3 Summary of Theoretical Results

The rigorous analysis of UQ for PDEs with random coefficients requires estimates for the error when discretisations in physical space (e.g. by finite differences) and probability space (e.g. by sampling techniques) are combined. The physical error estimates typically need to be probabilistic in form (e.g. estimates of expectation of the physical error). Such estimates are quite well-developed for elliptic PDEs—

see for example [9] but this question is almost untouched for the transport equation (or more specifically the NTE). We outline here some results which are proved in the forthcoming paper [25]. This paper proceeds by first giving an error analysis for (1), (2) with variable cross-sections, which is explicit in σ, σ_S, and then uses this to derive probabilistic error estimates for the spatial discretisation (9), (10).

The numerical analysis of the NTE (and related integro-differential equation problems such as radiative transfer) dates back at least as far as the work of Keller [30]. After a huge growth in the mathematics literature in the 1970s and 1980s, progress has been slower since. This is perhaps surprising, since discontinuous Galerkin (DG) methods have enjoyed a massive recent renaissance and the solution of the neutron transport problem was one of the key motivations behind the original introduction of DG [43]. Even today, an error analysis of the NTE with variable (even deterministic) cross-sections (with explicit dependence on the data) is still not available, even for the model case of mono-energetic 1D slab geometry considered here.

The fundamental paper on the analysis of the discrete ordinates method for the NTE is [42]. Here a full analysis of the combined effect of angular and spatial discretisation is given under the assumption that the cross-sections σ and σ_S in (4) are constant. The delicate relation between spatial and angular discretisation parameters required to achieve stability and convergence is described there. Later research e.g. [2, 3] produced analogous results for models of increasing complexity and in higher dimensions, but the proofs were mostly confined to the case of cross-sections that are constant in space. A separate and related sequence of papers (e.g. [35, 48], and [1]) allow for variation in cross-sections, but error estimates explicit in this data are not available there.

The results outlined here are orientated to the case when σ, σ_S have relatively rough fluctuations. As a precursor to attacking the random case, we first consider rough deterministic coefficients defined as follows. We assume that there is some partition of [0, 1] and that σ, σ_S are C^η functions on each subinterval of the partition (with $\eta \in (0, 1]$), but that σ, σ_S may be discontinuous across the break points. We assume that the mesh $\{x_j\}_{j=o}^M$ introduced in Sect. 2.2 resolves these break points. (Here C^η is the usual Hölder space of index η with norm $\|\cdot\|_\eta$.) We also assume that the source function $f \in C^\eta$.

When discussing the error when (9), (10) is applied to (1), (2), it is useful to consider the "pure transport" problem:

$$\mu \frac{\mathrm{d}u}{\mathrm{d}x} + \sigma u = g, \text{ with } u(0) = 0, \text{ when } \mu > 0 \text{ and } u(1) = 0 \text{ when } \mu < 0, \tag{18}$$

and with $g \in C$ a generic right-hand side (where $\mu \neq 0$ is now a parameter). Application of the Crank-Nicolson method (as in (9)) yields

$$\mu \left(\frac{U_j - U_{j-1}}{h_j} \right) + \sigma_{j-1/2} \left(\frac{U_j + U_{j-1}}{2} \right) = g_{j-1/2}\ , \quad \text{for } j = 1, \ldots, M\ , \tag{19}$$

with analogous boundary conditions, where, for any continuous function c, we use $c_{j-1/2}$ to denote $c(x_{j-1/2})$. Letting V^h denote the space of continuous piecewise linear functions with respect to the mesh $\{x_j\}$, (19) is equivalent to seeking a $u^h \in V^h$ (with nodal values U_j) such that

$$\int_{I_j} \left(\mu \frac{\mathrm{d}u^h}{\mathrm{d}x} + \widetilde{\sigma} u^h \right) = \int_{I_j} \widetilde{g}\,, \quad j = 1, \ldots, M, \quad \text{where} \quad I_j = (x_{j-1}, x_j),$$

and $\widetilde{c}$ denotes the piecewise constant function with respect to the grid $\{x_j\}$ which interpolates c at the mid-points of subintervals.

It is easy to show that both (18) and (19) have unique solutions and we denote the respective solution operators by $\mathscr{S}_\mu$ and $\mathscr{S}_\mu^h$, i.e.

$$u = \mathscr{S}_\mu g \quad \text{and} \quad u^h = \mathscr{S}_\mu^h g\ .$$

Bearing in mind the angular averaging process in (2) and (10), it is useful to then introduce the corresponding continuous and discrete spatial operators:

$$(\mathscr{K} g)(x) := \frac{1}{2} \int_{-1}^{1} \left(\mathscr{S}_\mu g\right)(x) \mathrm{d}\mu, \quad \text{and} \quad (\mathscr{K}^{h,N} g)(x) \;=\; \frac{1}{2} \sum_{|k|=1}^{N} w_k (\mathscr{S}_{\mu_k}^h g)(x)\ .$$

It is easy to see (and well known classically—e.g. [29]) that

$$(\mathscr{K} g)(x) = \frac{1}{2} \int_0^1 E_1(|\tau(x,y)|) g(y) \mathrm{d}y,$$

where E_1 is the exponential integral and the function $\tau(x,y) = \int_x^y \sigma$ is known as the optical path. In fact (even when σ is merely continuous), $\mathscr{K}$ is a compact Fredholm integral operator on a range of function spaces and $\mathscr{K}^{h,N}$ is a finite rank approximation to it. The study of these integral operators in the deterministic case is a classical topic, e.g. [45]. In the case of random σ, $\mathscr{K}$ is an integral operator with a random kernel which merits further investigation. Returning to (1), (2), we see readily that

$$\psi(x,\mu) = \mathscr{S}_\mu(\sigma_S \phi + f)(x), \quad \text{so that} \quad \phi \;=\; \mathscr{K}(\sigma_S \phi + f). \tag{20}$$

Moreover (9) and (10) correspond to a discrete analogue of (20) as follows. Introduce the family of functions $\psi_k^{h,N} \in V^h$, $|k| = 1, \ldots, N$, by requiring $\psi_k^{h,N}$ to have nodal values $\Psi_{k,j}$. Then set

$$\phi^{h,N} := \frac{1}{2} \sum_{|k|=1}^{N} w_k \psi_k^{h,N} \;\in V^h,$$

and it follows that (9) and (10) may be rewritten (for each $j = 1, \ldots, M$)

$$\int_{I_j} \left(\mu_k \frac{\mathrm{d}\psi_k^{h,N}}{\mathrm{d}x} + \widetilde{\sigma}\psi_k^{h,N} \right) = \int_{I_j} \widetilde{g^{h,N}} , \quad \text{where} \quad g^{h,N} = \sigma_S \phi^{h,N} + f .$$

and thus

$$\psi_k^{h,N} = \mathscr{S}_{\mu_k}^{h} \left(\sigma_S \phi^{h,N} + f \right), \quad \text{so that} \quad \phi^{h,N} = \mathscr{K}^{h,N}(\sigma_S \phi^{h,N} + f) . \tag{21}$$

The numerical analysis of (9) and (10) is done by analysing (the second equation in) (21) as an approximation of the second equation in (20). This is studied in detail in [42] for constant σ, σ_S. In [25] we discuss the variable case, obtaining all estimates explicitly in σ, σ_S. Elementary manipulation on (20) and (21) shows that

$$\phi - \phi^{h,N} = (I - \mathscr{K}^{h,N}\sigma_S)^{-1}(\mathscr{K} - \mathscr{K}^{h,N})(\sigma_S\phi + f), \tag{22}$$

and so

$$\|\phi - \phi^{h,N}\|_\infty \leq \|(I - \mathscr{K}^{h,N}\sigma_S)^{-1}\|_\infty \|(\mathscr{K} - \mathscr{K}^{h,N})(\sigma_S\phi + f)\|_\infty . \tag{23}$$

The error analysis in [25] proceeds by estimating the two terms on the right-hand side of (23) separately. We summarise the results in the lemmas below. To avoid writing down the technicalities (which will be given in detail in [25]), in the following results, we do not give the explicit dependence of the constants C_i, $i = 1, 2, \ldots,$ on the cross sections σ and σ_S. For simplicity we restrict our summary to the case when the right-hand side of (19) is the average of g over I_j (rather than the point value $f, \sigma, \sigma_{S \in C^2}$ without jumps and when $g_{j-1/2}$). The actual scheme (19) is then analysed by a perturbation argument, see [25].

Lemma 1 *Suppose N is sufficiently large and $h \log N$ is sufficiently small. Then*

$$\|(I - \mathscr{K}^{h,N}\sigma_S)^{-1}\|_\infty \leq C_1 , \tag{24}$$

where C_1 depends on σ and σ_S, but is independent of h and N.

Sketch of Proof The proof is obtained by first obtaining an estimate of the form (24) for the quantity $\|(I - \mathscr{K}\sigma_S)^{-1}\|_\infty$, and then showing that the perturbation $\|\mathscr{K} - \mathscr{K}^{h,N}\|_\infty$ is small, when N is sufficiently large and $h \log N$ is sufficiently small. (The constraint linking h and $\log N$ arises because the transport equation (1) has a singularity at $\mu = 0$.) The actual values of h, N which are sufficient to ensure that the bound (24) holds depend on the cross-sections σ, σ_S.

Lemma 2

$$\|(\mathscr{K} - \mathscr{K}^{h,N})(\sigma_S\phi + f)\|_\infty \leq \left(C_2\, h \log N + C_3\, h^\eta + C_4 \frac{1}{N} \right) \|f\|_\eta ,$$

where C_2, C_3, C_4 depend again on σ and σ_S, but are independent of h, N and f.

Sketch of Proof Introducing the semidiscrete operator:

$$(\mathscr{K}^N g)(x) \;=\; \frac{1}{2}\sum_{|k|=1}^{N} w_k(\mathscr{S}_{\mu_k} g)(x)$$

(corresponding to applying quadrature in angle but no discretisation in space), we then write $\mathscr{K} - \mathscr{K}^{h,N} = (\mathscr{K} - \mathscr{K}^N) + (\mathscr{K}^N - \mathscr{K}^{h,N})$ and consider, separately, the semidiscrete error due to quadrature in angle:

$$(\mathscr{K} - \mathscr{K}^N)(\sigma_S\phi + f) = \frac{1}{2}\left(\int_{-1}^{1}\psi(x,\mu)\mathrm{d}\mu - \sum_{|k|=1}^{N} w_k\psi(x,\mu_k)\right), \tag{25}$$

and the spatial error for a given N:

$$(\mathscr{K}^N - \mathscr{K}^{h,N})(\sigma_S\phi + f) = \frac{1}{2}\sum_{|k|=1}^{N} w_k\left(\mathscr{S}_{\mu_k} - \mathscr{S}^h_{\mu_k}\right)(\sigma_S\phi + f). \tag{26}$$

The estimate for (25) uses estimates for the regularity of ψ with respect to μ (which are explicit in the cross-sections), while (26) is estimated by proving stability of the Crank-Nicolson method and a cross-section-explicit bound on $\|\phi\|_\eta$.

Putting together Lemmas 1 and 2, we obtain the following.

Theorem 1 *Under the assumptions outlined above,*

$$\|\phi - \phi^{h,N}\|_\infty \;\le\; C_1\left(C_2\, h\log N + C_3\, h^\eta \;+\; C_4\,\frac{1}{N}\right)\|f\|_\eta\ .$$

Returning to the case when σ, σ_S are random functions, this theorem provides pathwise estimates for the error. In [25], these are turned into estimates in the corresponding Bochner space provided the coefficients C_i are bounded in probability space. Whether this is the case depends on the choice of the random model for σ, σ_S.

In particular, using the results in [9, §2], [23], it can be shown that $C_i \in L^p(\Omega)$, for all $1 \le p < \infty$, for the specific choices of σ and σ_S in Sect. 2. Hence, we have:

Corollary 1 *For all* $1 \le p < \infty$,

$$\|\phi - \phi^{h,N}\|_{L^p(\Omega, L^\infty(0,1))} \;\le\; C\left(h\log N + h^\eta \;+\; \frac{1}{N}\right)\|f\|_\eta\,,$$

where C *is independent of* h, N *and* f.

4 Modern Variants of Monte Carlo

Let $Q(\omega) \in \mathbb{R}$ denote a functional of ϕ or ψ representing a quantity of interest. We will focus on estimating $\mathbb{E}[Q]$, the expected value of Q. Since we are not specific about what functionals we are considering, this includes also higher order moments or CDFs of quantities of interest. The expected value is a high-dimensional integral and the goal is to apply efficient quadrature methods in high dimensions. We consider Monte Carlo type sampling methods.

As outlined above, to obtain samples of $Q(\omega)$ the NTE has to be approximated numerically. First, the random scattering cross section σ_S in (4) is sampled using the KL expansion of $\log \sigma_S$ in (8) truncated after d terms. The stochastic dimension d is chosen sufficiently high so that the truncation error is smaller than the other approximation errors. For each $n \in \mathbb{N}$, let $Z^n \in \mathbb{R}^d$ be a realisation of the multivariate Gaussian coefficient $Z := (Z_i)_{i=1,\ldots,d}$ in the KL expansion (8). Also, denote by $Q_h(Z^n)$ the approximation of the nth sample of Q obtained numerically using a spatial grid with mesh size h and $2N$ angular quadrature points. We assume throughout that $N \sim 1/h$, so there is a single discretisation parameter h.

We will consider various unbiased, sample-based estimators $\widehat{Q}_h$ for the expected value $\mathbb{E}[Q]$ and we will quantify the accuracy of each estimator by its mean square error (MSE) $e(\widehat{Q}_h)^2$. Since $\widehat{Q}_h$ is assumed to be an unbiased estimate of $\mathbb{E}[Q_h]$, i.e. $\mathbb{E}[\widehat{Q}_h] = \mathbb{E}[Q_h]$, the MSE can be expanded as

$$e(\widehat{Q}_h)^2 \;=\; \mathbb{E}\left[(\widehat{Q}_h - \mathbb{E}[Q])^2\right] \;=\; (\mathbb{E}\left[Q - Q_h\right])^2 \;+\; \mathbb{V}[\widehat{Q}_h]\,, \tag{27}$$

i.e., the squared bias due to the numerical approximation plus the sampling (or quadrature) error $\mathbb{V}[\widehat{Q}_h] = \mathbb{E}[(\widehat{Q}_h - \mathbb{E}[Q_h])^2]$. In order to compare computational costs of the various methods we will consider their ϵ-cost $\mathscr{C}_\epsilon$, that is, the number of floating point operations to achieve a MSE $e(\widehat{Q}_h)^2$ less than ϵ^2.

To bound the ϵ-cost for each method, we make the following assumptions on the discretisation error and on the average cost to compute a sample from Q_h:

$$\left|\mathbb{E}\left[Q - Q_h\right]\right| \;=\; \mathscr{O}(h^{\alpha})\,, \tag{28}$$

$$\mathbb{E}\left[\mathscr{C}(Q_h)\right] \;=\; \mathscr{O}(h^{-\gamma})\,, \tag{29}$$

for some constants $\alpha, \gamma > 0$. We have seen in Sect. 2 that (29) holds with γ between 2 and 3. The new theoretical results in Sect. 3 guarantee that (28) also holds for some $0 < \alpha \leq 1$. Whilst the results of Sect. 3 (and [25]) are shown to be sharp in some cases, the practically observed values for α in the numerical experiments here are significantly bigger, with values between 1.5 and 2.

In recent years, many alternative methods for high-dimensional integrals have emerged that use tensor product deterministic quadrature rules combined with sparse grid techniques to reduce the computational cost [4, 6, 17, 21, 26, 40, 49]. The efficiency of these approaches relies on high levels of smoothness of the parameter to output map and in general their cost may grow exponentially with the number of parameters (the *curse of dimensionality*). Such methods are not competitive with Monte Carlo type methods for problems with low smoothness in the coefficients, where large numbers of parameters are needed to achieve a reasonable accuracy. For example, in our later numerical tests we will consider problems in up to 3600 stochastic dimensions.

However, standard Monte Carlo methods are notoriously slow to converge, requiring thousands or even millions of samples to achieve acceptable accuracies. In our application, where each sample involves the numerical solution of an integro-differential equation this very easily becomes intractable. The novel Monte Carlo approaches that we present here, aim to improve this situation in two complementary ways. Quasi-Monte Carlo methods reduce the number of samples to achieve a certain accuracy dramatically by using deterministic ideas to find well distributed samples in high dimensions. Multilevel methods use the available hierarchy of numerical approximations to our integro-differential equation to shift the bulk of the computations to cheap, inaccurate coarse models while providing the required accuracy with only a handful of expensive, accurate model solves.

4.1 Standard Monte Carlo

The (standard) Monte Carlo (MC) estimator for $\mathbb{E}[Q]$ is defined by

$$\widehat{Q}_h^{MC} := \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} Q_h(Z^n) \,, \tag{30}$$

where N_{MC} is the number of Monte Carlo points/samples $Z^n \sim \mathscr{N}(0, I)$. The sampling error of this estimator is $\mathbb{V}[\widehat{Q}_h^{MC}] = \mathbb{V}[Q_h]/N_{MC}$.

A sufficient condition for the MSE to be less than ϵ^2 is for both the squared bias and the sampling error in (27) to be less than $\epsilon^2/2$. Due to assumption (28), a sufficient condition for the squared bias to be less than $\epsilon^2/2$ is $h \sim \epsilon^{1/\alpha}$. Since $\mathbb{V}[Q_h]$ is bounded with respect to $h \to 0$, the sampling error of $\widehat{Q}_h^{MC}$ is less than $\epsilon^2/2$ for $N_{MC} \sim \epsilon^{-2}$. With these choices of h and N_{MC}, it follows from Assumption (29) that the mean ϵ-cost of the standard Monte Carlo estimator is

$$\begin{aligned} \mathbb{E}\left[\mathscr{C}_\epsilon(\widehat{Q}_h^{MC})\right] &= \mathbb{E}\left[\sum_{n=1}^{N_{MC}} \mathscr{C}(Q_h(Z^n))\right] \\ &= N_{MC}\,\mathbb{E}\left[\mathscr{C}(Q_h)\right] \\ &= \mathscr{O}\left(\epsilon^{-2-\frac{\gamma}{\alpha}}\right) \,. \end{aligned} \tag{31}$$

Our aim is to find alternative methods that have a lower ϵ-cost.

4.2 Quasi-Monte Carlo

The first approach to reduce the ϵ-cost is based on using quasi-Monte Carlo (QMC) rules, which replace the random samples in (30) by carefully chosen deterministic samples and treat the expected value with respect to the d-dimensional Gaussian Z in (8) as a high-dimensional integral with Gaussian measure.

Initially interest in QMC points arose within number theory in the 1950's, and the theory is still at the heart of good QMC point construction today. Nowadays, the fast component-by-component construction (CBC) [41] provides a quick method for generating good QMC points, in very high-dimensions. Further information on the best choices of deterministic points and QMC theory can be found in e.g. [14, 15, 39, 46].

The choice of QMC points can be split into two categories; lattice rules and digital nets. We will only consider randomised rank-1 lattice rules here. In particular, given a suitable generating vector $z \in \mathbb{Z}^d$ and R independent, uniformly distributed random shifts $(\Delta_r)_{r=1}^R$ in $[0, 1]^d$, we construct $N_{QMC} = RP$ lattice points in the unit cube $[0, 1]^d$ using the simple formula

$$v^{(n)} = \text{frac}\left(\frac{nz}{P} + \Delta_r\right), \qquad n = 1, \ldots, P, \;\; r = 1, \ldots, R$$

where "frac" denotes the fractional part function applied componentwise and the number of random shifts R is fixed and typically small e.g. $R = 8, 16$. To transform the lattice points $v^n \in [0, 1]^d$ into "samples" $\widetilde{Z}^n \in \mathbb{R}^d$, $n = 1, \ldots, N_{QMC}$, of the multivariate Gaussian coefficients Z in the KL expansion (8) we apply the inverse cumulative normal distribution. See [24] for details.

Finally, the QMC estimator is given by

$$\widehat{Q}_h^{QMC} := \frac{1}{N_{QMC}} \sum_{n=1}^{N_{QMC}} Q_h(\widetilde{Z}^n) \, ,$$

Note that this is essentially identical in its form to the standard MC estimator (30), but crucially with deterministically chosen and then randomly shifted $\widetilde{Z}^n$. The random shifts ensure that the estimator is unbiased, i.e. $\mathbb{E}[\widehat{Q}_h^{QMC}] = \mathbb{E}[Q_h]$.

The bias for this estimator is identical to the MC case, leading again to a choice of $h \sim \varepsilon^{1/\alpha}$ to obtain a MSE of ε^2. Here the MSE corresponds to the mean square error of a randomised rank-1 lattice rule with P points averaged over the shift $\Delta \sim \mathscr{U}([0, 1]^d)$. In many cases, it can be shown that the quadrature error, i.e., the second term in (27), converges with $\mathscr{O}(N_{QMC}^{-1/2\lambda})$, with $\lambda \in (\frac{1}{2}, 1]$. That is, we can potentially achieve $\mathscr{O}(N_{QMC}^{-1})$ convergence for $\widehat{Q}_h^{QMC}$ as opposed to the $\mathscr{O}(N_{MC}^{-1/2})$ convergence for $\widehat{Q}_h^{MC}$. A rigorous proof of the rate of convergence requires detailed analysis of the quantity of interest (the integrand), in an appropriate weighted Sobolev space, e.g. [23]. Such an analysis is still an open question for this class of problems, and we do not attempt it here. Moreover, the generating vector z does in theory have to be chosen problem specific. However, standard generating vectors, such as those available at [31], seem to also work well (and better than MC samples). Furthermore, we note the recent developments in "higher-order nets" [12, 22], which potentially increase the convergence of QMC methods to $\mathscr{O}(N_{QMC}^{-q})$, for $q \geq 2$.

Given the improved rate of convergence of the quadrature error and fixing the number of random shifts to $R = 8$, it suffices to choose $P \sim \epsilon^{-2\lambda}$ for the quadrature error to be $\mathscr{O}(\varepsilon^2)$. Therefore it follows again from Assumption (29) that the ϵ-cost of the QMC method satisfies

$$\mathbb{E}_\Delta \left[\mathscr{C}_\epsilon(\widehat{Q}^{QMC}) \right] = \mathscr{O}\left(\epsilon^{-2\lambda - \frac{\gamma}{\alpha}} \right) . \tag{32}$$

When $\lambda \to \frac{1}{2}$, this is essentially a reduction in the ϵ-cost by a whole order of ϵ. In the case of non-smooth random fields, we typically have $\lambda \approx 1$ and the ϵ-cost grows with the same rate as that of the standard MC method. However, in our experiments and in experiments for diffusion problems [24], the absolute cost is always reduced.

4.3 Multilevel Methods

The main issue with the above methods is the high cost for computing the samples $\{Q_h(Z^{(n)})\}$, each requiring us to solve the NTE. The idea of the multilevel Monte Carlo (MLMC) method is to use a hierarchy of discrete models of increasing cost and accuracy, corresponding to a sequence of decreasing discretisation parameters $h_0 > h_1 > \ldots > h_L = h$. Here, only the most accurate model on level L is designed to give a bias of $\mathscr{O}(\epsilon)$ by choosing $h_L = h \sim \epsilon^{1/\alpha}$ as above. The bias of the other models can be significantly higher.

MLMC methods were first proposed in an abstract way for high-dimensional quadrature by Heinrich [28] and then popularised in the context of stochastic differential equations in mathematical finance by Giles [18]. MLMC methods were first applied in uncertainty quantification in [7, 10]. The MLMC method has quickly gained popularity and has been further developed and applied in a variety of other problems. See [19] for a comprehensive review. In particular, the multilevel approach is not restricted to standard MC estimators and can also be used in conjunction with QMC estimators [20, 32, 34] or with stochastic collocation [47]. Here, we consider multilevel variants of standard MC and QMC.

MLMC methods exploit the linearity of the expectation, writing

$$\mathbb{E}[Q_h] = \sum_{\ell=0}^{L} \mathbb{E}[\,Y_\ell\,] \,, \qquad \text{where } \; Y_\ell := Q_{h_\ell} - Q_{h_{\ell-1}} \; \text{ and } \; Q_{h_{-1}} := 0.$$

Each of the expected values on the right hand side is then estimated separately. In particular, in the case of a standard MC estimator with N_ℓ samples for the ℓth term, we obtain the MLMC estimator

$$\widehat{Q}_h^{MLMC} := \sum_{\ell=0}^{L} \widehat{Y}_\ell^{MC} = \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} Y_\ell(Z^{\ell,n}) . \tag{33}$$

Here, $\{Z^{\ell,n}\}_{n=1}^{N_\ell}$ denotes the set of i.i.d. samples on level ℓ, chosen independently from the samples on the other levels.

The key idea in MLMC is to avoid estimating $\mathbb{E}[Q_h]$ directly. Instead, the expectation $\mathbb{E}[\,Y_0] = \mathbb{E}[Q_{h_0}]$ of a possibly strongly biased, but cheap approximation of Q_h is estimated. The bias of this coarse model is then estimated by a sum of correction terms $\mathbb{E}[\,Y_\ell]$ using increasingly accurate and expensive models. Since the Y_ℓ represent small corrections between the coarse and fine models, it is reasonable to conjecture that there exists $\beta > 0$ such that

$$\mathbb{V}[\,Y_\ell] = \mathscr{O}(h_\ell^\beta) , \tag{34}$$

i.e., the variance of Y_ℓ decreases as $h_\ell \to 0$. This is verified for diffusion problems in [9]. Therefore the number of samples N_ℓ to achieve a prescribed accuracy on level ℓ can be gradually reduced, leading to a lower overall cost of the MLMC estimator. More specifically, we have the following cost savings:

- On the coarsest level, using (29), the cost per sample is reduced from $\mathscr{O}(h^{-\gamma})$ to $\mathscr{O}(h_0^{-\gamma})$. Provided $\mathbb{V}[Q_{h_0}] \approx \mathbb{V}[Q_h]$ and h_0 can be chosen independently of ϵ, the cost of estimating $\mathbb{E}[Q_{h_0}]$ to an accuracy of ε in (33) is reduced to $\mathscr{O}(\epsilon^{-2})$.
- On the finer levels, the number of samples N_ℓ to estimate $\mathbb{E}[\,Y_\ell]$ to an accuracy of ε in (33) is proportional to $\mathbb{V}[\,Y_\ell]\epsilon^{-2}$. Now, provided $\mathbb{V}[\,Y_\ell] = \mathscr{O}(h_\ell^\beta)$, for some $\beta > 0$, which is guaranteed if Q_{h_ℓ} converges almost surely to Q pathwise, then we can reduce the number of samples as $h_\ell \to 0$. Depending on the actual values of $\alpha,\ \beta$ and γ, the cost to estimate $\mathbb{E}[\,Y_L]$ on the finest level can, in the best case, be reduced to $\mathscr{O}(\epsilon^{-\gamma/\alpha})$.

The art of MLMC is to balance the number of samples across the levels to minimise the overall cost. This is a simple constrained optimisation problem to achieve $\mathbb{V}[\widehat{Q}_h^{MLMC}] \leq \epsilon^2/2$. As shown in [18], using the technique of Lagrange Multipliers, the optimal number of samples on level ℓ is given by

$$N_\ell = \left\lceil 2\epsilon^{-2} \left(\sum_{\ell=0}^{L} \sqrt{\mathbb{V}[\,Y_\ell]/\mathscr{C}_\ell} \right) \sqrt{\mathbb{V}[\,Y_\ell]\mathscr{C}_\ell} \right\rceil , \tag{35}$$

where $\mathscr{C}_\ell := \mathbb{E}\,[\mathscr{C}(Y_\ell)]$. In practice, it is necessary to estimate $\mathbb{V}[\,Y_\ell]$ and $\mathscr{C}_\ell$ in (35) from the computed samples, updating N_ℓ as the simulation progresses.

Using these values of N_ℓ it is possible to establish the following theoretical complexity bound for MLMC [10].

Theorem 2 *Let us assume that* (28), (34) *and* (29) *hold with* $\alpha, \beta, \gamma > 0$. *Then, with* $L \sim \log(\epsilon^{-1})$ *and with the choice of* $\{N_\ell\}_{l=0}^{L}$ *in* (35) *we have*

$$\mathbb{E}\left[\mathscr{C}_\epsilon(\widehat{Q}_{h_L}^{MLMC})\right] = \mathscr{O}\left(\epsilon^{-2-\max\left(0,\ \frac{\gamma-\beta}{\alpha}\right)}\right). \tag{36}$$

When $\beta = \gamma$, *then there is an additional factor* $\log(\epsilon^{-1})$.

Using lattice points $\widetilde{Z}^{\ell,n}$, as defined in Sect. 4.2, instead of the random samples $Z^{\ell,n}$ we can in the same way define a multilevel quasi-Monte Carlo (MLQMC) estimator

$$\widehat{Q}_h^{MLQMC} := \sum_{\ell=0}^{L} \widehat{Y}_\ell^{QMC} = \sum_{\ell=0}^{L} \frac{1}{\widetilde{N}_\ell} \sum_{n=1}^{\widetilde{N}_\ell} Y_\ell(\widetilde{Z}^{\ell,n}). \tag{37}$$

The optimal values for $\widetilde{N}_\ell$ can be computed in a similar way to those in the MLMC method. However, they depend strongly on the rate of convergence of the lattice rule and in particular on the value of λ which is difficult to estimate accurately. We will give a practically more useful approach below.

It is again possible to establish a theoretical complexity bound, cf. [32, 34].

Theorem 3 *Let us assume that* (28) *and* (29) *hold with* $\alpha, \gamma > 0$ *and that there exists* $\lambda \in (\frac{1}{2}, 1]$ *and* $\beta > 0$ *such that*

$$\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}] = \mathscr{O}\left(\widetilde{N}_\ell^{-1/\lambda} h_\ell^{\beta}\right). \tag{38}$$

Let the number of random shifts on each level be fixed to R and let $L \sim \log(\epsilon^{-1})$. *Then, there exists a choice of* $\{N_\ell\}_{l=0}^{L}$ *such that*

$$\mathbb{E}_\Delta\left[\mathscr{C}_\epsilon(\widehat{Q}_{h_L}^{MLQMC})\right] = \mathscr{O}\left(\epsilon^{-2\lambda-\max\left(0,\ \frac{\gamma-\beta\lambda}{\alpha}\right)}\right). \tag{39}$$

When $\beta\lambda = \gamma$, *then there is an additional factor* $\log(\epsilon^{-1})^{1+\lambda}$.

The convergence rate can be further improved by using higher order QMC rules [13], but we will not consider this here.

It can be shown, for the theoretically optimal values of N_ℓ, that there exists a constant C such that

$$\frac{\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]}{\mathscr{C}_\ell} = C, \tag{40}$$

independently of the level ℓ and of the value of λ (cf. [32, Sect. 3.3]). The same holds for MLMC. This leads to the following adaptive procedure to choose N_ℓ suggested in [20], which we use in our numerical experiments below instead of (35).

In particular, starting with an initial number of samples on all levels, we alternate the following two steps until $\mathbb{V}[\widehat{Q}_h^{MLMC}] \leq \epsilon^2/2$:

1. Estimate $\mathscr{C}_\ell$ and $\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]$ (resp. $\mathbb{V}[\widehat{Y}_\ell^{MC}]$).
2. Compute

$$\ell^* = \operatorname*{argmax}_{\ell=0}^{L} \left(\frac{\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]}{\mathscr{C}_\ell} \right)$$

and double the number of samples on level ℓ^*.

This procedure ensures that, on exit, (40) is roughly satisfied and the numbers of samples across the levels N_ℓ are quasi-optimal.

We use this adaptive procedure for both the MLMC and the MLQMC method. The lack of optimality typically has very little effect on the actual computational cost. Since the optimal formula (35) for MLMC also depends on estimates of $\mathscr{C}_\ell$ and $\mathbb{V}[Y_\ell]$, it sometimes even leads to a better performance. An additional benefit in the case of MLQMC is that the quadrature error in rank-1 lattice rules is typically lowest when the numbers of lattice points is a power of 2.

5 Numerical Results

We now present numerical results to confirm the gains that are possible with the novel multilevel and quasi-Monte Carlo method applied to our 1D NTE model (1)–(3). We assume that the scattering cross-section σ_S is a log-normal random field as described in Sect. 2.1 and that the absorption cross section is constant, $\sigma_A \equiv \exp(0.25)$. We assume no fission, $\sigma_F \equiv 0$, and a constant source term $f = \exp(1)$. We consider two cases, characterised by the choice of smoothness parameter ν in the Matérn covariance function (7). For the first case, we choose $\nu = 0.5$. This corresponds to the exponential covariance and in the following is called the "exponential field". For the second case, denoted the "Matérn field", we choose $\nu = 1.5$. The correlation length and the variance are $\lambda_C = 1$ and $\sigma_{var}^2 = 1$, respectively. The quantity of interest we consider is

$$Q(\omega) = \int_0^1 |\phi(x, \omega)| \mathrm{d}x \, . \tag{41}$$

For the discretisation, we choose a uniform spatial mesh with mesh width $h = 1/M$ and a quadrature rule (in angle) with $2N = M$ points. The KL expansion of $\log(\sigma_S)$ in (8) is truncated after d terms. We heuristically choose d to ensure that the error due to this truncation is negligible compared to the discretisation error. In particular, we choose $d = 8h^{-1}$ for the Matérn field and $d = 225h^{-1/2}$ for the exponential field, leading to a maximum of 2048 and 3600 KL modes, respectively, for the finest spatial resolution in each case. Even for such large numbers of KL modes, the sampling cost does not dominate because the randomness only exists in the (one) spatial dimension.

We introduce a hierarchy of levels $\ell = 0, \ldots, L$ corresponding to a sequence of discretisation parameters $h_\ell = 2^{-\ell} h_0$ with $h_0 = 1/4$, and approximate the quantity of interest in (41) by

$$Q_h(\omega) \;:=\; \frac{1}{M} \sum_{j=1}^{M} |\Phi_{j-1/2}(\omega)|.$$

To generate our QMC points we use an (extensible) randomised rank-1 lattice rule (as presented in Sect. 4.2), with $R = 8$ shifts. We use the generating vector `lattice-32001-1024-1048576.3600`, which is downloaded from [31].

5.1 A Hybrid Direct-Iterative Solver

To compute samples of the neutron flux and thus of the quantity of interest, we propose a hybrid version of the direct and the iterative solver for the Schur complement system (12) described in Sect. 2.3.

The cost of the iterative solver depends on the number K of iterations that we take. For each ω, we aim to choose K such that the L^2-error $\|\phi(\omega) - \phi^{(K)}(\omega)\|_2$ is less than ϵ. To estimate K we fix $h = 1/1024$ and $d = 3600$ and use the direct solver to compute ϕ_h for each sample ω. Let $\rho(\omega) := \|\sigma_S(\cdot, \omega)/\sigma(\cdot, \omega)\|_\infty$. For a sufficiently large number of samples, we then evaluate

$$\frac{\log\left(\left\|\phi_h(\omega) \;-\; \phi_h^{(K)}(\omega)\right\|_2\right)}{K \,\log\left(\rho(\omega)\right)}$$

and find that this quotient is less than log(0.5) in more than 99% of the cases, for $K = 1, \ldots, 150$, so that we can choose $C = 0.5$ in (17). We repeat the experiment also for larger values of h and smaller values of d to verify that this bound holds in at least 99% of the cases independently of the discretisation parameter h and of the truncation dimensions d.

Hence, a sufficient, a priori condition to achieve $\|\phi_h(\omega) - \phi_h^{(K)}(\omega)\|_2 < \epsilon$ in at least 99% of the cases is

$$K \;=\; K(\epsilon, \omega) \;=\; \max\left\{ 1, \left\lceil \frac{\log(2\epsilon)}{\log\left(\rho(\omega)\right)} \right\rceil \right\}, \tag{42}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. It is important to note that K is no longer a deterministic parameter for the solver (like M or N). Instead, K is a random variable that depends on the particular realisation of σ_S. It follows from (42), using the results in [9, §2], [23] as in Sect. 3, that $\mathbb{E}[K(\epsilon,\cdot)] = \mathcal{O}(\log(\epsilon))$ and $\mathbb{V}[K(\epsilon,\cdot)] = \mathcal{O}\left(\log(\epsilon)^2\right)$, with more variability in the case of the exponential field.

Recall from (13) and (15) that, in the case of $N = 2M$, the costs for the direct and iterative solvers are C_1M^3 and C_2KM^2, respectively. In our numerical experiments, we found that in fact $C_1 \approx C_2$, for this particular relationship between M and N. This motivates a third "hybrid" solver, presented in Algorithm 1, where the iterative solver is chosen when $K(\omega) < M$ and the direct solver when $K(\omega) \geq M$. This allows us to use the optimal solver for each particular sample.

We finish this section with a study of timings in seconds (here referred to as the cost) of the three solvers. In Fig. 1, we plot the average cost (over 2^{14} samples) divided by M_ℓ^3, against the level parameter ℓ. We observe that, as expected, the (scaled) expected cost of the direct solver is almost constant and the iterative solver is more efficient for larger values of M_ℓ. Over the range of values of M_ℓ considered in our experiments, a best fit for the rate of growth of the cost with respect to the discretisation parameter h_ℓ in (29) is $\gamma \approx 2.2$, for both fields. Thus our solver has a practical complexity of $\mathcal{O}(n^{1.1})$, where $n \sim M^2$ is the total number of degrees of freedom in the system.

Algorithm 1 Hybrid direct-iterative solver of (12), for one realisation

```
Require: Given $\sigma_S$, $\sigma$ and a desired accuracy $\epsilon$
  $K = \left\lceil\; log(2\epsilon) \;/\; log(\rho) \;\right\rceil$
  if $K < M$ then
    Solve using $K$ source iterations
  else
    Solve using the direct method
  end if
```

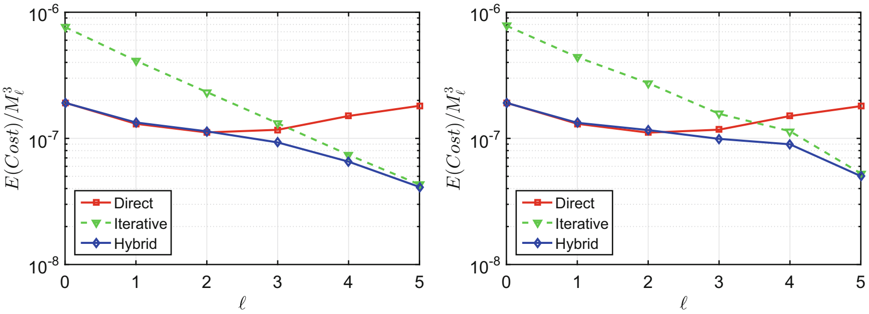


Fig. 1 Comparison of the average costs of the solvers (actual timings in seconds divided by M_ℓ^3) for the Matérn field (left) and for the exponential field (right)

5.2 A Priori Error Estimates

Studying the complexity theorems of Sect. 4, we can see that the effectiveness of the various Monte Carlo methods depends on the parameters α, β, γ and λ in (28), (29), (34) and (38). In this section, we will (numerically) estimate these parameters in order to estimate the theoretical computational cost for each approach.

We have already seen that $\gamma \approx 2.2$ for the hybrid solver. In Fig. 2, we present estimates of the bias $\mathbb{E}[Q - Q_{h_\ell}]$, as well as of the variances of Q_{h_ℓ} and of Y_ℓ, computed via sample means and sample variances over a sufficiently large set of samples. We only explicitly show the curves for the Matérn field. The curves for the exponential field look similar. From these plots, we can estimate $\alpha \approx 1.9$ and $\beta \approx 4.1$, for the Matérn field, and $\alpha \approx 1.7$ and $\beta \approx 1.9$, for the exponential field.

To estimate λ in (38), we need to study the convergence rate of the QMC method with respect to the number of samples N_{QMC}. This study is illustrated in Fig. 3. As expected, the variance of the standard MC estimator converges with $\mathscr{O}(N_{MC}^{-1})$. On the other hand, we observe that the variance of the QMC estimator converges

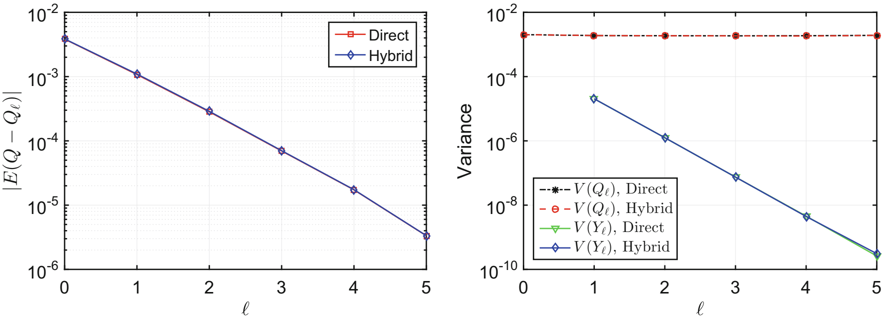


Fig. 2 Estimates of the bias due to discretisation errors (left) and of the variances of Q_{h_ℓ} and Y_ℓ (right), in the case of the Matérn field

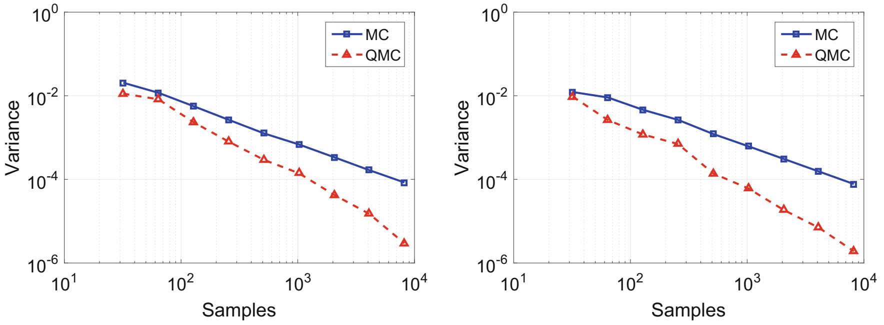


Fig. 3 Convergence of standard Monte Carlo and quasi-Monte Carlo estimators: Matérn field (left) and exponential field (right)

approximately with $\mathscr{O}(N_{QMC}^{-1.6})$ and $\mathscr{O}(N_{QMC}^{-1.4})$ (or $\lambda = 0.62$ and $\lambda = 0.71$) for the Matérn field and for the exponential field, respectively.

We summarise all the estimated rates in Table 1.

5.3 Complexity Comparison of Monte Carlo Variants

For a fair comparison of the complexity of the various Monte Carlo estimators, we now use the a priori bias estimates in Sect. 5.2 to choose a suitable tolerance ϵ_L for each choice of $h = h_L$. Let τ_ℓ be the estimated bias on level ℓ. Then, for each $L = 2, \ldots, 6$, we choose $h = h_L$ and $\epsilon_L := \sqrt{2}\,\tau_L$, and we plot in Fig. 4 the actual cost of each of the estimators described in Sect. 4 against the estimated bias on level L. The numbers of samples for each of the estimators are chosen such that $\mathbb{V}[\widehat{Q}_h] \le \epsilon_\ell^2/2$. The coarsest mesh size in the multilevel methods is always $h_0 = 1/4$. We can clearly see the benefits of the QMC sampling rule and of the multilevel variance reduction, and the excellent performance of the multilevel QMC estimator confirms that the two improvements are indeed complementary. As expected, the gains are more pronounced for the smoother (Matérn) field.

We finish by comparing the actual, observed ϵ-cost of each of the methods with the ϵ-cost predicted theoretically using the estimates for α, β, γ and λ in Sect. 5.2. Assuming a growth of the ϵ-cost proportional to ϵ^{-r}, for some $r > 0$, we compare in Table 2 estimated and actual rates r for all the estimators. Some of the estimated rates in Sect. 5.2 are fairly crude, so the good agreement between estimated and actual rates is quite impressive.

Table 1 Summary of estimated rates in (28), (29), (34) and (38)

	α	β	γ	λ
Matérn field	1.9	4.1	2.2	0.62
Exponential field	1.7	1.9	2.2	0.71

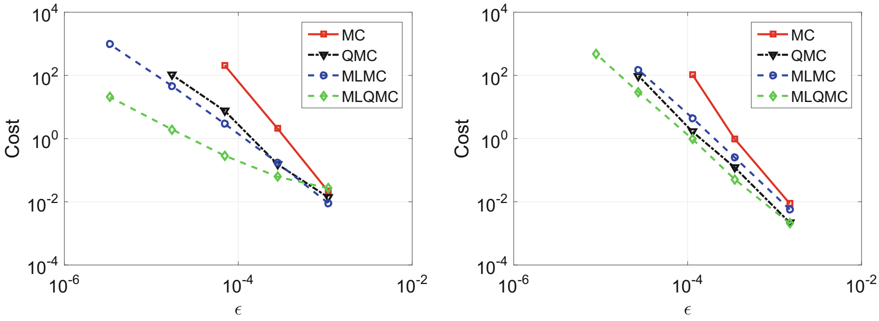


Fig. 4 Actual cost plotted against estimated bias on level L for standard Monte Carlo, QMC, multilevel MC and multilevel QMC: Matérn field (left) and exponential field (right)

Table 2 Comparison of the estimated theoretical and actual computational ϵ-cost rates, for different Monte Carlo methods, using the hybrid solver

	MC		QMC		MLMC		MLQMC	
Field	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual
Matérn	3.2	3.4	2.4	2.7	2.0	2.1	1.2	1.5
Exponential	3.3	3.6	2.7	2.4	2.2	2.5	1.9	1.9

6 Conclusions

To summarise, we have presented an overview of novel error estimates for the 1D slab geometry simplification of the Neutron Transport Equation, with spatially varying and random cross-sections. In particular, we consider the discrete ordinates method with Gauss quadrature for the discretisation in angle, and a diamond differencing scheme on a quasi-uniform grid in space. We represent the spatial uncertainties in the cross-sections by log-normal random fields with Matérn covariances, including cases of low smoothness. These error estimates are the first of this kind. They allow us to satisfy key assumptions for the variance reduction in multilevel Monte Carlo methods.

We then use a variety of recent developments in Monte Carlo methods to study the propagation of the uncertainty in the cross-sections, through to a non-linear functional of the scalar flux. We find that the Multilevel Quasi Monte Carlo method gives us significant gains over the standard Monte Carlo method. These gains can be as large as almost two orders of magnitude in the computational ϵ-cost for $\epsilon = 10^{-4}$.

As part of the new developments, we present a hybrid solver, which automatically switches between a direct or iterative method, depending on the rate of convergence of the iterative solver which varies from sample to sample. Numerically, we observe that the hybrid solver is almost an order of magnitude cheaper than the direct solver on the finest mesh, on the other hand the direct solver is almost an order of magnitude cheaper than the iterative solver on the coarsest mesh we considered.

We conclude that modern variants of Monte Carlo based sampling methods are extremely useful for the problem of Uncertainty Quantification in Neutron Transport. This is particularly the case when the random fields are non-smooth and a large number of stochastic variables are required for accurate modelling.

Acknowledgements We thank EPSRC and AMEC Foster Wheeler for financial support for this project and we particularly thank Professor Paul Smith (AMECFW) for many helpful discussions. Matthew Parkinson is supported by the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under project EP/L015684/1. This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath.

References

1. Allen, E.J., Victory Jr., H.D., Ganguly, K.: On the convergence of finite-differenced multigroup, discrete-ordinates methods for anisotropically scattered slab media. SIAM J. Numer. Anal. **26**, 88–106 (1989)
2. Asadzadeh, M.: A finite element method for the neutron transport equation in an infinite cylindrical domain. SIAM J. Numer. Anal. **35**, 1299–1314 (1998)
3. Asadzadeh, M., Thevenot, L.: On discontinuous Galerkin and discrete ordinates approximations for neutron transport equation and the critical eigenvalue. Nuovo Cimento C **33**, 21–29 (2010)
4. Ayres, D.A.F., Eaton, M.D.: Uncertainty quantification in nuclear criticality modelling using a high dimensional model representation. Ann. Nucl. Energy **80**, 379–402 (2015)
5. Ayres, D.A.F., Park, S., Eaton, M.D.: Propagation of input model uncertainties with different marginal distributions using a hybrid polynomial chaos expansion. Ann. Nucl. Energy **66**, 1–4 (2014)
6. Babuska, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**, 1005–1034 (2007)
7. Barth, A., Schwab, C., Zollinger, N.: Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. Numer. Math. **119**, 123–161 (2011)
8. Blake, J.C.H.: Domain decomposition methods for nuclear reactor modelling with diffusion acceleration. Ph.D. Thesis, University of Bath (2016)
9. Charrier, J., Scheichl, R., Teckentrup, A.L.: Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. SIAM J. Numer. Anal. **51**, 322–352 (2013)
10. Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.L.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. **14**, 3–15 (2011)
11. Dautray, R., Lions, J.L.: Mathematical Analysis and Numerical Methods for Science and Technology. Physical Origins and Classical Methods, vol. 1. Springer, Heidelberg (2012)
12. Dick, J., Kuo, F.Y., Le Gia, Q.T., Nuyens, D., Schwab, C.: Higher order QMC Petrov–Galerkin discretization for affine parametric operator equations with random field inputs. SIAM J. Numer. Anal. **52**, 2676–2702 (2014)
13. Dick, J., Kuo, F.Y., Le Gia, Q.T., Schwab, C.: Multi-level higher order QMC Galerkin discretization for affine parametric operator equations. SIAM J. Numer. Anal. **54**, 2541–2568 (2016)
14. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
15. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
16. Eiermann, M., Ernst, O.G., Ullmann, E.: Computational aspects of the stochastic finite element method. Comput. Vis. Sci. **10**, 3–15 (2007)
17. Fichtl, E.D., Prinja, A.K.: The stochastic collocation method for radiation transport in random media. J. Quant. Spectrosc. Radiat. Tran. **112**(4), 646–659 (2011)
18. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**, 607–617 (2008)
19. Giles, M.B.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)
20. Giles, M.B., Waterhouse, B.J.: Multilevel quasi-Monte Carlo path simulation. In: Advanced Financial Modelling. Radon Series on Computational and Applied Mathematics, pp. 165–181 (2009)
21. Gilli, L., Lathouwers, D., Kloosterman, J.L., van der Hagen, T.H.J.J., Koning, A.J., Rochman, D.: Uncertainty quantification for criticality problems using non-intrusive and adaptive polynomial chaos techniques. Ann. Nucl. Energy **56**, 71–80 (2013)
22. Goda, T., Dick, J.: Construction of interlaced scrambled polynomial lattice rules of arbitrary high order. Found. Comput. Math. **15**, 1245–1278 (2015)

23. Graham, I.G., Kuo, F.Y., Nichols, J.A., Scheichl, R., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. Numer. Math. **131**, 329–368 (2015)
24. Graham, I.G., Kuo, F.Y., Nuyens, D., Scheichl, R., Sloan, I.H.: Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. J. Comput. Phys. **230**, 3668–3694 (2011)
25. Graham, I.G., Parkinson, M.J., Scheichl, R.: Error analysis and uncertainty quantification for the heterogenous transport equation in slab geometry (2018, in preparation)
26. Gunzburger, M., Webster, C.G., Zhang, G.: Stochastic finite element methods for PDEs with random input data. Acta Numer. **23**, 521–650 (2014)
27. Haji-Ali, A.L., Nobile, F., Tempone, R.: Multi-index Monte Carlo: when sparsity meets sampling. Numer. Math. **132**, 767–806 (2016)
28. Heinrich, S.: Multilevel Monte Carlo methods. Lecture Notes in Computer Science, vol. 2179. Springer, Heidelberg (2001)
29. Kaper, H.G., Kellogg, R.B.: Asymptotic behavior of the solution of the integral transport equation in slab geometry. SIAM J. Appl. Math. **32**(1), 191–200 (1977)
30. Keller, H.B.: On the pointwise convergence of the discrete-ordinates method. SIAM J. Appl. Math. **8**, 560–567 (1960)
31. Kuo, F.Y.: http://web.maths.unsw.edu.au/~fkuo/lattice/index.html
32. Kuo, F.Y., Scheichl, R., Schwab, C., Sloan, I.H., Ullmann, E.: Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. Math. Comput. **86**, 2827–2860 (2017)
33. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficient. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
34. Kuo, F.Y., Schwab, C., Sloan, I.H.: Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. Found. Comput. Math. **15**, 411–449 (2015)
35. Larsen, E.W., Nelson, P.: Finite difference approximations and superconvergence for the discrete-ordinate equations in slab geometry. SIAM J. Numer. Anal. **19**, 334–348 (1982)
36. Lee, C.W., Lee, Y.O., Cho. Y.S.: Comparison of the nuclear data libraries in the shielding calculation for the accelerator facility of the Proton Engineering Frontier Project in Korea. In: International Conference on Nuclear Data for Science and Technology. EDP Sciences, Les Ulis (2007)
37. Lewis, E.E., Miller, W.F.: Computational Methods of Neutron Transport. Wiley, New York (1984)
38. Lord, G.J., Powell, C.E. Shardlow, T.: An Introduction to Computational Stochastic PDEs. Cambridge University Press, Cambridge (2014)
39. Niederreiter, H.: Quasi-Monte Carlo Methods. Wiley, New York (2010)
40. Nobile, F., Tempone, R., Webster, C.G.: An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**, 2411–2442 (2008)
41. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comput. **75**, 903–920 (2006)
42. Pitkaranta, J., Scott, L.R.: Error estimates for the combined spatial and angular approximations of the transport equation for slab geometry. SIAM J. Numer. Anal. **20**, 922–950 (1983)
43. Reed, W.H., Hill, T.R.: Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73–479. Los Alamos National Laboratory (1973)
44. Sanchez, R., McCormick, N.J.: Review of neutron transport approximations. Nucl. Sci. Eng. **80**, 481–535 (1982)
45. Sloan, I.H.: Error analysis for a class of degenerate-kernel methods. Numer. Math. **25**, 231–238 (1975)
46. Sloan, I.H., Wozniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complexity **14**, 1–33 (1998)

47. Teckentrup, A.L., Jantsch, P., Webster, C.G., Gunzburger, M.: A multilevel stochastic collocation method for partial differential equations with random input data. SIAM/ASA JUQ **3**, 1046–1074 (2015)
48. Victory Jr., H.D.: Convergence of the multigroup approximations for subcritical slab media and applications to shielding calculations. Adv. Appl. Math. **5**, 227–259 (1984)
49. Xiu, D., Karniadakis, G.E.: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**, 614–644 (2002)

On the Representation of Symmetric and Antisymmetric Tensors

Wolfgang Hackbusch

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Various tensor formats are used for the data-sparse representation of large-scale tensors. Here we investigate how symmetric or antisymmetric tensors can be represented. We mainly investigate the hierarchical format, but also the use of the canonical format is mentioned.

1 Introduction

We consider tensor spaces of huge dimension exceeding the capacity of computers. Therefore the numerical treatment of such tensors requires a special representation technique which characterises the tensor by data of moderate size. These representations (or formats) should also support operations with tensors. Examples of operations are the addition, the scalar product, the componentwise product (Hadamard product), and the matrix-vector multiplication. In the latter case, the 'matrix' belongs to the tensor space of Kronecker matrices, while the 'vector' is a usual tensor.

In certain applications the subspaces of symmetric or antisymmetric tensors are of interest. For instance, fermionic states in quantum chemistry require antisymmetry, whereas bosonic systems are described by symmetric tensors. The appropriate representation of (anti)symmetric tensors is seldom discussed in the literature. Of course, all formats are able to represent these tensors since they are particular examples of general tensors. However, the special (anti)symmetric format should exclusively produce (anti)symmetric tensors. For instance, the truncation procedure must preserve the symmetry properties.

W. Hackbusch (✉)
Max-Planck-Institut Mathematik in den Naturwissenschaften, Leipzig, Germany
e-mail: wh@mis.mpg.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_23

The formats in use are the r-term format (also called the canonical format), the subspace (or Tucker) format, and the hierarchical representation including the TT format. In the general case, the last format has turned out to be very efficient and flexible. We discuss all formats concerning application to (anti)symmetric tensors.

The r-term format is seemingly the simplest one, but has several numerical disadvantages. In Sect. 2 we discuss two different approaches to representing (anti)symmetric tensors. However, they inherit the mentioned disadvantages.

As explained in Sect. 3, the subspace (or Tucker) format is not helpful.

The main part of the paper discusses the question how the TT format can be adapted to the symmetry requirements. The analysis leads to unexpected difficulties. In contrast to the general case, the subspaces $\mathbf{U}_j$ involved in the TT format (see Sect. 4.3) have to satisfy conditions which are not easy to check. We can distinguish the following two different situations.

In the *first case* we want to construct the TT format with subspaces $\mathbf{U}_j$ not knowing the tensor $\mathbf{v}$ to be represented in advance. For instance we change $\mathbf{U}_j$ to obtain a variation of $\mathbf{v}$, or the dimension of $\mathbf{U}_j$ is reduced to obtain a truncation. In these examples, $\mathbf{v}$ is obtained as a result on the chosen $\mathbf{U}_j$. It turns out that the choice of $\mathbf{U}_j$ is delicate. If $\mathbf{U}_j$ is too small, no nontrivial tensors can be represented. On the other hand, $\mathbf{U}_j$ may contain a useless nontrivial part, i.e., $\mathbf{U}_j$ may be larger than necessary. The algebraic characterisation of the appropriate $\mathbf{U}_j$ is rather involved.

In the *second case*, we start from $\mathbf{v}$ and know the minimal subspaces $\mathbf{U}_j^{\min}(\mathbf{v})$ (cf. (9)). Then $\mathbf{U}_j^{\min}(\mathbf{v}) \subset \mathbf{U}_j$ is a sufficient condition. However, as soon as we want to truncate the tensor $\mathbf{v}$, its result $\mathbf{v}'$ must be determined from modified subspaces $\mathbf{U}_j'$ so that we return to the difficulties of the first case.

In Sect. 8 we describe the combination of the TT format and the ANOVA technique for symmetric tensors. This leads to a favourable method as long as the ANOVA degree is moderate.

Quite another approach for antisymmetric tensors is the so-called ‘second quantisation’ (cf. Legeza et al. [13, §2.3]) which does not fit into the following schemes.

1.1 Tensor Notation

1.1.1 Tensors Spaces

In the general case, vector spaces V_j $(1 \leq j \leq d)$ are given which determine the algebraic tensor space $\mathbf{V} := \bigotimes_{j=1}^{d} V_j$. The common underlying field of the vector spaces V_j is either $\mathbb{R}$ or $\mathbb{C}$. In the following we write $\mathbb{K}$ for either of the fields. In the particular case of

$$V_j = V \qquad \text{for all } 1 \leq j \leq d \tag{1}$$

we write $\mathbf{V} := \otimes^d V$. Set

$$D := \{1, \ldots, d\} \tag{2}$$

and consider any nonempty subset $\alpha \subset D$. We set

$$\mathbf{V}_\alpha := \bigotimes\nolimits_{j\in\alpha} V_j \,. \tag{3}$$

Note that $\mathbf{V} = \mathbf{V}_D$ is isomorphic to $\mathbf{V}_\alpha \otimes \mathbf{V}_{D\backslash\alpha}$.

We assume that V is a pre-Hilbert space with the scalar product $\langle\cdot,\cdot\rangle$. Then $\mathbf{V}$ and each $\mathbf{V}_\alpha$ is defined as a pre-Hilbert space with the induced scalar product uniquely defined by (cf. [10, Lemma 4.124])

$$\left\langle \bigotimes\nolimits_{j\in\alpha} v^{(j)}, \bigotimes\nolimits_{j\in\alpha} w^{(j)} \right\rangle = \prod\nolimits_{j\in\alpha} \left\langle v^{(j)}, w^{(j)} \right\rangle. \tag{4}$$

1.1.2 Functionals

Let $\varphi_\alpha \in \mathbf{V}'_\alpha$ be a linear functional. The same symbol φ_α is used for the linear map $\varphi_\alpha : \mathbf{V} \to \mathbf{V}_{D\backslash\alpha}$ defined by

$$\varphi_\alpha \left(\bigotimes\nolimits_{j=1}^{d} v^{(j)} \right) = \varphi_\alpha \left(\bigotimes\nolimits_{j\in\alpha} v^{(j)} \right) \bigotimes\nolimits_{j\in D\backslash\alpha} v^{(j)} \tag{5}$$

(it is sufficient to define a linear map by its action on elementary tensors, cf. [10, Remark 3.55]). In the case of (1) and $\varphi \in V'$ we introduce the following notation. The linear mapping $\varphi^{(k)} : \otimes^d V \to \otimes^{d-1} V$ is defined by

$$\varphi^{(k)} \left(\bigotimes\nolimits_{j=1}^{d} v^{(j)} \right) = \varphi\left(v^{(k)}\right) \bigotimes\nolimits_{j\neq k} v^{(j)}. \tag{6}$$

1.1.3 Permutations and (Anti)symmetric Tensor Spaces

A permutation $\pi \in P_d$ is a bijection of D onto itself. For $\nu, \mu \in D$, the permutation $\pi_{\nu\mu}$ is the transposition swapping the positions ν and μ. If $\nu = \mu$, $\pi_{\nu\mu}$ is the identity id. Let $\mathbf{V} = \otimes^d V$. Then the symbol of the permutation π is also used for the linear map $\pi : \mathbf{V} \to \mathbf{V}$ defined by

$$\pi \left(\bigotimes\nolimits_{j=1}^{d} v^{(j)} \right) = \bigotimes\nolimits_{j=1}^{d} v^{(\pi^{-1}(j))}.$$

Each permutation π is a (possibly empty) product of transpositions: $\pi = \pi_{\nu_1\mu_1} \circ \pi_{\nu_2\mu_2} \circ \ldots \circ \pi_{\nu_k\mu_k}$ with $\nu_i \neq \mu_i$ $(1 \leq i \leq k)$. The number k determines the parity ± 1 of the permutation: $\operatorname{sign}(\pi) = (-1)^k$.

A tensor $\mathbf{v} \in \otimes^d V$ is called *symmetric* if $\pi(\mathbf{v}) = \mathbf{v}$ for all permutations, and antisymmetric if $\pi(\mathbf{v}) = \operatorname{sign}(\pi)\mathbf{v}$. This defines the (anti)symmetric tensor space:

$$\mathbf{V}_{\mathrm{sym}} := \left\{\mathbf{v} \in \otimes^d V : \pi(\mathbf{v}) = \mathbf{v}\right\}, \tag{7a}$$

$$\mathbf{V}_{\mathrm{anti}} := \left\{\mathbf{v} \in \otimes^d V : \pi(\mathbf{v}) = \operatorname{sign}(\pi)\mathbf{v}\right\}. \tag{7b}$$

If the parameter d should be emphasised, we also write $\mathbf{V}_{\mathrm{sym}}^{(d)}$ and $\mathbf{V}_{\mathrm{anti}}^{(d)}$. Correspondingly, if $U \subset V$ is a subspace, the (anti)symmetric tensors in $\otimes^d U$ are denoted by $\mathbf{U}_{\mathrm{sym}}^{(d)}$, resp. $\mathbf{U}_{\mathrm{anti}}^{(d)}$. Another notation for $\mathbf{V}_{\mathrm{anti}}^{(d)}$ is $\bigwedge^d V$ using the exterior product $\wedge$.

Besides the well-known applications in physics (cf. [3]), symmetric and antisymmetric tensors occur in different mathematical fields.

The symmetric tensor space is related to multivariate polynomials which are homogenous of degree d, i.e., $p(\lambda x) = \lambda^d p(x)$. These polynomials are called *quantics* by Cayley [7]. If $n = \dim(V)$, the symmetric tensor space $\mathbf{V}_{\mathrm{sym}}^{(d)}$ is isomorphic to the vector space of n-variate quantics of degree d (cf. [10, §3.5.2]).

The antisymmetric spaces are connected with the Clifford algebra $C\ell_d$ of $\mathbb{R}^n$, which is isomorphic to the direct sum $\bigoplus_{j=1}^d \bigwedge^j \mathbb{R}^n$ (cf. Lounesto [14, Chap. 22]).

1.1.4 Properties

Since all permutations are products of transpositions $\pi_{i,i+1}$, the next remark follows.

Remark 1 A tensor $\mathbf{v} \in \otimes^d V$ is symmetric (resp. antisymmetric) if and only if $\pi(\mathbf{v}) = \pi_{i,i+1}(\mathbf{v})$ (resp. $\pi(\mathbf{v}) = -\pi_{i,i+1}(\mathbf{v})$) holds for all transpositions with $1 \leq i < d$.

Let $\mathbf{V} := \otimes^d V$. The linear maps

$$\mathscr{S} = \mathscr{S}_d := \frac{1}{d!} \sum_{\pi \in \mathbf{P}_d} \pi : \mathbf{V} \to \mathbf{V}, \; \mathscr{A} = \mathscr{A}_d := \frac{1}{d!} \sum_{\pi \in \mathbf{P}_d} \operatorname{sign}(\pi)\pi : \mathbf{V} \to \mathbf{V} \tag{8}$$

are projections onto $\mathbf{V}_{\mathrm{sym}}$ and $\mathbf{V}_{\mathrm{anti}}$, respectively (For a proof note that $\mathscr{S} = \mathscr{S}\pi$ and $\mathscr{A} = \operatorname{sign}(\pi)\mathscr{A}\pi$ so that the application of $\frac{1}{d!}\sum_{\pi \in \mathbf{P}_d}$ yields $\mathscr{S} = \mathscr{S}\mathscr{S}$ and $\mathscr{A} = \mathscr{A}\mathscr{A}$). $\mathscr{S}$ and $\mathscr{A}$ are called the *symmetrisation* and *alternation*, respectively.

Remark 2 Let $\varphi_{D\backslash\alpha} \in \mathbf{V}'_{D\backslash\alpha}$ be a functional[1] (no symmetry condition assumed). If $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}^{(D)}$ or $\mathbf{v} \in \mathbf{V}_{\mathrm{anti}}^{(D)}$, then $\varphi_{D\backslash\alpha}(\mathbf{v}) \in \mathbf{V}_{\mathrm{sym}}^{(\alpha)}$ or $\varphi_{D\backslash\alpha}(\mathbf{v}) \in \mathbf{V}_{\mathrm{anti}}^{(\alpha)}$, respectively.

The following expansion lemma will be used in the following.

[1] Compare the definition (5) with interchanged subsets α and $D\backslash\alpha$.

Lemma 1 *Let $\{u_1, \ldots, u_r\}$ be a basis of the subspace $U \subset V$. Any tensor $\mathbf{v} \in \otimes^k U$ can be written in the form*

$$\mathbf{v} = \sum_{\ell=1}^{r} \mathbf{v}_{[\ell]} \otimes u_\ell \qquad \textit{with } \mathbf{v}_{[\ell]} \in \otimes^{k-1} U.$$

Let $\{\varphi_1, \ldots, \varphi_r\} \subset U'$ be a dual basis of $\{u_1, \ldots, u_r\}$, i.e., $\varphi_i(u_j) = \delta_{ij}$. Then the tensors $\mathbf{v}_{[\ell]}$ are defined by $\mathbf{v}_{[\ell]} = \varphi_\ell(\mathbf{v})$.

A consequence of the last equation and Remark 2 is the following.

Remark 3 If $\mathbf{v} \in \otimes^k U$ is (anti-)symmetric, then so is $\mathbf{v}_{[\ell]} \in \otimes^{k-1} U$.

1.2 Minimal Subspaces

Given a tensor $\mathbf{v} \in \mathbf{V} = \bigotimes_{j \in D} V_j$ and a subset $\alpha \subset D$, the corresponding minimal subspace is defined by

$$\mathbf{U}_\alpha^{\min}(\mathbf{v}) := \left\{ \varphi_{D \backslash \alpha} \mathbf{v} : \varphi_{D \backslash \alpha} \in \mathbf{V}'_{D \backslash \alpha} \right\} \in \mathbf{V}_\alpha \tag{9}$$

(cf. (5); [10, §6]). $\mathbf{U}_\alpha^{\min}(\mathbf{v})$ is the subspace of smallest dimension with the property $\mathbf{v} \in \mathbf{U}_\alpha^{\min}(\mathbf{v}) \otimes \mathbf{V}_{D \backslash \alpha}$. The dual space $\mathbf{V}'_{D \backslash \alpha}$ in (9) may be replaced by $\bigotimes_{j \in D \backslash \alpha} V'_j$.

For a subset $\mathbf{V}_0 \subset \mathbf{V}$ we define $\mathbf{U}_\alpha^{\min}(\mathbf{V}_0) := \operatorname{span}\{\mathbf{U}_\alpha^{\min}(\mathbf{v}) : \mathbf{v} \in \mathbf{V}_0\}$.

Remark 4 Let $\emptyset \neq \beta \subsetneqq \alpha \subset D$ be nonempty subsets. Then $\mathbf{U}_\beta^{\min}(\mathbf{v}) = \mathbf{U}_\beta^{\min}(\mathbf{U}_\alpha^{\min}(\mathbf{v}))$.

A conclusion from Remark 2 is the following statement.

Conclusion 1 *If $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}$ [or $\mathbf{V}_{\mathrm{anti}}$], then $\mathbf{U}_\alpha^{\min}(\mathbf{v}) \subset \mathbf{V}_{\mathrm{sym}}^{(\alpha)}$ [or $\mathbf{U}_\alpha^{\min}(\mathbf{v}) \subset \mathbf{V}_{\mathrm{anti}}^{(\alpha)}$].*

2 *r*-Term Format for (Anti)symmetric Tensors

Let $r \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$. A tensor $\mathbf{v} \in \mathbf{V} = \otimes^d V$ can be represented in the *r-term format* (or canonical format) if there are $v_\nu^{(j)} \in V$ for $1 \le j \le d$ and $1 \le \nu \le r$ such that

$$\mathbf{v} = \sum_{\nu=1}^{r} \bigotimes_{j=1}^{d} v_\nu^{(j)}.$$

We recall that the smallest possible r in the above representation is called the *rank* of $\mathbf{v}$ and denoted by $\operatorname{rank}(\mathbf{v})$. The number r used above is called the *representation* *rank*. Since the determination of rank($\mathbf{v}$) is NP hard (cf. Håstad [11]) we cannot expect that $r \geq \operatorname{rank}(\mathbf{v})$ holds with an equal sign.

Two approaches to representing (anti)symmetric tensors by the r-term format are described in Sects. 2.1 and 2.2.

2.1 Indirect Representation

A symmetric tensor $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}$ may be represented by a general tensor $\mathbf{w} \in \mathbf{V}$ with the property $\mathscr{S}(\mathbf{w}) = \mathbf{v}$, where $\mathscr{S}$ ($\mathscr{A}$) is the symmetrisation (alternation) defined in (8). The representation of $\mathbf{w} \in \mathbf{V}$ uses the r-term format: $\mathbf{w} = \sum_{i=1}^{r} \bigotimes_{j=1}^{d} w_i^{(j)}$. This approach is proposed by Mohlenkamp, e.g., in [4]. For instance, $\mathbf{v} = a \otimes a \otimes b + a \otimes b \otimes a + b \otimes a \otimes a \in \mathbf{V}_{\mathrm{sym}}$ is represented by $\mathbf{w} = 3a \otimes a \otimes b$. This example indicates that $\mathbf{w}$ may be of a much simpler form than the symmetric tensor $\mathbf{v} = \mathscr{S}(\mathbf{w})$.

However, the cost (storage size) of the representation is only one aspect. Another question concerns the tensor operations. In the following we discuss the addition, the scalar product, and the matrix-vector multiplication.

The *addition* is easy to perform. By linearity of $\mathscr{S}$, the sum of $\mathbf{v}' = \mathscr{S}(\mathbf{w}')$ and $\mathbf{v}'' = \mathscr{S}(\mathbf{w}'')$ is represented by $\mathbf{w}' + \mathbf{w}''$. Similar in the antisymmetric case.

The *summation* within the r-term format does not require computational work, but increases the representation rank r. This leads to the question how to *truncate* $\mathbf{w} = \mathbf{w}' + \mathbf{w}''$ to a smaller rank. It is known that truncation within the r-term format is not an easy task. However, if one succeeds to split $\mathbf{w}$ into $\hat{\mathbf{w}} + \delta\mathbf{w}$, where $\hat{\mathbf{w}}$ has smaller rank and $\delta\mathbf{w}$ is small, this leads to a suitable truncation of $\mathbf{v} = \hat{\mathbf{v}} + \delta\mathbf{v}$ with $\hat{\mathbf{v}} = \mathscr{S}(\hat{\mathbf{w}})$, $\delta\mathbf{v} = \mathscr{S}(\delta\mathbf{w})$, since $\|\delta\mathbf{v}\| \leq \|\delta\mathbf{w}\|$ with respect to the Euclidean norm.

The computation of the *scalar product* $\langle \mathbf{v}', \mathbf{v}'' \rangle$ of $\mathbf{v}', \mathbf{v}''$ in $\mathbf{V}_{\mathrm{sym}}$ or $\mathbf{V}_{\mathrm{anti}}$ is more involved. In the antisymmetric case, $\langle \mathbf{v}', \mathbf{v}'' \rangle$ with $\mathbf{v}' = \mathscr{A}(\mathbf{w}')$, $\mathbf{v}'' = \mathscr{A}(\mathbf{w}'')$ and

$$\mathbf{w}' = \sum\nolimits_{i'} \bigotimes\nolimits_{j=1}^{d} w_{i'}^{\prime(j)}, \qquad \mathbf{w}'' = \sum\nolimits_{i''} \bigotimes\nolimits_{j=1}^{d} w_{i''}^{\prime\prime(j)}$$

can be written as the sum $\langle \mathbf{v}', \mathbf{v}'' \rangle = \sum_{i',i''} s_{i'i''}$ with the terms

$$s_{i'i''} := \left\langle \mathscr{A}\left(\bigotimes_{j=1}^{d} w_{i'}^{\prime(j)} \right), \mathscr{A}\left(\bigotimes_{j=1}^{d} w_{i''}^{\prime\prime(j)} \right) \right\rangle.$$

The latter product coincides with the determinant

$$s_{i'i''} = \det\left(\left(\langle w_{i'}^{\prime(\nu)}, w_{i''}^{\prime\prime(\mu)} \rangle \right)_{1 \leq \nu,\mu \leq d} \right)$$

(cf. Löwdin [15, (35)]). If the respective representation ranks of $\mathbf{v}'$ and $\mathbf{v}''$ are r' and r'', the cost amounts to $\mathscr{O}(r'r''d^3)$.

While in the antisymmetric case the determinant can be computed in polynomial time, this does not hold for the analogue in the symmetric case. Instead of the determinant one has to compute the permanent.[2] As proved by Valiant [17], its computation is NP hard. Hence the computation of the scalar product is only feasible for small d or in special situations.

Next we consider the *multiplication* of a symmetric Kronecker matrix $\mathbf{A} \in \mathbf{L}_{\mathrm{sym}} \subset \otimes^d L(V)$ (cf. Sect. 9) by a tensor $\mathbf{v} \in \mathbf{V}_{\mathrm{sym/anti}} \subset \otimes^d V$. $\mathbf{A}$ is represented by $\mathbf{B} \in \otimes^d L(V)$ via $\mathbf{A} = \mathscr{S}(\mathbf{B})$ and $\mathbf{B} = \sum_\nu \bigotimes_{j=1}^d B_\nu^{(j)}$, while $\mathbf{v} = \mathscr{S}(\mathbf{w})$ or $\mathbf{v} = \mathscr{A}(\mathbf{w})$ is represented by $\mathbf{w} = \sum_\mu \bigotimes_{j=1}^d w_\mu^{(j)}$. The property $\mathbf{v} \in \mathbf{V}_{\mathrm{sym/anti}}$ implies the respective property $\mathbf{A}\,\mathbf{v} \in \mathbf{V}_{\mathrm{sym/anti}}$. Unfortunately, $\mathbf{A}\,\mathbf{v}$ is not the (anti)symmetrisation of $\mathbf{Bw}$. Instead one may use (cf. Lemma 9)

$$\mathbf{A}\,\mathbf{v} = \mathscr{S}(\mathbf{A}\,\mathbf{w}) = \mathscr{S}(\mathbf{B}\,\mathbf{v}) \quad \text{or} \quad \mathbf{A}\,\mathbf{v} = \mathscr{A}(\mathbf{A}\,\mathbf{w}) = \mathscr{A}(\mathbf{B}\,\mathbf{v}), \text{ resp.}$$

However, this requires that either the symmetric tensor $\mathbf{A}$ or the (anti)symmetric tensor $\mathbf{v}$ must be constructed explicitly, which contradicts the intention of the indirect representation. Similarly, the Hadamard product $\mathbf{v}' \odot \mathbf{v}''$ and the convolution $\mathbf{v}' \star \mathbf{v}''$ are hard to perform within this format.

Conclusion 2 *The indirect representation is suited to antisymmetric tensors if only the addition and the scalar product is required. In the case of symmetric tensors, the computation of the scalar product is restricted to small d.*

Let $\mathbf{v} \in \mathbf{V}_{\mathrm{sym/anti}}$ be a tensor of rank r_v. The indirect representation uses the r_w-term representation of some $\mathbf{w} \in \mathbf{V}$. The gain is characterised by the ratio r_v/r_w where r_w is the smallest possible rank. r_v/r_w takes values from 1 to $d!$. According to Seigal (private communication, 2016), the generic reduction factor r_v/r_w approaches d for large $\dim(V)$. The proof uses the results of Abo–Vannieuwenhoven [1] and Abo et al. [2]. Note that low-rank tensors belong to the measure-zero set of non-generic tensors.

2.2 Direct Symmetric r-Term Representation

While the previous approach uses general (nonsymmetric) tensors, we now represent the symmetric tensors by an r-term representation involving only symmetric rank-1 tensors:

$$\mathbf{v} = \sum_{i=1}^{r} \alpha_i \otimes^d v_i \qquad \text{for suitable } r \in \mathbb{N}_0 \text{ and } v_i \in V,\ \alpha_i \in \mathbb{K} \tag{10}$$

[2] The permanent of $A \in \mathbb{R}^{d\times d}$ is $\mathrm{Perm}(A) = \sum_{\pi\in\mathbf{P}_d} \prod_{i=1}^d a_{i,\pi(i)}$.

(cf. [10, p. 65]).[3] The minimal r in this representation is called the *symmetric rank* of $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}$ and is denoted by $\mathrm{rank}_{\mathrm{sym}}(\mathbf{v})$. Details about symmetric tensors and the symmetric tensor rank are described, e.g., by Comon–Golub–Lim–Mourrain [8].

Since the symmetric rank is at least as large as the standard tensor rank, the required r may be large. A difficulty of the r-term format is caused by the fact that, in general, the set $\left\{\mathbf{v} \in \otimes^d V : \mathrm{rank}(v) \le r\right\}$ is not closed. The simplest counterexamples are symmetric tensors of the form $\lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(\otimes^3(\mathbf{v}+\varepsilon\mathbf{w}) - \otimes^3\mathbf{v}\right)$.[4] Therefore also the subset of the symmetric tensors (10) is not closed.

3 Subspace (Tucker) Format

Let $\mathbf{v} \in \mathbf{V}$ be the tensor to be represented. The subspace format (Tucker format) uses subspaces $U_j \subset V_j$ with the property $\mathbf{v} \in \bigotimes_{j=1}^d U_j$. In the (anti)symmetric case one can choose equal subspaces $U \subset V$ (set, e.g., $U = \bigcap_{j=1}^d U_j$). Let $\{u_1, \ldots, u_r\}$ be a basis of U. Then the explicit Tucker representation of $\mathbf{v}$ takes the form

$$\mathbf{v} = \sum_{i_1,\ldots,i_d=1}^{r} c_{i_1,\ldots,i_d} \bigotimes_{j=1}^{d} u_{i_j} \tag{11}$$

with the so-called core tensor $\mathbf{c} \in \bigotimes_{j=1}^d \mathbb{K}^r$. Obviously, $\mathbf{v}$ is (anti)symmetric if and only if $\mathbf{c}$ is so. Therefore the difficulty is shifted into the treatment of the core tensor. The representation (11) itself does not help to represent (anti)symmetric tensors. One may construct hybrid formats, using one of the other representations for $\mathbf{c}$.

4 Hierarchical Format

In the general case the hierarchical format is a very efficient and flexible representation (cf. [10, §§11–12]). Here we briefly describe the general setting, the TT variant, and first consequences for its application to (anti)symmetric tensors.

4.1 General Case

The recursive partition of the set $D = \{1, \ldots, d\}$ is described by a binary partition tree T_D. It is defined by the following properties: (a) $D \in T_D$ is the root; (b) the

[3] If $\mathbb{K} = \mathbb{C}$ or if d is odd, the factor α_i can be avoided since its d-th root can be combined with v_i.

[4] The described limit $\mathbf{x}$ satisfies $\mathrm{rank}_{\mathrm{sym}}(\mathbf{x}) \ge \mathrm{rank}(\mathbf{x}) = d$ (cf. Buczyński–Landsberg [6]), although it is the limit of tensors with symmetric rank 2.

singletons $\{1\}, \ldots, \{d\}$ are the leaves; (c) if $\alpha \in T_D$ is not a leaf, the sons $\alpha', \alpha'' \in T_D$ are disjoint sets with $\alpha = \alpha' \cup \alpha''$.

The hierarchical representation of a tensor $v \in \mathbf{V} = \bigotimes_{j=1}^{d} V_j$ is algebraically characterised by subspaces $\mathbf{U}_\alpha \subset \mathbf{V}_\alpha$ ($\mathbf{V}_\alpha$ defined in (3)) for all $\alpha \in T_D$ with

$$\mathbf{v} \in \mathbf{U}_D, \tag{12a}$$

$$\mathbf{U}_\alpha \subset \mathbf{U}_{\alpha'} \otimes \mathbf{U}_{\alpha''} \qquad (\alpha', \alpha'' \text{ sons of } \alpha), \text{ if } \alpha \text{ is not a leaf.} \tag{12b}$$

Let the dimensions of $\mathbf{U}_\alpha$ be $r_\alpha := \dim(\mathbf{U}_\alpha)$. Since $\mathbf{U}_D = \operatorname{span}(\mathbf{v})$ is sufficient, $r_D = 1$ is the general value.

4.2 Implementation

The subspaces $\mathbf{U}_\alpha$ are described by bases $\{\mathbf{b}_k^{(\alpha)} : k = 1, \ldots, r_\alpha\}$. For leaves $\alpha \in T_D$, the basis is stored explicitly. Otherwise, condition (12b) ensures that

$$\mathbf{b}_\ell^{(\alpha)} = \sum_{i=1}^{r_{\alpha'}} \sum_{j=1}^{r_{\alpha''}} c_{ij}^{(\alpha,\ell)} \mathbf{b}_i^{(\alpha')} \otimes \mathbf{b}_j^{(\alpha'')} \qquad (\alpha', \alpha'' \text{ sons of } \alpha). \tag{13}$$

Therefore it is sufficient to store the coefficients matrices $(c_{ij}^{(\alpha,\ell)})_{1 \le i \le r_{\alpha'}, 1 \le j \le r_{\alpha''}}$, as well as the vector $c^D \in \mathbb{K}^{r_D}$ for the final representation $\mathbf{v} = \sum_i c_i^D \mathbf{b}_i^{(D)}$ (cf. (12a)).

4.3 TT Variant

The TT format is introduced in Oseledets [16] (cf. [10, §12]). It is characterised by a linear tree T_D. That means that the non-leaf vertices $\alpha \in T_D$ are of the form $\alpha = \{1, \ldots, j\}$ with the sons $\alpha' = \{1, \ldots, j-1\}$ and $\alpha'' = \{j\}$. The embedding (12b) is $\mathbf{U}_{\{1,\ldots,j+1\}} \subset \mathbf{U}_{\{1,\ldots,j\}} \otimes U_{\{j\}}$.

Below we shall consider the case $\mathbf{V} = \otimes^d V$, i.e., $V_j = V$ is independent of j. Also their subspaces are independent of j and denoted by $U_{\{j\}} = U$. We abbreviate $\mathbf{U}_{\{1,\ldots,j\}}$ by $\mathbf{U}_j$ and denote its dimension by $r_j := \dim(\mathbf{U}_j)$, $r := r_1 = \dim(U)$ (note that $U = \mathbf{U}_1$). Now the nested inclusion (12b) becomes

$$\mathbf{U}_{j+1} \subset \mathbf{U}_j \otimes U. \tag{14}$$

Similarly, we rewrite $\mathbf{U}_{\{1,\ldots,j\}}^{\min}(\mathbf{v})$ as $\mathbf{U}_j^{\min}(\mathbf{v})$.

4.4 (Anti)symmetric Case

Conclusion 1 proves that (anti)symmetric tensors $\mathbf{v}$ lead to (anti)symmetric minimal subspaces: $\mathbf{U}_\alpha^{\min}(\mathbf{v}) \subset \mathbf{V}_{\mathrm{sym}}^{(\alpha)}$ or $\mathbf{U}_\alpha^{\min}(\mathbf{v}) \in \mathbf{V}_{\mathrm{anti}}^{(\alpha)}$, respectively.

The hierarchical representation (12a,b) of $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}^{(D)}$ should also use subspaces with the property $\mathbf{U}_\alpha \subset \mathbf{V}_{\mathrm{sym}}^{(\alpha)}$ (similar in the antisymmetric case).

The basic task is the determination of a basis $\mathbf{b}_\ell^{(\alpha)} \in \mathbf{U}_\alpha \subset \mathbf{V}_{\mathrm{sym}}^{(\alpha)}$ $(1 \leq \ell \leq r_\alpha := \dim(\mathbf{U}_\alpha))$ by suitable linear combinations of the tensors $\mathbf{b}_i^{(\alpha')} \otimes \mathbf{b}_j^{(\alpha'')}$. The assumptions $\mathbf{b}_i^{(\alpha')} \in \mathbf{V}_{\mathrm{sym}}^{(\alpha')}$ and $\mathbf{b}_j^{(\alpha'')} \in \mathbf{V}_{\mathrm{sym}}^{(\alpha'')}$ lead to a partial symmetry, but, in general, $\pi_{\nu\mu}(\mathbf{b}_\ell^{(\alpha)}) = \mathbf{b}_\ell^{(\alpha)}$ is not satisfied for $\nu \in \alpha'$ and $\mu \in \alpha''$.

Using (14) and symmetry, we conclude that

$$\mathbf{U}_j \subset \left(\otimes^j U\right) \cap \mathbf{V}_{\mathrm{sym}}^{(j)} = \mathbf{U}_{\mathrm{sym}}^{(j)}. \tag{15}$$

Remark 5

(a) Because of (15) we can restrict the vector space V in (7a,b) to U.
(b) If we want to represent the tensor $\mathbf{v}$, the subspace $\mathbf{U}_j$ must satisfy

$$\mathbf{U}_j^{\min}(\mathbf{v}) \subset \mathbf{U}_j.$$

4.5 Dimensions of $\mathbf{U}_j^{\min}$ in the (Anti)symmetric Case

The following statement shows that, in the case of antisymmetric tensors, the hierarchical approach becomes costly for high dimensions d. The simplest antisymmetric tensor is the antisymmetrisation of an elementary tensor:

$$\mathbf{a} := \mathscr{A}\left(\bigotimes_{j=1}^{d} u^{(j)}\right).$$

To ensure $\mathbf{a} \neq 0$, the vectors $u^{(j)}$ must be linearly independent. In that case the minimal subspace $\mathbf{U}_k^{\min}(\mathbf{v})$ is spanned by all tensors $\mathscr{A}\left(\bigotimes_{j=1}^{k} u^{(i_j)}\right)$ with $1 \leq i_1 < i_2 < \ldots < i_k \leq d$. There are $\binom{k}{d}$ tensors of this form. Since they are linearly independent, $\dim \mathbf{U}_k^{\min}(\mathbf{a}) = \binom{k}{d}$ follows. The sum $\sum_{k=1}^{d} \dim \mathbf{U}_k^{\min}(\mathbf{a})$ is $2^d - 1$. Hence, this approach cannot be recommended for large d.

The situation is different in the symmetric case, since the vectors in $\mathscr{S}(\bigotimes_{j=1}^{d} u^{(j)})$ need not be linearly independent. The next lemma uses the symmetric rank defined in Sect. 2.2.

Lemma 2 *All symmetric tensors* $\mathbf{v}$ *satisfy* $\dim \mathbf{U}_k^{\min}(\mathbf{v}) \leq \mathrm{rank}_{\mathrm{sym}}(\mathbf{v})$.

Proof Let $\mathbf{v} = \sum_{i=1}^{r} \alpha_i \otimes^d v_i$ with $r = \mathrm{rank}_{\mathrm{sym}}(\mathbf{v})$. Then all minimal subspaces $\mathbf{U}_k^{\min}(\mathbf{v})$ are contained in the r-dimensional space span $\left\{\otimes^k v_i : 1 \le i \le r\right\}$. □

The following symmetric tensor describes, e.g., the structure of the Laplace operator (a, b of the next example are the identity map and one-dimensional Laplacian, respectively).

Example 1 An important example is the symmetric tensor $\mathbf{v} := \mathscr{S}(\otimes^{d-1} a \otimes b)$, where $a, b \in V$ are linearly independent. In this case we have

$$\dim(\mathbf{U}_k^{\min}(\mathbf{v})) = 2 \qquad \text{for } 1 \le k < d.$$

More precisely, $\mathbf{U}_k^{\min}(\mathbf{v})$ is spanned by $\otimes^k a$ and $\mathscr{S}_k\big(b \otimes \big(\otimes^{k-1} a\big)\big)$.

5 TT Format for Symmetric Tensors

In the following we focus on the representation of symmetric tensors in the TT format (cf. Sect. 4.3). In principle, the same technique can be used for antisymmetric tensors (but compare Sect. 4.5).

In Sect. 5.4 we try to construct the space $\mathbf{U}_{j+1}$ from $\mathbf{U}_j$. This will lead to open questions in Sect. 5.4.6. If we start from $\mathbf{v}$ and the related minimal subspace $\mathbf{U}_j^{\min}(\mathbf{v})$, then an appropriate choice is $\mathbf{U}_j = \mathbf{U}_j^{\min}(\mathbf{v})$ (see Sect. 5.5).

5.1 The Space $\big(\mathbf{U}_j \otimes U\big) \cap \mathscr{S}\big(\mathbf{U}_j \otimes U\big)$ and the Principal Idea

We want to repeat the same construction of nested spaces as in (14). In contrast to the general case, we also have to ensure symmetry. By induction, we assume that $\mathbf{U}_j$ contains only symmetric tensors:

$$\mathbf{U}_j \subset \mathbf{V}_{\mathrm{sym}}^{(j)}. \tag{16}$$

On the one hand, the new space $\mathbf{U}_{j+1}$ should satisfy $\mathbf{U}_{j+1} \subset \mathbf{U}_j \otimes U$; on the other hand, symmetry $\mathbf{U}_{j+1} \subset \mathbf{V}_{\mathrm{sym}}^{(j+1)}$ is required. Together, $\mathbf{U}_{j+1} \subset \big(\mathbf{U}_j \otimes U\big) \cap \mathbf{V}_{\mathrm{sym}}^{(j+1)}$ must be ensured.

Remark 6 $\big(\mathbf{U}_j \otimes U\big) \cap \mathbf{V}_{\mathrm{sym}}^{(j+1)} = \big(\mathbf{U}_j \otimes U\big) \cap \mathscr{S}_{j+1}\big(\mathbf{U}_j \otimes U\big)$ holds with the symmetrisation $\mathscr{S}_{j+1}$ in (8).

Proof Let $\mathbf{v} \in \big(\mathbf{U}_j \otimes U\big) \cap \mathbf{V}_{\mathrm{sym}}^{(j+1)}$. Since $\mathbf{v} \in \mathbf{V}_{\mathrm{sym}}^{(j+1)}$, $\mathscr{S}(\mathbf{v}) = \mathbf{v}$ holds. Since $\mathbf{v} \in \mathbf{U}_j \otimes U$, $\mathbf{v} = \mathscr{S}(\mathbf{v}) \in \mathscr{S}\big(\mathbf{U}_j \otimes U\big)$ follows. This proves $\big(\mathbf{U}_j \otimes U\big) \cap \mathbf{V}_{\mathrm{sym}}^{(j+1)} \subset \big(\mathbf{U}_j \otimes U\big) \cap \mathscr{S}\big(\mathbf{U}_j \otimes U\big)$. The reverse inclusion follows from $\mathscr{S}\big(\mathbf{U}_j \otimes U\big) \subset \mathbf{V}_{\mathrm{sym}}^{(j+1)}$. □

This leads us to the condition

$$\mathbf{U}_{j+1} \subset \hat{\mathbf{U}}_{j+1} := \big(\mathbf{U}_j \otimes U\big) \cap \mathscr{S}\,\big(\mathbf{U}_j \otimes U\big) \tag{17}$$

for the choice of the next subspace $\mathbf{U}_{j+1}$.

It must be emphasised that, in general, $\mathscr{S}\,\big(\mathbf{U}_j \otimes U\big)$ is not a subspace of $\mathbf{U}_j \otimes U$. Example 5 will show nontrivial subspaces $\mathbf{U}_j$, U that may even lead to $\hat{\mathbf{U}}_{j+1} = \{0\}$.

To repeat the construction (13), we assume that there is a basis $\{\mathbf{b}_1^{(j)}, \ldots, \mathbf{b}_{r_j}^{(j)}\}$ of $\mathbf{U}_j$ and the basis $\{u_1, \ldots, u_r\}$ of U. Then the basis $\{\mathbf{b}_1^{(j+1)}, \ldots, \mathbf{b}_{r_{j+1}}^{(j+1)}\}$ of $\mathbf{U}_{j+1}$ can be constructed by (13) which now takes the form

$$\mathbf{b}_k^{(j+1)} = \sum_{\nu=1}^{r_j} \sum_{\mu=1}^{r} c_{\nu\mu}^{(k)} \mathbf{b}_\nu^{(j)} \otimes u_\mu \qquad (1 \le k \le r_{j+1}). \tag{18}$$

In order to check linear independence and to construct orthonormal bases, we also have to require that we are able to determine *scalar products*. Assuming by induction that the scalar products $\langle \mathbf{b}_{\nu'}^{(j)}, \mathbf{b}_{\nu''}^{(j)} \rangle$ and $\big\langle u_{\mu'}, u_{\mu''} \big\rangle$ are known, the value of $\langle \mathbf{b}_{k'}^{(j+1)}, \mathbf{b}_{k''}^{(j+1)} \rangle$ follows from (4). Therefore, we are able to form an orthonormal basis of $\mathbf{U}_{j+1}$.

To avoid difficulties with a too small intersection $\hat{\mathbf{U}}_{j+1}$, an alternative idea could be to choose the subspace $\mathbf{U}_{j+1}$ in $\mathscr{S}\,\big(\mathbf{U}_j \otimes U\big)$ and not necessarily in $\mathbf{U}_j \otimes U$. Then, instead of (18), we have $\mathbf{b}_k^{(j+1)} = \sum_{\nu,\mu} c_{\nu\mu}^{(k)} \mathscr{S}(\mathbf{b}_\nu^{(j)} \otimes u_\mu)$. This would be a very flexible approach, were it not for the fact that we need knowledge of the scalar products $\Big\langle s_{\nu\mu}^{(j)}, s_{\nu'\mu'}^{(j)} \Big\rangle$ for $s_{\nu\mu}^{(j)} := \mathscr{S}(\mathbf{b}_\nu^{(j)} \otimes u_\mu)$ (otherwise, an orthonormal basis $\{\mathbf{b}_k^{(j+1)}\}$ cannot be constructed). One finds that $\Big\langle s_{\nu\mu}^{(j)}, s_{\nu'\mu'}^{(j)} \Big\rangle = \frac{\delta_{\mu\mu'}}{j+1} \Big\langle \mathbf{b}_\nu^{(j)}, \mathbf{b}_{\nu'}^{(j)} \Big\rangle + \frac{j}{j+1} \Big\langle \mathbf{b}_{\nu,[\mu']}^{(j)}, \mathbf{b}_{\nu',[\mu]}^{(j)} \Big\rangle$ where the expression $\mathbf{b}_{\nu,[\mu']}^{(j)}$ is defined in Lemma 1. The scalar products $\Big\langle \mathbf{b}_{\nu,[\mu']}^{(j)}, \mathbf{b}_{\nu',[\mu]}^{(j)} \Big\rangle$ can be derived from $\Big\langle s_{\nu\mu,[\ell]}^{(j)}, s_{\nu'\mu',[\ell']}^{(j)} \Big\rangle$. This expression, however, requires the knowledge of $\Big\langle \mathbf{b}_{\nu,[\ell,\mu']}^{(j-1)}, \mathbf{b}_{\nu',[\ell',\mu]}^{(j-1)} \Big\rangle$ (concerning the subscript $[\ell, \mu']$ compare (22)). Finally, we need scalar products of the systems $\{\mathbf{b}_\nu^{(d)}\}$, $\{\mathbf{b}_\nu^{(d-1)}\}$, $\{\mathbf{b}_{\nu,[\mu]}^{(d-1)}\}$, $\{\mathbf{b}_{\nu,[\ell m]}^{(d-2)}\}, \ldots,$ $\{\mathbf{b}_{\nu,[\ell_1,\ell_2,\ldots,\ell_{j*}]}^{(j)}\}$ with $j^* = \min\{j, d-j\}, \ldots$. This leads to a data size increasing exponentially in d.

5.2 The Spaces $\big(\mathbf{U}_j \otimes U\big) \cap \mathscr{S}\,\big(\mathbf{U}_j \otimes U\big)$ and $\mathbf{U}_{j+1}^{\min}(\mathbf{v})$

Let $\mathbf{v} \in (\bigotimes^d V) \cap \mathbf{V}_{\mathrm{sym}}^{(d)}$ be the symmetric tensor which we want to represent. We recall the minimal subspaces defined in Sect. 1.2. According to the notation of the TT format, $\mathbf{U}_j^{\min}(\mathbf{v})$ is the space $\mathbf{U}_{\{1,\ldots,j\}}^{\min}(\mathbf{v}) \subset \otimes^j V$ defined in (9). The minimality property of $\mathbf{U}_j^{\min}(\mathbf{v})$ (cf. [10, §6]) implies that the subspaces U and $\mathbf{U}_j$ must satisfy

$$U \supset U_1^{\min}(\mathbf{v}), \qquad \mathbf{U}_j \supset \mathbf{U}_j^{\min}(\mathbf{v}); \tag{19}$$

otherwise $\mathbf{v}$ cannot be represented by the TT format.

The next theorem states that (19) guarantees that there is a suitable subspace $\mathbf{U}_j$ with $\hat{\mathbf{U}}_j \supset \mathbf{U}_j \supset \mathbf{U}_j^{\min}(\mathbf{v})$, so that the requirement (19) is also valid for $j+1$.

Theorem 1 *Let (19) be valid for* $\mathbf{v} \in \mathbf{U}_{\mathrm{sym}}^{(d)}$ *and let* $j < d$. *Then* $\hat{\mathbf{U}}_{j+1}$ *in (17) satisfies* $\hat{\mathbf{U}}_{j+1} \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$.

Proof We have $\mathbf{U}_j \otimes U \supset \mathbf{U}_j^{\min}(\mathbf{v}) \otimes U_1^{\min}(\mathbf{v})$. A general property of the minimal subspace is

$$\mathbf{U}_j^{\min}(\mathbf{v}) \otimes U_1^{\min}(\mathbf{v}) \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$$

(cf. [10, Proposition 6.17]). Since $\mathbf{U}_{j+1}^{\min}(\mathbf{v})$ is symmetric (cf. Conclusion 1), it follows that

$$\mathscr{S}\left(\mathbf{U}_j \otimes U\right) \supset \mathscr{S}(\mathbf{U}_j^{\min}(\mathbf{v}) \otimes U_1^{\min}(\mathbf{v})) \supset \mathscr{S}(\mathbf{U}_{j+1}^{\min}(\mathbf{v})) = \mathbf{U}_{j+1}^{\min}(\mathbf{v}).$$

This inclusion together with the previous inclusion $\mathbf{U}_j \otimes U \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$ yields the statement. □

So far, we could ensure that there exists a suitable subspace $\mathbf{U}_{j+1} \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$. Concerning the practical implementation, two questions remain:

(a) How can we find the subspace $\hat{\mathbf{U}}_{j+1} \subset \mathbf{U}_j \otimes U$?
(b) Given $\hat{\mathbf{U}}_{j+1}$, how can we ensure $\mathbf{U}_{j+1} \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$?

The next subsection yields a partial answer to the first question.

5.3 Criterion for Symmetry

According to Lemma 1, any $\mathbf{v} \in \mathbf{U}_j \otimes U$ is of the form

$$\mathbf{v} = \sum_{\ell=1}^{r} \mathbf{v}_{[\ell]} \otimes u_\ell \qquad (\mathbf{v}_{[\ell]} \in \mathbf{U}_j). \tag{20}$$

The mapping $\mathbf{v} \in \mathbf{V}^{(j+1)} \mapsto \mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)}$ can be iterated:

$$\mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)} \mapsto (\mathbf{v}_{[\ell]})_{[m]} = \mathbf{v}_{[\ell][m]} = \mathbf{v}_{[\ell,m]} \in \mathbf{V}^{(j-1)}.$$

In the case of $j = 1$, the empty product $\otimes^{j-1} V$ is defined as the field $\mathbb{K}$, i.e., $\mathbf{v}_{[\ell,m]}$ is a scalar.

Lemma 3 *A necessary and sufficient condition for* $\mathbf{v} \in \mathbf{V}^{(j+1)}_{\mathrm{sym}}$ *is*

$$\mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)}_{\mathrm{sym}} \quad \textit{and} \quad \mathbf{v}_{[\ell,m]} = \mathbf{v}_{[m,\ell]} \;\; \textit{for all} \;\; 1 \leq \ell, m \leq r. \tag{21}$$

Here, $\mathbf{v}_{[\ell]}$ *refers to (20), and* $\mathbf{v}_{[\ell,m]}$ *is the expansion term of* $\mathbf{v}_{[\ell]} \in \otimes^j V$.

Proof

(a) Assume $\mathbf{v} \in \mathbf{V}^{(j+1)}_{\mathrm{sym}}$. $\mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)}_{\mathrm{sym}}$ is stated in Remark 3. Applying the expansion to $\mathbf{v}_{[\ell]}$ in (20), we obtain

$$\mathbf{v} = \sum_{\ell,m=1}^{r} \mathbf{v}_{[\ell,m]} \otimes u_m \otimes u_\ell. \tag{22}$$

Note that the tensors $\{u_m \otimes u_\ell : 1 \leq \ell, m \leq r\}$ are linearly independent. Therefore, transposition $u_m \otimes u_\ell \mapsto u_\ell \otimes u_m$ and symmetry of $\mathbf{v}$ imply that $\mathbf{v}_{[\ell,m]} = \mathbf{v}_{[m,\ell]}$.

(b) Assume (21). Because of $\mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)}_{\mathrm{sym}}$, $\mathbf{v}$ is invariant under all transpositions $\pi_{i,i+1}$ for $1 \leq i < j$. Condition $\mathbf{v}_{[\ell,m]} = \mathbf{v}_{[m,\ell]}$ ensures that $\mathbf{v}$ is also invariant under the transposition $\pi_{j,j+1}$. This proves the symmetry of $\mathbf{v}$ (cf. Remark 1).

□

To apply this criterion to the construction of $\hat{\mathbf{U}}_{j+1} := \left(\mathbf{U}_j \otimes U\right) \cap \mathscr{S}\left(\mathbf{U}_j \otimes U\right)$, we search for a symmetric tensor (18) of the form

$$\mathbf{b} = \sum_{\nu=1}^{r_j} \sum_{\mu=1}^{r} c_{\nu\mu} \mathbf{b}^{(j)}_\nu \otimes u_\mu. \tag{23}$$

The tensor $\mathbf{b}$ corresponds to $\mathbf{v}_{[\ell]} := \sum_{\nu=1}^{r_j} c_{\nu\ell} \mathbf{b}^{(j)}_\nu$ in (20). $\mathbf{v}_{[\ell]} \in \mathbf{V}^{(j)}_{\mathrm{sym}}$ is satisfied because of (16). The condition $\mathbf{v}_{[\ell,m]} = \mathbf{v}_{[m,\ell]}$ in (21) becomes

$$\sum_{\nu=1}^{r_j} c_{\nu\ell} \mathbf{b}^{(j)}_{\nu,[m]} = \sum_{\nu=1}^{r_j} c_{\nu m} \mathbf{b}^{(j)}_{\nu,[\ell]}. \tag{24}$$

The tensors $\mathbf{b}^{(j)}_{\nu,[m]}$ and $\mathbf{b}^{(j)}_{\nu,[\ell]}$ belong to $\mathbf{U}_{j-1}$. The new (nontrivial) algebraic task is to find the set of coefficients $c_{\nu\mu}$ satisfying (24) for all $1 \leq \ell, m \leq r$.

Remark 7 The ansatz (23) has $rr_j - 1$ free parameters (one has to be subtracted because of the normalisation). Condition (24) describes $r(r-1)/2$ equations in the space $\mathbf{U}_{j-1}$ equivalent to $\frac{r(r-1)}{2} r_{j-1}$ scalar equations.

5.4 TT Symmetrisation

In the following approach we obtain a symmetric tensor in $\mathscr{S}\left(\mathbf{U}_j \otimes U\right)$, but not necessarily in $\mathbf{U}_j \otimes U$.

5.4.1 Symmetrisation Operator $\mathscr{S}_{j+1}$

In the present approach we directly symmetrise the tensors. The general symmetrisation map $\mathscr{S}$ consists of $d!$ terms. However, since $\mathbf{U}_j$ is already symmetric, tensors in $\mathbf{U}_j \otimes U$ can be symmetrised by only $j+1$ transpositions $\pi_{i,j+1}$.

Theorem 2 *Let* $\mathbf{v}^{(j)} \in \mathbf{V}^{(j)}_{\mathrm{sym}}$ *and* $w \in V$. *Applying*

$$\hat{\mathscr{S}}_{j+1} := \frac{1}{j+1}\sum_{i=1}^{j+1} \pi_{i,j+1}$$

to $\mathbf{v}^{(j)} \otimes w \in \mathbf{V}^{(j+1)}$ *yields a symmetric tensor:*

$$\mathbf{s} := \hat{\mathscr{S}}_{j+1}\left(\mathbf{v}^{(j)} \otimes w\right) \in \mathbf{V}^{(j+1)}_{\mathrm{sym}}.$$

Proof According to Remark 1, we have to show that $\pi_{k,k+1}\mathbf{s} = \mathbf{s}$ for all $1 \le k \le j$. First we consider the case of $k < j$. If $i \notin \{k, k+1\}$, we have $\pi_{k,k+1}\pi_{i,j+1} = \pi_{i,j+1}\pi_{k,k+1}$. For $i \in \{k, k+1\}$ we obtain

$$\pi_{k,k+1}\pi_{k,j+1} = \pi_{k+1,j+1}\pi_{k,k+1}, \qquad \pi_{k,k+1}\pi_{k+1,j+1} = \pi_{k,j+1}\pi_{k,k+1}.$$

This proves $\pi_{k,k+1}\hat{\mathscr{S}}_{j+1} = \hat{\mathscr{S}}_{j+1}\pi_{k,k+1}$ and

$$\pi_{k,k+1}\mathbf{s} = \hat{\mathscr{S}}_{j+1}\pi_{k,k+1}\left(\mathbf{v}^{(j)} \otimes w\right) = \hat{\mathscr{S}}_{j+1}\left(\pi_{k,k+1}\mathbf{v}^{(j)} \otimes w\right).$$

Symmetry of $\mathbf{v}^{(j)}$ implies $\pi_{k,k+1}\mathbf{v}^{(j)} = \mathbf{v}^{(j)}$ so that $\pi_{k,k+1}\mathbf{s} = \mathbf{s}$ is proved.

The remaining case is $k = j$. For $i < j$, the identity $\pi_{j,j+1}\pi_{i,j+1} = \pi_{i,j+1}\pi_{i,j}$ together with $\pi_{i,j}\mathbf{v}^{(j)} = \mathbf{v}^{(j)}$ implies $\pi_{j,j+1}\pi_{i,j+1}(\mathbf{v}^{(j)} \otimes w) = \pi_{i,j+1}(\mathbf{v}^{(j)} \otimes w)$. For $i \in \{j, j+1\}$ we obtain

$$\pi_{j,j+1}\pi_{j,j+1} = \mathrm{id} = \pi_{j+1,j+1}, \qquad \pi_{j,j+1}\pi_{j+1,j+1} = \pi_{j,j+1}\cdot\mathrm{id} = \pi_{j,j+1};$$

i.e., $\pi_{j,j+1}\left(\pi_{i,j+1} + \pi_{j+1,j+1}\right) = \pi_{i,j+1} + \pi_{j+1,j+1}$. Hence, also $\pi_{j,j+1}\mathbf{s} = \mathbf{s}$ is proved. □

Corollary 1 *The corresponding antisymmetrisation is obtained by*

$$\hat{\mathscr{A}}_d := \frac{1}{d}\sum_{i=1}^{d} (-1)^{d-i}\,\pi_{id}.$$

Although the symmetrisation ensures that $\mathbf{s} \in \mathscr{S}\left(\mathbf{U}_j \otimes U\right)$, there is no guaranty that $\mathbf{s} \in \mathbf{U}_j \otimes U$. Hence, whether $\mathbf{s} \in \hat{\mathbf{U}}_{j+1}$ holds or not is still open.

5.4.2 Expansion of s

Since $\mathbf{v}^{(j)} \otimes w \in \otimes^{j+1} U$, the symmetrisation $\mathbf{s} = \hat{\mathscr{S}}_{j+1}(\mathbf{v}^{(j)} \otimes w)$ also belongs to $\otimes^{j+1} U$. By Lemma 1 there is a representation $\mathbf{s} = \sum_{\ell=1}^{r} \mathbf{s}_{[\ell]} \otimes u_\ell$ with $\mathbf{s}_{[\ell]} \in \mathbf{U}_{\mathrm{sym}}^{(j)}$.

Lemma 4 *Let* $w = \sum_{\ell=1}^{r} c_\ell u_\ell$ *and* $\mathbf{v}^{(j)} \in \mathbf{U}_{\mathrm{sym}}^{(j)}$. *Then* $\mathbf{s} := \hat{\mathscr{S}}_{j+1}(\mathbf{v}^{(j)} \otimes w)$ *satisfies*

$$\mathbf{s} = \sum_{\ell=1}^{r} \mathbf{s}_{[\ell]} \otimes u_\ell \quad \textit{with } \mathbf{s}_{[\ell]} := \frac{1}{j+1}\left(c_\ell \mathbf{v}^{(j)} + \sum_{i=1}^{j} \pi_{i,j}\left(\mathbf{v}_{[\ell]}^{(j)} \otimes w\right)\right). \tag{25}$$

The latter sum $\sum_{i=1}^{j} \pi_{i,j}(\mathbf{v}_{[\ell]}^{(j)} \otimes w)$ *can be written as* $j\hat{\mathscr{S}}_j(\mathbf{v}_{[\ell]}^{(j)} \otimes w)$.

Proof Using $\pi_{j+1,j+1} = \mathrm{id}$, we obtain

$$(j+1)\,\mathbf{s} = \mathbf{v}^{(j)} \otimes w + \sum_{i=1}^{j} \pi_{i,j+1}(\mathbf{v}^{(j)} \otimes w) = \sum_{\ell=1}^{r} c_\ell \mathbf{v}^{(j)} \otimes u_\ell + \sum_{i=1}^{j} \pi_{i,j+1}(\mathbf{v}^{(j)} \otimes w).$$

Since $\pi_{i,j+1} = \pi_{i,j}\pi_{j,j+1}\pi_{i,j}$ for $i \le j$ and $\mathbf{v}^{(j)} = \sum_{\ell=1}^{r} \mathbf{v}_{[\ell]}^{(j)} \otimes u_\ell \in \mathbf{U}_{\mathrm{sym}}^{(j)}$, we have

$$\begin{aligned}
&\pi_{i,j+1}(\mathbf{v}^{(j)} \otimes w) = \pi_{i,j}\pi_{j,j+1}\left((\pi_{i,j}\mathbf{v}^{(j)}) \otimes w\right) = \pi_{i,j}\pi_{j,j+1}\left(\mathbf{v}^{(j)} \otimes w\right) \\
&= \pi_{i,j}\pi_{j,j+1}\sum_{\ell=1}^{r} \mathbf{v}_{[\ell]}^{(j)} \otimes u_\ell \otimes w = \pi_{i,j}\sum_{\ell=1}^{r} \mathbf{v}_{[\ell]}^{(j)} \otimes w \otimes u_\ell = \sum_{\ell=1}^{r}\left(\pi_{i,j}(\mathbf{v}_{[\ell]}^{(j)} \otimes w)\right) \otimes u_\ell.
\end{aligned}$$

Together we obtain $(j+1)\,\mathbf{s} = \sum_{\ell=1}^{r}\left(c_\ell \mathbf{v}^{(j)} + \sum_{\ell=1}^{r}\left(\pi_{i,j}(\mathbf{v}_{[\ell]}^{(j)} \otimes w)\right)\right) \otimes u_\ell$. □

The last equation explicitly provides the expansion of $\mathbf{s}$ defined in Lemma 1.

5.4.3 Scalar Products

The definition of $\mathbf{s} := \hat{\mathscr{S}}_{j+1}(\mathbf{v}^{(j)} \otimes w)$ seems a bit abstract, since (25) contains the permuted tensor which not necessarily belongs to $\mathbf{U}_j \otimes U$. Even in that case it is possible to determine the scalar products $\langle \mathbf{s}, \mathbf{b}_\nu^{(j)} \otimes u_\mu \rangle$ with the basis vectors $\mathbf{b}_\nu^{(j)} \otimes u_\mu$ of $\mathbf{U}_j \otimes U$. The first term in (25) yields

$$\left\langle \mathbf{v}^{(j)} \otimes u_\ell, \mathbf{b}_\nu^{(j)} \otimes u_\mu \right\rangle = \left\langle \mathbf{v}^{(j)}, \mathbf{b}_\nu^{(j)} \right\rangle \left\langle u_\ell, u_\mu \right\rangle.$$

By induction, we assume that the scalar product of $\mathbf{v}^{(j)} \in \mathbf{U}_j$ and $\mathbf{b}_\nu^{(j)}$ is known. Usually, the basis $\{u_\ell\}$ is chosen orthonormal so that $\left\langle u_\ell, u_\mu \right\rangle = \delta_{\ell\mu}$. The other terms yield the products

$$\left\langle \pi_{i,j} \left(\mathbf{v}_{[\ell]}^{(j)} \otimes w \otimes u_\ell \right), \mathbf{b}_\nu^{(j)} \otimes u_\mu \right\rangle = \left\langle \pi_{i,j} \left(\mathbf{v}_{[\ell]}^{(j)} \otimes w \right), \mathbf{b}_\nu^{(j)} \right\rangle \left\langle u_\ell, u_\mu \right\rangle.$$

Using the selfadjointness of $\pi_{i,j}$ and $\mathbf{b}_\nu^{(j)} \in \mathbf{V}_{\mathrm{sym}}^{(j)}$, we obtain

$$\begin{aligned} \left\langle \pi_{i,j} \left(\mathbf{v}_{[\ell]}^{(j)} \otimes w \right), \mathbf{b}_\nu^{(j)} \right\rangle &= \left\langle \mathbf{v}_{[\ell]}^{(j)} \otimes w, \pi_{i,j} \mathbf{b}_\nu^{(j)} \right\rangle = \left\langle \mathbf{v}_{[\ell]}^{(j)} \otimes w, \mathbf{b}_\nu^{(j)} \right\rangle \\ &= \left\langle \mathbf{v}_{[\ell]}^{(j)} \otimes w, \sum_{k=1}^{r} \mathbf{b}_{\nu,[k]}^{(j)} \otimes u_k \right\rangle = \sum_{k=1}^{r} \left\langle \mathbf{v}_{[\ell]}^{(j)}, \mathbf{b}_{\nu,[k]}^{(j)} \right\rangle \langle w, u_k \rangle . \end{aligned}$$

If $\{u_\ell\}$ is an orthogonal basis, $\langle w, u_k \rangle = c_k$ holds (cf. Lemma 4).

Remark 8 Let the bases $\{\mathbf{b}_\nu^{(j)} : 1 \leq \nu \leq r_j\}$ and $\{u_\ell : 1 \leq \ell \leq r\}$ be orthonormal. If $\mathbf{s} := \hat{\mathscr{S}}_{j+1}(\mathbf{v}^{(j)} \otimes w) \in \mathbf{U}_j \otimes U,$ the explicit representation is given by

$$\mathbf{s} = \sum_{\nu=1}^{r_j} \sum_{\mu=1}^{r} c_{\nu\mu} \mathbf{b}_\nu^{(j)} \otimes u_\mu \tag{26}$$

with coefficients $c_{\nu\mu} = \langle \mathbf{s}, \mathbf{b}_\nu^{(j)} \otimes u_\mu \rangle$, which are computable as explained above.

Even if $\mathbf{s} \notin \mathbf{U}_j \otimes U$, the right-hand side in (26) is computable and describes the orthogonal projection $P_{\mathbf{U}_j \otimes U}\mathbf{s}$ of $\mathbf{s}$ onto the space $\mathbf{U}_j \otimes U$.

The check whether $\mathbf{s}$ belongs to $\mathbf{U}_j \otimes U$ is equivalent to the check whether $P_{\mathbf{U}_j \otimes U}\mathbf{s}$ is symmetric (cf. Sect. 5.3), as stated next.

Criterion 1 $\mathbf{s} \in \mathbf{U}_j \otimes U$ *[and therefore also* $\mathbf{s} \in \hat{\mathbf{U}}_{j+1}$, *cf. (17)] holds if and only if* $P_{\mathbf{U}_j \otimes U}\mathbf{s} = \mathbf{s} \in \mathbf{V}_{\mathrm{sym}}^{(j+1)}$ *(implying* $P_{\mathbf{U}_j \otimes U}\mathbf{s} \in \hat{\mathbf{U}}_{j+1}$ *in the positive case).*

Proof

(a) Abbreviate $P_{\mathbf{U}_j \otimes U}$ by P. Let $\mathbf{s} \in \mathbf{U}_j \otimes U$. This implies $P\mathbf{s} = \mathbf{s}$. Since, by construction, $\mathbf{s}$ is symmetric, $P\mathbf{s} \in \mathbf{V}_{\mathrm{sym}}^{(j+1)}$ holds.

(b) Assume $P\mathbf{s} \in \mathbf{V}_{\mathrm{sym}}^{(j+1)}$. Because of $\mathbf{s} = P\mathbf{s} + (I - P)\mathbf{s}$, also $\mathbf{s}^\perp := (I - P)\mathbf{s} \in (\mathbf{U}_j \otimes U)^\perp$ is symmetric. The properties of projections show

$$\langle \mathbf{s}^\perp, \mathbf{s}^\perp \rangle = \langle (I - P)\mathbf{s}, (I - P)\mathbf{s} \rangle = \langle \mathbf{s}, (I - P)\mathbf{s} \rangle = \langle \mathscr{S}_{j+1}(\mathbf{v}^{(j)} \otimes w), \mathbf{s}^\perp \rangle.$$

Since $\mathscr{S}_{j+1}$ is selfadjoint and $\mathbf{s}^\perp$ is symmetric, we have

$$\langle \mathbf{s}^\perp, \mathbf{s}^\perp \rangle = \langle \mathbf{v}^{(j)} \otimes w, \mathscr{S}_{j+1}\mathbf{s}^\perp \rangle = \langle \mathbf{v}^{(j)} \otimes w, \mathbf{s}^\perp \rangle = 0$$

because of $\mathbf{v}^{(j)} \otimes w \in \mathbf{U}_j \otimes U$ and $\mathbf{s}^{\perp} \in \left(\mathbf{U}_j \otimes U\right)^{\perp}$. This proves $\mathbf{s}^{\perp} = 0$ and $\mathbf{s} = P_{\mathbf{U}_j \otimes U}\mathbf{s}$, i.e., $\mathbf{s} \in \mathbf{U}_j \otimes U$. □

5.4.4 Geometric Characterisation

Let $\{\mathbf{b}_{\nu}^{(j)} : 1 \le \nu \le r_j\}$ and $\{u_{\mu} : 1 \le \mu \le r\}$ be orthonormal bases of $\mathbf{U}_j \subset \mathbf{U}_{\mathrm{sym}}^{(j)}$ and U, respectively. $\mathbf{b}_{\nu,[\ell]}^{(j)}$ are the expansion terms: $\mathbf{b}_{\nu}^{(j)} = \sum_{\ell} \mathbf{b}_{\nu,[\ell]}^{(j)} \otimes u_{\ell}$. They give rise to the scalar products

$$B_{(\nu,\mu),(\nu',\mu')} := \left\langle \mathbf{b}_{\nu,[\mu']}^{(j)}, \mathbf{b}_{\nu',[\mu]}^{(j)} \right\rangle \qquad (1 \le \nu, \nu' \le r_j, 1 \le \mu, \mu' \le r).$$

Let $B \in \mathbb{K}^{I \times I}$ be the corresponding matrix, where $I = \{1, \ldots, r_j\} \times \{1, \ldots, r\}$. The orthonormality of $\{\mathbf{b}_{\nu}^{(j)}\}$ is equivalent to $\sum_{\ell} B_{(\nu,\ell),(\nu',\ell)} = \delta_{\nu,\nu'}$. Note that $B = B^{\mathrm{H}}$.

Consider the tensor $\mathbf{v} = \sum_{\nu=1}^{r_j} \sum_{\mu=1}^{r} c_{\nu\mu} \mathbf{b}_{\nu}^{(j)} \otimes u_{\mu}$. The normalisation $\|\mathbf{v}\| = 1$ gives $\sum_{\nu,\mu} |c_{\nu\mu}|^2 = 1$. The entries $c_{\nu\mu}$ define the vector $c \in \mathbb{K}^I$.

Theorem 3 *The spectrum of B is bounded by* 1. *The above defined tensor* $\mathbf{v}$ *is symmetric if and only if c is an eigenvector of B corresponding to the eigenvalue* 1.

Proof Let $\mathbf{s} = \mathscr{S}_{j+1}\mathbf{v}$. The projection property of $\mathscr{S}_{j+1}$ implies that $\langle \mathbf{v}, \mathbf{s} \rangle \le 1$. Criterion 1 states that $\mathbf{v}$ is symmetric (i.e., $\mathbf{v} = \mathbf{s}$) if and only if $\langle \mathbf{v}, \mathbf{s} \rangle = 1$. Calculating the scalar product according to Sect. 5.4.3 yields $(j+1)\langle \mathbf{v}, \mathbf{s} \rangle = 1 + j\,(Bc, c)$, where $(\cdot, \cdot)$ is the Euclidean product of $\mathbb{K}^I$. The inequality $\langle \mathbf{v}, \mathbf{s} \rangle \le 1$ shows that all eigenvalues of B are bounded by 1. The equality $\langle \mathbf{v}, \mathbf{s} \rangle = 1$ requires that $(Bc, c) = 1 = \max\{(Bc', c') : \|c'\| = 1\}$, i.e., c is the eigenvector with eigenvalue $\lambda = 1$. □

The questions from above take now the following form: (a) How can we ensure that 1 belongs to the spectrum of B, (b) what is the dimension of the corresponding eigenspace?

5.4.5 Examples

The following examples use tensors of order $d = 3$. The case $d = 2$ is too easy since tensors of $\otimes^2 U$ correspond to matrices via $\mathbf{v} = \sum_{\nu,\mu=1}^{r} c_{\nu\mu} u_{\nu} \otimes u_{\mu} \mapsto C := (c_{\nu\mu})_{\nu,\mu=1}^{r}$. Hence symmetric tensors $\mathbf{v}$ are characterised by symmetric matrices C.

In the following examples $u_1 = a, u_2 = b \in V$ are orthonormal vectors. A possible choice is $V = \mathbb{K}^2$.

Example 2 We want to represent the symmetric tensor $\mathbf{s} := a \otimes a \otimes a$. We use $U = \mathrm{span}\{a, b\}$ and the symmetric subspace $\mathbf{U}_2 := \mathrm{span}\{\mathbf{b}_1^{(2)}\} \subset \mathscr{S}(U \otimes U) \subset U \otimes U$ with $\mathbf{b}_1^{(2)} := a \otimes a$. Symmetrisation of $\mathbf{U}_2 \otimes U = \mathrm{span}\{a \otimes a \otimes a, a \otimes a \otimes b\}$

yields $\mathscr{S}(\mathbf{U}_2 \otimes U) = \operatorname{span}\{a \otimes a \otimes a, \frac{1}{3}(a \otimes a \otimes b + a \otimes b \otimes a + b \otimes a \otimes a)\}$. Obviously, $\mathscr{S}(\mathbf{U}_2 \otimes U)$ is not a subspace of $\mathbf{U}_2 \otimes U$.

The reason for $\mathscr{S}(\mathbf{U}_2 \otimes U) \not\subset \mathbf{U}_2 \otimes U$ in Example 2 may be seen in the choice of $U = \operatorname{span}\{a, b\}$. This space is larger than necessary: $U = U_1^{\min}(\mathbf{s}) = \operatorname{span}\{a\}$ is sufficient and this choice leads to $\mathscr{S}(\mathbf{U}_2 \otimes U) = \mathbf{U}_2 \otimes U$.

In the next example, U is chosen as $U_1^{\min}(\mathbf{s})$.

Example 3 We want to represent the symmetric tensor $\mathbf{s} := a \otimes a \otimes a + b \otimes b \otimes b$. We use $U = \operatorname{span}\{a, b\}$ and the symmetric subspace $\mathbf{U}_2 := \operatorname{span}\{\mathbf{b}_1^{(2)}, \mathbf{b}_2^{(2)}\} \subset \mathscr{S}(U \otimes U) \subset U \otimes U$ with $\mathbf{b}_1^{(2)} := a \otimes a$ and $\mathbf{b}_2^{(2)} := b \otimes b$. The tensor space $\mathbf{U}_2 \otimes U$ is spanned by $a \otimes a \otimes a,\ b \otimes b \otimes b,\ a \otimes a \otimes b,\ b \otimes b \otimes a$. The first two tensors are already symmetric. The symmetrisation of $a \otimes a \otimes b$ leads to a tensor which is not contained in $\mathbf{U}_2 \otimes U$. The same holds for the last tensor. Hence, $\mathscr{S}(\mathbf{U}_2 \otimes U) \not\subset \mathbf{U}_2 \otimes U$.

In Examples 2 and 3, we can omit the tensors $\mathbf{b}_i^{(2)} \otimes u_j$ whose symmetrisation does not belong to $\mathbf{U}_2 \otimes U$, and still obtain a subspace containing the tensor $\mathbf{s}$ to be represented. The latter statement is not true in the third example.

Example 4 We want to represent the symmetric tensor $\mathbf{s} := \otimes^3(a + b) + \otimes^3(a - b)$. We use $U = \operatorname{span}\{a, b\}$ and the symmetric subspace $\mathbf{U}_2 := \operatorname{span}\{\mathbf{b}_1^{(2)}, \mathbf{b}_2^{(2)}\} \subset \mathscr{S}(U \otimes U) \subset U \otimes U$ with $\mathbf{b}_1^{(2)} := \otimes^2(a + b)$ and $\mathbf{b}_2^{(2)} := \otimes^2(a - b)$. The tensor space $\mathbf{U}_2 \otimes U$ is spanned by four tensors $\mathbf{b}_i^{(2)} \otimes u_j$. For $i = j = 1$, we have $\mathbf{b}_1^{(2)} \otimes a = (a + b) \otimes (a + b) \otimes a$, whose symmetrisation does not belong to $\mathbf{U}_2 \otimes U$. The same holds for the other three tensors. Hence, $\mathscr{S}(\mathbf{U}_2 \otimes U) \not\subset \mathbf{U}_2 \otimes U$.

Note that the setting of Example 4 coincides with Example 3 when we replace the orthonormal basis $\{u_1 = a, u_2 = b\}$ with $\{u_1 = (a+b)/\sqrt{2}, u_2 = (a-b)/\sqrt{2}\}$.

The next example underlines the important role of condition $\mathbf{U}_j^{\min}(\mathbf{v}) \subset \mathbf{U}_j$.

Example 5 Let $\mathbf{U}_2 := \operatorname{span}\{\mathbf{b}_1^{(2)}\}$ with $\mathbf{b}_1^{(2)} := a \otimes b + b \otimes a$. A general tensor in $\mathbf{U}_2 \otimes U$ has the form $\mathbf{b}_1^{(2)} \otimes (\alpha a + \beta b)$. There is no symmetric tensor of this form, except the zero tensor ($\alpha + \beta = 0$). This shows that $\mathbf{U}_2$ is too small: there is no nontrivial symmetric tensor $\mathbf{v}$ with $\mathbf{U}_2^{\min}(\mathbf{v}) \subset \mathbf{U}_2$.

5.4.6 Open Questions About $\mathscr{S}(\mathbf{U}_j \otimes U) \cap \left(\mathbf{U}_j \otimes U\right)$

We repeat the definition $\hat{\mathbf{U}}_{j+1} := \mathscr{S}(\mathbf{U}_j \otimes U) \cap \left(\mathbf{U}_j \otimes U\right)$. The main questions are:

- What is the dimension of $\hat{\mathbf{U}}_{j+1}$, in particular, compared with $\dim(\mathbf{U}_j^{\min}(\mathbf{v}))$, if $\mathbf{v}$ is the tensor to be represented?
- Is there a constructive description of $\hat{\mathbf{U}}_{j+1}$?

The minimal set, which is needed for the construction of the tensors in $\hat{\mathbf{U}}_{j+1}$, is

$$\check{\mathbf{U}}_j := \sum\nolimits_{\mathbf{v}\in\hat{\mathbf{U}}_{j+1}} \mathbf{U}_j^{\min}(\mathbf{v}).$$

By definition, $\check{\mathbf{U}}_j \subset \mathbf{U}_j$ holds, but it is not obvious whether $\check{\mathbf{U}}_j = \hat{\mathbf{U}}_j$. This yields the next question:

- Does $\check{\mathbf{U}}_j = \mathbf{U}_j$ hold?

In the negative case, there is a direct sum $\mathbf{U}_j = \check{\mathbf{U}}_j \oplus \mathbf{Z}_j$, where $\mathbf{Z}_j \neq \{0\}$ contains symmetric tensors in $\mathbf{V}_{\mathrm{sym}}^{(j)}$ which cannot be continued to symmetric tensors in $\mathbf{V}_{\mathrm{sym}}^{(j+1)}$. Using $\mathbf{U}_j$ instead of $\check{\mathbf{U}}_j$ would be inefficient.

5.4.7 Answers for $d = 3$ and $r = 2$

The questions from above can be answered for the simple case of $d = 3$ (transfer from $d = 2$ to $d = 3$) and $r = 2$. Hence we have

$$\dim(U) = 2, \quad \mathbf{U}_1 = U, \quad \mathbf{U}_2 \subset \mathbf{U}_{\mathrm{sym}}^{(2)}$$

and have to investigate the space $\hat{\mathbf{U}}_3 := \mathscr{S}_3(\mathbf{U}_2 \otimes U) \cap (\mathbf{U}_2 \otimes U)$. We recall that $\check{\mathbf{U}}_2 = \sum_{\mathbf{w}\in\hat{\mathbf{U}}_3} \mathbf{U}_2^{\min}(\mathbf{w}) \subset \mathbf{U}_2$ is the smallest subspace of $\mathbf{U}_2$ with the property $\mathscr{S}_3(\check{\mathbf{U}}_2 \otimes U) \cap \left(\check{\mathbf{U}}_2 \otimes U\right) = \hat{\mathbf{U}}_3$. Hence, if $\dim(\mathbf{U}_2) > \dim(\check{\mathbf{U}}_2)$, $\mathbf{U}_2$ contains tensor which are useless for the construction of symmetric tensor in $\hat{\mathbf{U}}_3$.

The symmetric tensors $\mathbf{v} \in \mathbf{U}_2$ correspond to symmetric 2 × 2 matrices. Since $\dim(\mathbf{U}_{\mathrm{sym}}^{(2)}) = 3$, the following list of cases is complete. The general assumption of the following theorems is $\dim(U) = 2$.

Theorem 4 (Case $\dim(\mathbf{U}_2) = 1$**)** *Let* $\dim(\mathbf{U}_2) = 1$ *and* $\mathbf{U}_2 = \mathrm{span}\{\mathbf{b}_1\} \subset \mathbf{U}_{\mathrm{sym}}^{(2)}$. *If* $\mathrm{rank}(\mathbf{b}_1) = 1$ *then*

$$\check{\mathbf{U}}_2 = \mathbf{U}_2, \quad \dim(\hat{\mathbf{U}}_3) = 1;$$

otherwise we have $\mathrm{rank}(\mathbf{b}_1) = 2$ *and*

$$\check{\mathbf{U}}_2 = \{0\} \subset \mathbf{U}_2, \quad \hat{\mathbf{U}}_3 = \{0\}.$$

Proof Note that $\mathrm{rank}(\mathbf{b}_1) \leq \dim(U) = 2$. $\mathrm{rank}(\mathbf{b}_1) = 0$ is excluded because of $\mathbf{b}_1 = 0$ and the assumption that $\mathbf{U}_2 = \mathrm{span}\{\mathbf{b}_1\}$ is one-dimensional. Hence, $\mathrm{rank}(\mathbf{b}_1)$ only takes the values 1 and 2.

If $\operatorname{rank}(\mathbf{b}_1) = 1$, $\mathbf{b}_1 = a \otimes a'$ follows. Symmetry shows that $\mathbf{b}_1 = a \otimes a$ (possibly after changing the sign[5]). Then

$$\hat{\mathbf{U}}_3 = \operatorname{span}\{\mathbf{b}_1 \otimes a\} = \operatorname{span}\{a \otimes a \otimes a\}.$$

If $\operatorname{rank}(\mathbf{b}_1) = 2$, the general form of $\mathbf{w} \in \mathbf{U}_2 \otimes U$ is $\mathbf{w} = \mathbf{b}_1 \otimes (\xi a + \eta b)$. Assume $(\xi, \eta) \neq 0$. Then the only symmetric tensor of this form is $\mathbf{w} = \otimes^3(\xi a + \eta b)$, i.e., $\mathbf{b}_1 = (\xi a + \eta b) \otimes (\xi a + \eta b)$. The contradiction follows from $\operatorname{rank}(\mathbf{b}_1) = 1$. Hence $\xi = \eta = 0$ leads to the assertion.

The statements about $\check{\mathbf{U}}_2$ follow from the definition $\check{\mathbf{U}}_2 = \mathbf{U}_2^{\min}(\hat{\mathbf{U}}_3)$. □

Theorem 5 ($\dim(\mathbf{U}_2) = 2$) *Let* $\dim(\mathbf{U}_2) = 2$. *Then*

$$\check{\mathbf{U}}_2 = \mathbf{U}_2, \quad \dim(\hat{\mathbf{U}}_3) = 2.$$

The precise characterisation of $\hat{\mathbf{U}}_3 \subset \mathbf{U}_{\mathrm{sym}}^{(3)}$ *is given in the proof.*

Proof

(i) There are two linearly independent and symmetric tensors $\mathbf{b}_1, \mathbf{b}_2$ with $\mathbf{U}_2 = \operatorname{span}\{\mathbf{b}_1, \mathbf{b}_2\}$. Fixing linearly independent vectors $a, b \in U = \operatorname{span}\{a, b\}$, the tensors have the form

$$\begin{aligned}\mathbf{b}_1 &= \alpha\, a \otimes a + \beta\, b \otimes b + \gamma\ (a \otimes b + b \otimes a)\,,\\ \mathbf{b}_2 &= \alpha' a \otimes a + \beta' b \otimes b + \gamma' (a \otimes b + b \otimes a)\end{aligned}$$

In part (vi) we shall prove that $\dim(\hat{\mathbf{U}}_3) \leq 2$. The discussion of the cases 1–3 will show that $\dim(\hat{\mathbf{U}}_3) \geq 2$, so that $\dim(\hat{\mathbf{U}}_3) = 2$ follows.

(ii) Case 1: $\gamma = \gamma' = 0$. One concludes that $\mathbf{U}_2 = \operatorname{span}\{a \otimes a, b \otimes b\}$. Then the first case in Theorem 4 shows that $a \otimes a \otimes a$ and $b \otimes b \otimes b$ belong to $\hat{\mathbf{U}}_3$ so that $\dim(\hat{\mathbf{U}}_3) \geq 2$ and part (vi) prove

$$\hat{\mathbf{U}}_3 = \operatorname{span}\{a \otimes a \otimes a, b \otimes b \otimes b\}. \tag{27}$$

(iii) Case 2: $(\gamma, \gamma') \neq 0$. W.l.o.g. assume $\gamma \neq 0$. We introduce the matrices

$$M_\alpha := \begin{bmatrix} \alpha & \alpha' \\ \gamma & \gamma' \end{bmatrix}, \qquad M_\beta := \begin{bmatrix} \gamma & \gamma' \\ \beta & \beta' \end{bmatrix}.$$

[5] If $\mathbb{K} = \mathbb{C}$, the representation $\mathbf{b}_1 = a \otimes a$ holds in the strict sense, If $\mathbb{K} = \mathbb{R}$, either $\mathbf{b}_1 = a \otimes a$ or $\mathbf{b}_1 = -a \otimes a$ can be obtained. Since the purpose of $\mathbf{b}_1$ is to span the subspace, we may w.l.o.g. replace $\mathbf{b}_1 = -a \otimes a$ by $\mathbf{b}_1 = a \otimes a$.

Since $\gamma \neq 0$, both matrices have a rank ≥ 1. If $\operatorname{rank}(M_\alpha) = \operatorname{rank}(M_\beta) = 1$, (α, β, γ) and $(\alpha', \beta', \gamma')$ would be linearly dependent in contradiction to the linear independence of $\{\mathbf{b}_1, \mathbf{b}_2\}$. Hence, at least one matrix has rank 2 and is regular. W.l.o.g. we assume that $\operatorname{rank}(M_\alpha) = 2$ (otherwise interchange the roles of a and b).

(iv) For any $(A, B) \in \mathbb{K}^2$ the system

$$M_\alpha \begin{bmatrix} \xi \\ \eta \end{bmatrix} = M_\beta \begin{bmatrix} A \\ B \end{bmatrix} \tag{28}$$

can be solved for $(\xi, \eta) \in \mathbb{K}^2$. Then the tensor

$$\mathbf{w} := (A\mathbf{b}_1 + B\mathbf{b}_2) \otimes a + (\xi \mathbf{b}_1 + \eta \mathbf{b}_2) \otimes b$$

is symmetric and belongs to $\hat{\mathbf{U}}_3$. For a proof apply Lemma 3: $\mathbf{w} \in \mathbf{U}_{\mathrm{sym}}^{(3)}$ is equivalent to $\varphi_b\,(A\mathbf{b}_1 + B\mathbf{b}_2) = \varphi_a\,(\xi\mathbf{b}_1 + \eta\mathbf{b}_2)$, where the functionals defined by $\varphi_a(a) = \varphi_b(b) = 1$, $\varphi_a(b) = \varphi_b(a) = 0$ apply to the last argument. The latter equation is equivalent to (28).

Let (ξ, η) be the solution of (28) for $(A, B) = (1, 0)$, while (ξ', η') is the solution for $(A, B) = (0, 1)$. Hence we have found a two-dimensional subspace

$$\operatorname{span}\{\mathbf{b}_1 \otimes a + (\xi\mathbf{b}_1 + \eta\mathbf{b}_2) \otimes b,\ \mathbf{b}_2 \otimes a + \left(\xi'\mathbf{b}_1 + \eta'\mathbf{b}_2\right) \otimes b\} \subset \hat{\mathbf{U}}_3. \tag{29}$$

(v) In both cases (27) and (29) the minimal subspace $\check{\mathbf{U}}_2 = \mathbf{U}_2^{\min}(\hat{\mathbf{U}}_3)$ coincides with $\mathbf{U}_2$.

(vi) For an indirect proof of $\dim(\hat{\mathbf{U}}_3) \leq 2$ assume $\dim(\hat{\mathbf{U}}_3) \geq 3$. Let $\varphi_a : \hat{\mathbf{U}}_3 \to \mathbf{U}_2^{\min}(\hat{\mathbf{U}}_3) = \mathbf{U}_2$ be the mapping $\mathbf{w} = \mathbf{v}_1 \otimes a + \mathbf{v}_2 \otimes b \mapsto \mathbf{v}_1$. Since $\dim(\hat{\mathbf{U}}_3) > \dim(\mathbf{U}_2)$, there is some $\mathbf{w} \in \hat{\mathbf{U}}_3, \mathbf{w} \neq 0$ with $\varphi_a(\mathbf{w}) = 0$. This implies $\mathbf{w} = \mathbf{v}_2 \otimes b$ and therefore, by symmetry, $\mathbf{w} = b \otimes b \otimes b$ up to a nonzero factor. Similarly, there are an analogously defined functional φ_b and $\mathbf{w} \in \hat{\mathbf{U}}_3, \mathbf{w} \neq 0$ with $\varphi_b(\mathbf{w}) = 0$ proving $a \otimes a \otimes a \in \hat{\mathbf{U}}_3$. From $a \otimes a \otimes a, b \otimes b \otimes b \in \hat{\mathbf{U}}_3$ we conclude that $\check{\mathbf{U}}_2 := \mathbf{U}_2^{\min}(\hat{\mathbf{U}}_3) \supset \operatorname{span}\{a \otimes a, b \otimes b\}$. Then $\check{\mathbf{U}}_2 \subset \mathbf{U}_2$ and $\dim(\mathbf{U}_2) = 2$ prove $\dim(\hat{\mathbf{U}}_3) \leq 2$.

□

Theorem 6 ($\dim(\mathbf{U}_2) = 3$) *If* $\dim(\mathbf{U}_2) = 3$, $\mathbf{U}_2$ *coincides with space* $\mathbf{U}_{\mathrm{sym}}^{(2)}$ *of all symmetric tensors in* $U \otimes U$ *and generates all tensors in* $\mathbf{U}_{\mathrm{sym}}^{(3)}$*:*

$$\check{\mathbf{U}}_2 = \mathbf{U}_2 = \mathbf{U}_{\mathrm{sym}}^{(2)}, \quad \hat{\mathbf{U}}_3 = \mathbf{U}_{\mathrm{sym}}^{(3)} \ \text{with}\ \dim(\hat{\mathbf{U}}_3) = 4.$$

Proof The statements follow from $\dim(\mathbf{U}_{\mathrm{sym}}^{(2)}) = 3$. □

5.5 *Direct Use of* $\mathbf{U}_j^{\min}(\mathbf{v})$

Statement (19) emphasises the important role of the minimal subspace $\mathbf{U}_j^{\min}(\mathbf{v})$.

5.5.1 Case of Known $\mathbf{U}_j^{\min}(\mathbf{v})$

If the minimal subspaces $\mathbf{U}_j^{\min}(\mathbf{v})$ of a symmetric tensor $\mathbf{v} \in \mathbf{U}_{\text{sym}}^{(d)}$ are given, the above problems disappear. In this case we may define $\mathbf{U}_j := \mathbf{U}_j^{\min}(\mathbf{v})$. This ensures that

$$\check{\mathbf{U}}_j = \mathbf{U}_j \quad \text{and} \quad \hat{\mathbf{U}}_{j+1} \supset \mathbf{U}_{j+1}^{\min}(\mathbf{v})$$

(cf. Theorem 19).

If we want to be able to represent all tensors of a subspace $\mathbf{V}_0 = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbf{U}_{\text{sym}}^{(d)}$, we may use

$$\mathbf{U}_j := \mathbf{U}_j^{\min}(\mathbf{V}_0) = \sum_{\nu=1}^{k} \mathbf{U}_j^{\min}(\mathbf{v}_\nu).$$

Lemma 6 shows that $\mathbf{U}_j$ satisfies (32).

Next we consider the case that $\mathbf{U}_j^{\min}(\mathbf{v})$ is not given explicitly, but can be determined by symmetrisation.

5.5.2 Case of $\mathbf{v} = \mathscr{S}(\mathbf{w})$

As in Sect. 2.1 we assume that the symmetric tensor $0 \neq \mathbf{v} \in \mathbf{U}_{\text{sym}}^{(d)}$ is the symmetrisation $\mathscr{S}(\mathbf{w})$ of a known tensor $\mathbf{w} \in \otimes^d V$. Unlike in Sect. 2.1, we assume that $\mathbf{w}$ is given in the TT format with minimal subspaces[6] $\mathbf{U}_j^{\min}(\mathbf{w})$. The obvious task is to transfer $\mathbf{U}_j^{\min}(\mathbf{w})$ into $\mathbf{U}_j^{\min}(\mathbf{v}) = \mathbf{U}_j^{\min}(\mathscr{S}(\mathbf{w}))$.

We solve this problem by induction on $d = 1, \ldots$. The proof also defines an recursive algorithm.

For $d = 1$ nothing is to be done since $\mathbf{v} = \mathscr{S}(\mathbf{w}) = \mathbf{w}$. Formally, $\mathbf{U}_0^{\min}(\mathbf{v})$ is the field $\mathbb{K}$ and $U_1 \subset \mathbb{K} \otimes U$ corresponds to (14).

The essential part of the proof and of the algorithm is the step from $d - 1$ to d.

Lemma 5 *Let $\mathscr{S}_{[j]} : \otimes^d V \to \otimes^d V$ $(1 \leq j \leq d)$ be the symmetrisation operator $\mathscr{S}_{d-1}$ in (8) applied to the directions $D \backslash \{j\}$ (cf. (2)). Using the transpositions π_{1j},*

[6] In the case of hierarchical tensor representations it is easy to reduce the subspaces to the minimal ones by introducing the HOSVD bases (cf. Hackbusch [10, §11.3.3]).

π_{j1}, the explicit definition is

$$\mathscr{S}_{[j]} = \pi_{1j} \; (id \otimes \mathscr{S}_{d-1}) \; \pi_{j1} \, .$$

Then the symmetrisation operator $\mathscr{S}_d$ is equal to

$$\mathscr{S}_d = \frac{1}{d} \sum_{j=1}^{d} \mathscr{S}_{[j]} \, \pi_{dj} \, .$$

This lemma proves the next result.

Conclusion 3 *Let $\mathbf{w} \in \otimes^d U$ and $\varphi \in U'$. Then*

$$\varphi^{(d)}(\mathscr{S}_d \mathbf{w}) = \frac{1}{d} \mathscr{S}_{d-1} \sum_{j=1}^{d} \varphi^{(j)}(\mathbf{w})$$

holds with $\varphi^{(j)}$ defined in (6).

The $\{1, \ldots, d-1\}$-plex rank r_{d-1} of $\mathbf{x} \in \otimes^d U$ introduced by Hitchcock [12] is the smallest r_{d-1} with $\mathbf{x} = \sum_{\nu=1}^{r_{d-1}} \mathbf{x}_\nu \otimes y_\nu$ ($\mathbf{x}_\nu \in \otimes^{d-1} U$, $y_\nu \in U$). For instance, this representation is the result of the HOSVD representation (cf. [10, §11.3.3]). The minimal subspace $\mathbf{U}_{d-1}^{\min}(\mathbf{x})$ is the span of $\{\mathbf{x}_\nu : 1 \le \nu \le r_{d-1}\}$.

Alternatively, choose the standard basis $\{u_\nu : 1 \le \nu \le r\}$ of U and the representation $\mathbf{x} = \sum_{\nu=1}^{r} \mathbf{z}_\nu \otimes u_\nu$ together with the dual basis $\{\varphi_\nu\}$ of $\{u_\nu\}$. Then the tensors $\mathbf{z}_\nu = \varphi_\nu(\mathbf{x})$ may be linearly dependent so that some of them may be omitted. Let $\mathbf{x}_\nu$ ($1 \le \nu \le r_{d-1}$) be the remaining ones: $\mathbf{U}_{d-1}^{\min}(\mathbf{x}) = \operatorname{span}\{\mathbf{x}_\nu : 1 \le \nu \le r_{d-1}\}$.

Remark 4 states that

$$\mathbf{U}_{d-2}^{\min}(\mathbf{x}) = \mathbf{U}_{d-2}^{\min}(\mathbf{U}_{d-1}^{\min}(\mathbf{x})) = \sum_{\nu=1}^{r_{d-1}} \mathbf{U}_{d-2}^{\min}(\mathbf{x}_\nu). \tag{30}$$

This allows us to determine the minimal subspaces recursively.

Let the TT representation of $\mathbf{w} \in \otimes^d U$ be given. The TT format is also called the matrix product representation since $\mathbf{w} \in \otimes^d U$ can be written as

$$\mathbf{w} = \sum_{k_1, k_2, \ldots, k_{d-1}} v_{1,k_1}^{(1)} \otimes v_{k_1,k_2}^{(2)} \otimes v_{k_2,k_3}^{(3)} \otimes \cdots \otimes v_{k_{d-1},1}^{(d)} \, ,$$

where the vectors $v_{k_j,k_{j+1}}^{(j)} \in U$ are data available from the TT representation. k_j varies in an index set I_j with $\#I_j = \dim(\mathbf{U}_j^{\min}(\mathbf{w}))$. The tensor $\varphi^{(j)}(\mathbf{w}) \in \otimes^{d-1} U$ takes the matrix product form

$$\varphi^{(j)}(\mathbf{w}) = \sum_{k_{j-1},k_j} \varphi(v_{k_{j-1},k_j}^{(j)}) \sum_{\substack{k_1,\ldots,k_{j-2},\\ k_{j+1},\ldots,k_{d-1}}} v_{1,k_1}^{(1)} \otimes \cdots \otimes v_{k_{j-2},k_{j-1}}^{(j-1)} \otimes v_{k_j,k_{j+1}}^{(j+1)} \otimes \cdots \otimes v_{k_{d-1},1}^{(d)} .$$

These tensors can be added within the TT format: $\mathbf{w}_\nu := \sum_{=1}^{d} \varphi_\nu^{(j)}(\mathbf{w}) \in \otimes^{d-1} U$. We conclude that

$$\mathbf{U}_{d-1}^{\min}(\mathbf{v}) = \mathbf{U}_{d-1}^{\min}(\mathscr{S}_d \mathbf{w}) = \operatorname{span}\{\mathscr{S}_{d-1}\mathbf{w}_\nu : 1 \leq \nu \leq r_{d-1}\}.$$

According to (30) the next minimal subspace $\mathbf{U}_{d-2}^{\min}(\mathbf{v})$ can be written as $\sum_\nu \mathbf{U}_{d-2}^{\min}(\mathscr{S}_{d-1}\mathbf{w}_\nu)$ so that we can apply the inductive hypothesis.

6 Operations

The computation of scalar products is already discussed. Next we investigate the tensor addition.

Assume that $\mathbf{v}'$ and $\mathbf{v}''$ are two symmetric tensors represented by subspaces $\mathbf{U}_k'$ and $\mathbf{U}_k''$ and corresponding bases. For $k = 1$, we use the notation $U' := \mathbf{U}_1'$ and $U'' := \mathbf{U}_1''$. By assumption, we have

$$\mathbf{U}_{k+1}^{\min}(\mathbf{v}') \subset \mathbf{U}_{k+1}' \subset \mathscr{S}_{k+1}(\mathbf{U}_k' \otimes U') \cap \left(\mathbf{U}_k' \otimes U'\right), \tag{31a}$$

$$\mathbf{U}_{k+1}^{\min}(\mathbf{v}'') \subset \mathbf{U}_{k+1}'' \subset \mathscr{S}_{k+1}(\mathbf{U}_k'' \otimes U'') \cap \left(\mathbf{U}_k'' \otimes U''\right). \tag{31b}$$

Lemma 6 *The sum* $\mathbf{s} := \mathbf{v}' + \mathbf{v}''$ *can be represented by the subspaces*

$$\mathbf{U}_k := \mathbf{U}_k' + \mathbf{U}_k'', \qquad U := U' + U''.$$

These spaces satisfy again the conditions

$$\mathbf{U}_{k+1}^{\min}(\mathbf{s}) \subset \mathbf{U}_{k+1} \subset \mathscr{S}_{k+1}(\mathbf{U}_k \otimes U) \cap (\mathbf{U}_k \otimes U). \tag{32}$$

Proof The inclusions (31a,b) imply

$$\mathbf{U}_{k+1}^{\min}(\mathbf{v}') + \mathbf{U}_{k+1}^{\min}(\mathbf{v}'') \subset \mathbf{U}_{k+1} = \mathbf{U}_{k+1}' + \mathbf{U}_{k+1}''$$
$$\subset \left(\mathscr{S}_{k+1}(\mathbf{U}_k' \otimes U') \cap \left(\mathbf{U}_k' \otimes U'\right)\right) + \left(\mathscr{S}_{k+1}(\mathbf{U}_k'' \otimes U'') \cap \left(\mathbf{U}_k'' \otimes U''\right)\right).$$

Since $\mathbf{U}_{k+1}^{\min}(\mathbf{s}) \subset \mathbf{U}_{k+1}^{\min}(\mathbf{v}') + \mathbf{U}_{k+1}^{\min}(\mathbf{v}'')$, the first part of (32) follows: $\mathbf{U}_{k+1}^{\min}(\mathbf{s}) \subset \mathbf{U}_{k+1}$. The inclusion $\mathbf{U}_k' \otimes U' \subset \mathbf{U}_k \otimes U$ implies that

$$\mathscr{S}_{k+1}(\mathbf{U}_k' \otimes U') \cap \left(\mathbf{U}_k' \otimes U'\right) \subset \mathscr{S}_{k+1}(\mathbf{U}_k \otimes U) \cap (\mathbf{U}_k \otimes U).$$

The analogous statement for $\mathscr{S}_{k+1}(\mathbf{U}_k'' \otimes U'') \cap \left(\mathbf{U}_k'' \otimes U''\right)$ yields

$$\left(\mathscr{S}_{k+1}(\mathbf{U}_k' \otimes U') \cap \left(\mathbf{U}_k' \otimes U'\right)\right) + \left(\mathscr{S}_{k+1}(\mathbf{U}_k'' \otimes U'') \cap \left(\mathbf{U}_k'' \otimes U''\right)\right)$$
$$\subset \mathscr{S}_{k+1}(\mathbf{U}_k \otimes U) \cap (\mathbf{U}_k \otimes U).$$

Hence also the second part of (32) is proved. □

The computation of the orthonormal basis of $\mathbf{U}_k$ is performed in the order $k = 1, 2, \ldots$. As soon as orthonormal bases of $\mathbf{U}_k$ and U are given, the orthonormal basis of $\mathbf{U}_{k+1}$ can be determined.

Since the spaces $\mathbf{U}_k = \mathbf{U}'_k + \mathbf{U}''_k$ may be larger than necessary, a truncation is advisable.

7 Truncation

The standard truncation procedure uses the SVD. Formally, we have for any tensor $\mathbf{u} \in \mathbf{V}$ and any $k \in \{1, \ldots, d-1\}$ a singular value decomposition

$$\mathbf{u} = \sum_{\nu=1}^{r_u} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu, \tag{33}$$

where $\{\mathbf{v}_\nu : 1 \le \nu \le r_u\} \subset \otimes^k V$ and $\{\mathbf{w}_\nu : 1 \le \nu \le r_u\} \subset \otimes^{d-k} V$ are orthonormal systems and $\{\sigma_1 \ge \sigma_2 \ge \ldots\}$ are the singular values. The usual approach is to choose some $s < r_u$ and to define the tensor

$$\hat{\mathbf{u}} = \sum_{\nu=1}^{s} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$$

which can be represented with subspaces of lower dimension.

In the case of symmetric tensors $\mathbf{u} \in \mathbf{V}_{\mathrm{sym}}^{(d)}$, we have $\mathbf{v}_\nu \in \mathbf{V}_{\mathrm{sym}}^{(k)}$ and $\mathbf{w}_\nu \in \mathbf{V}_{\mathrm{sym}}^{(d-k)}$. However, the standard truncation cannot be used since there is no guarantee that the truncated tensor $\hat{\mathbf{u}} = \sum_{\nu=1}^{s} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$ again belongs to $\mathbf{V}_{\mathrm{sym}}^{(d)}$.

7.1 *Truncation for $k = 1$ and $k = d - 1$*

In the cases $k = 1$ and $k = d - 1$, the truncation can be performed as follows. For $k = 1$, the standard truncation $\mathbf{u} \mapsto \hat{\mathbf{u}}$ can be written as $\hat{\mathbf{u}} = (P \otimes \mathbf{I})\,\mathbf{u}$, where $P : V \to V$ is the orthogonal projection onto the subspace

$$\hat{U} := \operatorname{span}\{\mathbf{v}_\nu : 1 \le \nu \le s\} \subset V,$$

while $\mathbf{I} = \otimes^{d-1} I$ is the identity on $\otimes^{d-1} V$.

In the symmetric case, we need a symmetric mapping. If we delete the components $\{\mathbf{v}_\nu : \nu > s\}$ in the first direction, the same must be done in the other directions. The corresponding mapping is the orthogonal projection

$$\mathbf{P} := \otimes^d P$$

onto the subspace $\hat{\mathbf{U}}_{\mathrm{sym}}^{(d)} \subset \mathbf{V}_{\mathrm{sym}}^{(d)}$. Note that the error $\mathbf{u} - \mathbf{Pu}$ does not only consist of the omitted terms $\sum_{\nu=s+1}^{r_u} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$, but also of $\sum_{\nu=1}^{s} \sigma_\nu \mathbf{v}_\nu \otimes \left(\mathbf{I} - \otimes^{d-1} P\right) \mathbf{w}_\nu$.

In the case of $k = d-1$, $\mathbf{w}_\nu$ belongs to V and the analogous construction can be performed. If $d = 3$, the cases $k = 1$ and $k = d-1$ cover all possible ones.

7.2 *Open Problem for* $1 < k < d-1$

If $d > 3$, there are integers k with $1 < k < d-1$. Then both $\mathbf{v}_\nu$ and $\mathbf{w}_\nu$ in (33) are tensors of order ≥ 2. It is not obvious how the analogue of the previous mapping $\mathbf{P}$ could look like. The advantage of a symmetric projection $\mathbf{P}$ would be the existence of a tensor $\mathbf{v}' = \mathbf{Pv} \in \mathbf{U}_{\mathrm{sym}}$. Define $\mathbf{w}' := \sum_{\nu=1}^{s} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$ and $\mathbf{w}'' := \sum_{\nu=s+1}^{r_u} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$ by SVD. Assume that $\mathbf{Pw}' \neq 0$ while $\mathbf{Pw}'' = 0$. Then $\mathbf{Pv} = \mathbf{P}\mathscr{S}(\mathbf{w}' + \mathbf{w}'') = \mathscr{S}(\mathbf{Pw}' + \mathbf{Pw}'') = \mathscr{S}(\mathbf{Pw}')$ (cf. Lemma 9) does not vanish, i.e., $\mathbf{v}' \neq 0$ ensures the existence of a nontrivial subspaces $\mathbf{U}_j^{\min}(\mathbf{v}')$ $(1 \leq j \leq d)$.

Let $P_k : \mathbf{V}_{\mathrm{sym}}^{(k)} \to \mathbf{V}_{\mathrm{sym}}^{(k)}$ be the orthogonal projection onto $\mathrm{span}\{\mathbf{v}_\nu : 1 \leq \nu \leq s\}$, so that $\mathbf{P}_k := P_k \otimes (\otimes^{d-k} I)$ maps $\mathbf{u}$ to the SVD truncation $\hat{\mathbf{u}} = \sum_{\nu=1}^{s} \sigma_\nu \mathbf{v}_\nu \otimes \mathbf{w}_\nu$. The symmetrisation $\mathbf{P} = \mathscr{S}(\mathbf{P}_k)$ defines $\mathbf{u}' := \mathbf{Pu} \in \mathbf{U}_{\mathrm{sym}}^{(d)}$. Since $\langle \mathbf{u}, \mathbf{u}' \rangle = \langle \mathbf{u}, \mathbf{Pu} \rangle = \langle \mathbf{u}, \mathscr{S}(\mathbf{P}_k)\mathbf{u} \rangle = \langle \mathbf{u}, \mathscr{S}(\mathbf{P}_k\mathbf{u}) \rangle = \langle \mathscr{S}(\mathbf{u}), \mathbf{P}_k\mathbf{u} \rangle = \langle \mathbf{u}, \mathbf{P}_k\mathbf{u} \rangle = \langle \hat{\mathbf{u}}, \hat{\mathbf{u}} \rangle$, the tensor $\mathbf{u}'$ does not vanish. However, it is not obvious that $\dim(\mathbf{U}_k^{\min}(\mathbf{u}')) \leq s$ holds, as intended by the truncation.

A remedy is a follows. Assume that the (not truncated) tensor uses the subspaces $\mathbf{U}_j$ satisfying (17). Let the SVD for index k reduce $\mathbf{U}_k$ to $\mathbf{U}_k'$. Since $\mathbf{U}_k' \subset \mathbf{U}_k$, $\mathbf{U}_k'$ still belongs to $\mathbf{U}_{\mathrm{sym}}^{(k)}$. Moreover

$$\hat{\mathbf{U}}_{k+1}' := \left(\mathbf{U}_k' \otimes U\right) \cap \mathscr{S}\left(\mathbf{U}_k' \otimes U\right) \subset (\mathbf{U}_k \otimes U) \cap \mathscr{S}(\mathbf{U}_k \otimes U) = \hat{\mathbf{U}}_{k+1}$$

guarantees the existence of a subspace $\mathbf{U}_{k+1}' \subset \hat{\mathbf{U}}_{k+1}'$ so that the construction can be continued. However, may it happen that $\mathbf{U}_k'$ is too small and $\hat{\mathbf{U}}_{k+1}' = \{0\}$ holds?

8 Combination with ANOVA

As already mentioned in [10, §17.4.4] the ANOVA[7] representation has favourable properties in connection with symmetric tensors. The ANOVA technique is briefly described in Sect. 8.1 (see also Bohn–Griebel [5, §4.3]). The ANOVA approximation uses terms with low-dimensional minimal subspaces. Therefore the combination with the TT format is an efficient approach.

[7] ANOVA abbreviates ‘analysis of variance’.

8.1 ANOVA Technique

Let $0 \neq e \in V$ be a special element. In the case of multivariate functions, e may be the constant function $e(x) = 1$. In the case of mappings, e is the identity mapping. Define

$$E := \operatorname{span}\{e\}, \qquad \overset{\circ}{V} := E^{\perp}. \tag{34}$$

The choice of the orthogonal complement $E^{\perp}$ simplifies later computations. Theoretically, it would be sufficient to choose $\overset{\circ}{V}$ such that there is a direct sum

$$V = E \oplus \overset{\circ}{V}.$$

The space $\overset{\circ}{V}$ gives rise to the symmetric tensor space $\overset{\circ}{\mathbf{V}}{}^{(d)}_{\mathrm{sym}}$.

We introduce the following notation of symmetric tensors in $\mathbf{V}^{(d)}_{\mathrm{sym}}$ generated by tensors from $\overset{\circ}{\mathbf{V}}{}^{(k)}_{\mathrm{sym}}$:

$$S(\mathbf{v}; d) := \mathscr{S}_d(\mathbf{v} \otimes (\otimes^{d-k} e)) \qquad \text{for } \mathbf{v} \in \overset{\circ}{\mathbf{V}}{}^{(k)}_{\mathrm{sym}} \text{ with } 1 \leq k \leq d. \tag{35}$$

The tensors in (35) form the following subspaces of $\mathbf{V}^{(d)}_{\mathrm{sym}}$:

$$\overset{\circ}{\mathbf{V}}_0 := \otimes^d E, \qquad \overset{\circ}{\mathbf{V}}_k := \mathscr{S}_d(\overset{\circ}{\mathbf{V}}{}^{(k)}_{\mathrm{sym}} \otimes (\otimes^{d-k} E)) \quad \text{for } 1 \leq k \leq d.$$

Lemma 7 *If (34) holds, there is an orthogonal decomposition* $\mathbf{V}^{(d)}_{\mathrm{sym}} = \bigoplus_{j=0}^{d} \overset{\circ}{\mathbf{V}}_j$.

Proof Let $k > \ell$, $\mathbf{v} \in \overset{\circ}{\mathbf{V}}_k$, $\mathbf{w} \in \overset{\circ}{\mathbf{V}}_\ell$. Tensor $\mathbf{v}$ can be written as a sum of elementary tensors $\mathbf{v}_\nu = \bigotimes_{j=0}^{d} v_\nu^{(j)}$ containing k vectors $v_\nu^{(j)} \in \overset{\circ}{V}$. Correspondingly, $\mathbf{w}$ is a sum of $\mathbf{w}_\mu = \bigotimes_{j=0}^{d} w_\mu^{(j)}$ with $d - \ell$ vectors $w_\mu^{(j)} = e$. Because of $\ell < k$, there must be some j with $v_\nu^{(j)} \in \overset{\circ}{V}$ orthogonal to $w_\mu^{(j)} = e$. Hence $\langle \mathbf{v}_\nu, \mathbf{w}_\mu \rangle = 0$ holds for all pairs implying $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. □

The ANOVA representation of a (symmetric) tensor $\mathbf{v} \in \mathbf{V}^{(d)}_{\mathrm{sym}}$ is the sum

$$\mathbf{v} = \sum_{k=0}^{L} \mathbf{v}_k \qquad \text{with } \mathbf{v}_k \in \overset{\circ}{\mathbf{V}}_k \text{ for some } 0 \leq L \leq d. \tag{36}$$

We call L the ANOVA degree.

Remark 9

(a) The motivation of ANOVA is to obtain an approximation (36) for a relative small degree L.

(b) Let $\mathbf{v}_{\mathrm{ex}} = \sum_{k=0}^{d} \mathbf{v}_k$ be the exact tensor. In order to approximate $\mathbf{v}_{\mathrm{ex}}$ by $\mathbf{v}$ from (36) we have to assume that the terms $\mathbf{v}_k$ are (rapidly) decreasing.

8.2 Representation

We assume that (36) holds with $\mathbf{v}_k = S(\mathbf{x}_k; d)$. The tensors $\mathbf{x}_k \in \mathring{\mathbf{V}}^{(k)}_{\mathrm{sym}}$ are given together with their minimal subspaces $\mathbf{U}_k^{\min}(\mathbf{x}_k)$ as described in Sect. 5.5.

We generalise the notation $S(\cdot; d)$ in (35) to subspaces:

$$S(\mathbf{X}; d) = \operatorname{span}\{S(\mathbf{x}; d) : \mathbf{x} \in \mathbf{X}\} \quad \text{for } \mathbf{X} \subset \mathring{\mathbf{V}}^{(k)}_{\mathrm{sym}}.$$

If $\mathbf{X} = \operatorname{span}\{\mathbf{x}_\nu : 1 \le \nu \le N\}$, we have $S(\mathbf{X}; d) = \operatorname{span}\{S(\mathbf{x}_\nu; d) : 1 \le \nu \le N\}$.

We remark that $\mathbf{U}_d^{\min}(\mathbf{v}) = \operatorname{span}\{\mathbf{v}\}$ for $\mathbf{v} \in \otimes^d V$.

Lemma 8 *Let* $\mathbf{v}_k \in \mathring{\mathbf{V}}_k$. *Then*

$$U_j^{\min}(S(\mathbf{v}_k; d)) = \sum_{\mu=\min\{j,k-j\}}^{\max\{k,d-j\}} S(U_\mu^{\min}(\mathbf{v}_k); j). \tag{37}$$

Proof By definition of $U_j^{\min}(\cdot)$, functionals in $d-j$ directions are to be applied. Let μ be associated with $\mathbf{v}_k$ and $d-j-\mu$ with $\otimes^{d-k} e$. This leads to the inequalities $0 \le \mu \le k$ and $0 \le d-j-\mu \le d-k$. Together they imply $\min\{j, k-j\} \le \mu \le \max\{k, d-j\}$.

Consider the case of $j = d-1$. Applying φ with $\varphi(e) = 1$ and $\varphi(v) = 0$ $(v \in \mathring{V})$, we obtain $\varphi^{(d)}(S(\mathbf{v}_k; d)) = S(\mathbf{v}_k; d-1)$, provided that $k < d$. On the other hand, φ with $\varphi(e) = 0$ yields $\varphi^{(d)}(S(\mathbf{v}_k; d)) = S(\varphi(\mathbf{v}_k); d-1)$. This proves

$$\mathbf{U}_{d-1}^{\min}(S(\mathbf{v}_k; d)) = S(\mathbf{v}_k; d-1) + S(\mathbf{U}_{k-1}^{\min}(\mathbf{v}_k); d-1).$$

Since $\operatorname{span}\{S(\mathbf{v}_k; d-1)\} = S(\mathbf{U}_k^{\min}(\mathbf{v}_k); d-1)$, this coincides with (37) for $j = d-1$. For the other j apply Remark 4 recursively. □

The ANOVA tensor is the sum $\mathbf{v} = \sum_k \mathbf{v}_k$. As $\mathbf{U}_j^{\min}(\mathbf{a}+\mathbf{b}) \subset \mathbf{U}_j^{\min}(\mathbf{a}) + \mathbf{U}_j^{\min}(\mathbf{b})$, we obtain from (37) that

$$\mathbf{U}_j^{\min}\left(S\left(\sum\nolimits_k \mathbf{v}_k; d\right)\right) \subset \sum\nolimits_{k,\mu} S(\mathbf{U}_\mu^{\min}(\mathbf{v}_k); j).$$

The dimension of the right-hand side may be larger than $\mathbf{U}_j^{\min}(\mathbf{v})$, but here we want to separate the spaces E and $\mathring{U} \subset \mathring{V}$. For instance, the tensor $\mathbf{v} = (a+e) \otimes (a+e)$ has the one-dimensional minimal subspace $U_1^{\min}(\mathbf{v}) = \operatorname{span}\{a+e\}$, but here we use $E + \mathring{U}$ with $\mathring{U} = \operatorname{span}\{a\}$.

8.3 Example

Consider an expression of the form $\sum_i S(a_i'; d) + S(\mathbf{b}; d)$, where $a_i' \in V$ and $\mathbf{b} \in \mathbf{V}_{\mathrm{sym}}^{(2)}$. We assume that $\mathbf{b}$ can be approximated by $\sum_k \left(b_k' \otimes c_k' + c_k' \otimes b_k'\right)$. Orthogonalisation of a_i', b_k', c_k' with respect to some e yields the vectors a_i, b_k, c_k and the ANOVA form

$$\mathbf{v} = \alpha S(1; d) + \sum_{i=1}^{N_1} S(x_i; d) + \sum_{k=1}^{N_2} S(a_k \otimes b_k + b_k \otimes a_k; d), \tag{38}$$

where x_i represents a_i and multiples of b_k and c_k. Note that $S(1; d) = \otimes^d e$. In the following we give the details for

$$\mathbf{v} = \alpha S(1; d) + S(x; d) + S(a \otimes b + b \otimes a; d).$$

The combined ANOVA-TT format uses the spaces

$$\begin{aligned}
j = 1 &: U_1 = \mathrm{span}\{e, x, a, b\}, \qquad (39)\\
j = 2 &: \mathbf{U}_2 = \mathrm{span}\{S(1; 2), S(x; 2), S(a; 2), S(b; 2), S(a \otimes b + b \otimes a, 2)\},\\
&\vdots\\
j < d &: \mathbf{U}_j = \mathrm{span}\{S(1; j), S(x; j), S(a; j), S(b; j), S(a \otimes b + b \otimes a, j)\},\\
j = d &: \mathbf{U}_d = \mathrm{span}\{\mathbf{v}\}.
\end{aligned}$$

The essential recursive definition (18) of the basis reads as follows:

$$\begin{aligned}
S(1; j) &= S(1; j-1) \otimes e,\\
S(x; j) &= S(x; j-1) \otimes e + S(1; j-1) \otimes x,\\
S(a; j),\ S(b; j) &: \text{ analogously},\\
S(a \otimes b + b \otimes a, j) &= S(a \otimes b + b \otimes a, j-1) \otimes e\\
&\quad + S(a; j-1) \otimes b + S(b; j-1) \otimes a.
\end{aligned}$$

The final step follows from

$$\begin{aligned}
\mathbf{v} =& \alpha S(1; d-1) + S(x; d-1) \otimes e + S(1; d-1) \otimes x\\
&+ S(a \otimes b + b \otimes a, d-1) \otimes e + S(a; d-1) \otimes b + S(b; d-1) \otimes a.
\end{aligned}$$

Remark 10 The terms

$$\alpha S(1;d) \quad \text{and} \quad \sum_{k=1}^{N_2} S(a_k \otimes b_k + b_k \otimes a_k; d)$$

in (38) lead to $3N_2 + 1$ basis vectors in $\mathbf{U}_j$. Let N_0 vectors x_i be linearly independent of $\operatorname{span}\{a_k, b_k : 1 \leq k \leq N_2\}$. Then $\sum_{i=1}^{N_1} S(x_i; d)$ requires N_0 additional basis vectors in $\mathbf{U}_j$.

8.4 Operations

Concerning scalar product one can exploit Lemma 7: $\langle S(\mathbf{v}; d), S(\mathbf{w}; d)\rangle = 0$ for $\mathbf{v} \in \mathring{\mathbf{V}}_k$, $\mathbf{w} \in \mathring{\mathbf{V}}_\ell$ with $k \neq \ell$. If the basis of $\mathbf{U}_j$ should be orthonormalised, it is sufficient to orthonormalise only the contributions $S(\mathbf{v}_\nu; j)$ for all $\mathbf{v}_\nu \in \mathring{\mathbf{V}}_k$ separately (cf. (39)).

One can possibly use that for tensors $\mathbf{v}_\nu, \mathbf{w}_\nu \in \mathring{\mathbf{V}}_k$, $k \leq j$, the scalar product $\langle S(\mathbf{v}; j), S(\mathbf{w}; j)\rangle$ is a fixed multiple of $\langle \mathbf{v}, \mathbf{w}\rangle$, provided that $\langle e, e\rangle = 1$:

$$\langle S(\mathbf{v}; j), S(\mathbf{w}; j)\rangle = \frac{j!}{k!} \langle \mathbf{v}, \mathbf{w}\rangle .$$

In principle, the operations within the TT format are as usual. However, one has to take care that the result is again of the ANOVA form.

As an example we consider the Hadamard product $\odot$ (pointwise product) for multivariate functions. For the standard choice that e is the constant function with value 1, we have $e \odot e = e$ (and $a \odot e = e \odot a = a$ for any a). If $\mathbf{v}$ is of the form (36) with $L = L_v$, while $\mathbf{w}$ corresponds to $L = L_w$, the product $\mathbf{z} := \mathbf{v} \odot \mathbf{w}$ satisfies (36) with degree $L_z = \min\{d, L_v + L_w\}$. Enlarging L increases the storage cost and the arithmetic cost of operations involving $\mathbf{z}$. A truncation $L_z \mapsto L_z' < L_z$ could be helpful, provided that omitted terms are small. Here we need that $\mathbf{z}$ satisfies the assumption of Remark 9b.

Let $Z = L(V)$ be the space of linear maps of V into V. Another example is the multiplication of an operator (Kronecker matrix) $\mathbf{A} = \sum_{k=0}^{L_A} \mathbf{A}_k \in \mathbf{Z}_{\mathrm{sym}}^{(d)}$ and a tensor $\mathbf{v} = \sum_{k=0}^{L_v} \mathbf{v}_k \in \mathbf{V}_{\mathrm{sym}}^{(d)}$. Let the identity be the special element I of Z (replacing e in the general description). This guarantees $Ie = e$. Again $\mathbf{w} := \mathbf{A}\mathbf{v}$ is of the form (36) with $L_w = \min\{d, L_A + L_v\}$. Only under the assumption that all maps in $U_1^{\min}(\mathbf{A})$ possess e as an eigenvector, we have $L_w = L_v$.

9 Operators, Kronecker Matrices

If V is a matrix space, the corresponding tensor space contains Kronecker matrices. More generally, linear operators on multivariate functions can be described by tensors. In this case, there is an operator $\mathbf{A}$ and a vector $\mathbf{v}$ as well as the product $\mathbf{Av}$. For completeness we list the relations between (anti)symmetric operators and (anti)symmetric tensors $\mathbf{v}$. The proofs are obvious and therefore omitted.

The expression $\pi\mathbf{A}$ used below means the application of the permutation π to the tensor $\mathbf{A} \in \otimes^d L(V) \subset L(\mathbf{V})$, where $L(V)$ are the linear maps from V into itself. Concerning $\otimes^d L(V) \subset L(\mathbf{V})$ compare [10, Proposition 3.49].

Lemma 9 *Let* $\mathbf{V} = \otimes^d V$, $\mathbf{A} : \mathbf{V} \to \mathbf{V}$ *a linear map and* π *a permutation. Then the action of* $\pi\mathbf{A}$ *can be expressed by the action of* $\mathbf{A}$*:*

$$(\pi\mathbf{A})\,\mathbf{u} = \pi\left(\mathbf{A}\left(\pi^{-1}\mathbf{u}\right)\right) \qquad \textit{for all } \mathbf{u} \in \mathbf{V}.$$

If $\mathbf{A}$ *is symmetric then* $\pi(\mathbf{Au}) = \mathbf{A}(\pi\mathbf{u})$. *If* $\mathbf{u} \in \mathbf{V}_{\mathrm{sym}}$ *then* $(\mathscr{S}(\mathbf{A}))\mathbf{u} = \mathscr{S}(\mathbf{Au})$. *If* $\mathbf{A} : \mathbf{V} \to \mathbf{V}$ *is symmetric then* $\mathbf{A}\mathscr{S}(\mathbf{u}) = \mathscr{S}(\mathbf{Au})$.

The last statement implies that if $\mathbf{A} : \mathbf{V} \to \mathbf{V}$ is symmetric, then $\mathbf{A} : \mathbf{V}_{\mathrm{sym}} \to \mathbf{V}_{\mathrm{sym}}$ and $\mathbf{A} : \mathbf{V}_{\mathrm{anti}} \to \mathbf{V}_{\mathrm{anti}}$.

The adjoint of $\mathbf{A}$ is denoted by $\mathbf{A}^*$, i.e., $\langle \mathbf{Au}, \mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{A}^*\mathbf{v}\rangle$. Any permutation satisfies $\pi^* = \pi^{-1}$ and $(\pi\mathbf{A})^* = \pi\mathbf{A}^*$. In particular, permutations of selfadjoint operators are again selfadjoint.

References

1. Abo, H., Vannieuwenhoven, N.: Most secant varieties of tangential varieties to Veronese varieties are nondefective. Trans. Am. Math. Soc. **370**, 393–420 (2018)
2. Abo, H., Ottoviani, G., Peterson, C.: Non-defectivity of Grassmannians of planes. J. Algebr. Geom. **21**, 1–20 (2012)
3. Bach, V., Delle, L. (eds.): Many-Electron Approaches in Physics, Chemistry and Mathematics. Springer, Cham (2014)
4. Beylkin, G., Mohlenkamp, M.J., Pérez, F.: Approximating a wavefunction as an unconstrained sum of Slater determinants. J. Math. Phys. **49**, 032107 (2008)
5. Bohn, B., Griebel, M.: An adaptive sparse grid approach for time series prediction. In: Garcke, J., Griebel, M. (eds.): Sparse Grids and Applications, Bonn, May 2011. Lecture Notes in Computational Science and Engineering, vol. 88, pp. 1–30. Springer, Berlin (2013)
6. Buczyński, J., Landsberg, J.M.: On the third secant variety. J. Algebr. Comb. **40**, 475–502 (2014)
7. Cayley, A.: An introductory memoir on quantics. Philos. Trans. R. Soc. Lond. **144**, 245–258 (1854)
8. Comon, P., Golub, G.H., Lim, L.H., Mourrain, B.: Symmetric tensors and symmetric tensor rank. SIAM J. Matrix Anal. Appl. **30**, 1254–1279 (2008)
9. Garcke, J., Griebel, M. (eds.): Sparse Grids and Applications, Bonn, May 2011. Lecture Notes in Computational Science and Engineering, vol. 88. Springer, Berlin (2013)

10. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. SSCM, vol. 42. Springer, Berlin (2012)
11. Håstad, J.: Tensor rank is NP-complete. J. Algorithms **11**, 644–654 (1990)
12. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. J. Math. Phys. **6**, 164–189 (1927)
13. Legeza, O., Rohwedder, T., Schneider, R., Szalay, S.: Tensor product approximation (DMRG) and coupled cluster method in quantum chemistry. In: Bach, V., Delle, L. (eds.): Many-Electron Approaches in Physics, Chemistry and Mathematics, pp. 53–76. Springer, Cham (2014)
14. Lounesto, P.: Clifford Algebras and Spinors, 2nd edn. Cambridge University Press, Cambridge (2001)
15. Löwdin, P.O.: Quantum theory for many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction. Phys. Rev. **97**(6), 1474–1489 (1955)
16. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**, 2295–2317 (2011)
17. Valiant, L.G.: The complexity of computing the permanent. Theor. Comput. Sci. **8**, 189–201 (1979)

Direct and Inverse Results on Bounded Domains for Meshless Methods via Localized Bases on Manifolds

Thomas Hangelbroek, Francis J. Narcowich, Christian Rieger, and Joseph D. Ward

Dedicated to Ian H. Sloan on the occasion of his 80th Birthday

Abstract This article develops direct and inverse estimates for certain finite dimensional spaces arising in kernel approximation. Both the direct and inverse estimates are based on approximation spaces spanned by local Lagrange functions which are spatially highly localized. The construction of such functions is computationally efficient and generalizes the construction given in Hangelbroek et al. (Math Comput, 2017, in press) for restricted surface splines on $\mathbb{R}^d$. The kernels for which the theory applies includes the Sobolev-Matérn kernels for closed, compact, connected, C^∞ Riemannian manifolds.

1 Introduction

This article investigates both direct estimates and inverse inequalities for certain finite dimensional spaces of functions. These spaces are spanned by either Lagrange or local Lagrange functions generated by certain positive definite or conditionally positive definite kernels.

While the topics of direct and inverse theorems for kernel-based approximation spaces have been considered in the boundary-free setting by a number of authors (see [10, 16–18, 20] as a partial list), the results for such theorems on compact

T. Hangelbroek
Department of Mathematics, University of Hawaii – Manoa, Honolulu, HI, USA
e-mail: hangelbr@math.hawaii.edu

F. J. Narcowich · J. D. Ward (✉)
Department of Mathematics, Texas A&M University, College Station, TX, USA
e-mail: fnarc@math.tamu.edu; jward@math.tamu.edu

C. Rieger
Institut für Numerische Simulation, Universität Bonn, Bonn, Germany
e-mail: rieger@ins.uni-bonn.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_24

domains is less well developed. The main results in this article pertain to inverse estimates (Sect. 5.3) and direct theorems (Sect. 6) for certain kernel based approximation spaces on compact domains in a fairly general setting.

The primary focus of this article pertains to certain positive definite kernels defined on a closed, compact, connected, C^∞ Riemannian manifold, which will be denoted by $\mathbb{M}$ throughout the sequel. We restrict to this setting; inverse theorems in the Euclidean space setting were recently given in [15].

Rather than dealing with the standard finite dimensional kernel spaces $S(X) = \operatorname{span}_{\xi\in X} k(\cdot,\xi)$, where $k(\cdot,\cdot)$ is a positive definite kernel and X, the set of centers, is a suitably chosen finite set of points, we will consider subspaces of $S(X)$ that are generated by Lagrange functions $\{\chi_\xi : \xi \in X\}$, which, for certain kernels, are highly localized. These subspaces are designed to deal with problems involving a compact domain $\Omega \subsetneq \mathbb{M}$, where Ω is subject to some mild restrictions discussed in Sect. 5.1.

Specifically, we look at spaces of the form $V_\Xi = \operatorname{span}_{\xi\in\Xi}\chi_\xi$, where χ_ξ is a Lagrange function for X, which is assumed to be suitably dense in a neighborhood of Ω, and Ξ is a subset of X. An important feature, perhaps unusual for RBF and kernel approximation, is that the centers X used to *construct* the Lagrange functions $\{\chi_\xi : \xi \in X\}$ and centers Ξ *defining* the function spaces V_Ξ, do not always coincide, because V_Ξ comprises only the Lagrange functions associated with $\xi \in \Xi$. The spaces V_Ξ differ slightly from $S(X)$ and are important for obtaining inverse estimates over Ω. We will discuss these spaces in Sect. 5 and provide inverse estimates in Theorem 3.

We also consider, in Sect. 4, locally (and efficiently) constructed functions b_ξ, which we call *local* Lagrange functions. These have properties similar to the χ_ξ's and also to those used in [6]. In Theorem 4, we give inverse estimates for $\tilde{V}_\Xi = \operatorname{span}_{\xi\in\Xi} b_\xi$.

1.1 Overview and Outline

In Sect. 2, a basic explanation and background on the manifolds and kernels used in this article will be given.

The direct and inverse theorems in this paper are associated with two approximation spaces V_Ξ and $\tilde{V}_\Xi$. In Sect. 3, we introduce the Lagrange basis (the functions which form a basis for the space V_Ξ) associated with the kernels described in Sect. 2.2. Such Lagrange functions are known to have stationary exponential decay (this notion is introduced in Sect. 3.1). To illustrate the power of these highly localized bases, we finish the section by providing estimates that control the Sobolev norm (i.e. $W_2^\sigma(\mathbb{M})$) of a function in V_Ξ by the l_2 norm on the Lagrange coefficients. That is, for $s = \sum_{\xi\in\Xi} a_\xi \chi_\xi$ we show

$$\|s\|_{W_2^\sigma(\mathbb{M})} \le Ch^{d/2-\sigma}\|(a_\xi)_{\xi\in\Xi}\|_{\ell^2(\Xi)}.$$

This estimate is a crucial first step for the inverse estimates.

Section 4 introduces the other stable basis considered in this paper: the local Lagrange basis, which generates the space $\tilde{V}_\Xi$. Unlike the Lagrange functions, the local Lagrange bases will be shown to be computationally efficient while enjoying many of the same (or similar) properties of the Lagrange bases. The local Lagrange bases, which generate the spaces $\tilde{V}_\Xi$, provide the focal point of this paper. We first give sufficient conditions to prove existence and stability of such a basis, given Lagrange functions with stationary exponential decay. The section culminates with Theorem 2 which states that there is a constant C so that for any $s = \sum_{\xi\in\Xi} a_\xi b_\xi$

$$\|s\|_{W_2^\sigma(\mathbb{M})} \leq Ch^{d/2-\sigma}\|(a_\xi)_{\sigma\in\Xi}\|_{\ell^2(\Xi)}$$

holds.

Section 5 provides lower stability estimates (i.e. bounding $\|s\|_{L_2}$ below in terms of the coefficients $\|(a_\xi)_{\xi\in\Xi}\|_{\ell^2}$) for elements of either V_Ξ or $\tilde{V}_\Xi$. Section 5.3 presents the complete Sobolev inverse estimates for both the spaces V_Ξ and $\tilde{V}_\Xi$ in Theorems 3 and 4 respectively.

Finally in Sect. 6 the direct theorems are given. More specifically, both spaces V_Ξ and $\tilde{V}_\Xi$ are shown to provide approximation orders for functions of varying smoothness. For a continuous function f with no known additional orders of smoothness, Theorem 5 shows that both the interpolant $I_\Xi f$ or the quasi-interpolant $Q_\Xi f$ approximate f pointwise at a rate comparable to the pointwise modulus of continuity $\omega(f, Kh|\ln h|, x_0)$ where

$$\omega(f,t,x_0) := \max_{|x-x_0|\leq t} |f(x) - f(x_0)|.$$

These are the first pointwise estimates of their kind for RBF approximation schemes.

The next result applies to smoother functions f. For a point set Ξ_e which is quasi-uniform over the manifold $\mathbb{M}$ and given kernel κ_m, we show that the smoothness of f is captured in the estimate

$$\text{dist}_{p,\mathbb{M}}(f, S(\Xi)) \leq Ch^\sigma\|f\|_{B^\sigma_{p,\infty}},\ 1 \leq p \leq \infty,\ 0 < \sigma \leq 2m,$$

where the Besov space $B^\sigma_{p,\infty}(\mathbb{M})$ is defined in (27).

Our final result shows that optimal L_∞ approximation rates, when approximating a smooth function f on Ω can be obtained from data sites contained in a set "slightly larger" than Ω. The result illustrates the local nature of the bases $\{\chi_\xi\}$ or $\{b_\xi\}$.

Let $f \in C^k(\Omega)$ and let $f_e \in C^k(\mathbb{M})$ be a smooth extension of f to $\mathbb{M}$, i.e., $f_e|_\Omega = f|_\Omega$. Let $\mathscr{S} = \{x \in \mathbb{M}\backslash\Omega,\ \text{dist}(x,\Omega) \leq Kh\log h^{-1}\}$ and Ξ a discrete quasi-uniform set contained in $\Omega \cup \mathscr{S}$ with fill distance h. Finally let Ξ_e be a quasi-uniform extension of Ξ to all of $\mathbb{M}$ as given in Lemma 2. Also let κ_m be a kernel as described in Sect. 2.2 with associated spaces

$$\tilde{V}_{\Xi_e} = \text{span}_{\xi\in\Xi_e}\{b_\xi\} \text{ and } \tilde{V}_\Xi = \text{span}_{\xi\in\Xi}\{b_\xi\}.$$

The result then states that $\mathrm{dist}_{\infty,\Omega}(f,\tilde{V}_{\Xi}) \sim \mathrm{dist}_{\infty,\mathbb{M}}(f_e,\tilde{V}_{\Xi_e})$—that is they are within constant multiples of each other. The upshot is that there are several results on estimating $\mathrm{dist}_{\infty,\mathbb{M}}(f_e,\tilde{V}_{\Xi_e})$.

2 Background: Manifolds and Kernels

2.1 The Manifold $\mathbb{M}$

As noted above, throughout this article $\mathbb{M}$ is assumed to be a closed, compact, connected, C^∞ Riemannian manifold. The metric for $\mathbb{M}$, in local coordinates $(x^1,\cdots,x^d)$, will be denoted by $g_{j,k}$ and the volume element by $\mathrm{d}\mu = \sqrt{\det(g_{j,k})}\mathrm{d}x^1\cdots\mathrm{d}x^d$. Such manifolds have the following properties:

1. *Geodesic completeness*. $\mathbb{M}$ is geodesically complete, by the Hopf-Rinow Theorem [7, Section 7.2]. Thus, $\mathbb{M}$ is a metric space with the distance $\mathrm{dist}(x,y)$ between $x,y \in \mathbb{M}$ given by the length of the shortest geodesic joining x and y. The *diameter* of $\mathbb{M}$, which is finite by virtue of the compactness of $\mathbb{M}$, will be denoted by $\mathrm{d}_{\mathbb{M}}$. The *injectivity radius* $\mathrm{r}_{\mathbb{M}}$, [7, p. 271], which is the infimum of the radius of the smallest ball on which geodesic normal coordinates are non singular, is positive and finite. Of course, $\mathrm{r}_{\mathbb{M}} \le \mathrm{d}_{\mathbb{M}}$.
2. *L_p embeddings*. For $\Omega \subset \mathbb{M}$, we define $\mathrm{vol}(\Omega) = \int_\Omega \mathrm{d}\mu$. In addition, with respect to $d\mu$, the inner product $\langle\cdot,\cdot\rangle$ and all L_p norms are defined in the usual way, and these standard embeddings hold:

$$L_p(\mathbb{M}) \subset L_q(\mathbb{M}) \text{ for } 1 \le q \le p \le \infty$$

3. *Bounded geometry*. $\mathbb{M}$ has bounded geometry [4, 19], which means that $\mathbb{M}$ has a positive injectivity radius and that derivatives of the Riemannian metric are bounded (see [12, Section 2] for details). This fact already implies the Sobolev embedding theorem, as well as a smooth family of local diffeomorphisms (uniform metric isomorphisms), [12, (2.6)], which induce a family of metric isomorphisms [12, Lemma 3.2] between Sobolev spaces on $\mathbb{M}$ and on $\mathbb{R}^d$.
4. *Volume comparisons*. Denote the (geodesic) ball centered at $x \in \mathbb{M}$ and having radius r by $B(x,r)$, where $0 < r \le \mathrm{d}_{\mathbb{M}}$. There exist constants $0 < \alpha_{\mathbb{M}} < \beta_{\mathbb{M}} < \infty$ so that, for all $0 < r \le \mathrm{d}_{\mathbb{M}}$,

$$\alpha_{\mathbb{M}} r^d \le \mathrm{vol}(B(x,r)) \le \beta_{\mathbb{M}} r^d. \tag{1}$$

This inequality requires the volume comparison theorem of Bishop and Gromov [8, 11]. See Sect. 7 for a proof and explicit estimates on $\alpha_{\mathbb{M}}$ and $\beta_{\mathbb{M}}$.

2.1.1 Point Sets

Given a set $D \subset \mathbb{M}$ and a finite set $X \subset D$, we define its *fill distance* (or *mesh norm*) h and the *separation radius* q to be:

$$h(X, D) := \sup_{x \in D} \operatorname{dist}(x, X) \qquad \text{and} \qquad q(X) := \frac{1}{2} \inf_{\xi, \zeta \in X, \xi \neq \zeta} \operatorname{dist}(\xi, \zeta). \tag{2}$$

The *mesh ratio* $\rho := h(X, D)/q(X)$ measures the uniformity of the distribution of X in D. If ρ is bounded, then we say that the point set X is quasi-uniformly distributed (in D), or simply that X is quasi-uniform.

We remark that for quasi-uniform X and any $\xi \in X$, we have, as a consequence of (1), this useful inequality:

Lemma 1 *Let $h = h(X, \mathbb{M})$ and let $f : [0, \infty) \to [0, \infty)$ be decreasing and satisfy the following: There is a continuous function $g : [0, \infty) \to [0, \infty)$ such that $f(xh) \leq g(x)$ and that $x^{d-1} g(x)$ is decreasing for $x \geq 1$, and is integrable on* $[0, \infty)$. *Then*

$$\sum_{\zeta \neq \xi \in X} f(\operatorname{dist}(\zeta, \xi)) \leq \frac{2^d d \, \rho^d \beta_{\mathbb{M}}}{\alpha_{\mathbb{M}}} \int_0^\infty g(r) r^{d-1} \mathrm{d}r. \tag{3}$$

Proof Divide $\mathbb{M}$ into $N \approx \mathrm{d}_{\mathbb{M}}/h$ annuli $\boldsymbol{a}_n$, with center ξ and inner and outer radii $(n-1)h$ and nh, $n \geq 2$. The cardinality of centers in each annulus $\boldsymbol{a}_n$ is approximately

$$\#\boldsymbol{a}_n \approx \frac{\operatorname{vol}(B(\xi, nh)) - \operatorname{vol}(B(\xi, (n-1)h))}{\operatorname{vol}(B(\xi, q))}.$$

By (1), we see that

$$\#\boldsymbol{a}_n \approx \frac{\beta_{\mathbb{M}}}{\alpha_{\mathbb{M}}} \frac{(nh)^d - ((n-1)h)^d}{q^d} \leq \frac{d\beta_{\mathbb{M}}}{q^d \alpha_{\mathbb{M}}} h^d n^{d-1} = \frac{d\rho^d \beta_{\mathbb{M}}}{\alpha_{\mathbb{M}}} n^{d-1}$$

By the assumption that f is decreasing, we have, using $n^{d-1} \leq 2^d (n-1)^{d-1}$, $n \geq 2$,

$$\sum_{\zeta \in \boldsymbol{a}_n} f(\operatorname{dist}(\xi, \zeta)) \leq \frac{d\rho^d \beta_{\mathbb{M}}}{\alpha_{\mathbb{M}}} f((n-1)h) n^{d-1} \leq \frac{2^d d \, \rho^d \beta_{\mathbb{M}}}{\alpha_{\mathbb{M}}} g((n-1))(n-1)^{d-1}.$$

Since $g(x)x^{d-1}$ is decreasing, we have $\sum_{n=2}^N g((n-1))(n-1)^{d-1} \leq \int_0^\infty g(r) r^{d-1} \mathrm{d}r$. This and the previous inequality then imply (3). □

Given D and $X \subset D$, we wish to find an extension $\widetilde{X} \supset X$ so that the separation radius is not decreased and the fill distance is controlled.

Lemma 2 *Suppose $X \subset D \subset \mathbb{M}$ has fill distance $h(X, D) = h$ and separation radius $q(X) = q$. Then there is a finite set $\widetilde{X}$ so that $\widetilde{X} \cap D = X$, $q(\widetilde{X}) = \min(q, h/2)$ and $h(\widetilde{X}, \mathbb{M}) = h$.*

Proof We extend X by taking $Z = \mathbb{M} \setminus \bigcup_{\xi \in X} B(X, h)$. Cover Z by a maximal ϵ-net with $\epsilon = h$ as follows.

Consider the set of discrete subsets $\mathscr{D} = \{D \subset Z \mid h(D, Z) = h, q(D) = h/2\}$. This is a partially ordered set under $\subset$ and therefore has a maximal element D^* by Zorn's lemma. This maximal element must satisfy $q(D^*) = h/2$ (since it's in $\mathscr{D}$) and must cover Z (if $x \in Z \setminus \bigcup_{z \in D^*} B(z, h)$ then D^* is not maximal). It follows that $\widetilde{X} = X \cup D^*$ has fill distance $h(\widetilde{X}, \mathbb{M}) = h$ and $q(\widetilde{X}) = \min(q, h/2)$. □

2.1.2 Sobolev Spaces

We can define Sobolev spaces in a number of equivalent ways. In this article, we focus on $W_p^\tau(\Omega)$, where $\tau \in \mathbb{N}$ and $1 \leq p < \infty$. For $p = \infty$, we make use of the short hand notation (usual for approximation theory) $W_\infty^\tau = C^\tau$ (i.e., substituting the L_∞ Sobolev space by the Hölder space).

Our definition is the one developed in [2], by using the covariant derivative operator. This permits us to correctly define Sobolev norms and semi-norms on domains. Namely,

$$\|f\|_{W_p^\tau(\Omega)} = \left(\sum_{k=0}^{\tau} \int_\Omega (\langle \nabla^k f, \nabla^k f \rangle_x)^{p/2} \mathrm{d}x \right)^{1/p}.$$

See [2, Chapter 2], [12, Section 3] or [19, Chapter 7] for details.

Here bounded geometry means that $\mathbb{M}$ has a positive injectivity radius and that derivatives of the Riemannian metric are bounded (see [12, Section 2] for details). This fact already implies the Sobolev embedding theorem, as well as a smooth family of local diffeomorphisms (uniform metric isomorphisms), [12, (2.6)], which induce a family of metric isomorphisms [12, Lemma 3.2] between Sobolev spaces on $\mathbb{M}$ and on $\mathbb{R}^d$.

2.2 *Sobolev-Matérn Kernels*

The kernels we consider in this article are positive definite. Much of the theory extends to kernels that are conditionally positive definite; for a discussion, see [14].

A positive definite kernel $k : \mathbb{M} \times \mathbb{M} \to \mathbb{R}$ satisfies the property that for every finite set $X \subset \mathbb{M}$, the collocation matrix

$$\mathrm{K}_X := (k(\xi, \zeta))_{\xi, \zeta \in X}$$

is strictly positive definite.

If $\tau > d/2$, then $W_2^\tau(\mathbb{M})$ is a reproducing kernel Hilbert space, and its kernel is positive definite. Conversely, every continuous positive definite kernel is the reproducing kernel for a Hilbert space of continuous functions $\mathscr{N}(k)$ on $\mathbb{M}$.

The positive definite kernels we consider in this article are the Sobolev-Matérn kernels, which are reproducing kernels for the Sobolev space $W_2^m(\mathbb{M})$. These were introduced in [12]; we will denote them by κ_m. They are also the fundamental solution of the elliptic differential operator, $\mathscr{L} = \sum_{j=0}^m (\nabla^j)^* \nabla^j$ of order $2m$. This fact, although not used directly, is a key fact used to establish the stationary energy decay estimates considered in Sect. 3.1.

For finite $X \subset \mathbb{M}$ we define $S(X) := \operatorname{span}_{\xi \in X} k(\cdot, \xi)$. The guaranteed invertibility of K_X is of use in solving interpolation problems—given $\boldsymbol{y} \in \mathbb{R}^X$, one finds $\boldsymbol{a} \in \mathbb{R}^X$ so that $\mathrm{K}_X \boldsymbol{a} = \boldsymbol{y}$. It follows that $\sum_{\xi \in X} a_\xi k(\cdot, \xi)$ is the unique interpolant to $(\xi, y_\xi)_{\xi \in X}$ in $S(X)$. It is also the case that $\sum_{\xi \in X} a_\xi k(\cdot, \xi)$ is the interpolant to $(\xi, y_\xi)_{\xi \in X}$ with minimum $\mathscr{N}(k)$ norm.

3 Lagrange Functions and First Bernstein Inequalities

In this section we introduce the Lagrange functions, which are a localized basis generated by the kernel κ_m. After this we give our first class of Bernstein estimates, valid for linear combinations of Lagrange functions.

3.1 Lagrange Functions

For a positive definite kernel k and a finite $X \subset \mathbb{M}$, there exists a family of uniquely defined functions $(\chi_\xi)_{\xi \in X} \subset S(X)$ that satisfy $\chi_\xi(\zeta) = \delta(\xi, \zeta)$ for all $\zeta \in X$, and have the representation $\chi_\xi = \sum_{\eta \in X} A_{\eta,\xi} k(\cdot, \eta)$. The χ_ξ's are the *Lagrange functions* associated with X; they are easily seen to form a basis for $S(X)$.

The $A_{\eta,\xi}$'s can be expressed in a useful way, in terms of an inner product. Let $\langle \cdot, \cdot \rangle_{\mathscr{N}(k)}$ denote the inner product for the reproducing Hilbert space $\mathscr{N}(k)$. Because $\chi_\xi \in \mathscr{N}(k)$, we have that $\langle \chi_\xi, \kappa_m(\cdot, \eta) \rangle_{\mathscr{N}(k)} = \chi_\xi(\eta)$. Representing a second χ_ζ by $\chi_\zeta = \sum_{\eta \in X} A_{\eta,\zeta} k(\cdot, \eta)$, we obtain

$$\big\langle \chi_\xi, \chi_\zeta \big\rangle_{\mathscr{N}(k)} = \big\langle \chi_\xi, \sum_{\eta \in X} A_{\eta,\zeta} k(\cdot, \eta) \big\rangle_{\mathscr{N}(k)} = \sum_{\eta \in X} A_{\eta,\zeta} \chi_\xi(\eta) = A_{\xi,\zeta}. \tag{4}$$

If $k = \kappa_m : \mathbb{M} \times \mathbb{M} \to \mathbb{R}$ is a Sobolev-Matérn kernel, then, by virtue of κ_m being a reproducing kernel for $\mathscr{N}(\kappa_m) \approx W_2^m(\mathbb{M})$, we can make the following "bump estimate" on the $A_{\eta,\xi}$'s. Consider a C^∞ function $\psi_{\xi,q} : \mathbb{M} \to [0, 1]$ that is compactly supported in $B(\xi, q)$ and that satisfies $\psi_{\xi,q}(\xi) = 1$. Moreover, the condition on $\operatorname{supp}(\psi_{\xi,q})$ implies that, for any $\zeta \in X$, $\psi_{\xi,q}(\zeta) = \delta(\xi, \zeta)$. Because

$\psi_{\xi,q} \in W_2^m \approx \mathscr{N}(\kappa_m)$ and χ_ξ is the minimum norm interpolant to $\zeta \to \delta(\xi,\zeta)$, we have that

$$\|\chi_\xi\|_{\mathscr{N}(\kappa_m)} \le \|\psi_{\xi,q}\|_{\mathscr{N}(\kappa_m)} \le C\|\psi_{\xi,q}\|_{W_2^m(\mathbb{M})} \le Cq^{\frac{d}{2}-m}.$$

As a consequence, the $A_{\xi,\zeta}$'s are uniformly bounded:

$$|A_{\xi,\zeta}| = |\langle \chi_\xi, \chi_\zeta\rangle_{\mathscr{N}(\kappa_m)}| \le Cq^{d-2m}. \tag{5}$$

This bound is rather rough, and can be substantially improved. In fact, when X is sufficiently dense in $\mathbb{M}$, there exist constants C, and $\nu > 0$, which depend on κ_m (see[9]), so that the coefficient bound

$$|A_{\xi,\zeta}| = |\langle \chi_\xi, \chi_\zeta\rangle_{\mathscr{N}(\kappa_m)}| \le Cq^{d-2m}\exp\left(-\nu\frac{\operatorname{dist}(\xi,\zeta)}{h}\right) \tag{6}$$

holds. The proof of this estimate is, *mutatis mutandis*, the same as that for [9, Eqn. 5.6].

Under the same hypotheses, we have the spatial decay of the Lagrange function:

$$|\chi_\xi(x)| \le C\rho^{m-d/2}\exp\left(-\mu\frac{\operatorname{dist}(x,\xi)}{h}\right), \tag{7}$$

with $\mu = 2\nu$. Both (7) and (6) are consequences of the zeros estimate [14, (A.15)] on $\mathbb{M}$ and a more basic estimate,

$$\|\chi_\xi\|_{W_2^m(\mathbb{M}\setminus B(\xi,R))} \le Cq^{d/2-m}\exp\left(-\mu\frac{R}{h}\right) \tag{8}$$

which we call an energy estimate. When (8) holds for a system of Lagrange functions, we say it exhibits *stationary exponential decay of order m.*

Stationary decay of order m was demonstrated for Lagrange functions generated by Sobolev-Matérn kernels on compact Riemannian manifolds in [12]. (Specifically, these results are found in [12, Corollary 4.4] for (8) and in [12, Proposition 4.5] for (7).) Similar bounds hold for Lagrange functions associated with other kernels, both positive definite and conditionally positive definite, as discussed in [14] and [15].

We stress that to get estimates (8), (7) and (6), the point set X must be dense in $\mathbb{M}$. This is clearly problematic when we consider behavior over $\Omega \subsetneq \mathbb{M}$ and $X \subset \Omega$ (which is a focus of this article). To handle this, for a given point set we require the dense, quasi-uniform extension to $\mathbb{M}$ that was developed in Lemma 2.

3.2 Bernstein Type Estimates for (Full) Lagrange Functions

We develop partial Bernstein inequalities for functions of the form $s = \sum_{\xi \in X} a_\xi \chi_\xi$. Our goal is to control Sobolev norms $\|s\|_{W_2^\sigma}$ by the $\ell_2(X)$ norm on the coefficients: $\|\boldsymbol{a}\|_{\ell_2(X)}$. We have the following theorem.

Theorem 1 *If X is sufficiently dense in Ω and $0 \le \sigma \le m$, then there exists $C < \infty$ such that*

$$\Big\| \sum_{\xi \in X} a_\xi \chi_\xi \Big\|_{W_2^\sigma(\Omega)} \le C \rho^m h^{d/2-\sigma} \|\boldsymbol{a}\|_{\ell_2(X)}. \tag{9}$$

Proof Since $\Omega \subseteq \mathbb{M}$ and $W_2^\sigma(\Omega) \subseteq W_2^\sigma(\mathbb{M})$, we only need to prove the result for $\Omega = \mathbb{M}$. In addition, we can replace X with $\widetilde{X}$, the extension of X to $\mathbb{M}$, whose existence was shown in Lemma 2. The point is that once the result is shown true for $\widetilde{X}$, we just restrict a_ξ's to $\xi \in X$, setting $a_\xi = 0$ for $\xi \in \widetilde{X} \setminus X$.

To begin, we use (6) to observe that $\chi_\xi \in W_2^m(\mathbb{M})$, whence we obtain

$$\begin{aligned}
\Big\| \sum_{\xi \in \widetilde{X}} a_\xi \chi_\xi \Big\|_{W_2^m(\mathbb{M})}^2 &\le C \Big\| \sum_{\xi \in \widetilde{X}} a_\xi \chi_\xi \Big\|_{\mathcal{N}(\kappa_m)}^2 \\
&= C \sum_{\xi \in \widetilde{X}} \sum_{\zeta \in \widetilde{X}} |a_\xi||a_\zeta| \big|\langle \chi_\xi, \chi_\zeta \rangle_{\mathcal{N}(\kappa_m)}\big| \\
&\le C q^{d-2m} \sum_{\xi \in \widetilde{X}} \sum_{\zeta \in \widetilde{X}} |a_\xi||a_\zeta| e^{-\nu \frac{\mathrm{dist}(\xi,\zeta)}{h}} \\
&\le C q^{d-2m} \Bigg(\sum_{\xi \in \widetilde{X}} |a_\xi|^2 + \sum_{\xi \in \widetilde{X}} \sum_{\zeta \in \widetilde{X}, \zeta \ne \xi} |a_\xi||a_\zeta| e^{-\nu \frac{\mathrm{dist}(\xi,\zeta)}{h}} \Bigg).
\end{aligned}$$

From this we have $\| \sum_{\xi \in \widetilde{X}} a_\xi \chi_\xi \|_{W_2^m(\mathbb{M})} \le C q^{d/2-m} \Big(\|\boldsymbol{a}\|_{\ell_2(\widetilde{X})} + (II)^{1/2} \Big)$. We focus on the off-diagonal part II. Since each term appears twice, we can make the estimate

$$\begin{aligned}
\sum_{\xi \in \widetilde{X}} \sum_{\zeta \ne \xi} |a_\xi||a_\zeta| e^{-\nu \frac{\mathrm{dist}(\xi,\zeta)}{h}} &\le \sum_{\xi \in \widetilde{X}} \sum_{\zeta \in \widetilde{X}, \zeta \ne \xi} |a_\xi|^2 e^{-\nu \frac{\mathrm{dist}(\xi,\zeta)}{h}} \\
&\le C \rho^d \left(\int_0^\infty e^{-\nu r} r^{d-1} \mathrm{d}r \right) \sum_{\xi \in \widetilde{X}} |a_\xi|^2.
\end{aligned}$$

The first inequality uses the estimate $|a_\xi||a_\zeta| \le \frac{1}{2}(|a_\xi|^2 + |a_\zeta|^2)$. The second inequality follows from (3). We have demonstrated that

$$\Big\| \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi \Big\|_{W_2^m(\mathbb{M})} \le C\rho^{d/2} q^{d/2-m} \|\boldsymbol{a}\|_{\ell_2(\widetilde{X})} \le C\rho^m h^{d/2-m}\, \|\boldsymbol{a}\|_{\ell_2(\widetilde{X})}\,. \tag{10}$$

On the other hand, using (7) we have

$$\begin{aligned}\Big\| \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi \Big\|^2_{L_2(\mathbb{M})} &\le \sum_{\xi\in\widetilde{X}} \sum_{\zeta\neq\xi} |a_\xi||a_\zeta||\langle \chi_\xi, \chi_\zeta\rangle_2| \\ &\le C\rho^{2m-d} \sum_{\xi\in\widetilde{X}} \sum_{\zeta\in\widetilde{X},\zeta\neq\xi} |a_\xi||a_\zeta| \int_{\mathbb{M}} e^{-2\nu\frac{\mathrm{dist}(x,\xi)}{h}} e^{-2\nu\frac{\mathrm{dist}(x,\zeta)}{h}} \mathrm{d}x.\end{aligned}$$

The integral can be estimated over two disjoint regions (the part of $\mathbb{M}$ closer to ξ and the part closer to ζ) to obtain

$$\begin{aligned}\Big\| \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi \Big\|^2_{L_2(\mathbb{M})} &\le C\rho^{2m-d} h^d \sum_{\xi\in\widetilde{X}} \sum_{\zeta\in\widetilde{X},\zeta\neq\xi} |a_\xi||a_\zeta| e^{-\nu\frac{\mathrm{dist}(\xi,\zeta)}{h}} \\ &\le C\rho^{2m} h^d \sum_{\xi\in\widetilde{X}} |a_\xi|^2.\end{aligned}$$

The second inequality repeats the estimate used to bound $\big\| \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi \big\|_{W_2^m(\mathbb{M})}$. It follows that

$$\Big\| \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi \Big\|_{L_2(\mathbb{M})} \le C\rho^m h^{d/2} \big\|\boldsymbol{a}\big\|_{\ell_2(\widetilde{X})}. \tag{11}$$

Define the operator $V : \ell_2(\widetilde{X}) \to W_2^m(\mathbb{M}) : \boldsymbol{a} \mapsto \sum_{\xi\in\widetilde{X}} a_\xi \chi_\xi$. We interpolate between (10) and (11), using the fact that $W_2^\sigma(\mathbb{M}) = B^\sigma_{2,2}(\mathbb{M}) = [L_2(\mathbb{M}), W_2^m(\mathbb{M})]_{\frac{\sigma}{m},2}$ (cf. [19]). As noted at the start, this implies the result for $\Omega \subseteq \mathbb{M}$ and $X \subset \Omega$. □

4 Local Lagrange Functions

We now consider locally constructed basis functions. We employ a small set of centers from X to construct "local" Lagrange functions: For each $\xi \in X$, we define

$$\Upsilon(\xi) := \{\zeta \in X \mid \mathrm{dist}(\zeta,\xi) \le Kh|\log h|\},$$

where $K > 0$ is a parameter used to adjust the number of points in $\Upsilon(\xi)$.

We define the *local Lagrange function* b_ξ at ξ to be the Lagrange function for $\Upsilon(\xi)$. We will call $\Upsilon(\xi)$ the *footprint* of b_ξ. Of course, $b_\xi \in S(\Upsilon(\xi))$. The choice of the parameter K depends on the constants appearing in the stationary exponential decay (8), the conditions we place on the manifold $\mathbb{M}$ and the rate at which we wish b_ξ to have decay away from ξ.

K may be chosen so that for a prescribed J, which depends linearly on K and other parameters, we can ensure that $\|\chi_\xi - b_\xi\|_{L_\infty(\mathbb{M})} = \mathscr{O}(h^J)$ holds. (See (17).)

The main goal of this section is to provide Sobolev estimates on the difference between locally constructed functions b_ξ and the analogous (full Lagrange) functions χ_ξ. As in [9] the analysis of this new basis is considered in two steps. First, an intermediate basis function $\widetilde{\chi}_\xi$ is constructed and studied: the *truncated Lagrange function*. These functions employ the same footprint as b_ξ (i.e., they are members of $S(\Upsilon(\xi))$) but their construction is global rather than local. This topic is considered in Sect. 4.1. Then, a comparison is made between the truncated Lagrange function and the local Lagrange function. In Sect. 4.2, we will show that the error between local and truncated Lagrange functions is controlled by the size of the coefficients in the expansion of $b_\xi - \widetilde{\chi}_\xi$ in the standard (kernel) basis for $S(\Upsilon(\xi))$.

4.1 Truncated Lagrange Functions

For a (full) Lagrange function $\chi_\xi = \sum_{\zeta\in X} A_{\xi,\zeta} k(\cdot,\zeta) \in S(X)$ on the point set X, the truncated Lagrange function $\widetilde{\chi}_\xi = \sum_{\zeta\in\Upsilon(\xi)} A_{\xi,\zeta} k(\cdot,\zeta)$ is a function in $S(\Upsilon(\xi))$ obtained by removing the $A_{\xi,\zeta}$'s for ζ not in $\Upsilon(\xi)$. The cost of truncating can be measured using the norm of the omitted coefficients (the tail).

Lemma 3 *Let $\mathbb{M}$ be as in Sect. 2.1 and let κ_m be a Sobolev-Matérn kernel generating $\{\chi_\xi : \xi \in X\}$. Suppose $X \subset \mathbb{M}$ has fill distance $0 < h \le h_0$ and separation radius $q > 0$. Let $K > (4m - 2d)/\nu$ and for each $\xi \in X$, let $\Upsilon(\xi) = \{\zeta \in X \mid \operatorname{dist}(\xi,\zeta) \le Kh|\log h|\}$. Then*

$$\sum_{\zeta\in X\setminus\Upsilon(\xi)} |A_{\xi,\zeta}| \le C\rho^{2m} h^{K\nu/2+d-2m}.$$

Proof The inequality (6) guarantees that

$$\begin{aligned}\sum_{\zeta\in X\setminus\Upsilon(\xi)} |A_{\xi,\zeta}| &\le Cq^{d-2m} \sum_{\operatorname{dist}(\zeta,\xi)\ge Kh|\log h|} \exp\left(-\nu\frac{\operatorname{dist}(\xi,\zeta)}{h}\right)\\ &\le Cq^{-2m} \int_{y\in\mathbb{M}\setminus B(\xi,Kh|\log h|)} \exp\left(-\nu\frac{\operatorname{dist}(\xi,y)}{h}\right) \mathrm{d}y\\ &\le Cq^{-2m} \int_{Kh|\log h|}^{\infty} \exp\left(-\nu\frac{r}{h}\right) r^{d-1}\mathrm{d}r.\end{aligned}$$

A simple way[1] to estimate this involves splitting $\nu = \nu/2 + \nu/2$ and writing

$$\sum_{\zeta \in X \setminus \Upsilon(\xi)} |A_{\xi,\zeta}| \le Ch^d q^{-2m} \left(\int_{K|\log h|}^{\infty} r^{d-1} \exp\left(-K|\log h|\frac{\nu}{2}\right) \exp\left(-r\frac{\nu}{2}\right) \mathrm{d}r \right)$$
$$\le Ch^d q^{-2m} h^{K\nu/2}.$$

The lemma follows. □

Standard properties of reproducing Hilbert kernels imply that, because $\mathbb{M}$ is a compact metric space, $\kappa_m(x,y)$ is continuous on $\mathbb{M} \times \mathbb{M}$. Consequently, $\kappa_m(x,x) = \|\kappa_m(\cdot,x)\|^2_{\mathscr{N}(\kappa_m)}$ is uniformly bounded in x. Moreover, since $\mathscr{N}(\kappa_m)$ and $W_2^m(\mathbb{M})$ are norm equivalent, there is a constant Γ such that

$$\sup_{x \in \mathbb{M}} \|\kappa_m(\cdot,x)\|_{W_2^m(\mathbb{M})} \le C \sup_{x \in \mathbb{M}} \|\kappa_m(\cdot,x)\|_{\mathscr{N}(\kappa_m)} \le \Gamma_{\kappa_m}.$$

From Lemma 3 and the inequality above, we have that

$$\|\chi_\xi - \widetilde{\chi}_\xi\|_{W_2^m(\mathbb{M})} \le \Gamma_{\kappa_m} \sum_{\zeta \in X \setminus \Upsilon(\xi)} |A_{\xi,\zeta}| \le C\Gamma_{\kappa_m} \rho^{2m} h^{K\nu/2-2m+d}. \qquad (12)$$

Applying the Sobolev embedding theorem then yields the result below.

Proposition 1 *Let κ_m the Sobolev-Matérn kernel, with $m > d/2$. Then, if $1 \le p < \infty$ and $\sigma \le m - (\frac{d}{2} - \frac{d}{p})_+$, or if $p = \infty$ and $0 \le \sigma < m - d/2$, we have*

$$\|\chi_\xi - \widetilde{\chi}_\xi\|_{W_p^\sigma(\mathbb{M})} \le C\Gamma_{\kappa_m} \rho^{2m} h^{K\nu/2+d-2m}, \quad C = C_{\sigma,m,p}. \qquad (13)$$

In particular, if $p = \infty$ and $\sigma = 0$, we have

$$\|\chi_\xi - \widetilde{\chi}_\xi\|_{L_\infty(\mathbb{M})} \le C_m \Gamma_{\kappa_m} \rho^{2m} h^{K\nu/2+d-2m}. \qquad (14)$$

Proof This follows from (12) by applying the Sobolev embedding theorem to $\|\chi_\xi - \widetilde{\chi}_\xi\|_{W_p^\sigma(\mathbb{M})}$. □

4.2 Local Lagrange Function Distance Estimates

In this section, we consider bounding the distance between b_ξ and χ_ξ and also b_ξ and $\widetilde{\chi}_\xi$, using Sobolev norms. The argument we will use is essentially the one used on the sphere in [9].

[1] The integral can be done exactly. However, we don't need to do that here.

By construction, both b_ξ and $\widetilde{\chi}_\xi$ are in $S(\Upsilon(\xi))$, and thus $b_\xi - \widetilde{\chi}_\xi \in S(\Upsilon(\xi))$ is, too. Hence, $b_\xi - \widetilde{\chi}_\xi = \sum_{\zeta\in\Upsilon(\xi)} a_\zeta \kappa_m(\cdot,\zeta)$. Let $\boldsymbol{a} := (a_\zeta)_{\zeta\in\Upsilon(\xi)}$ and $\boldsymbol{y} = (b_\xi - \widetilde{\chi}_\xi)|_{\Upsilon(\xi)}$. where $\boldsymbol{a}$ and $\boldsymbol{y}$ are related by $\mathrm{K}_{\Upsilon(\xi)}\boldsymbol{a} = \boldsymbol{y}$.

We can write $\boldsymbol{y}$ another way. Since b_ξ is a Lagrange function for $\Upsilon(\xi)$, we have that $b_\xi(\zeta) = \delta_{\xi,\zeta}$ when $\zeta \in \Upsilon(\xi)$. However, because χ_ξ is a Lagrange function for all X, it also satisfies $\chi_\xi(\zeta) = \delta_{\xi,\zeta}$, $\zeta \in \Upsilon(\xi)$. Consequently, $\boldsymbol{y} = (\chi_\xi - \widetilde{\chi}_\xi)|_{\Upsilon(\xi)}$.

Using this form of $\boldsymbol{y}$ we have that $\|\boldsymbol{y}\|_1 \le (\#\Upsilon(\xi))\|\boldsymbol{y}\|_\infty \le (\#\Upsilon(\xi))\|\chi_\xi - \widetilde{\chi}_\xi\|_{L_\infty(\mathbb{M})}$. From (14) and the bound $\#\Upsilon(\xi) \le C\rho^d|\log h|^d$, we arrive at

$$\|\boldsymbol{y}\|_1 \le C\rho^{2m+d}h^{K\nu/2+d-2m}|\log h|^d.$$

The matrix $(\mathrm{K}_{\Upsilon(\xi)})^{-1}$ has entries $(A_{\zeta,\eta})_{\zeta,\eta\in\Upsilon(\xi)}$. These can be estimated by (5): $|A_{\zeta,\eta}| \le Cq^{d-2m}$. It follows that $(\mathrm{K}_{\Upsilon(\xi)})^{-1}$ has ℓ_1 matrix norm

$$\left\| \left(\mathrm{K}_{\Upsilon(\xi)}\right)^{-1} \right\|_{1\to 1} \le C(\#\Upsilon(\xi))q^{d-2m} \le C\rho^{2m}|\log h|^d h^{d-2m}.$$

This and the bound on $\|\boldsymbol{y}\|_1$ above imply that

$$\|\boldsymbol{a}\|_1 \le \left\| \left(\mathrm{K}_{\Upsilon(\xi)}\right)^{-1} \right\|_{1\to 1}\|\boldsymbol{y}\|_1 \le C\rho^{4m+d}|\log h|^{2d}h^{K\nu/2+2d-4m}. \tag{15}$$

Under the conditions in Proposition 1, $b_\xi - \widetilde{\chi}_\xi$ is in $W_p^\sigma(\mathbb{M})$, as are the $\kappa_m(\cdot,\zeta)$'s. Consequently, $\|b_\xi - \widetilde{\chi}_\xi\|_{W_p^\sigma(\mathbb{M})} \le \|\boldsymbol{a}\|_1 \max_{z\in\mathbb{M}} \|\kappa_m(\cdot,z)\|_{W_p^\sigma(\mathbb{M})} \le \Gamma_{\kappa_m}\|\boldsymbol{a}\|_1$. Using the triangle inequality, the bound in (15), and the estimate above, we have the following result:

Lemma 4 *Let $\mathbb{M}$ be as in Sect. 2.1 and let κ_m be a Sobolev-Matérn kernel. Then, we have, for $0 \le \sigma \le m - (d/2 - d/p)_+$, or with $p = \infty$ and $0 \le \sigma < m - d/2$,*

$$\left\|b_\xi - \chi_\xi\right\|_{W_p^\sigma(\mathbb{M})} \le C\Gamma_{\kappa_m}\rho^{4m+d}h^{K\nu/2+2d-4m}|\log h|^{2d},\ C = C_{m,p,\sigma} \tag{16}$$

We remark that $|\log h|^{2d} \le Ch^{-1}$, so that either by finding a sufficiently small h^*, so that this holds for $h < h^*$, or by increasing the constant, or both we have

$$\left\|b_\xi - \chi_\xi\right\|_{W_p^\sigma(\mathbb{M})} \le C\rho^{4m+d}h^{K\nu/2+2d-4m-1}. \tag{17}$$

4.3 Bernstein Type Estimate for Local Lagrange Functions

In this section we discuss the local Lagrange functions b_ξ generated by κ_m and the centers X. We develop partial Bernstein inequalities, where for functions of the form $s = \sum_{\xi\in X} a_\xi b_\xi$ the norms $\|s\|_{W_2^\sigma}$ are controlled by an ℓ_2 norm on the coefficients: $\|\boldsymbol{a}\|_{\ell_2(X)}$.

We will now obtain estimates similar to (9) for the expansion $\sum_{\xi\in X} a_\xi b_\xi$. In contrast to the full Lagrange basis, which is globally decaying, we have a family of functions $(b_\xi)_{\xi\in X}$ whose members are uniformly small (on compact sets), but do not necessarily decay (at least not in a stationary way).

Theorem 2 *Suppose X is sufficiently dense in Ω. Assume $K\nu+d-4m-1 \geq d/2-\sigma$. Then there is C, depending on the constants appearing in* (1) *and* (7) *so that*

$$\Big\| \sum_{\xi\in X} a_\xi b_\xi \Big\|_{W_2^\sigma(\Omega)} \leq C_{\mathbb{M}} \rho^{4m+2d} h^{d/2-\sigma} \|\boldsymbol{a}\|_{\ell_2(X)}. \tag{18}$$

Proof As in the case of Theorem 1, because $\Omega \subseteq \mathbb{M}$, we only have to prove the result for $\mathbb{M}$. We start with the basic splitting

$$s := \sum_{\xi\in X} a_\xi b_\xi = \Big(\sum_{\xi\in X} a_\xi \chi_\xi\Big) + \Big(\sum_{\xi\in X} a_\xi (b_\xi - \chi_\xi)\Big) =: G + B.$$

Applying the Sobolev norm gives $\|s\|^2_{W_2^\sigma(\mathbb{M})} \leq \|G\|^2_{W_2^\sigma(\mathbb{M})} + \|B\|^2_{W_2^\sigma(\mathbb{M})}$. From (9), we have $\|G\|_{W_2^\sigma(\mathbb{M})} \leq C\rho^m h^{d/2-\sigma} \|\boldsymbol{a}\|_{\ell_2(X)}$.

We now restrict our focus to B. For $|\alpha| \leq m$, Hölder's inequality ensures that $\| \sum_{\xi\in X} a_\xi \nabla^\alpha (b_\xi - \chi_\xi)\|_x \leq \big(\sum_{\xi\in X} |a_\xi|^2\big)^{1/2} \big(\sum_{\xi\in X} \|\nabla^\alpha (b_\xi - \chi_\xi)\|_x^2\big)^{1/2}$. Here we have used, for a rank α-covariant tensor field F (i.e., a smooth section of the vector bundle of rank α covariant tensors), the norm on the fiber at x given by the Riemannian metric, i.e., $\| F\|_x$ is the norm of the tensor $F(x)$.

Therefore, for $0 \leq \sigma \leq m$,

$$\begin{aligned}
\|B\|_{W_2^\sigma(\mathbb{M})} &\leq \|\boldsymbol{a}\|_{\ell_2(X)} \Big\| \sum_{\xi\in X} (b_\xi - \chi_\xi) \Big\|_{W_2^\sigma(\mathbb{M})} \\
&\leq \|\boldsymbol{a}\|_{\ell_2(X)} \sum_{\xi\in X} \big\| (b_\xi - \chi_\xi) \big\|_{W_2^\sigma(\mathbb{M})} \\
&\leq \|\boldsymbol{a}\|_{\ell_2(X)} (\#X) \max_{\xi\in X} \big\| (b_\xi - \chi_\xi) \big\|_{W_2^\sigma(\mathbb{M})}
\end{aligned}$$

The inequality $\|B\|_{W_2^\sigma(\mathbb{M})} \leq C\rho^{4m+2d} h^{K\nu/2+d-4m-1} \|\boldsymbol{a}\|_{\ell_2(X)}$ follows by applying Lemma 4, and the fact that $\#X \leq C\rho^d h^{-d}$. Inequality (18) follows, which completes the proof. □

5 Stability Results and Inverse Inequalities

In this section we consider finite dimensional spaces of the form $V_\Xi = \mathrm{span}_{\xi\in\Xi} \chi_\xi$ and $\widetilde{V}_\Xi = \mathrm{span}_{\xi\in\Xi} b_\xi$, using the Lagrange and local Lagrange functions considered in Sects. 3.2 and 4.3. We note that the localized functions χ_ξ and b_ξ are indexed by a dense set of centers $X \subset \mathbb{M}$, but the spaces V_Ξ and $\widetilde{V}_\Xi$ are constructed using a restricted set of centers $\Xi = X \cap \Omega$, corresponding to the centers located inside $\Omega \subset \mathbb{M}$, which the underlying region over which we take the L_2 norm.

5.1 The Domain Ω

We now consider a compact region $\Omega \subset \mathbb{M}$. This presents two challenges.

The first concerns the density of point sets $\Xi \subset \Omega$. Unless $\Omega = \mathbb{M}$, the given set Ξ does not itself satisfy the density condition $h(\Xi, \mathbb{M}) < h_0$. For this, we need a larger set $X \subset \mathbb{M}$ with points lying outside of Ω (in fact, when working with local Lagrange functions b_ξ, it suffices to consider $X \subset \{x \in \mathbb{M} \mid \operatorname{dist}(x, \Omega) < Kh|\log h|\}$). This assumption is in place to guarantee decay of the basis functions. It would be quite reasonable to be "given" initially only the set $\Xi \subset \Omega$ and to use this to construct X. Lemma 2 demonstrates that it is possible to extend a given set of centers $X \subset \Omega$ in a controlled way to obtain a dense subset of $\mathbb{M}$.

The second challenge concerns the domain Ω. Previously we have not needed to make extra assumptions about such a region, but for estimates relating $\|\boldsymbol{a}\|_{\ell_2}$ and $\|\sum_\xi a_\xi b_\xi\|_{L_2(\mathbb{M})}$ or $\|\sum_\xi a_\xi \chi_\xi\|_{L_2(\mathbb{M})}$, the boundary becomes slightly more important. Fortunately, the extra assumption we make on Ω is quite mild—it is given below in Assumption 1.

For the remainder of the article, we assume $\Omega \subset \mathbb{M}$ satisfies the Boundary Regularity condition and $\Xi \subset \Omega$ is finite. We utilize the extended point set $\widetilde{\Xi}$ from Lemma 2; this gives rise to the family $(\chi_\xi)_{\xi \in \widetilde{\Xi}}$. With this setup, we define

$$V_\Xi := \operatorname{span}_{\xi \in \Xi} \chi_\xi \text{ (Full Lagrange) and } \widetilde{V}_\Xi := \operatorname{span}_{\xi \in \Xi} b_\xi \text{ (Local Lagrange).}$$

We note that $V_\Xi \subset S(\widetilde{\Xi})$, while $\widetilde{V}_\Xi \subset S(\widetilde{\Xi} \cap \{x \in \mathbb{M} \mid \operatorname{dist}(x, \Omega) \le Kh|\log h|\}) \subset S(\widetilde{\Xi})$. A property of $\mathbb{M}$, in force throughout the article, is the following.

Assumption 1 (Boundary Regularity) *There exists a constant* $0 < \alpha_\Omega$ *for which the following holds: for all* $x \in \Omega$ *and all* $r \le \mathrm{d}_{\mathbb{M}}$,

$$\alpha_\Omega r^d \le \operatorname{vol}(B(x, r) \cap \Omega).$$

Note that this holds when Ω satisfies an interior cone condition.

5.2 Stability of Full and Local Lagrange Functions on Ω

In this section we show that the synthesis operators $\boldsymbol{a} \mapsto \sum_{\xi \in \Xi} a_\xi \chi_\xi$ and $\boldsymbol{a} \mapsto \sum_{\xi \in \Xi} a_\xi b_\xi$ are bounded above and below from $\ell_p(\Xi)$ to $L_p(\Omega)$.

In addition to the pointwise and coefficient decay (namely (7) and (6)) stemming from (8), we can employ the following uniform equicontinuity property of the Lagrange functions. There is $0 < \epsilon \le 1$ so that

$$|\chi_\xi(x) - \chi_\xi(y)| \le C \left[\frac{\operatorname{dist}(x, y)}{q}\right]^\epsilon \tag{19}$$

with constant C depending only on ϵ, the mesh ratio $\rho = h/q$, and the constants in (8). This follows from the energy estimate (8) and a zeros estimate [14, Corollary A.15], and the embedding $C^\epsilon(\mathbb{M}) \subset W_2^m(\mathbb{M})$ where $0 < \epsilon < m - d/2$. We refer the interested reader to [13, Lemma 7.2] for details.

Proposition 2 *Let $\Omega \subseteq \mathbb{M}$ be a compact domain satisfying Assumption 1. Then for the Lagrange functions corresponding to κ_m, there exist constants $c, C > 0$ and $q_0 > 0$, so that for $q < q_0$, for $1 \le p \le \infty$ and for all functions in V_Ξ,*

$$c \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} \le q^{-d/p} \| \sum_{\xi\in\Xi} a_\xi \chi_\xi \|_{L_p(\Omega)} \le C \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} \, . \tag{20}$$

If, in addition $K\nu/2 + 2d - 4m - 2 =: \varepsilon > 0$, with K chosen sufficiently large, then

$$\frac{c}{2} \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} \le q^{-d/p} \| \sum_{\xi\in\Xi} a_\xi b_\xi \|_{L_p(\Omega)} \le \frac{3C}{2} \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} \, . \tag{21}$$

Proof We begin with the case in which $\Omega = \mathbb{M}$ and $s = \sum_{\xi\in\Xi} a_\xi \chi_\xi \in V_\Xi$. Then (20) follows directly from [13, Proposition 3.10]. In particular, we note that the boundary regularity assumption guarantees that $\mathbb{M}$ satisfies [13, Assumption 2.1]. The family of functions $(\chi_\xi)_{\xi\in\Xi}$ fulfills the three requirements on $(v_\xi)_{\xi\in\Xi}$.

1. They are Lagrange functions on Ξ (this is [13, Assumption 3.3]),
2. The decay property given in (7) guarantees that [13, Assumption 3.4] holds (with $\mathrm{r}_{\mathbb{M}} = \operatorname{diam}(\mathbb{M})$,
3. The equicontinuity assumption [13, Assumption 3.5] is a consequence of the Hölder property (19).

The case $\Omega \ne \mathbb{M}$ is more difficult, and the proof too long to be given here. It may be carried out by following the proofs of [15, Lemma B.1] and [15, Lemma B.6], with appropriate modifications.

To establish (21), we begin by using (17), with $K\nu/2 + 2d - 4m - 2 := \varepsilon > 0$ and $\sigma = 0$, to obtain $\|\chi_\xi - b_\xi\|_{L_p(\Omega)} \le \|\chi_\xi - b_\xi\|_{L_p(\mathbb{M})} \le C'\rho^{4m+d}h^{1+\varepsilon}$. From this, the triangle inequality, and $\sum_{\xi\in\Xi} |a_\xi| \le (\#X)^{1-1/p}\|a\|_{\ell_p} \le C'q^{-d(1-1/p)}$, we have that, for q_0 sufficiently small,

$$q^{-d/p} \| \sum_{\xi\in\Xi} a_\xi (\chi_\xi - b_\xi) \|_{L_p(\Omega)} \le C' \rho^{4m+d-\varepsilon-1} q_0^\varepsilon \|a\|_{\ell_p}$$

Again applying the triangle inequality and employing (20), we arrive at

$$\begin{aligned} c(1 - C'\rho^{4m+d-\varepsilon-1} q_0^{\varepsilon}) \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} &\le q^{-d/p} \| \sum_{\xi \in \Xi} a_\xi b_\xi \|_{L_p(\Omega)} \\ &\le C(1 + C'\rho^{4m+d-\varepsilon-1} q_0^{\varepsilon}) \, \|\boldsymbol{a}\|_{\ell_p(\Xi)} \, . \end{aligned}$$

Next, taking $q_0 < 1$, and (by increasing K if necessary) $q_0^{\varepsilon} \le \frac{\rho^{-4m-d+\varepsilon+1}}{2C'}$, and using these in the previous inequality results in (21). □

5.3 Inverse Inequalities for Full and Local Lagrange Functions on Ω

At this point we can prove the inverse inequality for both full and local Lagrange functions. We start with the full Lagrange functions.

Theorem 3 *Let $\Omega \subseteq \mathbb{M}$ be a compact domain satisfying Assumption 1. Then for the Lagrange functions corresponding to κ_m, there exist constants $C > 0$ and $h_0 > 0$, so that for $h < h_0$ if $\Xi \subset \Omega$ has fill distance h, mesh ratio ρ, and $\widetilde{\Xi} \subset \mathbb{M}$ is a suitable extension of Ξ (for instance, the one given by Lemma 2) then $V_\Xi \subset W_2^m(\Omega)$ and for all $s = \sum_{\xi \in \Xi} a_\xi \chi_\xi \in V_\Xi$ and for $0 \le \sigma \le m$, we have*

$$\|s\|_{W_2^\sigma(\Omega)} \le C\rho^{m+d/2} h^{-\sigma} \|s\|_{L_2(\Omega)}.$$

Proof From (9), we have $\|s\|_{W_2^\sigma(\Omega)} \le C\rho^m h^{d/2-\sigma} \left\|\boldsymbol{a}\right\|_{\ell_2(X)}$, and from (20), with $p = 2$ and $q = h/\rho$, we have $c \, \|\boldsymbol{a}\|_{\ell_2(\Xi)} \le h^{-d/2} \rho^{d/2} \| \sum_{\xi \in \Xi} a_\xi \chi_\xi \|_{L_p(\Omega)}$. Combining the two inequalities completes the proof. □

The proof for the local version is the same, except that we use (18) and (21).

Theorem 4 *Let $\Omega \subset \mathbb{M}$ be a compact domain satisfying Assumption 1. Then for the local Lagrange functions corresponding to κ_m, with K sufficiently large, we have that here exists a constant $h_0 > 0$, so that for $h < h_0$ if $\Xi \subset \Omega$ has fill distance h, mesh ratio ρ, and $\widetilde{\Xi} \subset \mathbb{M}$ is a suitable extension of Ξ (for instance, the one given by Lemma 2) then for all $s = \sum_{\xi \in \Xi} a_\xi b_\xi \in \widetilde{V}_\Xi$ the following holds for all $0 \le \sigma \le m$,*

$$\|s\|_{W_2^\sigma(\Omega)} \le C\rho^{m+d/2} h^{-\sigma} \|s\|_{L_2(\Omega)}.$$

6 Implications for Quasi-Interpolation and Approximation

At this point, we are able to state several results that satisfactorily answer questions concerning interpolation, quasi-interpolation, and approximation properties of the spaces V_{Ξ} and $\tilde{V}_{\Xi}$. Some of these results have appeared previously in more restrictive settings while other results, such as pointwise error estimates for quasi-interpolation of continuous functions, are entirely new.

The first result is that the Lebesgue constant for interpolation is uniformly bounded. For the setting considered here (compact Riemannian manifolds and Sobolev-Matérn kernels), this has been proven in [12].

Proposition 3 (Lebesgue Constant, [12, Theorem 4.6]) *Suppose that $m > \frac{d}{2}$. For a sufficiently dense set $\Xi \subset \mathbb{M}$ with mesh ratio ρ, the Lebesgue constant $\Lambda := \sup_{\alpha\in\mathbb{M}} \sum_{\xi\in\Xi} |\chi_{\xi}(\alpha)|$, associated with the Sobolev-Matérn kernel κ_m, is bounded by a constant depending only on m, ρ, and $\mathbb{M}$.*

We remark that the key to proving this result is the pointwise exponential decay of the Lagrange function χ_{ξ}, as given in (7). The same kind of bound also holds for local Lagrange functions. This can be shown by using the "perturbation" technique employed to prove (21).

Similar results hold for other kernels on specific compact manifolds [14]. In the case where the manifold is not compact, one typically is more interested in Lagrange functions based on finite point sets which are quasi-uniform with respect to a compact subset $\Omega \subset \mathbb{M}$. Nevertheless, a similar pointwise decay estimate for Lagrange functions holds for that setting as well [15, Inequality 3.5].

There are two kinds of stability associated with the spaces V_{Ξ} and $\widetilde{V}_{\Xi}$. The first concerns basis stability. In Proposition 2, we showed that both local and full Lagrange bases were very stable.

The second kind of stability, which was established in [13], concerns the L_p norm of the L_2 projector. Let $W : \mathbb{C}^{\#\Xi} \to V(\kappa_m, \Xi) := V_{\Xi}$ be a "synthesis operator" so $W : (a_{\xi})_{\xi\in\Xi} \to \sum_{\xi\in\Xi} a_{\xi} v_{\xi}$ for a basis $(v_{\xi})_{\xi\in\Xi}$ of V_{Ξ}. Likewise, let $W^* : L_1(\mathbb{M}) \to \mathbb{C}^{\#\Xi}$ be its formal adjoint $W^* : f \to (\langle f, v_{\xi}\rangle)|_{\xi\in\Xi}$. The L_2 projector is then

$$T_{\Xi} := W(W^*W)^{-1}W^* : L_1(\mathbb{M}) \to V_{\Xi} \tag{22}$$

in the sense that when $f \in L_2(\mathbb{M})$, $T_{\Xi}f$ is the best L_2 approximant to f from V_{Ξ}.

The L_2 norm of this projector is one—because it is orthogonal—while the L_p and $L_{p'}$ norms are equal because it is self-adjoint. Thus to estimate its L_p operator norm $(1 \le p \le \infty)$ it suffices to estimate its L_{∞} norm.

Proposition 4 (Least Squares Projector, [13, Theorem 5.1]) *For the Sobolev-Matérn kernels, for all $1 \le p \le \infty$, the L_p norm of the L_2 projector T_{Ξ} is bounded by a constant depending only on $\mathbb{M}$, ρ and κ_m.*

For applications, the local Lagrange functions $\{b_\xi\}_{\xi\in\Xi}$ are substantially more computationally efficient than the full Lagrange functions. Nevertheless the bases $\{b_\xi\}_{\xi\in\Xi}$ and the space $\tilde{V}_\Xi = \mathrm{span}_{\xi\in\Xi} b_\xi$, under appropriate assumptions, enjoy essentially all the key properties as $\{\chi_\xi\}_{\xi\in\Xi}$ and V_Ξ do.

In particular, Proposition 1 shows that the spaces V_Ξ and $\tilde{V}_\Xi$ can be quite close in Hausdorff distance and that the bases $\{b_\xi\}_{\xi\in\Xi}$ are slight perturbations of the bases $\{\chi_\xi\}_{\xi\in\Xi}$ even on compact subsets of the manifold. For the compact Riemannian manifold setting, under appropriate assumptions, the set $\{b_\xi\}_{\xi\in\Xi}$ is L_p stable and each b_ξ has pointwise polynomial decay of high order. This can be shown in the same way as in [9, Thm 6.5].

A method to implement approximation from the space $\tilde{V}_\Xi$ is by means of the quasi-interpolation operator

$$Q_\Xi f := \sum_{\xi\in\Xi} f(\xi) b_\xi.$$

The quasi-interpolation operator provides L_∞ convergence estimates at the same asymptotic rate as the interpolation operator. Indeed

$$\begin{aligned}|I_\Xi f(x) - Q_\Xi f(x)| &\le \sum_{\xi\in\Xi} |b_\xi(x) - \chi_\xi(x)||f(\xi)| \\ &\le C(\#\Xi)\|f\|_{L\infty(\mathbb{M})} \max_{\xi\in\Xi} \|b_\xi - \chi_\xi\|_{L\infty(\mathbb{M})}.\end{aligned}$$

where Lemma 4 guarantees that $\|b_\xi - \chi_\xi\|_{L\infty(\mathbb{M})}$ is as small as one likes depending on the "footprint" of b_ξ. Moreover the operators provide optimal L_∞ approximation orders when the Lebesgue constant is uniformly bounded (see Proposition 3). So, for example, it is shown in [14, Cor 5.9] that restricted surface spline interpolation satisfies $\|I_\Xi f - f\|_{L\infty(\mathbb{M})} \le Ch^\sigma$ for $f \in C^{2m}(\mathbb{S}^2)$ when $\sigma = 2m$ and $f \in B^\sigma_{\infty,\infty}(\mathbb{S}^2)$ for $\sigma \le 2m$. Thus Q_Ξ inherits the same rate of approximation.

The quasi-interpolation operator also provides two more useful approximation properties. The first deals with pointwise error estimates for continuous functions. In the early 1990s, Brown [3] showed that, for several classes of RBFs, if the density parameter h_Ξ decreased to zero for point sets Ξ in compact Ω, then

$$\mathrm{dist}_\infty(f, V_\Xi) \to 0$$

for any continuous function f. The argument given was nonconstructive. The next result gives pointwise error estimates when approximating an arbitrary continuous function f on $\mathbb{M}$ in terms of its modulus of continuity. The result is reminiscent of a similar one for univariate splines.

6.1 Approximation Rates Based on Local Smoothness

Recall that the global modulus of continuity of f is $\omega(f,t) := \max_{|x-y|\le t}|f(x) - f(y)|$, the modulus of continuity at x_0 is $\omega(f,t,x_0) := \max_{|x-x_0|\le t}|f(x) - f(x_0)|$ and Λ is the Lebesgue constant. The constants K and J are discussed in Sect. 4.

Theorem 5 *Assume the conditions and notation of Theorem 2 and Proposition 3 hold. Then for each $x_0 \in \mathbb{M}$, $f \in C(\mathbb{M})$ with $\|f\|_{L_\infty(\mathbb{M})} = 1$, the following hold.*

i) $|f(x_0) - I_\Xi f(x_0)| \le \max\{\Lambda\omega(f, Kh|\log h|), 2h^{J-1}\}$
ii) $\|f - I_\Xi f\|_{L_\infty(\mathbb{M})} \le \Lambda(K+1)\omega(f, h|\log h|)$
iii) $|f(x_0) - Q_\Xi f(x_0)| \le \max\{\Lambda\omega(f, Kh|\log h|, x_0), h^{J-2}\}$
iv) $\|f - Q_\Xi f\|_{L_\infty(\mathbb{M})} = O(\|f - I_\Xi f\|_{L_\infty(\mathbb{M})})$

Proof Note that

$$|f(x_0) - I_\Xi f(x_0)| \le \sum_{\xi\in\Xi}\big|f(x_0) - f(\xi)\big||\chi_\xi(x_0)| + C\|f\|_{L\infty(\mathbb{M})}|1 - \sum_{\xi\in\Xi}\chi_\xi(x_0)|.$$

By Corollary 1 below, $|1 - \sum_{\xi\in\Xi}\chi_\xi(x)| = \mathscr{O}(h^{2m})$. Let $B_{x_0} := B(x_0, Kh\ln h^{-1})$ and $B_{x_0}^C := \mathbb{M}\backslash B_{x_0}$. Then

$$\begin{aligned}
|f(x_0) - I_\Xi f(x_0)| \le & \sum_{\xi\in B_{x_0}\cap\Xi} |f(\xi) - f(x_0)|\,|\chi_\xi(x_0)| \\
& + \sum_{\xi\in B_{x_0}^C\cap\Xi} |f(x_0) - f(\xi)|\,|\chi_\xi(x_0)| + C\|f\|_{L_\infty(\mathbb{M})}h^{2m} \\
\le & \max_{\xi\in B_{x_0}\cap\Xi} |f(\xi) - f(x_0)| \sum_{\xi\in B_{x_0}\cap\Xi} |\chi_\xi(x_0)| \\
& \qquad + C\|f\|_{L\infty(\mathbb{M})}\Big(\sum_{\xi\in B_{x_0}^C\cap\Xi}\big(1 + \tfrac{\mathrm{dist}(x_0,\xi)}{h}\big)^{-J} + h^{2m}\Big) \\
\le & \Lambda\omega(f, Kh|\log h|)(x_0) + C\max(h^{J-1}, h^{2m})\|f\|_{L_\infty(\mathbb{M})}.
\end{aligned}$$

The second inequality follows from $\omega(f, Kt) \le (K+1)\omega(f,t)$ and the fact that if $\omega(f,t)/t \to 0$ as $t \to 0$, then f is a constant [5]. Inequality iii) follows from

$$\begin{aligned}
|f(x_0) - Q_\Xi f(x_0)| &\le |f(x_0) - I_\Xi f(x_0)| + |I_\Xi f(x_0) - Q_\Xi f(x_0)| \\
&\le |f(x_0) - I_\Xi f(x_0)| + \sum_{\xi\in\Xi} |f(\xi)|\,|\chi_\xi(x_0) - b_\xi(x_0)| \\
&\le |f(x_0) - I_\Xi f(x_0)| + \|f\|_{L\infty(\mathbb{M})}\sum_{\xi\in\Xi}\|\chi_\xi - b_\xi\|_\infty
\end{aligned}$$

$$\le \Lambda\omega(f, Kh|\log h|)(x_0) + Ch^J(\#\Xi)$$
$$\le \Lambda\omega(f, Kh|\log h|)(x_0) + C\rho^d h^{J-d}.$$

The last inequality is clear. □

We remark that the pointwise estimate in the first inequality above requires only continuity at a single point, and boundedness elsewhere.

6.2 Rates for Functions with Higher Smoothness

By the global boundedness of the Lebesgue constant, we know that interpolation is "near-best". Similarly, by Theorem 5 (iv), quasi-interpolation is near-best as well. In this subsection, we establish precise rates of decay $\mathrm{dist}_\infty(f, S(\Xi))$. This is established using an approximation scheme similar to the one employed in [6]—it uses the fact that the kernel is a fundamental solution for $\mathscr{L} = \sum_{j=0}^m (\nabla^j)^* \nabla^j$ (pointed out in Sect. 2.2) to obtain the identity $f(x) = \int_{\mathbb{M}} \mathscr{L}f(\alpha)\kappa_m(x,\alpha)\mathrm{d}\alpha$ for $f \in C^{2m}(\mathbb{M})$. As in [6], for every $\alpha \in \mathbb{M}$, we use a modified kernel $\tilde{\kappa}(x,\alpha)$ constructed from Ξ by taking $\tilde{\kappa}(\cdot,\alpha) \in S(\Xi)$, with coefficients depending continuously on α. We may then replace κ_m by $\tilde{\kappa}$ in the reproduction formula for f.

For $\alpha \in \mathbb{M}$, define Ξ_α as follows:

$$\Xi_\alpha := \begin{cases} \Xi \cup \{\alpha\}, & \mathrm{dist}(\alpha, \Xi) \ge h/2 \\ \Xi \cup \{\alpha\} \setminus \{\xi^*\}, & \mathrm{dist}(\alpha, \Xi) \le h/2 \end{cases}$$

where ξ^* is the nearest point of Ξ to α. For this point set, we have $h(\Xi_\alpha, \mathbb{M}) \le 3h/2$ and $q(\Xi_\alpha) \ge \min(q, h/2)$.

For every $\alpha \in \mathbb{M}$, we consider the Lagrange function $\lambda_\alpha \in S(\Xi_\alpha)$ centered at α. We can express this Lagrange function as $\lambda_\alpha = \sum_{\xi \in \Xi_\alpha} A_{\alpha,\xi}\kappa_m(\cdot,\xi)$. Let $a(\xi,\alpha) := -A_{\alpha,\xi}/A_{\alpha,\alpha}$ for $\xi \in \Xi_\alpha \setminus \{\alpha\}$. The approximation scheme is given by way of the operator

$$S_\Xi f := \sum_{\xi \in \Xi} c_\xi \kappa_m(\cdot,\xi)$$

with $c_\xi = \int_{\mathbb{M}} \mathscr{L}f(\alpha)a(\xi,\alpha)\mathrm{d}\alpha$.

This works because the kernel κ_m used in the reproduction of smooth f can be replaced by a modified kernel $\tilde{\kappa}(x,\alpha) = \sum_{\xi \in \Xi_\alpha, \xi \ne \alpha} a(\xi,\alpha)\kappa_m(\cdot,\xi)$, which is a linear combination of the original kernel sampled from Ξ_α. We measure the difference of the two kernels as:

$$\mathrm{err}(x,\alpha) := \kappa_m(x,\alpha) - \tilde{\kappa}(x,\alpha) = \kappa_m(x,\alpha) - \sum_{\substack{\xi \in \Xi_\alpha \\ \xi \ne \alpha}} a(\xi,\alpha)\kappa_m(\cdot,\xi) = \frac{1}{A_{\alpha,\alpha}}\lambda_\alpha(x).$$

To further control this error, we estimate $|A_{\alpha,\alpha}|$ from below. We do this by applying the zeros lemma for balls [14] on the set $B(\alpha, Mh)$ (for a sufficiently large constant M—a constant which depends only on $\mathbb{M}$ and m). Thus, we have

$$\begin{aligned}|\lambda_\alpha(\alpha)| &\le \|\lambda_\alpha\|_{L_\infty(B(\alpha,Mh))} \le C\big(Mh\big)^{m-d/2}\|\chi_\alpha\|_{W_2^m(\mathbb{M})}\\ &= Ch^{m-d/2}|\langle\chi_\alpha,\chi_\alpha\rangle|^{1/2}\end{aligned} \tag{23}$$

Replacing $\chi_\alpha(\alpha)$ with 1 and $\langle\chi_\alpha,\chi_\alpha\rangle$ with $|A_{\alpha,\alpha}|$, we have a lower bound for $|A_{\alpha,\alpha}|$. Namely, there is a constant $C > 0$ depending only on $m, \mathbb{M}$ so that

$$|A_{\alpha,\alpha}| \ge Ch^{d-2m} \tag{24}$$

Combining (23), (24) and the pointwise decay rates for the Lagrange functions, we obtain the bound

$$|\mathrm{err}(x,\alpha)| \le C\rho^{m-d/2}h^{2m-d}e^{-\nu\left(\frac{\mathrm{dist}(x,\alpha)}{h}\right)}. \tag{25}$$

At this point, we have the following result for approximation of smooth functions.

Theorem 6 *For $1 \le p < \infty$ and $f \in W_p^{2m}(\mathbb{M})$, or $f \in C^{2m}(\mathbb{M})$ when $p = \infty$, there is a constant $C < \infty$ depending only on m and $\mathbb{M}$ so that*

$$\|f - S_\Xi f\|_{L_p(\mathbb{M})} \le Ch^{2m}\|f\|_{W_p^{2m}(\mathbb{M})}$$

Proof The L_p error $\|f - S_\Xi f\|_p$ is controlled by the norm of the integral operator $\mathrm{Err} : g \mapsto \int_{\mathbb{M}} g(\alpha)|\mathrm{err}(\cdot,\alpha)|\mathrm{d}\alpha$, which has non-negative kernel $|\mathrm{err}(x,\alpha)|$. Indeed, we have $|f(x) - S_\Xi f(x)| \le \int_{\mathbb{M}} |\mathscr{L}f(\alpha)|\,|\mathrm{err}(x,\alpha)|\,\mathrm{d}\alpha$, so

$$\|f - S_\Xi f\|_{L_p(\mathbb{M})} \le \|\mathscr{L}f\|_{L_p(\mathbb{M})}\|\mathrm{Err}\|_{L_p\to L_p}.$$

We estimate the norm of this operator on L_1 and L_∞—the L_p result then follows by interpolation. In other words, we estimate $\|\mathrm{Err}\|_{1\to1} \le \max_{\alpha\in\mathbb{M}}\int_{\mathbb{M}}|\mathrm{err}(x,\alpha)|\mathrm{d}x$ and $\|\mathrm{Err}\|_{\infty\to\infty} \le \max_{x\in\mathbb{M}}\int_{\mathbb{M}}|\mathrm{err}(x,\alpha)|\mathrm{d}\alpha$. Using (25) and symmetry, both are bounded by

$$C\rho^{m-d/2}h^{2m-d}\max_{\alpha\in\mathbb{M}}\int_{\mathbb{M}} e^{-\nu\left(\frac{\mathrm{dist}(x,\alpha)}{h}\right)}\mathrm{d}x \le C\rho^{m-d/2}h^{2m} \tag{26}$$

and the theorem follows. □

A result for lower smoothness is also possible. Let us define the Besov space $B^\sigma_{p,\infty}(\mathbb{M})$ as a real interpolation space between $L_p(\mathbb{M})$ and $W_p^{2m}(\mathbb{M})$. Let $B^\sigma_{p,\infty}(\mathbb{M})$

be the set of (equivalence classes) of functions $f \in L_p(\mathbb{M})$ for which the expression

$$\|f\|_{B^\sigma_{p,\infty}(\mathbb{M})} := \sup_{t>0} t^{-\sigma/2m} \inf_{g \in W_p^{2m}(\mathbb{M})} \left(\|f-g\|_{L_p(\mathbb{M})} + t\|g\|_{W_p^{2m}(\mathbb{M})}\right) \tag{27}$$

is finite. (When $p = \infty$, we replace $L_p(\mathbb{M})$ by $C(\mathbb{M})$ and $W_p^{2m}(\mathbb{M})$ by $C^{2m}(\mathbb{M})$.) That this is a Banach space and the above is a norm can be found in [1, 5] or [19]. We note in particular that [19] shows this definition is equivalent to other standard, intrinsic constructions of Besov spaces on manifolds, and relates these to the Sobolev scale and other families of smoothness spaces. Of special interest is the case of the Hölder spaces with fractional exponent: $C^\sigma(\mathbb{M}) = B^\sigma_{\infty,\infty}(\mathbb{M})$.

Theorem 7 *Let $f \in B^\sigma_{p,\infty}(\mathbb{M})$ for $1 \le p \le \infty$ and $0 < \sigma \le 2m$. Then we have*

$$\mathrm{dist}_{p,\mathbb{M}}(f, S(\Xi)) \le Ch^\sigma \|f\|_{B^\sigma_{p,\infty}(\mathbb{M})}.$$

Proof The follows from a standard K-functional argument, by splitting $f = g + (f-g)$, with $g \in W_p^{2m}(\mathbb{M})$ (or $C^{2m}(\mathbb{M})$) and $f-g \in L_p(\mathbb{M})$ (or $C(\mathbb{M})$). In particular, for $h > 0$, set $t = h^{2m}$. and find g so that

$$\|f-g\|_{L_p(\mathbb{M})} + t\|g\|_{W_p^{2m}(\mathbb{M})} \le 2t^{\sigma/2m}\|f\|_{B^\sigma_{p,\infty}(\mathbb{M})}.$$

This ensures that

$$\|f-g\|_{L_p(\mathbb{M})} \le 2h^\sigma \|f\|_{B^\sigma_{p,\infty}(\mathbb{M})} \quad \text{and} \quad \|g\|_{W_p^{2m}(\mathbb{M})} \le 2h^{\sigma-2m}\|f\|_{B^\sigma_{p,\infty}(\mathbb{M})}.$$

Finally, we take $S_\Xi g$ as our approximant to f, obtaining the desired result by applying the triangle inequality and Theorem 6. □

A drawback of the previous results in this section is that the approximation scheme S_Ξ is not easy to implement. The good news is that the stability of the schemes I_Ξ, Q_Ξ and T_Ξ imply that these operators inherit the same convergence rate. This is a consequence of the Lebesgue constants being bounded (Proposition 3) and the small error between I_Ξ and Q_Ξ.

Corollary 1 *There exists a constant $C > 0$ so that for $0 < \sigma \le 2m$ and $f \in C^\sigma(\mathbb{M})$, we have*

$$\|f - I_\Xi f\|_{L_\infty(\mathbb{M})} \le Ch^\sigma \|f\|_{C^\sigma(\mathbb{M})} \text{ and} \|f - Q_\Xi f\|_{L_\infty(\mathbb{M})} \le Ch^\sigma \|f\|_{C^\sigma(\mathbb{M})}.$$

For $0 < \sigma \le 2m$, $1 \le p \le \infty$ and $f \in B^\sigma_{p,\infty}(\mathbb{M})$ or $f \in W_p^{2m}(\mathbb{M})$ when $\sigma = 2m$, we have

$$\|f - T_\Xi f\|_{L_p(\mathbb{M})} \le Ch^\sigma \|f\|_{B^\sigma_{p,\infty}(\mathbb{M})},$$

where T_Ξ is the least-squares projector defined in (22).

6.3 Approximation on Bounded Regions

As a final note, we observe that the approximation power of spaces $S(X)$ on $\mathbb{M}$, where X is dense in $\mathbb{M}$, extends to the setting of approximation over a compact domain $\Omega \subset \mathbb{M}$ having a Lipschitz boundary and satisfying Assumption 1, using $\tilde{V}_{\Xi}$, with Ξ dense in the union of Ω and an "annulus" around Ω.

Our final result shows that optimal L_∞ approximation rates, when approximating a smooth function f on Ω, can be obtained from data sites either inside or "close" to Ω. The result illustrates the local nature of the basis $\{b_\xi\}$.

Let $f \in C^\sigma(\Omega)$, where $\sigma > 0$ is an integer, and let $\tilde{f} \in C^\sigma(\mathbb{M})$ be a smooth extension of f to $\mathbb{M}$, i.e., $\tilde{f}|_\Omega = f|_\Omega$. Suppose that $\mathscr{A} = \{x \in \mathbb{M}\backslash\Omega : \operatorname{dist}(x,\Omega) \le Kh|\log h|\}$ and that Ξ is a finite set contained in $\Omega \cup \mathscr{A}$, with fill distance h. In addition, let $\widetilde{\Xi}$ be a quasi-uniform extension of Ξ to all of $\mathbb{M}$, as given in Lemma 2. Finally, let κ_m be a kernel as described in Sect. 2.2 with associated spaces

$$\tilde{V}_\Xi = \operatorname{span}_{\xi\in\Xi} b_\xi \quad\text{and}\quad \tilde{V}_{\widetilde{\Xi}} = \operatorname{span}_{\xi\in\widetilde{\Xi}} b_\xi .$$

Theorem 8 *If* $\sigma \le 2m$, *then* $\operatorname{dist}_{\infty,\Omega}(f,\tilde{V}_X) \sim \operatorname{dist}_{\infty,\mathbb{M}}(f_e,\tilde{V}_{\widetilde{X}}) \le \begin{cases} Ch^\sigma \|f\|_{C^\sigma(\Omega)} \\ Ch^\sigma \|f_e\|_{C^\sigma(\mathbb{M})}. \end{cases}$

Proof By the global boundedness of the Lebesgue constant Λ, we know that interpolation is near-best approximation. Similarly, by Theorem 5(iv), quasi-interpolation is near-best approximation as well. Hence, with $J = K\nu/2 - 2m + d$, we have that

$$\begin{aligned}
\max_{x\in\Omega} |f(x) - \sum_{\xi\in\Xi} \tilde{f}(\xi) b_\xi(x)| &\le \max_{x\in\Omega} |f(x) - \sum_{\xi\in\Xi_e} f_e(\xi) b_\xi(x)| \\
&\quad + \max_{x\in\Omega} \sum_{\xi\in\widetilde{\Xi}\backslash\Xi} |\tilde{f}(\xi) b_\xi(x)| \\
&\le \max_{x\in\Omega} |f(x) - \sum_{\xi\in\widetilde{\Xi}} \tilde{f}(\xi) b_\xi(x)| + \|f\|_\infty \sum_{\xi\in\widetilde{\Xi}\backslash\Xi} \left(1 + \tfrac{\operatorname{dist}(x_0,\xi)}{h}\right)^{-J} \\
&\le \max_{x\in\Omega} |f(x) - \sum_{\xi\in\widetilde{\Xi}} \tilde{f}(\xi) b_\xi(x)| + \|f\|_\infty h^{J-1}
\end{aligned}$$

where h^J can be chosen small compared to the first term, because, by (12), the parameter K in J can be chosen large enough for this to happen. The theorem then follows from Corollary 1. □

We remark that one only needs to have *local* information in a small annulus outside Ω to obtain full approximation order. Moreover, as previously discussed, approximation order on manifolds is often known.

7 Volume Comparisons

Proposition 5 *We assume that* $\mathbb{M}$ *is a closed, compact, connected d-dimensional* C^∞ *Riemannian manifold. There exist constants* $0 < \alpha_\mathbb{M} < \beta_\mathbb{M} < \infty$ *so that any ball* $B(x,r)$ *satisfies*

$$\alpha_\mathbb{M} r^d \le \mathrm{vol}(B(p,r)) \le \beta_\mathbb{M} r^d \tag{28}$$

for all $0 \le r \le \mathrm{d}_\mathbb{M}$.

Proof By Property 3, $\mathbb{M}$ has bounded geometry, so the Ricci curvature, Ric, is bounded below. Hence, there is a $k \in \mathbb{R}$ such that Ric $\ge (d-1)k$. Let $\mathbb{M}_k^d$ denote one of the canonical manifolds (sphere, $\mathbb{R}^d$, hyperbolic space) having constant sectional curvature k. In addition, let $p \in \mathbb{M}$, $\tilde{p} \in \mathbb{M}_k^d$, $V_r := \mathrm{vol}(B(p,r))$ and $V_r^k := \mathrm{vol}(B^k(\tilde{p},r))$. The Bishop-Gromov Comparison theorem states that the ratio V_r/V_r^k is non increasing and, as $r \downarrow 0$, $V_r/V_r^k \to 1$, no matter which $p, \tilde{p}$ are chosen. Since Ric may become negative, we can handle all of the cases at once by assuming that $k < 0$, which means that $\mathbb{M}_k^d$ is a hyperbolic space.

The model that we take for $\mathbb{M}_k^d$ will be the Poincaré ball, so that $\mathbb{M}_k^d = \{x = (x^1, \dots, x^d) \in \mathbb{R}^d \mid \|x\|_2^2 < -4/k\}$. Let $A := 1 + (k/4)\|x\|_2^2$. In these coordinates, the Riemannian metric is given by $g_{jk} = \delta_{jk}/A^2$; equivalently, $ds^2 = \sum_{j=1}^d (dx^j)^2/A^2$. We want to introduce geodesic normal coordinates, centered at $x^j = 0, j = 1, \dots, d$. Let $t \ge 0$ and set $x^j = \frac{2}{\sqrt{|k|}} \tanh(\sqrt{|k|}t/2)\xi^j$, where $\xi = (\xi^1, \dots, \xi^d) \in \S^{d-1}$. A straightforward computation shows that

$$ds^2 = dt^2 + \frac{1}{|k|}\sinh^2(\sqrt{|k|}t)ds^2_{\S^{d-1}}, \tag{29}$$

where t the length of the geodesic joining the origin to $x^j = t\xi^j$. It follows that the volume element in these coordinates is

$$d\mu_k = \frac{1}{\sqrt{|k|}^{d-1}} \sinh^{d-1}(\sqrt{|k|}t)dtd\mu_{\S^{d-1}}, \tag{30}$$

and, consequently,

$$V_r^k = \mathrm{vol}(B^k(\tilde{p},r)) = \frac{1}{\sqrt{|k|}^{d-1}}\omega_{d-1}\int_0^r \sinh^{d-1}(\sqrt{|k|}t)dt. \tag{31}$$

We will need bounds on V_r^k for $r \le R$, where R is fixed. These are easy to obtain, since $1 \le \frac{\sinh(x)}{x} \le \frac{\sinh(X)}{X}$ for all $0 \le x \le X$. Just take $X = \sqrt{|k|}R$:

$$1 \le \left(\frac{\sinh(\sqrt{|k|}t)}{\sqrt{|k|}t}\right)^{d-1} \le \left(\frac{\sinh(\sqrt{|k|}R)}{\sqrt{|k|}R}\right)^{d-1} := \beta_{d,k,R}.$$

Multiplying both sides by $\omega_{d-1}t^{d-1}$ and integrating results in this inequality:

$$\frac{\omega_{d-1}}{d}r^d \leq \int_0^r \omega_{d-1}\left(\frac{\sinh(\sqrt{|k|}t)}{\sqrt{|k|}t}\right)^{d-1} t^{d-1}dt \leq \beta_{d,k,R}\frac{\omega_{d-1}}{d}r^d.$$

Using this in (31) results in

$$\frac{\omega_{d-1}}{d}r^d \leq V_r^k \leq \beta_{d,k,R}\frac{\omega_{d-1}}{d}r^d. \tag{32}$$

We can now employ the Bishop-Gromov Theorem to obtain (1). Since V_r/V_r^k is non increasing and tends to 1 as $r \downarrow 0$, we have that $V_r \geq V_r^k \geq \frac{\omega_{d-1}}{d}r^d$. Also, we have that $V_r/V_r^k \leq V_{\mathrm{d}_{\mathbb{M}}}/V_{\mathrm{d}_{\mathbb{M}}}^k$. Thus, $V_r \leq \big(V_{\mathrm{d}_{\mathbb{M}}}/V_{\mathrm{d}_{\mathbb{M}}}^k\big)V_r^k$. Employing (32) in conjunction with these inequalities yields

$$\frac{\omega_{d-1}}{d}r^d \leq V_r \leq \beta_{d,k,\mathrm{d}_{\mathbb{M}}}\frac{\omega_{d-1}}{d}\big(V_{\mathrm{d}_{\mathbb{M}}}/V_{\mathrm{d}_{\mathbb{M}}}^k\big)r^d.$$

We want to refine this. To do that, we begin by observing that $\overline{B(p,\mathrm{d}_{\mathbb{M}})} = \mathbb{M}$, because no point in $\overline{B(p,\mathrm{d}_{\mathbb{M}})}$ is at a distance from p greater than the diameter $\mathrm{d}_{\mathbb{M}}$; thus, $V_{\mathrm{d}_{\mathbb{M}}} = \mathrm{vol}(\mathbb{M})$. Next, by (32), $V_{\mathrm{d}_{\mathbb{M}}}^k \geq \frac{\omega_{d-1}}{d}\mathrm{d}_{\mathbb{M}}^d$. Finally, using these in the inequality above yields

$$\frac{\omega_{d-1}}{d}r^d \leq V_r \leq \beta_{d,k,\mathrm{d}_{\mathbb{M}}}\mathrm{d}_{\mathbb{M}}^{-d}\mathrm{vol}(\mathbb{M})r^d,$$

from which (1) follows with $\alpha_{\mathbb{M}} = \frac{\omega_{d-1}}{d}$ and $\beta_{\mathbb{M}} = \beta_{d,k,\mathrm{d}_{\mathbb{M}}}\mathrm{d}_{\mathbb{M}}^{-d}\mathrm{vol}(\mathbb{M})$. □

Acknowledgements Thomas Hangelbroek is supported by grant DMS-1413726 from the National Science Foundation. Francis Narcowich and Joseph Ward are supported by grant DMS-1514789 from the National Science Foundation. Christian Rieger is supported by (SFB) 1060 of the Deutsche Forschungsgemeinschaft.

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Pure and Applied Mathematics (Amsterdam), vol. 140, 2nd edn. Elsevier/Academic, Amsterdam (2003)
2. Aubin, T.: Nonlinear analysis on manifolds. Monge-Ampère equations. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 252. Springer, New York (1982)
3. Brown, A.L.: Uniform Approximation by Radial Basis Functions. In: Light, W.A. (ed.) Advances in Numerical Analysis, vol. II, pp. 203–206. Oxford University Press, Oxford (1992)
4. Cheeger, J., Gromov, M., Taylor, M.: Finite propagation speed, kernel estimates for functions of the Laplace operator, and the geometry of complete Riemannian manifolds. J. Differ. Geom. **17**(1), 15–53 (1982)

5. DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 303. Springer, Berlin (1993)
6. Devore, R., Ron, A.: Approximation using scattered shifts of a multivariate function. Trans. Am. Math. Soc. **362**(12), 6205–6229 (2010)
7. do Carmo, M.P.: Riemannian Geometry. Mathematics: Theory and Applications. Birkhäuser, Boston, MA (1992). Translated from the second Portuguese edition by Francis Flaherty
8. Eschenburg, J.H.: Comparison theorems and hypersurfaces. Manuscripta Math. **59**(3), 295–323 (1987)
9. Fuselier, E., Hangelbroek, T., Narcowich, F.J., Ward, J.D., Wright, G.B.: Localized bases for kernel spaces on the unit sphere. SIAM J. Numer. Anal. **51**(5), 2538–2562 (2013)
10. Griebel, M., Rieger, C., Zwicknagl, B.: Multiscale approximation and reproducing kernel Hilbert space methods. SIAM J. Numer. Anal. **53**(2), 852–873 (2015)
11. Grove, K.: Metric differential geometry. In: Differential Geometry (Lyngby, 1985). Lecture Notes in Mathematics, vol. 1263, pp. 171–227. Springer, Berlin (1987)
12. Hangelbroek, T., Narcowich, F.J., Ward, J.D.: Kernel approximation on manifolds I: bounding the Lebesgue constant. SIAM J. Math. Anal. **42**(4), 1732–1760 (2010)
13. Hangelbroek, T., Narcowich, F.J., Sun, X., Ward, J.D.: Kernel approximation on manifolds II: the L_∞ norm of the L_2 projector. SIAM J. Math. Anal. **43**(2), 662–684 (2011)
14. Hangelbroek, T., Narcowich, F.J., Ward, J.D.: Polyharmonic and related kernels on manifolds: interpolation and approximation. Found. Comput. Math. **12**(5), 625–670 (2012)
15. Hangelbroek, T., Narcowich, F.J., Rieger, C., Ward, J.D.: An inverse theorem for compact Lipschitz regions in R^d using localized kernel bases. Math. Comput. (2017). https://doi.org/10.1090/mcom/3256
16. Mhaskar, H.N., Narcowich, F.J., Prestin, J., Ward, J.D.: L^p Bernstein estimates and approximation by spherical basis functions. Math. Comput. **79**(271), 1647–1679 (2010)
17. Narcowich, F.J., Ward, J.D., Wendland, H.: Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. Constr. Approx. **24**(2), 175–186 (2006)
18. Rieger, C.: Sampling inequalities and applications. Ph.D. thesis, Universität Göttingen (2008)
19. Triebel, H.: Theory of Function Spaces. II. Monographs in Mathematics, vol. 84. Birkhäuser, Basel (1992)
20. Ward, J.P.: L^p Bernstein inequalities and inverse theorems for RBF approximation on $\mathbb{R}^d$. J. Approx. Theory **164**(12), 1577–1593 (2012)

A Discrete Collocation Method for a Hypersingular Integral Equation on Curves with Corners

Thomas Hartmann and Ernst P. Stephan

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract This paper is devoted to the approximate solution of a hypersingular integral equation on a closed polygonal boundary in $\mathbb{R}^2$. We propose a fully discrete method with a trial space of trigonometric polynomials, combined with a trapezoidal rule approximation of the integrals. Before discretization the equation is transformed using a nonlinear (mesh grading) parametrization of the boundary curve which has the effect of smoothing out the singularities at the corners and yields fast convergence of the approximate solutions. The convergence results are illustrated with some numerical examples.

1 Introduction

In this paper we consider the hypersingular integral equation

$$Wu(z) := \frac{1}{\pi} \frac{\partial}{\partial n_z} \int_{\Gamma} \frac{\partial}{\partial n_\xi} \log |z - \xi| u(\xi) \, ds_\xi = f(z) \tag{1}$$

with additional constraint

$$\int_{\Gamma} u(\xi) \, ds_\xi = 0. \tag{2}$$

T. Hartmann
Hochschule Ulm, Ulm, Germany
e-mail: hartmann@hs-ulm.de

E. P. Stephan (✉)
Institut für Angewandte Mathematik, Leibniz Universität Hannover, Hannover, Germany
e-mail: stephan@ifam.uni-hannover.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_25

where Γ denotes the boundary of a simply connected bounded domain Ω in $\mathbb{R}^2$. Equation (1) arises in solving the Neumann problem for Laplace's equation on Ω, using a double layer potential ansatz. Hypersingular integral equations arise in the context of acoustic wave scattering by thin screens and elastic wave scattering by cracks. Plane elastostatic crack problems can be described and solved in terms of hypersingular integral equations, where the crack opening displacements are the unknown functions. The results can be used to derive the crack tip stress intensity factors, which are essential in fracture analysis. For the derivation of these equations we refer to [1] and [12]. More applications of hypersingular integral equations can be found in [13].

In the paper [6] results on stability and optimal convergence for spline collocation methods of arbitrarily high order for Symm's weakly singular integral equation are available for polygonal Γ. Here the mesh grading transformation method has been applied to obtain a rapidly convergent numerical method (see also [10, 14]) Analogous results for a fully discrete version of the method in [6] are obtained in [7]. A discrete collocation method using trigonometric polynomials as trial functions is investigated for Symm's integral equation in [8] and high order of convergence can be achieved provided the given function f is smooth. In this paper we extend this method to a hypersingular integral equation.

The paper is organized as follows. In Sect. 2 we introduce the mesh grading transformation and give a complete ellipticity and solvability analysis of the mesh grading transformed hypersingular integral equation in an appropriate Sobolev space setting. In Sect. 3 we introduce a trigonometric collocation method and present an error analysis. In Sect. 4 we introduce and analyze a corresponding discrete collocation method. We show that both methods converge with a rate as high as justified by the (finite) order of mesh grading and the regularity of the given data. Following the spirit of the results of [6] and [8] we only prove stability if the approximate solution is cut off by zero over some number of intervals near each corner. This is typical when Mellin convolution operators are discretized (see also [9]). This modification of the collocation method seems not to be needed in practice. The numerical examples presented in Sect. 5 show that the method with no cut-offs appears to be perfectly stable.

2 Properties of the Transformed Integral Equation

We assume Γ is an (infinitely) smooth curve with the exception of a corner at a point x_0. In the neighbourhood of x_0 the curve Γ should consist of two straight lines intersecting with an interior angle ω. We make this restriction, because we apply Mellin techniques (see [4, 9]) The extension to curves with more than one corner is straightforward, see [6, 7].

From [4] we know that the operator W in (1) is bijective from $\overset{\circ}{H}{}^{1/2}(\Gamma) := \{u \in H^{1/2}(\Gamma) \mid \int_\Gamma u = 0\}$ to $H^{-1/2}(\Gamma)$ where $H^{1/2}(\Gamma)$ is the trace space and $H^{-1/2}(\Gamma)$ is its dual.

In the following we use for the interior angle the representation $\omega = (1-\chi)\pi$, i.e. we have $0 < |\chi| < 1$. As in [8] we rewrite (1) using an appropriate nonlinear parametrization $\gamma : [0,1] \to \Gamma$ which varies more slowly than the arc-length parametrization in the vicinity of x_0. We take a parametrization $\gamma_0 : [0,1] \to \Gamma$ such that $\gamma_0(0) = \gamma_0(1) = x_0$ and $|\gamma_0'(s)| > 0$ for all $0 < s < 1$. With a grading exponent $q \in \mathbb{N}$ and selecting a function ν such that

$$\nu \in C^\infty[0,1], \quad \nu(0) = 0, \quad \nu(1) = 1, \quad \nu'(s) > 0, \quad 0 < s < 1, \tag{3}$$

we define the mesh grading transformation

$$\gamma(s) = \gamma_0(\tilde{\omega}(s)), \tag{4}$$

where

$$\tilde{\omega}(s) = \frac{\nu^q(s)}{\nu^q(s) + \nu^q(1-s)}.$$

The parametrization γ we have chosen is graded with exponent q near the corner. Using the change of variables $x = \gamma(s)$, $\xi = \gamma(\sigma)$, and multiplying by $|\gamma'(s)|$ Eq. (1) becomes

$$Kw(s) := \frac{1}{\pi}\frac{\partial}{\partial n_x}\int_0^1 \frac{\partial}{\partial n_\xi} \log|x-\xi|\;|\gamma'(\sigma)||\gamma'(s)|w(\sigma)\,d\sigma = g(s),$$
$$s \in [0,1], \tag{5}$$

where

$$w(\sigma) = u(\gamma(\sigma)), \qquad g(s) = |\gamma'(s)|f(\gamma(s)). \tag{6}$$

As shown in Theorem 2 below the solution w of the transformed equation (5) may be made as smooth as desired on $[0,1]$ provided f is smooth and the grading exponent is sufficiently large, and w can be optimally approximated using trigonometric polynomials as basis functions.

Now let us look more closely at the behaviour of K near the corner. Without loss of generality we assume that the corner is located at $x_0 = 0$. Then the parametrization (4) (possibly after rotation) takes the form with $\varepsilon > 0$, sufficiently small,

$$\gamma(s) = \begin{cases} (-s)^q(-\cos\chi\pi, \sin\chi\pi) & s \in (-\varepsilon, 0) \\ s^q(1,0) & s \in (0,\varepsilon) \end{cases}. \tag{7}$$

In the following lemma we give a representation of the operator K and its kernel. These results will be used below to show that the operator K is bijective between suitably chosen spaces (Theorem 1). Furthermore the representation (11) will be used in the proof of Theorem 3.

Lemma 1 *Let ψ be a 1-periodic non-negative cut-off function such that $\psi \equiv 1$ in some neighbourhood of $x_0 = 0$. Then we have*

$$\psi K \psi = \psi C \psi + E,$$

where E is compact from $\overset{\circ}{H}{}^1(\mathbb{R})$ to $\overset{\circ}{H}{}^0(\mathbb{R})$, and where C is given by ($v \in \overset{\circ}{H}{}^1(\mathbb{R})$):

$$Cv(s) = \int_{-\infty}^{\infty} c(s,\sigma)v(\sigma)\, d\sigma \tag{8}$$

with

$$c(s,\sigma) = \frac{1}{\pi}\begin{cases} \frac{q^2\sigma^{q-1}s^{q-1}}{(\sigma^q - s^q)^2} & s > 0, \sigma > 0 \\ \frac{q^2\sigma^{q-1}(-s)^{q-1}[\sigma^{2q}\cos\chi\pi + (-s)^{2q}\cos\chi\pi + 2\sigma^q(-s)^q]}{[\sigma^{2q} + 2\sigma^q(-s)^q\cos\chi\pi + (-s)^{2q}]^2} & s < 0, \sigma > 0 \\ \frac{q^2(-\sigma)^{q-1}s^{q-1}[(-\sigma)^{2q}\cos\chi\pi + s^{2q}\cos\chi\pi + 2(-\sigma)^q s^q]}{[(-\sigma)^{2q} + 2(-\sigma)^q s^q\cos\chi\pi + s^{2q}]^2} & s > 0, \sigma < 0 \\ \frac{q^2(-\sigma)^{q-1}(-s)^{q-1}}{((-\sigma)^q - (-s)^q)^2} & s < 0, \sigma < 0 \end{cases}. \tag{9}$$

Further there holds $c(s,\sigma) = D_\sigma \widetilde{k}(s,\sigma)$ with

$$\widetilde{k}(s,\sigma) = \frac{q}{\pi}\begin{cases} -\frac{s^{q-1}}{\sigma^q - s^q} & s > 0, \sigma > 0 \\ \frac{(-s)^{2q-1} + (-s)^{q-1}\sigma^q\cos\chi\pi}{\sigma^{2q} + 2\sigma^q(-s)^q\cos\chi\pi + (-s)^{2q}} & s < 0, \sigma > 0 \\ -\frac{s^{2q-1} + s^{q-1}(-\sigma)^q\cos\chi\pi}{\sigma^{2q} + 2(-\sigma)^q s^q\cos\chi\pi + (-s)^{2q}} & s > 0, \sigma < 0 \\ \frac{(-s)^{q-1}}{(-\sigma)^q - (-s)^q} & s < 0, \sigma < 0 \end{cases} \tag{10}$$

and $c(s,\sigma) = D_\sigma D_s \widehat{k}(s,\sigma)$ with

$$\widehat{k}(s,\sigma) = \frac{1}{\pi}\begin{cases} \log|\sigma^q - s^q| & s > 0, \sigma > 0 \\ \log|\sigma^{2q} + 2\sigma^q(-s)^q\cos\chi\pi + (-s)^{2q}| & s < 0, \sigma > 0 \\ \log|\sigma^{2q} + 2(-\sigma)^q s^q\cos\chi\pi + (-s)^{2q}| & s > 0, \sigma < 0 \\ \log|(-\sigma)^q - (-s)^q| & s < 0, \sigma < 0 \end{cases}. \tag{11}$$

Locally we have $\psi C D^{-1}\psi = -\psi\widetilde{K}\psi$ with

$$\widetilde{K}v(s) = \int_{-\infty}^{\infty} \widetilde{k}(s,\sigma)v(\sigma)\, d\sigma, \tag{12}$$

where $D^{-1}\psi$ denotes the antiderivative of ψ.

Proof The kernel $k(s,\sigma)$ of K in (5) can be written as $k(s,\sigma) = k_1(s,\sigma) + k_2(s,\sigma)$ with [11]

$$k_1(s,\sigma) := -\frac{1}{\pi}\frac{\langle n(x), n(\xi)\rangle}{|x-\xi|^2}|\gamma'(s)|\,|\gamma'(\sigma)|,$$

$$k_2(s,\sigma) := \frac{2}{\pi}\frac{\langle x-\xi, n(x)\rangle\langle x-\xi, n(\xi)\rangle}{|x-y|^4}|\gamma'(s)|\,|\gamma'(\sigma)|.$$

Using

$$n(\gamma(s)) = -i\gamma'(s)/|\gamma'(s)| \tag{13}$$

and (7), we derive (9) which yields (10) and (11).

The assertion (12) can be seen by integration by parts as follows.

$$CD^{-1}v(s) = \int_{-\infty}^{\infty} c(s,\sigma)D_\sigma^{-1}v(\sigma)\,d\sigma = \int_{-\infty}^{\infty} D_\sigma\widetilde{k}(s,\sigma)D_\sigma^{-1}v(\sigma)\,d\sigma$$

$$= -\int_{-\infty}^{\infty} \widetilde{k}(s,\sigma)v(\sigma)\,d\sigma = -\widetilde{K}v(s).$$

□

In the proof of Theorem 1 we will use results from [6] concerning the operator $\widetilde{K}$ in (12). It was shown in [6] that $\widetilde{K}$ acting on $L^2(\mathbb{R})$ can be identified with 2-by-2-matrices of Mellin convolution operators acting on $L^2(\mathbb{R}^+) \times L^2(\mathbb{R}^+)$. The calculation of the symbol of the Mellin convolution operator and using localisation techniques give more precise mapping properties of $\widetilde{K}$. Before we can use the results of [6] to analyse the operator K in (5) we introduce some more notation.

Let $I = (0,1)$ and let $H^t(I), t \in \mathbb{R}$, be the usual Sobolev space of 1-periodic functions (distributions) on the real line. Its norm is given by

$$\|v\|_t^2 = |\hat{v}(0)|^2 + \sum_{m\neq 0} |m|^{2t}|\hat{v}(m)|^2,$$

where the Fourier coefficients of v are defined by

$$\hat{v}(m) = (v, e^{i2\pi ms}) = \int_0^1 v(s)e^{-i2\pi ms}\,ds.$$

Further we define the Sobolev space

$$\overset{\circ}{H}{}^s(I) := \{u \in H^s(I) \mid \int_0^1 u(s)ds = 0\}.$$

In order to investigate the mapping properties of K in $\overset{\circ}{H}{}^s(I)$ we introduce

$$Aw(s) = -2\int_0^1 \log|2e^{-1/2}\sin(\pi(s-\sigma))|w(\sigma)\,d\sigma \tag{14}$$

$$= -2\int_0^1 \log|2\sin(\pi(s-\sigma))|w(\sigma)\,d\sigma + \int_0^1 w(\sigma)\,d\sigma =: Vw(s) + \mathscr{J}w.$$

The operator A takes the form [2]

$$Av(s) = \sum_{m\in\mathbb{Z}} \frac{\widehat{v}(m)}{\max(1,|m|)} e^{i2\pi ms}, \tag{15}$$

and that A is an isomorphism of $H^t(I)$ to $H^{t+1}(I)$ for any real t. Its inverse is given by

$$A^{-1} = -\mathscr{H}D + \mathscr{J} = -D\mathscr{H} + \mathscr{J}, \tag{16}$$

where $Dv(s) = v'(s)$ is the periodic differentiation operator, $\mathscr{J}v(s) = \widehat{v}(0)$ and $\mathscr{H}$ denotes the Hilbert transform

$$\mathscr{H}v(s) = -\int_0^1 \cot(\pi(s-\sigma))v(\sigma)\,d\sigma, \tag{17}$$

and we have

$$\mathscr{H}v(s) = 0 \qquad \text{for} \qquad v(s) = const. \tag{18}$$

Using the Fourier coefficients we obtain the representation

$$A^{-1}v(s) = \sum_{m\in\mathbb{Z}} \max(1,|m|)\widehat{v}(m)e^{i2\pi ms}. \tag{19}$$

Further we need the well-known formulas

$$DVw(s) = \mathscr{H}w(s), \tag{20}$$

where V is defined in (14), and

$$\mathscr{H}^2 = -I + \mathscr{J}. \tag{21}$$

We decompose the hypersingular integral operator in (5) as

$$Kw = A^{-1}w + Bw = g \tag{22}$$

with

$$Bw(s) = Kw(s) - A^{-1}w(s) = \int_0^1 b(s,\sigma)w(\sigma)\, d\sigma. \tag{23}$$

The kernel function b is 1-periodic in both variables and C^∞ for $0 < s, \sigma < 1$, but different from the case of a smooth curve Γ it has fixed singularities at the four corners of the square $[0, 1] \times [0, 1]$.

First we derive from (16), (21) and (18):

$$A^{-2} = (-D\mathscr{H} + \mathscr{J})A^{-1} = -D\mathscr{H}(-\mathscr{H}D + \mathscr{J}) + \mathscr{J}A^{-1}$$

$$= D\mathscr{H}^2D - D\mathscr{H}\mathscr{J} + \mathscr{J}A^{-1} = -D^2 + D\mathscr{J}D + \mathscr{J}A^{-1} = -D^2 + \mathscr{J}A^{-1},$$

and, using (19) and the fact that

$$\mathscr{J}\mathscr{A}^{-1}v = 0 \qquad \text{for} \qquad v \in \overset{\circ}{H}{}^s(I), \tag{24}$$

we have

$$A^{-2}v = -D^2v \qquad \text{for} \qquad v \in \overset{\circ}{H}{}^s(I). \tag{25}$$

Thus we obtain from (22) with $M := A^{-1}B$ and $e := A^{-1}g$ for all $w \in \overset{\circ}{H}{}^s(I)$:

$$A^{-1}Kw = A^{-1}(A^{-1} + B)w = -D^2w + A^{-1}Bw = -D^2w + Mw = e \tag{26}$$

Theorem 1 *The operators $A^{-1}K : \overset{\circ}{H}{}^1(I) \to \overset{\circ}{H}{}^{-1}(I)$ and $K : \overset{\circ}{H}{}^1(I) \to \overset{\circ}{H}{}^0(I)$ are continuously invertible, and we have the strong ellipticity estimate*

$$\Re\langle (A^{-1}K + T_1)v, v\rangle \geq C\,\|Dv\|_0^2 \qquad v \in \overset{\circ}{H}{}^1(I), \tag{27}$$

with some compact operator T_1 mapping from $\overset{\circ}{H}{}^1(I)$ to $\overset{\circ}{H}{}^{-1}(I)$ and a constant $C > 0$.

Proof First we consider KD^{-1}. This operator (up to a compact perturbation) can be represented by a Mellin convolution operator $\widetilde{K}$ near the corner (cf. Lemma 1). From [6] we know that the Mellin symbol of $\widetilde{K}$ does not vanish. Hence a regularizer of $\widetilde{K}$ as well as of KD^{-1} can be given. Since D^{-1} is an invertible mapping from $\overset{\circ}{H}{}^0(I)$ to $\overset{\circ}{H}{}^1(I)$ the operator K is a Fredholm operator. As in [6] a homotopy argument yields that the index of K is independent of the grading parameter q. Let $w \in \overset{\circ}{H}{}^1(I)$ satisfy $Kw = 0$. Using the definition of w (6) we obtain $u(s) = w(\gamma^{-1}(s))$ with $Wu = 0$ where W is given in (1). Therefore u is constant. Thus w is constant as well and therefore $K : \overset{\circ}{H}{}^1(I) \to \overset{\circ}{H}{}^0(I)$ is invertible. Since A^{-1} is an isomorphism of $H^s(I)$

onto $H^{s-1}(I)$ as well as of $\overset{\circ}{H}^s(I)$ onto $\overset{\circ}{H}^{s-1}(I)$ we derive from the invertibility of K the invertibility of $A^{-1}K$.

In order to obtain the strong ellipticity (27) we decompose $A^{-1}K$ with (25) and (16) and we use that for any $v \in \overset{\circ}{H}^s(I)$ we have $A^{-1} = -D\mathscr{H} = -\mathscr{H}D$ and $D^{-1}Dv = v$. Hence

$$A^{-1}K = A^{-2} + A^{-1}(K - A^{-1}) = -D^2 - D\mathscr{H}(K - A^{-1})$$

$$= -D^2 - D\mathscr{H}(K + \mathscr{H}D) = -D^2 - D\mathscr{H}(KD^{-1} + \mathscr{H})D.$$

From Lemma 1 we derive

$$KD^{-1} = -\widetilde{K} + E$$

with a compact operator E from $\overset{\circ}{H}^0(I)$ onto $\overset{\circ}{H}^0(I)$. Hence we have for $v \in \overset{\circ}{H}^1(I)$

$$\Re\langle A^{-1}Kv, v\rangle = \Re\langle(-D^2 - D\mathscr{H}(KD^{-1} + \mathscr{H})D)v, v\rangle$$

$$= \Re\langle(-D^2 - D\mathscr{H}(\mathscr{H} - \widetilde{K} + E)D)v, v\rangle = \Re\langle(I + \mathscr{H}(\mathscr{H} - \widetilde{K} + E))Dv, Dv\rangle,$$

where we have used that the adjoint operator to D is given by $-D$. From [6] we recall there exists a compact operator $T : \overset{\circ}{H}^0(I) \to \overset{\circ}{H}^1(I)$ and a constant $C > 0$ with

$$\Re\langle(I + \mathscr{H}(\mathscr{H} - \widetilde{K}) + T)Dv, Dv\rangle \geq C\,\|Dv\|_0^2, \qquad v \in \overset{\circ}{H}^1(I). \tag{28}$$

Hence setting $T_1 := D\mathscr{H}ED - DTD$ finishes the proof of the strong ellipticity. □

For the solution of the integral equation (5) we have the following regularity result.

Theorem 2 *Let $l \in \mathbb{N}$ and f in (1) be smooth enough. Then the solution w of (5) satisfies $w \in \overset{\circ}{H}^l(I)$, if the grading exponent q in (4) is chosen with $q > (l-1/2)(1+|\chi|)$.*

Proof From [4] we know that the solution u of (1) has near the vertex x_0 (which is identified with $x = 0$) the expansion $u \sim x^\alpha +$ higher order terms, where $\alpha = \frac{1}{1+|\chi|}$ and $\omega = (1-\chi)\pi$ denotes the interior angle of Γ at x_0. Hence due to (6) and (7) the solution w of (5) satisfies

$$|D_\sigma^m w(\sigma)| \leq C\,\sigma^{q\alpha - m}.$$

Choosing $q > (l - 1/2)(1 + |\chi|)$ we have with a suitable $\epsilon > 0$

$$|D_\sigma^m w(\sigma)| \leq C\,\sigma^{l-1/2+\epsilon-m}. \tag{29}$$

Thus w is contained in $H^l(I)$ since $D_\sigma^l w \in L^2(I)$. □

The following result is obtained by applying Lemma 1 and will be used in the error analysis in Sect. 3.

Theorem 3 *On each compact subset of* $\mathbb{R}\times\mathbb{R}\setminus(\mathbb{Z}\times\mathbb{Z})$, *the derivatives* $D_s^i D_\sigma^m b(s,\sigma)$ *of order* $i+m \leq q-2$ *of the kernel* $b(s,\sigma)$ *in (23) are bounded and 1-periodic. Moreover, for* $s,\sigma \in [-1/2,1/2]\setminus\{0\}$ *we have the estimates*

$$|D_s^i D_\sigma^m b(s,\sigma)| \leq C\,(|s|+|\sigma|)^{-i-m-2}, \quad 0 \leq i+m+2 \leq q. \tag{30}$$

Proof Using Lemma 1 we have for the kernel of K

$$k(s,\sigma) = D_s D_\sigma \widehat{k}(s,\sigma) + l(s,\sigma),$$

where $l(s,\sigma)$ is a smooth function. On the other hand for the kernel of $-D\mathscr{H}$ we have with (16) and (17)

$$-D_s \cot(\pi(s-\sigma)) = -D_s D_\sigma \log|\sin(\pi(s-\sigma))|.$$

Hence the kernel of B given in (23) satisfies with (11):

$$b(s,\sigma) = D_s D_\sigma \frac{\widehat{k}(s,\sigma)}{\log|\sin(\pi(s-\sigma))|} + l(s,\sigma), \tag{31}$$

where $l(s,\sigma)$ is a smooth function. For example, for $s>0, \sigma>0$ the representation of $\hat{k}(.,.)$ in (11) yields

$$b(s,\sigma) = D_s D_\sigma \log\left|\frac{\sigma^q - s^q}{\sin\pi(s-\sigma)}\right|,$$

which satisfies (30) by application of Theorem 2.3 in [8] or [7]. □

3 Trigonometric Collocation

Introduce the collocation points

$$s_j = jh, \quad j \in \mathbb{Z}, \quad j \neq 0 \bmod 2n+1\,, \quad \text{where} \quad h := 1/(2n+1), \tag{32}$$

and let $\mathscr{T}_h^0$ denote the space of trigonometric polynomials of degree $\leq n$ with the standard basis

$$\varphi_k(s) = e^{i2\pi ks}, \quad |k| \leq n, k \neq 0. \tag{33}$$

Then, for any continuous 1-periodic function v with $\int_0^1 v(s)ds = 0$, the interpolatory projection $Q_h v$ onto $\mathscr{T}_h^0$ is well defined by

$$(Q_h v)(s_j) = v(s_j), \quad j = 1, \dots, 2n, \tag{34}$$

and satisfies [3]

$$\|v - Q_h v\|_t \leq C h^{r-t} \|v\|_r, \quad v \in H^r(I), \text{ for } r > 1/2, \quad r \geq t \geq 0. \tag{35}$$

Using the basis (33), the projection Q_h is given by

$$Q_h v(s) = \sum_{k=-n,k\neq 0}^{n} \alpha_k \varphi_k(s), \quad \alpha_k := h \sum_{j=-n}^{n} v(s_j)\overline{\varphi_k(s_j)};$$

see [2] or [15, Kap. 2.3].

The collocation method for (5) consists of solving for $w_h \in \mathscr{T}_h^0$

$$K w_h(s_j) = g(s_j), \qquad j = 1, \dots, 2n.$$

This can be written using (22) and the interpolatory projection as

$$Q_h(A^{-1} + B)w_h = Q_h g, \qquad w_h \in \mathscr{T}_h^0,$$

and since Q_h commutes with A^{-1} on $\mathscr{T}_h^0$, we have the equivalent form

$$(A^{-1} + Q_h B)w_h = Q_h g, \qquad w_h \in \mathscr{T}_h^0. \tag{36}$$

Due to the derivation of (26) from (22) we rewrite (36) as the second kind equation

$$A^{-1}(A^{-1} + Q_h B) = -D^2 + R_h A^{-1} B + \mathscr{J} A^{-1}.$$

With (24) and $R_h := A^{-1} Q_h A$ the collocation method (36) is equivalent to

$$(-D^2 + R_h A^{-1} B)w_h = R_h e, \qquad w_h \in \mathscr{T}_h^0. \tag{37}$$

Here R_h is a well defined projection operator of $\mathring{H}^r(I)$ onto $\mathscr{T}_h^0$ which satisfies

$$\|v - R_h v\|_t \leq C h^{r-t} \|v\|_r, \quad v \in \mathring{H}^r(I), \quad \text{for } r > -1/2, \quad r \geq t \geq -1. \tag{38}$$

Furthermore we have for $v \in \mathring{H}^r(I)$ the estimate

$$\|R_h v\|_{-1} \leq \|R_h v - v\|_{-1} + \|v\|_{-1} \leq C h \|v\|_0 + \|v\|_{-1}. \tag{39}$$

It is well known that for Mellin convolution operators one proves only stability for a slightly modified method. For $\tau > 0$ sufficiently small, we introduce the truncation T_τ of the 1-periodic extension

$$T_\tau v(s) = \begin{cases} 0, & s \in (-\tau, \tau), \\ v(s), & s \in (-1/2, -\tau) \cup (\tau, 1/2). \end{cases} \tag{40}$$

Now we consider instead of (37) the modified collocation method

$$(-D^2 + R_h A^{-1} B T_{i^* h}) w_h = R_h e, \qquad w_h \in \mathscr{T}_h^0. \tag{41}$$

where i^* is a fixed integer independent of h. If $i^* = 0$ then (41) coincides with (37).

Lemma 2

1. For $v \in \mathscr{T}_h^0$ there holds

$$\|\sigma^{-1} v(\sigma)\|_0 \leq C \; \|Dv\|_0. \tag{42}$$

2. Let $w \in H^l(I)$ solve $Kw = g$, then there holds for $(j = 0, \ldots, l)$:

$$\|\sigma^{-j}(I - T_{i^* h}) w\|_0 \leq C h^{l-j} \|w\|_l. \tag{43}$$

The constant C in 1. and 2. is independent of h.

Proof To show assertion 1. we write $v \in \mathscr{T}_h^0$ as a linear combination of the set of functions $\{\sin(2\pi k\sigma), \cos(2\pi k\sigma) | \; k = 1, \ldots, n\}$. Integration by parts gives

$$\|\sigma^{-1} \sin(2\pi k\sigma)\|_0^2 = \int_0^1 \sigma^{-2} \sin^2(2\pi k\sigma) \; d\sigma = 2\pi k \int_0^1 \sigma^{-1} \sin(4\pi k\sigma) \; d\sigma$$

$$\leq -8(\pi k)^2 \int_0^1 \log \sigma \; d\sigma = 8(\pi k)^2,$$

which yields with

$$\|D_\sigma \sin(2\pi k\sigma)\|_0 = C k, \qquad k = 1, \ldots, n.$$

the desired relation (42) for $v = \sin 2\pi k\sigma$. For $\cos(2\pi k\sigma)$ we can estimate similarly, and the proof of (42) is complete.

To show 2. we find with (29)

$$\|\sigma^{-j}(I - T_{i^* h}) w\|_0^2 = \int_{-i^* h}^{i^* h} \sigma^{-2j} w^2(\sigma) d\sigma \leq C \int_{-i^* h}^{i^* h} \sigma^{-2j} |\sigma|^{2l-1+2\epsilon} d\sigma \leq C h^{2l-2j}.$$

(For $j = 0$ this estimate is proved in [6].) □

The analysis of the collocation method depends heavily on the stability of a "finite section" approximation $-D^2 + A^{-1}BT_\tau$ to the operator $A^{-1}K$ (cf. (26)) which is given in the next lemma. We follow [6].

Lemma 3 *There exists $C > 0$ and $\tau_0 > 0$ such that for all $q \geq 1$*

$$\|(-D^2 + A^{-1}BT_\tau)v\|_{-1} \geq C\|Dv\|_0, \tag{44}$$

for all $v \in \mathring{H}^1(I)$ and $0 < \tau \leq \tau_0$ with T_τ as in (40).

Proof In order to derive the lemma we follow Theorem 6 of [6]. which gives the lemma by replacing $I + A^{-1}(K - A)$ and $v \in H^0$, considered in [6], by $-D^2 + A^{-1}B$ and $v \in \mathring{H}^1(I)$, considered here. With the decomposition

$$v \to (T_\tau v, (I - T_\tau)v)^T$$

the map $v \to (-D^2 + A^{-1}BT_\tau)v$ is represented by the matrix operator

$$\begin{pmatrix} T_\tau v \\ (I - T_\tau)v \end{pmatrix} \to \begin{pmatrix} T_\tau(-D^2 + A^{-1}B)T_\tau & 0 \\ (I - T_\tau)A^{-1}BT_\tau & -D^2 \end{pmatrix} \quad \begin{pmatrix} T_\tau v \\ (I - T_\tau)v \end{pmatrix}.$$

Its inverse is represented by

$$\begin{pmatrix} T_\tau v \\ (I - T_\tau)v \end{pmatrix} \to \begin{pmatrix} (T_\tau(-D^2 + A^{-1}B)T_\tau)^{-1} & 0 \\ -D^{-2}(I - T_\tau)A^{-1}B(T_\tau(-D^2 + A^{-1}B)T_\tau)^{-1} & D^{-2} \end{pmatrix} \quad \begin{pmatrix} T_\tau v \\ (I - T_\tau)v \end{pmatrix}.$$

Lemma 3 is derived by using the representations given above in the proof of Theorem 6 of [6]. □

By the following theorem the (modified) collocation method (41) converges with optimal order in the H^1 norm.

Theorem 4 *Let $q \geq 2$ and suppose that i^* is sufficiently large. Then there holds*

1. *The method (41) is stable, that is the estimate*

$$\|(-D^2 + R_hA^{-1}BT_{i^*h})v\|_{-1} \geq C\,\|Dv\|_0 \geq C\,\|v\|_1, \qquad v \in \mathscr{T}_h^0 \tag{45}$$

   holds for all h sufficiently small, where C is independent of h and v.
2. *If, in addition, the hypothesis of Theorem 2 holds, then (41) has a unique solution for all h sufficiently small and there holds the estimate*

$$\|w - w_h\|_1 \leq C\,h^{l-1} \tag{46}$$

   with C a constant which depends on i^* and w but is independent of h.

Proof With the triangle inequality we have

$$\begin{aligned}&\|(-D^2 + R_hA^{-1}BT_{i^*h})v\|_{-1}\\&\quad\geq \|(-D^2 + A^{-1}BT_{i^*h})v\|_{-1} - \|(I - R_h)A^{-1}BT_{i^*h}v\|_{-1.}\end{aligned} \tag{47}$$

In order to estimate the second term on the right hand side of (47) we use the fact that $I - R_h$ annihilates the constants and obtain from (16) and (38)

$$\|(I - R_h)A^{-1}BT_{i^*h}v\|_{-1} \leq \|(I - R_h)\mathscr{H}DBT_{i^*h}v\|_{-1} \leq C\,h\|DBT_{i^*h}v\|_0. \tag{48}$$

Using Theorem 3 yields

$$\begin{aligned}|DBT_{i^*h}v(s)| &= \left|\int_{J_{i^*h}} D_sb(s,\sigma)v(\sigma)\,d\sigma\right|\\&\leq C\int_{J_{i^*h}} (|s| + |\sigma|)^{-3}|v(\sigma)|\,d\sigma\\&\leq C/(i^*h)\int_{J_{i^*h}} \frac{|\sigma|^2}{(|s| + |\sigma|)^3}|\sigma^{-1}v(\sigma)|\,d\sigma,\end{aligned}$$

where $J_{i^*h} = (-1/2, -i^*h) \cup (i^*h, 1/2)$. Taking L^2 norms and using the fact that the integral operator with Mellin convolution kernel $\frac{|\sigma|^2}{(|s|+|\sigma|)^3}$ is bounded on $H^0(0,\infty)$ gives with (42)

$$\|DBT_{i^*h}v\|_0 \leq C/(i^*h)\|\sigma^{-1}v(\sigma)\|_0 \leq C/(i^*h)\|Dv\|_0. \tag{49}$$

Combining (48) and (49) gives

$$\|(I - R_h)A^{-1}BT_{i^*h}v\|_{-1} \leq C/i^*\|Dv\|_0. \tag{50}$$

Now choose i^* in such a way, that C/i^* with C given in (50) is smaller than C given in (44). Then combining (50) and (44) yields (45). Note that the norms $\|Dv\|_0$ and $\|v\|_1$ are equivalent for $v \in \mathscr{T}_h^0$.

To prove the error estimate (46) we note that

$$\|w - w_h\|_1 \leq \|(I - R_h)w\|_1 + \|w_h - R_hw\|_1 \tag{51}$$

and estimate the first term in (51) by (38). To estimate the second term we derive an auxiliary equation which can be composed from (26) and (41):

$$-R_hD^2w + R_hMw = R_he = (-D^2 + R_hMT_{i^*h})w_h. \tag{52}$$

Thus for the second term in (51) we obtain with (44) and (52):

$$\begin{aligned}
\|w_h - R_h w\|_1 &\leq C\,\|(-D^2 + R_h M T_{i^*h})(w_h - R_h w)\|_{-1} \\
&= C\,\|R_h(-D^2 + M)w + D^2 R_h w - R_h M T_{i^*h} R_h w\|_{-1} \qquad (53) \\
&= C\,\| - R_h D^2 w + D^2 w - D^2 w + D^2 R_h w + R_h M w - R_h M T_{i^*h} w \\
&\quad + R_h M T_{i^*h} w - R_h M T_{i^*h} R_h w\|_{-1} \\
&\leq C\,\|(I - R_h)D^2 w\|_{-1} + C\,\|D^2(I - R_h)w\|_{-1} \\
&\quad + C\,\|R_h M(I - T_{i^*h})w\|_{-1} + C\,\|R_h M T_{i^*h}(I - R_h)w\|_{-1}. \qquad (54)
\end{aligned}$$

With (38) we have for the first and second term:

$$\|(I - R_h)D^2 w\|_{-1} \leq C\,h^{l-1}\|D^2 w\|_{l-2} \leq C\,h^{l-1}\|w\|_l,$$

$$\|D^2(I - R_h)w\|_{-1} \leq \|(I - R_h)w\|_1 \leq C\,h^{l-1}\|w\|_l,$$

and further with (39) and the invertibility of A^{-1}:

$$\begin{aligned}
\|R_h M(I - T_{i^*h})w\|_{-1} &\leq C\,h\|M(I - T_{i^*h})w\|_0 + \|M(I - T_{i^*h})w\|_{-1} \\
&\leq C\,h\|DB(I - T_{i^*h})w\|_0 + c\|B(I - T_{i^*h})w\|_0. \qquad (55)
\end{aligned}$$

For the first term on the right hand side of (55) we have the estimate:

$$\begin{aligned}
DB(I - T_{i^*h})w(s) &= \int_0^1 D_s b(s,\sigma)(I - T_{i^*h})w(\sigma)\,d\sigma \\
&\leq \int_0^1 \frac{1}{(|s| + |\sigma|)^3}|(I - T_{i^*h})w(\sigma)|\,d\sigma \\
&\leq \int_0^1 \frac{|\sigma|^2}{(|s| + |\sigma|)^3}|\sigma^{-2}(I - T_{i^*h})w(\sigma)|\,d\sigma.
\end{aligned}$$

Taking L^2 norms and using (43) yields

$$\|DB(I - T_{i^*h})w\|_0 \leq C\,\|\sigma^{-2}(I - T_{i^*h})w(\sigma)\|_0 \leq C\,h^{l-2}\|w\|_l.$$

The second term in (55) can be estimated alike. For the fourth term in (54) we have

$$\begin{aligned}
\|R_h M T_{i^*h}(I - R_h)w\|_{-1} &\leq C\,h\|M T_{i^*h}(I - R_h)w\|_0 + \|M T_{i^*h}(I - R_h)w\|_{-1} \\
&\leq C\,h\|DB T_{i^*h}(I - R_h)w\|_0 + \|B T_{i^*h}(I - R_h)w\|_0 \qquad (56)
\end{aligned}$$

and further

$$\begin{aligned} D_sBT_{i^*h}(I-R_h)w(s) &= \int_{J_{i^*h}} D_sb(s,\sigma)(I-R_h)w(\sigma)\ d\sigma \\ &\leq \int_{J_{i^*h}} \frac{1}{(|s|+|\sigma|)^3}|(I-R_h)w(\sigma)|\ d\sigma \\ &\leq C/(i^*h)^2 \int_{J_{i^*h}} \frac{|\sigma|^2}{(|s|+|\sigma|)^3}|(I-R_h)w(\sigma)|\ d\sigma. \end{aligned}$$

Again, taking L^2 norms and using (38) we obtain:

$$\|DBT_{i^*h}(I-R_h)w\|_0 \leq C\,h^{-2}\|(I-R_h)w(\sigma)\|_0 \leq C\,h^{l-2}\|w\|_l.$$

The second term on the right hand side in (56) can be estimated in the same way and hence (46) is proved. □

Using the solution w_h of (36) we can derive an approximation u_h to the solution of the original equation (1). Due to (6) we define

$$u_h(\gamma(\sigma)) := w_h(\sigma) - w_{h0}, \tag{57}$$

where the constant w_{h0} is given by

$$w_{h0} := \frac{\int_0^1 w_h(s)|\gamma'(s)|\ ds}{\int_0^1 |\gamma'(s)|\ ds}.$$

This definition insures that the additional constraint $\int_\Gamma u_h = 0$ in (1) is satisfied.

4 Discrete Collocation

For a fully discrete version of the collocation method (36), introduce the quadrature points

$$\sigma_j = jh + h/2, \quad j \in \mathbb{Z}, \quad \text{where} \quad h := 1/(2n+1). \tag{58}$$

To evaluate the integral

$$I(v) = \int_0^1 v(\sigma)\ d\sigma$$

for a 1-periodic continuous function v, we approximate it by a trapezoidal rule

$$I_h(v) = h \sum_{j=0}^{2n} v(\sigma_j).$$

Due to Theorem 3 the kernel b of the operator B is bounded only on compact subsets of $[0,1] \times [0,1]$, but has singular behaviour at the corners of $[0,1] \times [0,1]$, like r^{-2}, where r denotes the distance to the corner. Therefore we regularize the hypersingular integral by subtracting $v(0)$. After regularization the integral is defined as a Cauchy singular integral. Now the integral operator B in (36) is approximated by

$$B_h v(s) := I_h(b(s,\cdot)(v(\cdot) - v(0)) = h \sum_{j=0}^{2n} b(s,\sigma_j)(v(\sigma_j) - v(0)). \tag{59}$$

Now the discrete collocation method for (5) is defined by replacing B with B_h in (36) and consists in solving for $w_h \in \mathscr{T}_h^0$

$$A^{-1} w_h(s_j) + B_h w_h(s_j) = g(s_j), \qquad j = 1, \dots, 2n.$$

The discrete collocation method can be written in the form

$$(A^{-1} + Q_h B_h) w_h = Q_h g, \qquad w_h \in \mathscr{T}_h^0. \tag{60}$$

To obtain a linear system for finding w_h, let

$$w_h(s) = \sum_{k=-n}^{n} \alpha_k \varphi_k(s)$$

and calculate the coefficients α_k from (see (19)):

$$\sum_{k=-n, k\neq 0}^{n} \alpha_k \left[\varphi_k(s_j)|k| + (B_h \varphi_k)(s_j) \right] = g(s_j), \qquad j = 1, \dots, 2n. \tag{61}$$

Similarly to [8] our convergence analysis follows the same lines as in the previous section: instead of (60) we consider the modified method

$$(A^{-1} + Q_h B_h T_{i^* h}) w_h = Q_h g, \qquad w_h \in \mathscr{T}_h^0. \tag{62}$$

Using the projection R_h (62) can be written as

$$(-D^2 + R_h M_h T_{i^* h}) w_h = R_h e, \qquad w_h \in \mathscr{T}_h^0, \tag{63}$$

where $M_h = A^{-1}B_h$. In our analysis we use the following standard estimate for the trapezoidal rule. This lemma is also used in [7] and goes back to [5].

Lemma 4 *Let $l \in \mathbb{N}$, and suppose that v has 1-periodic continuous derivatives of order $< l$ on $\mathbb{R}$ and that $D^l v$ is integrable on $(0, 1)$. Then*

$$|I(v) - I_h(v)| \le C\, h^l \int_0^1 |D^l v(\sigma)| d\sigma,$$

where C does not depend on v and h.

Theorem 5

1. *Let $q \ge 2$, and suppose $i^* \ge 1$ is sufficiently large Then the estimate*
$$\|(-D^2 + R_h M_h T_{i^*h})v\|_{-1} \ge C\, \|Dv\|_0 \ge C\, \|v\|_1, \qquad v \in \mathscr{T}_h^0 \tag{64}$$
holds for all h sufficiently small, where C is independent of v and h.
2. *If the hypothesis of Theorem 2 holds and hence we have $w \in \mathring{H}^l(I)$, then*
$$\|w - w_h\|_1 \le c\, h^{l-1} \tag{65}$$
with a constant c independent of h.

Proof Let $v \in \mathscr{T}_h^0$. Due to the decomposition

$$\|(-D^2 + R_h M_h T_{i^*h})v\|_{-1} \ge \|(-D^2 + R_h M T_{i^*h})v\|_{-1} - \|R_h(M - M_h)T_{i^*h}v\|_{-1}$$

and (45) it is sufficient to verify that for each $\epsilon \ge 0$ there exists i^* such that

$$\|R_h(M - M_h)T_{i^*h}v\|_{-1} \le \epsilon \|Dv\|_0. \tag{66}$$

With (39) we have

$$\|R_h(M - M_h)T_{i^*h}v\|_{-1} \le C\, h\|D(B - B_h)T_{i^*h}v\|_0 + C\, \|(B - B_h)T_{i^*h}v\|_0 \tag{67}$$

and further using Lemma 4 and Theorem 3

$$C\, h|D(B - B_h)T_{i^*h}v(s)| + C\, |(B - B_h)T_{i^*h}v(s)|$$
$$\le Ch^2 \int_{J_{i^*h}} |D_s b(s,\sigma)||D_\sigma v(\sigma)|\, d\sigma + C\, h^2 \int_{J_{i^*h}} |D_\sigma D_s b(s,\sigma)|\, |v(\sigma)|\, d\sigma$$
$$+ C\, h \int_{J_{i^*h}} |D_\sigma b(s,\sigma)|\, |v(\sigma)|\, d\sigma + C\, h \int_{J_{i^*h}} |b(s,\sigma)|\, |D_\sigma v(\sigma)|\, d\sigma$$

$$\leq C h^2 \int_{J_{i^*h}} \frac{1}{(|s|+|\sigma|)^3} \, |D_\sigma v(\sigma)| \, d\sigma + C h^2 \int_{J_{i^*h}} \frac{1}{(|s|+|\sigma|)^4} \, |v(\sigma)| \, d\sigma$$

$$+ C h \int_{J_{i^*h}} \frac{1}{(|s|+|\sigma|)^3} \, |v(\sigma)| \, d\sigma + C h \int_{J_{i^*h}} \frac{1}{(|s|+|\sigma|)^2} \, |D_\sigma v(\sigma)| \, d\sigma \qquad (68)$$

$$\leq C/(i^*)^2 \int_{J_{i^*h}} \frac{|\sigma|^2}{(|s|+|\sigma|)^3} \, |D_\sigma v(\sigma)| \, d\sigma + C/(i^*)^2 \int_{J_{i^*h}} \frac{|\sigma|^3}{(|s|+|\sigma|)^4} \, |\sigma^{-1} v(\sigma)| \, d\sigma$$

$$+ C/i^* \int_{J_{i^*h}} \frac{|\sigma|^2}{(|s|+|\sigma|)^3} \, |\sigma^{-1} v(\sigma)| \, d\sigma + C/i^* \int_{J_{i^*h}} \frac{|\sigma|}{(|s|+|\sigma|)^2} \, |D_\sigma v(\sigma)| \, d\sigma.$$

Hence

$$ch \|D(B - B_h) T_{i^*h} v\|_0 + C \|(B - B_h) T_{i^*h} v\|_0 \leq C/i^* \|Dv\|_0 + C/i^* \|\sigma^{-1} v(\sigma)\|_0.$$

For i^* sufficiently large this yields with (42) the desired estimate (66) and hence the stability of the discrete collocation method.

To estimate the error we proceed as in Theorem 4:

$$\|w - w_h\|_1 \leq \|(I - R_h) w\|_1 + \|w_h - R_h w\|_1,$$

where the first term is due to (38) of order h^{l-1}. Using (64) we have:

$$\|w_h - R_h w\|_1 \leq C \, \|(-D^2 + R_h M_h T_{i^*h})(w_h - R_h w)\|_{-1}$$

$$\leq C \, \|(-D^2 + R_h M_h T_{i^*h}) w_h - (-D^2 + R_h M_h T_{i^*h}) R_h w\|_{-1}.$$

Combining (63) and (26), we have an auxiliary equation like (52) but with M_h instead of M on the right hand side. This yields

$$\begin{aligned} \|w_h - R_h w\|_1 &\leq C \, \|R_h(-D^2 + M) w - (-D^2 + R_h M_h T_{i^*h}) R_h w\|_{-1} \\ &\leq \|R_h(-D^2 + M) w - (-D^2 + R_h M T_{i^*h}) R_h w\|_{-1} \\ &\quad + \|R_h(M - M_h) T_{i^*h} R_h w\|_{-1}. \end{aligned} \qquad (69)$$

For the first term on the right hand side of (69) the estimate is given in Theorem 4, see (53). For the second term we have:

$$\begin{aligned} \|R_h(M - M_h) T_{i^*h} R_h w\|_{-1} &\leq \|R_h(M - M_h) T_{i^*h} w\|_{-1} \\ &\quad + \|R_h(M - M_h) T_{i^*h}(I - R_h) w\|_{-1}. \end{aligned} \qquad (70)$$

With (39) we derive for the first term in (70):

$$\|R_h(M-M_h)T_{i^*h}w\|_{-1} \leq C\,\|(B-B_h)T_{i^*h}w\|_0 + C\,h\|D(B-B_h)T_{i^*h}w\|_0. \tag{71}$$

With Lemma 4 we have

$$\begin{aligned}|D_s(B-B_h)T_{i^*h}w(s)| &\leq C\,h^{l-2}\int_{J_{i^*h}}\sum_{m=0}^{l-2}|D_sD_\sigma^m b(s,\sigma)D_\sigma^{l-m-2}w(\sigma)|\,d\sigma\\ &\leq C\,h^{l-2}\int_{J_{i^*h}}\sum_{m=0}^{l-2}\frac{1}{(|s|+|\sigma|)^{m+3}}|D_\sigma^{l-m-2}w(\sigma)|\,d\sigma\\ &\leq C\,h^{l-2}\int_{J_{i^*h}}\sum_{m=0}^{l-2}\frac{|\sigma|^{m+2}}{(|s|+|\sigma|)^{m+3}}|\sigma^{-m-2}D_\sigma^{l-m-2}w(\sigma)|\,d\sigma.\end{aligned}$$

Taking L^2 norms yields

$$\|D_s(B-B_h)T_{i^*h}w(s)\|_0 \leq C\,h^{l-2}\|\sigma^{-m-2}D_\sigma^{l-m-2}w(\sigma)\|_0, \tag{72}$$

and with (29) there exists $\epsilon > 0$ arbitrary such that

$$|\sigma^{-m-2}D_\sigma^{l-m-2}w(\sigma)| \leq c|\sigma^{-m-2}\sigma^{l-1/2+\epsilon-l+m+2}| \leq C\,|\sigma^{-1/2+\epsilon}|.$$

Hence the L^2 norm on the right hand side in (72) is bounded. Putting together (72) and (71) gives the estimate for the first term in (70). To estimate the second term in (70) we substitute $v := (I-R_h)w$ and argue similarly to (68).

$$C\,h|D(B-B_h)T_{i^*h}v(s)| + C\,|(B-B_h)T_{i^*h}v(s)|$$

$$\begin{aligned}&\leq C\int_{J_{i^*h}}\frac{|\sigma|^2}{(|s|+|\sigma|)^3}\,|D_\sigma v(\sigma)|\,d\sigma + C\,h^{-1}\int_{J_{i^*h}}\frac{|\sigma|^3}{(|s|+|\sigma|)^4}\,|v(\sigma)|\,d\sigma\\ &+C\,h^{-1}\int_{J_{i^*h}}\frac{|\sigma|^2}{(|s|+|\sigma|)^3}\,|v(\sigma)|\,d\sigma + C\int_{J_{i^*h}}\frac{|\sigma|}{(|s|+|\sigma|)^2}\,|D_\sigma v(\sigma)|\,d\sigma.\end{aligned}$$

Hence

$$\|R_h(M-M_h)T_{i^*h}v\|_{-1} \leq C\,h^{-1}\|v\|_0 + C\,\|Dv\|_0 \leq C\,h^{-1}\|v\|_0 + C\,\|v\|_1.$$

Substituting $v := (I-R_h)w$ and using (38) gives

$$\|R_h(M-M_h)T_{i^*h}(I-R_h)w\|_{-1} \leq C\,h^{-1}\|(I-R_h)w\|_0 + C\,\|(I-R_h)w\|_1 \leq C\,h^{l-1}\|w\|_l.$$

Thus the error estimate (65) is proved. □

5 Numerical Results

Here we present a numerical example for the solution of Eq. (1) when Γ is the boundary of a "teardrop-shaped" region with a simple corner at the origin which corresponds to $s = 0$ and $s = 1$. The parametrization of Γ is given by

$$\gamma_0(s) = \sin \pi s(\cos(1-\chi)\pi s, \sin(1-\chi)\pi s)^T, \quad s \in [0,1], \quad \chi \in (0,1)$$

with exterior angle $(1+\chi)\pi$ at the origin. As right hand side in (5) we use

$$g(s) = 5s^4(1-s)^4(1-2s).$$

(Note: $g(s) = D_s(s^5(1-s)^5)$). Now the integral of g on $[0,1]$ vanishes and so does the integral of f in (1) on Γ, where f corresponds to g by (6). Further this choice of g insures that f in (1) is sufficiently smooth. For the numerical implementation we use $\nu(s) = s$ in (4). In our numerical experiments below, no modification of the collocation method was necessary to insure stability and throughout we have set $i^* = 0$. Since the exact solution of (5) is unknown, we have the error $\|w - w_h\|_1$ approximated by $\|w^* - w_h\|_1$, where w^* is the solution of the discrete collocation method with $2 * 512$ trial functions. Empirically determined convergence rates are given in columns headed "EOC" in Table 1 for $\chi = 0.76$. Those numbers demonstrate the improvement of the convergence order for increasing values of the grading exponent q as expected from Theorem 5.

Table 1 H^1-errors of the transformed density, $\chi = 0.76$

	$q=1$		$q=2$		$q=3$		$q=5$	
n	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC
30	$1.235-5$		$2.35-5$		$1.29-5$		$5.62-6$	
		0.19		0.70		1.28		2.54
40	$1.170-5$		$1.92-5$		$8.94-6$		$2.70-6$	
		0.16		0.70		1.27		2.50
50	$1.129-5$		$1.64-5$		$6.74-6$		$1.55-6$	
		0.14		0.70		1.26		2.47
60	$1.101-5$		$1.45-5$		$5.35-6$		$9.85-7$	
		0.12		0.71		1.27		2.45
70	$1.081-5$		$1.29-5$		$4.40-6$		$6.75-7$	
		0.10		0.73		1.27		2.44
80	$1.068-5$		$1.17-5$		$3.71-6$		$4.87-7$	
		0.09		0.75		1.28		2.42
90	$1.056-5$		$1.08-5$		$3.19-6$		$3.66-7$	
		0.07		0.77		1.30		2.42
100	$1.048-5$		$9.91-6$		$2.78-6$		$2.84-7$	

Table 2 H^1-errors of the transformed density, $\chi = 0.30$

	$q = 1$		$q = 2$		$q = 3$		$q = 5$	
n	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC	$\|w_h - w^*\|_1$	EOC
30	$4.29 - 6$		$7.71 - 7$		$1.22 - 7$		$1.37 - 8$	
		0.20		1.09		1.78		3.43
40	$4.05 - 6$		$5.65 - 7$		$7.29 - 8$		$5.09 - 9$	
		0.22		1.08		1.79		3.42
50	$3.86 - 6$		$4.44 - 7$		$4.89 - 8$		$2.37 - 9$	
		0.24		1.07		1.80		3.41
60	$3.69 - 6$		$3.65 - 7$		$3.52 - 8$		$1.28 - 9$	
		0.26		1.07		1.80		3.40
70	$3.55 - 6$		$3.09 - 7$		$2.67 - 8$		$7.56 - 10$	
		0.28		1.07		1.81		3.39
80	$3.42 - 6$		$2.68 - 7$		$2.09 - 8$		$4.81 - 10$	
		0.30		1.07		1.81		3.38
90	$3.30 - 6$		$2.36 - 7$		$1.69 - 8$		$3.27 - 10$	
		0.32		1.07		1.82		3.38
100	$3.19 - 6$		$2.11 - 7$		$1.40 - 8$		$2.26 - 10$	

Note that Theorem 5 predicts the convergence rates 0.07, 0.64, 1.20 and 2.34 for $q = 1, 2, 3, 5$, respectively. For a second experiment we chose $\chi = 0.30$ and give the results in Table 2. In this case one would expect the convergence rates 0.27, 1.04, 1.81 and 3.35 corresponding to $q = 1, 2, 3, 5$, respectively.

References

1. Ang, W.-T.: Hypersingular Integral Equations in Fracture Analysis. Elsevier Science and Technology, Amsterdam (2013)
2. Atkinson, K.E.: A discrete Galerkin method for first kind integral equations with a logarithmic kernel. J. Integral Equ. Appl. **1**, 343–363 (1988)
3. Atkinson, K.E., Sloan, I.H.: The numerical solution of first-kind logarithmic-kernel integral equations on smooth open arcs. Math. Comput. **56**, 119–139 (1991)
4. Costabel, M., Stephan, E.P.: The normal derivative of the double layer potential on polygons and Galerkin approximation. Appl. Anal. **16**, 205–228 (1983)
5. Davis, P.J., Rabinowitz, P.: Numerical Integration. Blaisdell, Waltham (1967)
6. Elschner, J., Graham, I.G.: An optimal order collocation method for first kind boundary integral equations on polygons. Numer. Math. **70**, 1–31 (1995)
7. Elschner, J., Graham, I.G.: Quadrature methods for Symm's integral equation on polygons. IMA J. Numer. Anal. **17**, 643–664 (1997)
8. Elschner, J., Stephan, E.P.: A discrete collocation method for Symm's integral equation on curves with corners. J. Comput. Appl. Math. **75**, 131–146 (1996)
9. Elschner, J., Jeon, Y., Sloan, I.H., Stephan, E.P.: The collocation method for mixed boundary value problems in domains with curved polygonal boundaries. Numer. Math. **76**, 355–381 (1997)

10. Jeon, Y.: A quadrature for the Cauchy singular integral equations. J. Integral Equ. Appl. **7**, 425–461 (1995)
11. Kieser, R., Kleemann, B., Rathsfeld A.: On a full discretization scheme for a hypersingular integral equation over smooth curves. Zeitschrift für Analysis und ihre Anwendungen **11**, 385–396 (1992)
12. Krishnasamy, G., Schmerr, L.W., Rudolphi, T.J., Rizzo F.J.: Hypersingular boundary integral equations: some applications in acoustic and elastic wave scattering. J. Appl. Mech. **57**, 404–414 (1990)
13. Lifanov, I.K., Poltavskiii, L.N., Vainikko, G.M.: Hypersingular Integral Equations and Their Applications. CRC Press, New York (2003)
14. Prössdorf, S., Rathsfeld, A.: Quadrature methods for strongly elliptic singular integral equations on an interval. In: Operator Theory: Advances and Applications, vol. 41, pp. 435–471. Birkhäuser, Basel (1989)
15. Prössdorf, S., Silbermann, B.: Projektionsverfahren und die näherungsweise Lösung singulärer Gleichungen. Teubner, Leipzig (1977)

On the Complexity of Parametric ODEs and Related Problems

Stefan Heinrich

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We present an iterative Monte Carlo procedure to solve initial value problems for systems of ordinary differential equations depending on a parameter. It is based on a multilevel Monte Carlo algorithm for parametric indefinite integration. As an application, we also obtain a respective method for solving almost linear first order partial differential equations. We also consider deterministic algorithms.

We study the convergence and, in the framework of information-based complexity, the minimal errors and show that the developed algorithms are of optimal order (in some limit cases up to logarithmic factors). In this way we extend recent complexity results on parametric ordinary differential equations. Moreover, we obtain the complexity of almost linear first-order partial differential equations, which has not been analyzed before.

1 Introduction

Monte Carlo (one-level) methods for integrals depending on a parameter were first considered in [8]. Multilevel Monte Carlo methods for parametric integration were developed in [15], where the problem was studied for the first time in the frame of information-based complexity theory (IBC). These investigations were continued in [3], where the complexity of parametric indefinite integration was studied for the first time.

Recently there arose considerable interest in the numerical solution of various parametric problems, also in connection with random partial differential equations, see [1, 7, 19, 20] and references therein. Deterministic methods for solving para-

S. Heinrich (✉)
Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany
e-mail: heinrich@informatik.uni-kl.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_26

metric initial value problems for systems of ordinary differential equations (ODEs) were first considered in [9].

The study of ODEs in IBC was begun in [16] for the deterministic case and in [17, 18] for the stochastic case. The complexity in the randomized setting was further studied in [2, 12, 13]. The complexity of parametric initial value problems for systems of ODEs was investigated in [4] and [5], where multilevel Monte Carlo algorithms for this problem were developed and shown to be of optimal order.

Here we study function classes satisfying a weaker Lipschitz condition than those considered in [4]. This is needed for the applications to the complexity analysis for almost linear partial differential equations (PDEs). Moreover, we present a new approach to solve initial value problems for ODEs depending on a parameter. We develop an iterative Monte Carlo procedure, based on a multilevel algorithm for parametric indefinite integration. This leads to an iterative multilevel Monte Carlo method for solving almost linear first order PDEs. We also consider deterministic algorithms.

We prove convergence rates, determine the minimal errors in the framework of IBC, and show that the developed algorithms are of optimal order (in some limit cases up to logarithmic factors). In this way we extend recent complexity results of [4] on parametric ordinary differential equations. Moreover, the complexity of almost linear first-order partial differential equation is determined, a topic, which has not been considered before.

The paper is organized as follows. In Sect. 2 we provide the needed notation. In Sects. 3 and 4 we recall the algorithms from [3] on Banach space valued and parametric indefinite integration, respectively, and improve some convergence results. Section 5 contains the main results. Based on the results of Sect. 4 we study the iterative solution of initial value problems for parametric ordinary differential equations. Finally, in Sect. 6 we apply the results of Sect. 5 to the analysis of the complexity of almost linear partial differential equations.

2 Preliminaries

Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. For a Banach space X the norm is denoted by $\|\ \|_X$, the closed unit ball by B_X, the identity mapping on X by I_X, and the dual space by X^*. The Euclidean norm on $\mathbb{R}^d$ $(d \in \mathbb{N})$ is denoted by $\|\ \|_{\mathbb{R}^d}$. Given another Banach space Y, we let $\mathscr{L}(X, Y)$ be the space of bounded linear mappings $T : X \to Y$ endowed with the canonical norm. If $X = Y$, we write $\mathscr{L}(X)$ instead of $\mathscr{L}(X, X)$. We assume all considered Banach spaces to be defined over the field of reals $\mathbb{R}$.

Concerning constants, we make the convention that the same symbol $c, c_1, c_2, \dots$ may denote different constants, even in a sequence of relations. Furthermore, we use the following notation: For nonnegative reals $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ we write $a_n \preceq b_n$ if there are $c > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $a_n \leq cb_n$. We also write $a_n \asymp b_n$ if simultaneously $a_n \preceq b_n$ and $b_n \preceq a_n$. If not specified, the function log means $\log_2$.

Given a set $D \subseteq \mathbb{R}^d$ which is the closure of an open set, and a Banach space X, we define $C^r(D,X)$ to be the space of all functions $f : D \to X$ which are r-times continuously differentiable in the interior of D and which together with their derivatives up to order r are bounded and possess continuous extensions to all of D. This space is equipped with the norm

$$\|f\|_{C^r(D,X)} = \sup_{|\alpha|\le r,\, s\in D} \left\| \frac{\partial^{|\alpha|} f(s)}{\partial s^\alpha} \right\|_X$$

with $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ and $\alpha = |\alpha_1| + \cdots + |\alpha_d|$. If $r = 0$ then we also write $C(D,X)$, and if $X = \mathbb{R}$ then we also write $C^r(D)$ and $C(D)$.

The type 2 constant of a Banach space X is denoted by $\tau_2(X)$. We refer to [21] as well as to the introductions in [3, 4] for this notion and related facts. Let $X \otimes Y$ be the algebraic tensor product of Banach spaces X and Y. For $z = \sum_{i=1}^n x_i \otimes y_i \in X \otimes Y$ define

$$\lambda(z) = \sup_{u\in B_{X^*},\, v\in B_{Y^*}} \left| \sum_{i=1}^n \langle x_i, u\rangle \, \langle y_i, v\rangle \right|.$$

The injective tensor product $X \otimes_\lambda Y$ is defined as the completion of $X \otimes Y$ with respect to the norm λ. Definitions and background on tensor products can be found in [6, 22]. Let us mention, in particular, the canonical isometric identification

$$C(D,X) = X \otimes_\lambda C(D) \tag{1}$$

for compact $D \subset \mathbb{R}^d$. We also note that for Banach spaces X_1, X_2, Y_1, Y_2 and operators $T_1 \in \mathscr{L}(X_1, Y_1)$, $T_2 \in \mathscr{L}(X_2, Y_2)$, the algebraic tensor product $T_1 \otimes T_2 : X_1 \otimes X_2 \to Y_1 \otimes Y_2$ extends to a bounded linear operator $T_1 \otimes T_2 \in \mathscr{L}(X_1 \otimes_\lambda X_2, Y_1 \otimes_\lambda Y_2)$ with

$$\|T_1 \otimes T_2\|_{\mathscr{L}(X_1\otimes_\lambda X_2, Y_1\otimes_\lambda Y_2)} = \|T_1\|_{\mathscr{L}(X_1,Y_1)} \|T_2\|_{\mathscr{L}(X_2,Y_2)}. \tag{2}$$

Let $Q = [0,1]^d$. For $r, m \in \mathbb{N}$ we let $P_m^{r,d,X} \in \mathscr{L}(C(Q,X))$ be d-tensor product Lagrange interpolation of degree r, composite with respect to the partition of $Q = [0,1]^d$ into m^d subcubes of sidelength m^{-1}. Thus, $P_m^{r,d,X}$ interpolates on Γ_{rm}^d, where $\Gamma_k^d = \left\{\frac{i}{k} : 0 \le i \le k\right\}^d$ for $k \in \mathbb{N}$. If $X = \mathbb{R}$, we write $P_m^{r,d}$. Note that in the sense of (1) we have $P_m^{r,d,X} = I_X \otimes P_m^{r,d}$. Furthermore, there are constants $c_1, c_2 > 0$ such that for all Banach spaces X and all m

$$\| P_m^{r,d,X} \|_{\mathscr{L}(C(Q,X))} \le c_1, \quad \sup_{f\in B_{C^r(Q,X)}} \|f - P_m^{r,d,X} f\|_{C(Q,X)} \le c_2 m^{-r}. \tag{3}$$

This is well-known in the scalar case, for the easy extension to Banach spaces see [3].

3 Banach Space Valued Indefinite Integration

Let X be a Banach space and let the indefinite integration operator be given by

$$S_0^X : C([0,1],X) \to C([0,1],X), \quad (S_0^X f)(t) = \int_0^t f(\tau)d\tau \quad (t \in [0,1]).$$

First we recall the Monte Carlo method from Section 4 of [14], here for integration domain $[0,1]$. Given $n \in \mathbb{N}$, we define $t_i = \frac{i}{n}$ $(0 \le i \le n)$. Let $\xi_i : \Omega \to [t_i, t_{i+1}]$ be independent uniformly distributed random variables on a probability space $(\Omega, \Sigma, \mathbb{P})$. For$f \in C(Q,X)$ and $\omega \in \Omega$ we define $g_\omega : \Gamma_n^1 \to \mathbb{R}$ by

$$g_\omega(t_i) = \frac{1}{n} \sum_{0 \le j < i} f(\xi_j(\omega)) \quad (0 \le i \le n).$$

Let $r \in \mathbb{N}_0$. If $r = 0$, we set

$$A_{n,\omega}^{0,0,X} f := P_n^{1,1,X} g_\omega,$$

and if $r \ge 1$,

$$A_{n,\omega}^{0,r,X} f = S_0^X(P_n^{r,1,X} f) + A_{n,\omega}^{0,0,X}(f - P_n^{r,1,X} f). \tag{4}$$

We write S_0 and $A_{n,\omega}^{0,r}$ if $X = \mathbb{R}$. Observe that in the sense of identification (1) we have

$$S_0^X = I_X \otimes S_0, \quad A_{n,\omega}^{0,r,X} = I_X \otimes A_{n,\omega}^{0,r}, \tag{5}$$

moreover, $A_{n,\omega}^{0,r,X} \in \mathscr{L}(C([0,1],X))$ and, since $g_\omega(0) = 0$,

$$\left(A_{n,\omega}^{0,r,X} f\right)(0) = 0 \quad (r \in \mathbb{N}_0). \tag{6}$$

We need the following result which complements Proposition 2 in [3].

Proposition 1 *Let $r \in \mathbb{N}_0$. Then there are constants $c_1, c_2 > 0$ such that for all Banach spaces X, $n \in \mathbb{N}$, $\omega \in \Omega$, and$f \in C([0,1],X)$ we have*

$$\|S_0^X f - A_{n,\omega}^{0,r,X} f\|_{C([0,1],X)} \le c_1 \|f\|_{C([0,1],X)} \tag{7}$$

$$(\mathbb{E}\,\|S_0^X f - A_{n,\omega}^{0,r,X} f\|_{C([0,1],X)}^2)^{1/2} \le c_2 \tau_2(X) n^{-1/2} \|f\|_{C([0,1],X)}. \tag{8}$$

Proof Relation (7) directly follows from the definitions. By Proposition 2 in [3], there is a constant $c > 0$ such that for all Banach spaces X, $n \in \mathbb{N}$, and $f \in C([0,1],X)$ we have

$$\left(\mathbb{E}\,\|S_0^X f - A_{n,\omega}^{0,0,X} f\|_{C([0,1],X)}^2\right)^{1/2} \le c\tau_2(X) n^{-1/2} \|f\|_{C([0,1],X)}. \tag{9}$$

This is the case $r = 0$ of (8). Now assume $r \geq 1$. Then (3), (4), and (9) give

$$\begin{aligned}
&\left(\mathbb{E}\,\|S_0^X f - A_{n,\omega}^{0,r,X} f\|^2_{C([0,1],X)}\right)^{1/2} \\
&= \left(\mathbb{E}\,\|S_0^X(f - P_n^{r,1,X} f) - A_{n,\omega}^{0,0,X}(f - P_n^{r,1,X} f)\|^2_{C([0,1],X)}\right)^{1/2} \\
&\leq c\tau_2(X) n^{-1/2} \|f - P_n^{r,1,X} f\|_{C([0,1],X)} \leq c\tau_2(X) n^{-1/2} \|f\|_{C([0,1],X)}.
\end{aligned}$$

□

We shall further study the multilevel procedure developed in [3]. Let $(T_l)_{l=0}^\infty \subset \mathscr{L}(X)$. For convenience we introduce the following parameter set

$$\mathscr{M} = \left\{\left(l_0, l_1, (n_{l_0})_{l=l_0}^{l_1}\right) : \, l_0, l_1 \in \mathbb{N}_0, \, l_0 \leq l_1, \, (n_{l_0})_{l=l_0}^{l_1} \subset \mathbb{N}\right\}. \tag{10}$$

For $\mu \in \mathscr{M}$ we define an approximation $A_{\mu,\omega}^{0,r}$ to S_0^X as follows:

$$A_{\mu,\omega}^{0,r} = T_{l_0} \otimes A_{n_{l_0},\omega}^{0,r} + \sum_{l=l_0+1}^{l_1} (T_l - T_{l-1}) \otimes A_{n_l,\omega}^{0,0}, \tag{11}$$

where the tensor product is understood in the sense of (1). We assume that the random variables $A_{n_{l_0},\omega}^{0,r}$ and $\left(A_{n_l,\omega}^{0,0}\right)_{l=l_0+1}^{l_1}$ are independent. We have

$$A_{\mu,\omega}^{0,r} \in \mathscr{L}(C([0,1],X)).$$

Denote

$$X_l = \mathrm{cl}_X(T_l(X)) \quad (l \in \mathbb{N}_0) \tag{12}$$

$$X_{l-1,l} = \mathrm{cl}_X((T_l - T_{l-1})(X)) \quad (l \in \mathbb{N}), \tag{13}$$

where cl_X denotes the closure in X. In particular, X_l and $X_{l-1,l}$ are endowed with the norm induced by X. The following result complements Proposition 3 in [3].

Proposition 2 *There is a constant $c > 0$ such that for all Banach spaces X and operators $(T_l)_{l=0}^\infty$ as above, for all $\mu \in \mathscr{M}$*

$$\begin{aligned}
&\sup_{f \in B_{C([0,1],X)}} \left(\mathbb{E}\,\|S_0^X f - A_{\mu,\omega}^{0,r} f\|^2_{C([0,1],X)}\right)^{1/2} \\
&\leq \|I_X - T_{l_1}\|_{\mathscr{L}(X)} + c\tau_2(X_{l_0})\|T_{l_0}\|_{\mathscr{L}(X)} n_{l_0}^{-1/2} \\
&\quad + c \sum_{l=l_0+1}^{l_1} \tau_2(X_{l-1,l}) \|(T_l - T_{l-1})\|_{\mathscr{L}(X)} n_l^{-1/2}.
\end{aligned} \tag{14}$$

Proof Denote

$$R_l = T_l \otimes I_{C([0,1])} \in \mathscr{L}(C([0,1],X)). \tag{15}$$

From (5) and (11) we get

$$A^{0,r}_{\mu,\omega} = A^{0,r,X}_{n_{l_0},\omega} R_{l_0} + \sum_{l=l_0+1}^{l_1} A^{0,0,X}_{n_l,\omega}(R_l - R_{l-1}). \tag{16}$$

To prove (14), let $f \in B_{C([0,1],X)}$. Then by (16),

$$\begin{aligned} &\|S^X_0 f - A^{0,r}_{\mu,\omega} f\|_{C([0,1],X)} \\ &\le \|S^X_0 f - S^X_0 R_{l_1} f\|_{C([0,1],X)} + \|S^X_0 R_{l_0} f - A^{0,r,X}_{n_{l_0},\omega} R_{l_0} f\|_{C([0,1],X_{l_0})} \\ &\quad + \left\| \sum_{l=l_0+1}^{l_1} \left(S^X_0 (R_l - R_{l-1}) f - A^{0,0,X}_{n_l,\omega}(R_l - R_{l-1}) f \right) \right\|_{C([0,1],X_{l-1,l})}. \end{aligned} \tag{17}$$

We have, using (2),

$$\begin{aligned} &\|S^X_0 f - S^X_0 R_{l_1} f\|_{C([0,1],X)} \le \|S^X_0\|_{\mathscr{L}(C([0,1],X))} \|f - R_{l_1} f\|_{C([0,1],X)} \\ &\le \|I_X - T_{l_1}\|_{\mathscr{L}(X)} \|f\|_{C([0,1],X)} \le \|I_X - T_{l_1}\|_{\mathscr{L}(X)}. \end{aligned} \tag{18}$$

Furthermore, by Proposition 1,

$$\begin{aligned} &\mathbb{E} \left(\|S^X_0 R_{l_0} f - A^{0,r,X}_{n_{l_0},\omega} R_{l_0} f\|^2_{C([0,1],X_{l_0})} \right)^{1/2} \\ &\le c\tau_2(X_{l_0}) \|T_{l_0}\|_{\mathscr{L}(X)} n_{l_0}^{-1/2}. \end{aligned} \tag{19}$$

For $l_0 < l \le l_1$ we obtain

$$\begin{aligned} &\mathbb{E} \left(\|S^X_0 (R_l - R_{l-1}) f - A^{0,0,X}_{n_l,\omega}(R_l - R_{l-1}) f\|^2_{C([0,1],X_{l-1,l})} \right)^{1/2} \\ &\le c\tau_2(X_{l-1,l}) \|T_l - T_{l-1}\|_{\mathscr{L}(X)} n_l^{-1/2}. \end{aligned} \tag{20}$$

The combination of (17)–(20) yields the result. □

4 Parametric Indefinite Integration

Let $d \in \mathbb{N}$, $Q = [0,1]^d$. The indefinite parametric integration operator $S_1 : C(Q \times [0,1]) \to C(Q \times [0,1])$ is given by

$$(S_1 f)(s,t) = \int_0^t f(s,\tau)d\tau \quad (s \in Q, t \in [0,1]).$$

This problem is related to the Banach space case from the previous section as follows. With $X = C(Q)$ we have the identifications

$$C(Q \times [0,1]) = C([0,1], X), \quad S_1 = S_0^{C(Q)}.$$

Let $r_1 = \max(r,1)$ and define for $l \in \mathbb{N}_0$

$$T_l = P_{2^l}^{r_1,d} \in \mathscr{L}(C(Q)). \tag{21}$$

By (3),

$$\|T_l\|_{\mathscr{L}(C(Q))} \le c_1, \quad \|J - T_l J\|_{\mathscr{L}(C^r(Q),C(Q))} \le c_2 2^{-rl}, \tag{22}$$

where $J : C^r(Q) \to C(Q)$ is the embedding. For $\mu = \left(l_0, l_1, (n_{l_0})_{l=l_0}^{l_1}\right) \in \mathscr{M}$ the algorithm $A_{\mu,\omega}^{0,r}$ defined in (11) takes the following form. For $f \in C(Q \times [0,1])$

$$\begin{aligned} A_{\mu,\omega}^{1,r} f = {} & P_{2^{l_0}}^{r_1,d}\left(\left(A_{n_{l_0},\omega}^{0,r}(f_s)\right)_{s \in \Gamma_{r_1 2^{l_0}}^d}\right) \\ & + \sum_{l=l_0+1}^{l_1} \left(P_{2^l}^{r_1,d} - P_{2^{l-1}}^{r_1,d}\right)\left(\left(A_{n_l,\omega}^{0,0}(f_s)\right)_{s \in \Gamma_{r_1 2^l}^d}\right), \end{aligned} \tag{23}$$

where for $s \in Q$ we used the notation $f_s = f(s, \cdot)$. Then

$$\operatorname{card}\left(A_{\mu,\omega}^{1,r}\right) \le c \sum_{l=l_0}^{l_1} n_l 2^{dl} \quad (\omega \in \Omega), \tag{24}$$

where $\operatorname{card}\left(A_{\mu,\omega}^{1,r}\right)$ denotes the cardinality of algorithm $A_{\mu,\omega}^{1,r}$, that is, the number of function values used in algorithm $A_{\mu,\omega}^{1,r}$ (see also the general remarks before Theorem 1 below). Moreover, we have $A_{\mu,\omega}^{1,r} \in \mathscr{L}(C(Q \times [0,1]))$ and it follows from (6) that

$$\left(A_{\mu,\omega}^{1,r} f\right)(s,0) = 0 \quad (s \in Q). \tag{25}$$

We also consider the following subset $\mathscr{M}_0 \subset \mathscr{M}$ corresponding to one-level algorithms

$$\mathscr{M}_0 = \{\mu \in \mathscr{M} : \mu = (l_0, l_0, n_{l_0})\} \tag{26}$$

thus, for $\mu_0 \in \mathscr{M}_0$,

$$A^{1,r}_{\mu_0,\omega} f = P^{r_1,d}_{2^{l_0}} \left(\left(A^{0,r}_{n_{l_0},\omega}(f_s) \right)_{s \in \Gamma^d_{r_1 2^{l_0}}} \right). \tag{27}$$

Parts of the following result (relations (30) and (33)) were shown in [3, Proposition 4]. We prove that algorithm $A^{1,r}_{\mu,\omega}$ simultaneously satisfies the estimates (32) and (33). The former is crucial for the stability analysis of the iteration in Sect. 5. We note that, due to the multilevel structure of $A^{1,r}_{\mu,\omega}$ relation (32) is not trivial (a trivial estimate would be $c\log(n+1)$). Some of the choices of multilevel parameters from [3] are not suitable to obtain both estimate simultaneously. So here we provide modified choices and verify the needed estimates for them, still using the analysis of [3].

Proposition 3 *Let $r \in \mathbb{N}_0$, $d \in \mathbb{N}$. There are constants $c_1, \dots, c_6 > 0$ such that the following hold. For each $n \in \mathbb{N}$ there is a $\mu_0(n) \in \mathscr{M}_0$ such that for all $\omega \in \Omega$*

$$\operatorname{card}\left(A^{1,r}_{\mu_0(n),\omega}\right) \le c_1 n \tag{28}$$

$$\sup_{f \in B_{C(Q\times[0,1])}} \|S_1 f - A^{1,r}_{\mu_0(n),\omega} f\|_{C(Q\times[0,1])} \le c_2 \tag{29}$$

$$\sup_{f \in B_{C^r(Q\times[0,1])}} \|S_1 f - A^{1,r}_{\mu_0(n),\omega} f\|_{C(Q\times[0,1])} \le c_3 n^{-\frac{r}{d+1}}. \tag{30}$$

Moreover, for each $n \in \mathbb{N}$ there is a $\mu(n) \in \mathscr{M}$ such that

$$\max_{\omega\in\Omega} \operatorname{card}\left(A^{1,r}_{\mu(n),\omega}\right) \le c_4 n \tag{31}$$

$$\sup_{f \in B_{C(Q\times[0,1])}} \left(\mathbb{E}\, \|S_1 f - A^{1,r}_{\mu(n),\omega} f\|^2_{C(Q\times[0,1])}\right)^{1/2} \le c_5 \tag{32}$$

$$\sup_{f \in B_{C^r(Q\times[0,1])}} \left(\mathbb{E}\, \|S_1 f - A^{1,r}_{\mu(n),\omega} f\|^2_{C(Q\times[0,1])}\right)^{1/2} \le c_6 n^{-\gamma_1} (\log(n+1))^{\gamma_2} \tag{33}$$

with

$$\gamma_1 = \begin{cases} \frac{r+1/2}{d+1} & \text{if} \quad \frac{r}{d} > \frac{1}{2} \\ \frac{r}{d} & \text{if} \quad \frac{r}{d} \le \frac{1}{2} \end{cases} \qquad \gamma_2 = \begin{cases} \frac{1}{2} & \text{if} \quad \frac{r}{d} > \frac{1}{2} \\ 2 & \text{if} \quad \frac{r}{d} = \frac{1}{2} \\ \frac{r}{d} & \text{if} \quad \frac{r}{d} < \frac{1}{2}. \end{cases} \tag{34}$$

Proof Let $n \in \mathbb{N}$, put

$$l^* = \left\lceil \frac{\log_2(n+1)}{d} \right\rceil, \quad l_0 = \left\lfloor \frac{d}{d+1} l^* \right\rfloor, \quad n_{l_0} = 2^{d(l^*-l_0)} \tag{35}$$

and $\mu_0(n) = (l_0, l_0, n_{l_0})$. For this choice relations (28) and (30) were shown in [3]. Relation (29) readily follows from (3), (7) of Proposition 1, and (27).

To prove (31)–(34), let $\mu(n) = \left(l_0, l_1, (n_{l_0})_{l=l_0}^{l_1}\right) \in \mathscr{M}$, with l_0 and n_{l_0} given by (35), and l_1, $(n_{l_0})_{l=l_0+1}^{l_1}$ to be fixed later on. For brevity we denote for $\varrho \in \mathbb{N}_0$,

$$E_\varrho(\mu(n)) := \sup_{f \in B_{C^\varrho(Q\times[0,1])}} \left(\mathbb{E}\, \| S_1 f - A^{1,r}_{\mu(n),\omega} f \|^2_{C(Q\times[0,1])} \right)^{1/2}.$$

We show that for $\varrho \in \{0, r\}$

$$E_\varrho(\mu(n)) \le c 2^{-\varrho l_1} + c(l_0+1)^{1/2} n_{l_0}^{-\varrho-1/2} + c \sum_{l=l_0+1}^{l_1} (l+1)^{1/2} 2^{-\varrho l} n_l^{-1/2}. \tag{36}$$

By (63) in [3], this holds for $\varrho = r$. It remains to prove the corresponding estimate for $r \ge 1$, $\varrho = 0$. By (12) and (21)

$$X_l = P^{r_1,d}_{2^l}(C(Q)), \tag{37}$$

therefore $X_{l-1} \subseteq X_l$ and, by (13), also $X_{l-1,l} \subseteq X_l$ for $l \ge 1$. As shown in [3],

$$\tau_2(X_{l-1,l}) \le \tau_2(X_l) \le c(l+1)^{1/2}. \tag{38}$$

We conclude from (14) of Proposition 2, (22), and (38) that

$$E_0(\mu(n)) \le c + c(l_0+1)^{1/2} n_{l_0}^{-1/2} + c \sum_{l=l_0+1}^{l_1} (l+1)^{1/2} n_l^{-1/2}, \tag{39}$$

which shows (36) for $\varrho = 0$.

From (35) we conclude

$$d(l^* - l_0) \ge \frac{dl^*}{d+1} \ge l_0,$$

thus,

$$n_{l_0}^{-\varrho-1/2} = 2^{-(\varrho+1/2)d(l^*-l_0)} \le 2^{-\varrho l_0 - d(l^*-l_0)/2} = 2^{-\varrho l_0} n_{l_0}^{-1/2}.$$

This means that we can include the middle term in (36) into the sum, which gives

$$E_\varrho(\mu(n)) \le c2^{-\varrho l_1} + c\sum_{l=l_0}^{l_1}(l+1)^{1/2}2^{-\varrho l}n_l^{-1/2} \quad (\varrho \in \{0, r\}). \tag{40}$$

If $r > d/2$, we set

$$\gamma = \frac{(r+1/2)d}{r(d+1)}, \quad l_1 = \lceil \gamma l^* \rceil .$$

Then

$$\frac{d}{d+1} < \gamma < 1.$$

Indeed, the left hand inequality is obvious, while the right-hand inequality is a consequence of the assumption $r > d/2$. With (35) it follows that

$$l_0 \le l_1 \le l^*.$$

We choose a $\delta > 0$ in such a way that

$$r - \delta/2 > d/2, \tag{41}$$

$$\delta\left(\gamma - \frac{d}{d+1}\right) < d(1-\gamma) \tag{42}$$

and put

$$n_l = \left\lceil 2^{d(l^*-l)-\delta(l-l_0)} \right\rceil \quad (l = l_0+1, \dots, l_1).$$

From (40)–(42) and (35) we obtain

$$\begin{aligned} E_r(\mu(n)) &\le c2^{-rl_1} + c(l^*+1)^{1/2}\sum_{l=l_0}^{l_1} 2^{-rl_0-(r-\delta/2)(l-l_0)-d(l^*-l)/2} \\ &\le c2^{-\frac{(r+1/2)d}{d+1}l^*} + c(l^*+1)^{1/2}2^{-rl_0-d(l^*-l_0)/2} \\ &\le c(l^*+1)^{1/2}2^{-\frac{(r+1/2)d}{d+1}l^*} \le cn^{-\frac{r+1/2}{d+1}}(\log(n+1))^{1/2}. \end{aligned}$$

Furthermore, using (40) and (42),

$$\begin{aligned} E_0(\mu(n)) &\le c + c(l^* + 1)^{1/2} \sum_{l=l_0}^{l_1} 2^{\delta(l-l_0)/2 - d(l^*-l)/2} \\ &\le c + c(l^* + 1)^{1/2} 2^{\delta(l_1-l_0)/2 - d(l^*-l_1)/2} \\ &\le c + c(l^* + 1)^{1/2} 2^{\delta\left(\gamma - \frac{d}{d+1}\right)\frac{l^*}{2} - d(1-\gamma)\frac{l^*}{2}} \le c. \end{aligned}$$

By (24) the number of function values fulfills

$$\operatorname{card}\left(A^{1,r}_{\mu(n),\omega}\right) \le c \sum_{l=l_0}^{l_1} n_l 2^{dl} \le c2^{dl_1} + c \sum_{l=l_0}^{l_1} 2^{dl^* - \delta(l-l_0)} \le cn. \tag{43}$$

This proves (31)–(34) for $r > d/2$.

If $r = d/2$, we set $l_1 = l^*$, put

$$n_l = \max\left(2^{d(l^*-l)}, \left\lceil (l^* + 1)2^{d(l^*-l)/2} \right\rceil\right) \quad (l = l_0 + 1, \dots, l_1)$$

and get from (35) and (40),

$$\begin{aligned} E_r(\mu(n)) &\le c2^{-rl^*} + c(l^* + 1)^{1/2} \sum_{l=l_0}^{l^*} 2^{-rl - d(l^*-l)/2} \\ &\le c(l^* + 1)^{3/2} 2^{-dl^*/2} \le cn^{-1/2}(\log(n + 1))^{3/2}, \end{aligned} \tag{44}$$

$$E_0(\mu(n)) \le c + c(l^* + 1)^{1/2} \sum_{l=l_0}^{l_1} n_l^{-1/2} \le c + c \sum_{l=l_0}^{l^*} 2^{-d(l^*-l)/4} \le c. \tag{45}$$

The cardinality satisfies

$$\begin{aligned} \operatorname{card}\left(A^{1,r}_{\mu(n),\omega}\right) &\le c \sum_{l=l_0}^{l^*} n_l 2^{dl} \le c2^{dl^*} + c \sum_{l=l_0}^{l^*} \left(2^{dl^*} + (l^* + 1)2^{d(l^*+l)/2}\right) \\ &\le c(l^* + 1)2^{dl^*} \le cn\log(n + 1). \end{aligned} \tag{46}$$

Transforming $n\log(n + 1)$ into n in relations (44)–(46) proves (31)–(34) for this case.

Finally, if $r < d/2$, we set

$$l_1 = l^* - \left\lceil d^{-1}\log_2(l^* + 1)\right\rceil, \tag{47}$$

choose a $\delta > 0$ in such a way that

$$(d-\delta)/2 > r \tag{48}$$

and put

$$n_l = \left\lceil 2^{d(l^*-l)-\delta(l_1-l)} \right\rceil \quad (l = l_0+1,\dots,l_1). \tag{49}$$

This is the same choice as in the respective case of the proof of Proposition 4 in [3]. Clearly, there is a constant $c > 0$ such that $l_0 \le l_1$ for $n \ge c$. For $n < c$ the statements (32) and (33) are trivial. It was shown in [3] that with the choice above $\operatorname{card}\big(A^{1,r}_{\mu(n),\omega}\big) \le cn$ and that relation (33) holds. Arguing similarly, we derive from (35), (36), (47), and (49) for the case $\varrho = 0$

$$\begin{aligned} E_0(\mu(n)) &\le c + c(l^*+1)^{1/2} 2^{-d(l^*-l_0)/2} \\ &\quad + c(l^*+1)^{1/2} \sum_{l=l_0+1}^{l_1} 2^{-d(l^*-l_1)/2-(d-\delta)(l_1-l)/2} \\ &\le c + c(l^*+1)^{1/2} 2^{-d(l^*-l_1)/2} \\ &\le c + c(l^*+1)^{1/2} 2^{-(\log_2(l^*+1))/2} \le c, \end{aligned}$$

which is (32). □

5 Fixed Point Iteration for Parametric ODEs

Here we apply the above results to the following problem. Let $d, q \in \mathbb{N}$, $r \in \mathbb{N}_0$, $Q = [0,1]^d$, and let $C^r_{\mathrm{Lip}}(Q \times [0,1] \times \mathbb{R}^q, \mathbb{R}^q)$ be the space of functions $f \in C^r(Q \times [0,1] \times \mathbb{R}^q, \mathbb{R}^q)$ satisfying for $s \in Q$, $t \in [0,1]$, $z_1, z_2 \in \mathbb{R}^q$

$$|f|_{\mathrm{Lip}} := \sup_{s\in Q, t\in[0,1], z_1 \ne z_2 \in \mathbb{R}^q} \frac{\|f(s,t,z_1) - f(s,t,z_2)\|_{\mathbb{R}^q}}{\|z_1 - z_2\|_{\mathbb{R}^q}} < \infty. \tag{50}$$

The space $C^r_{\mathrm{Lip}}(Q \times [0,1] \times \mathbb{R}^q, \mathbb{R}^q)$ is endowed with the norm

$$\|f\|_{C^r_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)} = \max\left(\|f\|_{C^r(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}, |f|_{\mathrm{Lip}}\right). \tag{51}$$

If $r = 0$, we also write $C_{\mathrm{Lip}}(Q \times [0,1] \times \mathbb{R}^q, \mathbb{R}^q)$. We consider the numerical solution of initial value problems for systems of ODEs depending on a parameter $s \in Q$

$$\frac{\partial u(s,t)}{\partial t} = f(s,t,u(s,t)) \quad (s \in Q, t \in [0,1]) \tag{52}$$

$$u(s,0) = u_0(s) \quad (s \in Q) \tag{53}$$

with $f \in C_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)$ and $u_0 \in C(Q,\mathbb{R}^q)$. A function $u : Q\times[0,1] \to \mathbb{R}^q$ is called a solution if for each $s \in Q$, $u(s,t)$ is continuously differentiable as a function of t and (52)–(53) are satisfied. Due to the assumptions on f and u_0 the solution exists, is unique, and belongs to $C(Q\times[0,1],\mathbb{R}^q)$. Let the solution operator

$$S_2 : C_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)\times C(Q,\mathbb{R}^q) \to C(Q\times[0,1],\mathbb{R}^q)$$

be given by $S_2(f,u_0) = u$, where $u = u(s,t)$ is the solution of (52)–(53). Furthermore, fix $\kappa > 0$ and let

$$F_2^r(\kappa) = \kappa B_{C^r_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)} \times \kappa B_{C^r(Q,\mathbb{R}^q)}. \tag{54}$$

Classical results on the regularity with respect to t and the parameter s (see, e.g., [24]) give

$$\sup_{(f,u_0)\in F_2^r(\kappa)} \|S_2(f,u_0)\|_{C^r(Q\times[0,1],\mathbb{R}^q)} \le c. \tag{55}$$

Now let $f \in C_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)$ and $u_0 \in C(Q,\mathbb{R}^q)$. We rewrite (52)–(53) in the equivalent form

$$u(s,t) = u_0(s) + \int_0^t f(s,\tau,u(s,\tau))d\tau \quad (s\in Q, t\in[0,1]). \tag{56}$$

Let $m \in \mathbb{N}$ and $t_i = im^{-1}$ $(i = 0,\dots,m)$. We solve (56) and thus (52)–(53) in m steps on the intervals $[t_i,t_{i+1}]$ $(i = 0,\dots,m-1)$. Let

$$S_{1,i} : C(Q\times[t_i,t_{i+1}],\mathbb{R}^q) \to C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)$$

be the q-dimensional version of the solution operator of parametric indefinite integration on $[t_i,t_{i+1}]$, i.e., for $g \in C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)$

$$\big(S_{1,i}g\big)(s,t) = \left(\int_{t_i}^t g_l(s,\tau)d\tau\right)_{l=1}^q \quad (t\in[t_i,t_{i+1}]), \tag{57}$$

where g_l are the components of g. Let $A^{1,r}_{\mu,i,\omega}$ be algorithm $A^{1,r}_{\mu,\omega}$ from (23), scaled to $[t_i,t_{i+1}]$ and applied to each component of g, that is

$$\big(A^{1,r}_{\mu,i,\omega}g\big)(s,t) = \Big(m^{-1}\big(A^{1,r}_{\mu,\omega}g_l^*\big)(s,m(t-t_i))\Big)_{l=1}^q \quad (t\in[t_i,t_{i+1}]), \tag{58}$$

with

$$g_l^*(s,\tau) = g_l(s,t_i+m^{-1}\tau) \quad (\tau\in[0,1]). \tag{59}$$

Let

$$\mathscr{N} = \left\{\left(m, M, k, (\mu_j)_{j=0}^{k-1}\right) : m, M, k \in \mathbb{N},\ (\mu_j)_{j=0}^{k-1} \subset \mathscr{M}\right\} \tag{60}$$

$$\mathscr{N}_0 = \left\{\left(m, M, k, (\mu_j)_{j=0}^{k-1}\right) \in \mathscr{N} : (\mu_j)_{j=0}^{k-1} \subset \mathscr{M}_0\right\} \subset \mathscr{N}, \tag{61}$$

where $\mathscr{M}$ and $\mathscr{M}_0$ were defined in (10) and (26), respectively, and let $r_1 = \max(r, 1)$. For $\nu = \left(m, M, k, (\mu_j)_{j=0}^{k-1}\right) \in \mathscr{N}$ define $u_{0,0} = P_M^{r_1,d} u_0$ and for $i = 0, \ldots, m-1, j = 0, \ldots, k-1, s \in Q$ the iteration

$$u_{i,j+1}(s,t) = u_{i,0}(s) + (A^{1,r}_{\mu_j,i,\omega} g_{ij})(s,t) \quad (t \in [t_i, t_{i+1}]), \tag{62}$$

where

$$g_{ij}(s,t) = f(s,t,u_{ij}(s,t)) \quad (t \in [t_i, t_{i+1}]), \tag{63}$$

and

$$u_{i+1,0}(s) = u_{ik}(s, t_{i+1}) \qquad (t \in [t_{i+1}, t_{i+2}],\ i \le m-2). \tag{64}$$

We assume that the involved random variables $(A^{1,r}_{\mu_j,i,\omega})_{i,j=0}^{m-1,k-1}$ are independent. Furthermore, for $s \in Q$, $t \in [0,1]$ put

$$u_\nu(s,t) = \begin{cases} u_{ik}(s,t) & \text{if} \quad t \in [t_i, t_{i+1}),\ i \le m-2 \\ u_{m-1,k}(s,t) & \text{if} \quad t \in [t_{m-1}, t_m] \end{cases} \tag{65}$$

$$A^{2,r}_{\nu,\omega}(f, u_0) = u_\nu. \tag{66}$$

Clearly, $u_{ij} \in C(Q \times [t_i, t_{i+1}], \mathbb{R}^q)$. Moreover, it follows from (58) and (25) that

$$(A^{1,r}_{\mu_j,i,\omega} g_{ij})(s, t_i) = 0,$$

and therefore (62) yields

$$u_{ik}(s, t_i) = u_{i,0}(s) = u_{i-1,k}(s, t_i) \quad (1 \le i \le m-1,\ s \in Q),$$

hence $v \in C(Q \times [0,1], \mathbb{R}^q)$. Next we give error and stability estimates for $A^{2,r}_{\nu,\omega}$.

Proposition 4 *Let $r \in \mathbb{N}_0$, $d, q \in \mathbb{N}$, $\kappa > 0$. Then there are constants $c_1, \ldots, c_6 > 0$ such that the following hold. For each $n \in \mathbb{N}$ there is a $\nu_0(n) \in \mathscr{N}_0$ such that for all $\omega \in \Omega$*

$$\operatorname{card}\left(A^{2,r}_{\nu_0(n),\omega}\right) \le c_1 n \tag{67}$$

$$\sup_{(f,u_0) \in F_2^r(\kappa)} \|S_2(f,u_0) - A^{2,r}_{\nu_0(n),\omega}(f,u_0)\|_{C(Q\times[0,1],\mathbb{R}^q)} \le c_2 n^{-\frac{r}{d+1}} \tag{68}$$

and for $(f, u_0), (\tilde{f}, \tilde{u}_0) \in F_2^0(\kappa)$

$$\begin{aligned}&\|A^{2,r}_{\nu_0(n),\omega}(f, u_0) - A^{2,r}_{\nu_0(n),\omega}(\tilde{f}, \tilde{u}_0)\|_{C(Q\times[0,1],\mathbb{R}^q)} \\ &\le c_3\big(\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)} + \|u_0 - \tilde{u}_0\|_{C(Q,\mathbb{R}^q)}\big). \end{aligned} \tag{69}$$

Moreover, for each $n \in \mathbb{N}$ *there is a* $\nu(n) \in \mathscr{N}$ *such that*

$$\max_{\omega\in\Omega} \operatorname{card}\left(A^{2,r}_{\nu(n),\omega}\right) \le c_4 n, \tag{70}$$

$$\begin{aligned}&\sup_{(f,u_0)\in F_2^r(\kappa)} \left(\mathbb{E}\, \|S_2(f, u_0) - A^{2,r}_{\nu(n),\omega}(f, u_0)\|^2_{C(Q\times[0,1],\mathbb{R}^q)}\right)^{1/2} \\ &\qquad\le c_5 n^{-\gamma_1}(\log(n + 1))^{\gamma_2}, \end{aligned} \tag{71}$$

with γ_1 *and* γ_2 *given by* (34)*, and for* $(f, u_0), (\tilde{f}, \tilde{u}_0) \in F_2^0(\kappa)$

$$\begin{aligned}&\left(\mathbb{E}\, \|A^{2,r}_{\nu(n),\omega}(f, u_0) - A^{2,r}_{\nu(n),\omega}(\tilde{f}, \tilde{u}_0)\|^2_{C(Q\times[0,1],\mathbb{R}^q)}\right)^{1/2} \\ &\le c_6\big(\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)} + \|u_0 - \tilde{u}_0\|_{C(Q,\mathbb{R}^q)}\big). \end{aligned} \tag{72}$$

Proof We prove (70)–(72), the proof of (67)–(69) is analogous, just simpler. For the sake of brevity we set

$$\varepsilon(n) = n^{-\gamma_1}(\log(n + 1))^{\gamma_2}. \tag{73}$$

By Proposition 3 and (57)–(59) there are constants $c, c(1), c(2) > 0$ and a sequence $(\mu(n))_{n=1}^{\infty} \subset \mathscr{M}$ such that for $m, n \in \mathbb{N}$

$$\max_{\omega\in\Omega} \operatorname{card}\left(A^{1,r}_{\mu(n),i,\omega}\right) \le cn, \tag{74}$$

for $f \in C(Q \times [t_i, t_{i+1}], \mathbb{R}^q)$

$$\left(\mathbb{E}\, \|S_{1,i}f - A^{1,r}_{\mu(n),i,\omega}f\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \le c(1)m^{-1}\|f\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \tag{75}$$

and for $f \in C^r(Q \times [t_i, t_{i+1}], \mathbb{R}^q)$

$$\begin{aligned}&\left(\mathbb{E}\, \|S_{1,i}f - A^{1,r}_{\mu(n),i,\omega}f\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\ &\qquad\le c(2)m^{-1}\varepsilon(n)\|f\|_{C^r(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}. \end{aligned} \tag{76}$$

In the rest of the proof we reserve the notation $c(1)$ and $c(2)$ for the constants in (75) and (76). We need the following stability property, which is a consequence of (75) and the linearity of $A_{\mu(n),i,\omega}$: For $f_1, f_2 \in C(Q \times [t_i, t_{i+1}], \mathbb{R}^q)$

$$\begin{aligned}
&\left(\mathbb{E}\, \|A^{1,r}_{\mu(n),i,\omega} f_1 - A^{1,r}_{\mu(n),i,\omega} f_2\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \|S_{1,i}(f_1 - f_2)\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \\
&\quad + \left(\mathbb{E}\, \|S_{1,i}(f_1 - f_2) - A^{1,r}_{\mu(n),i,\omega}(f_1 - f_2)\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le (c(1) + 1)m^{-1} \|f_1 - f_2\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}. \qquad (77)
\end{aligned}$$

We choose $m \in \mathbb{N}$ in such a way that

$$\theta := (c(1) + 1)m^{-1}\kappa \le 1/2. \qquad (78)$$

Now we fix $n \in \mathbb{N}$ and define

$$M = \lceil n^{1/d} \rceil, \quad k = \left\lfloor \frac{\gamma_1 \log_2 n + \log_2 m}{-\log_2 \theta} \right\rfloor + 1, \qquad (79)$$

$$n_j = \left\lceil n\theta^{\frac{k-1-j}{\gamma_1+1}} \right\rceil \quad (j = 0, \dots, k-1), \qquad (80)$$

and set

$$\nu(n) = \left(m, M, k, \mu(n_j)_{j=0}^{k-1}\right).$$

Then the cardinality of algorithm $A^{2,r}_{\nu(n),\omega}$ satisfies

$$\operatorname{card}\left(A^{2,r}_{\nu(n),\omega}\right) \le cM^d + cm \sum_{j=0}^{k-1} n_j \le cn + cm \sum_{j=0}^{k-1} \left\lceil n\theta^{\frac{k-1-j}{\gamma_1+1}} \right\rceil \le cn$$

(note that by the choice (78), m is just a constant). This shows (70).

Next we prove the error estimate (71). Let $(f, u_0) \in F_2^r(\kappa)$. By (3) and (34)

$$\|u(\,\cdot\,, 0) - u_{0,0}\|_{C(Q,\mathbb{R}^q)} = \left\|u_0 - P_M^{r_1,d} u_0\right\|_{C(Q,\mathbb{R}^q)} \le cn^{-r/d} \le cn^{-\gamma_1}. \qquad (81)$$

Setting

$$g(s, t) = f(s, t, u(s, t)), \qquad (82)$$

we get from (55)

$$\|g\|_{C^r(Q\times[0,1],\mathbb{R}^q)} \le c. \qquad (83)$$

Moreover, (63) implies

$$\|g - g_{ij}\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \le \kappa \|u - u_{ij}\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}. \tag{84}$$

We have

$$u(s,t) = u(s,t_i) + (S_{1,i}g)(s,t) \quad (s \in Q, t \in [t_i, t_{i+1}]).$$

We estimate, using (83), (84), (76), and (77)

$$\begin{aligned}
&\left(\mathbb{E}\, \|u - u_{i,j+1}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \left(\mathbb{E}\, \left\|u(\,\cdot\,,t_i) + S_{1,i}g - u_{i,0} - A^{1,r}_{\mu(n_j),i,\omega} g_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} + \left(\mathbb{E}\, \left\|S_{1,i}g - A^{1,r}_{\mu(n_j),i,\omega} g\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\quad + \left(\mathbb{E}\,\mathbb{E}\, \left(\left\|A^{1,r}_{\mu(n_j),i,\omega} g - A^{1,r}_{\mu(n_j),i,\omega} g_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \Big| u_{ij}\right)\right)^{1/2} \\
&\le \left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} + \left(\mathbb{E}\, \left\|S_{1,i}g - A^{1,r}_{\mu(n_j),i,\omega} g\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\quad + (c(1)+1)m^{-1} \left(\mathbb{E}\, \|g - g_{ij}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} + c(1)m^{-1}\varepsilon(n_j) \\
&\quad + \theta \left(\mathbb{E}\, \|u - u_{ij}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}.
\end{aligned} \tag{85}$$

We get from (85) by recursion over j

$$\begin{aligned}
&\left(\mathbb{E}\, \|u - u_{ik}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} \sum_{j=0}^{k-1} \theta^j + c(1)m^{-1} \sum_{j=0}^{k-1} \theta^j \varepsilon(n_{k-j-1}) \\
&\quad + \theta^k \left(\mathbb{E}\, \|u - u_{i,0}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\
&\le \theta^k \|u - u(\,\cdot\,,t_i)\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} + \left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} \sum_{j=0}^{k} \theta^j \\
&\quad + c(1)m^{-1} \sum_{j=0}^{k-1} \theta^j \varepsilon(n_{k-j-1}).
\end{aligned} \tag{86}$$

By (55) and (79),

$$\theta^k \, \|u - u(\,\cdot\,, t_i)\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \le c\theta^k \le cm^{-1}n^{-\gamma_1}. \tag{87}$$

Moreover, (78) implies

$$\sum_{j=0}^{k} \theta^j = \left(1 + \theta \frac{1-\theta^k}{1-\theta}\right) \le 1 + 2\theta. \tag{88}$$

Finally, using (80) and (73), we obtain

$$\begin{aligned} \sum_{j=0}^{k-1} \theta^j \varepsilon(n_{k-j-1}) &= \sum_{j=0}^{k-1} \theta^j n_{k-j-1}^{-\gamma_1} (\log(n_{k-j-1} + 1))^{\gamma_2} \\ &\le \sum_{j=0}^{k-1} \theta^j n^{-\gamma_1} \theta^{-\frac{\gamma_1 j}{\gamma_1+1}} (\log(n+1))^{\gamma_2} \\ &= n^{-\gamma_1} \log(n+1))^{\gamma_2} \sum_{j=0}^{k-1} \theta^{\frac{j}{\gamma_1+1}} \le cn^{-\gamma_1} \log(n+1))^{\gamma_2}. \end{aligned} \tag{89}$$

Combining (86)–(89), we conclude

$$\begin{aligned} &\left(\mathbb{E}\, \|u - u_{ik}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \\ &\le cm^{-1}n^{-\gamma_1}(\log(n+1))^{\gamma_2} + (1+2\theta)\left(\mathbb{E}\, \|u(\,\cdot\,,t_i) - u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}. \end{aligned} \tag{90}$$

In particular, taking into account (64), (78), and (81), we obtain by recursion over i,

$$\begin{aligned} &\left(\mathbb{E}\, \|u(\,\cdot\,,t_{i+1}) - u_{i+1,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} \\ &\le cm^{-1}n^{-\gamma_1}(\log(n+1))^{\gamma_2} \sum_{l=0}^{i} (1+2\theta)^l + (1+2\theta)^{i+1} \|u(\,\cdot\,,0) - u_{0,0}\|_{C(Q,\mathbb{R}^q)} \\ &\le c(1+2\theta)^m n^{-\gamma_1}(\log(n+1))^{\gamma_2} \le cn^{-\gamma_1}(\log(n+1))^{\gamma_2}. \end{aligned}$$

Inserting this into (90), we get

$$\left(\mathbb{E}\, \|u - u_{ik}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2} \le cn^{-\gamma_1}(\log(n+1))^{\gamma_2}$$

and hence,

$$\begin{aligned}&\left(\mathbb{E}\,\|u - A^{2,r}_{\nu(n),\omega}(f,u_0)\|^2_{C(Q\times[0,1],\mathbb{R}^q)}\right)^{1/2}\\ &= \left(\mathbb{E}\,\|u - u_{\nu(n)}\|^2_{C(Q\times[0,1],\mathbb{R}^q)}\right)^{1/2} \le cmn^{-\gamma_1}(\log(n+1))^{\gamma_2} \le cn^{-\gamma_1}(\log(n+1))^{\gamma_2},\end{aligned}$$

which is (71).

Finally we prove the stability (72) of algorithm $A^{2,r}_{\nu(n),\omega}$. Let $(f,u_0),(\tilde f,\tilde u_0) \in F^0_2(\kappa)$, let $\tilde u_{ij}$ and $\tilde g_{ij}$ be defined analogously to (62)–(64) and set

$$g^*_{ij}(s,t) = f(s,t,\tilde u_{ij}(s,t)) \qquad (s \in Q, t \in [t_i,t_{i+1}]). \tag{91}$$

From (62) we get

$$\begin{aligned}&\left(\mathbb{E}\,\|u_{i,j+1} - \tilde u_{i,j+1}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\ &\le \left(\mathbb{E}\,\|u_{i,0} - \tilde u_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}\\ &\quad + \left(\mathbb{E}\,\left\|A^{1,r}_{\mu(n_j),i,\omega}g_{ij} - A^{1,r}_{\mu(n_j),i,\omega}g^*_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\ &\quad + \left(\mathbb{E}\,\left\|A^{1,r}_{\mu(n_j),i,\omega}g^*_{ij} - A^{1,r}_{\mu(n_j),i,\omega}\tilde g_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}.\end{aligned} \tag{92}$$

We have by (82) and (91)

$$g_{ij}(s,t) - g^*_{ij}(s,t) = f(s,t,u_{ij}(s,t)) - f(s,t,\tilde u_{ij}(s,t)),$$

hence

$$\|g_{ij} - g^*_{ij}\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \le \kappa\|u_{ij} - \tilde u_{ij}\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}.$$

It follows from (77) that

$$\begin{aligned}&\left(\mathbb{E}\,\left\|A^{1,r}_{\mu(n_j),i,\omega}g_{ij} - A^{1,r}_{\mu(n_j),i,\omega}g^*_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\ &= \left(\mathbb{E}\,\mathbb{E}\,\left(\left\|A^{1,r}_{\mu(n_j),i,\omega}g_{ij} - A^{1,r}_{\mu(n_j),i,\omega}g^*_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\,\Big|(u_{ij},\tilde u_{ij})\right)\right)^{1/2}\\ &\le (c(1)+1)m^{-1}\left(\mathbb{E}\,\|g_{ij} - g^*_{ij}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\ &\le (c(1)+1)m^{-1}\kappa\left(\mathbb{E}\,\left\|u_{ij} - \tilde u_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}.\end{aligned} \tag{93}$$

Similarly,

$$g_{ij}^*(s,t) - \tilde{g}_{ij}(s,t) = f(s,t,\tilde{u}_{ij}(s,t)) - \tilde{f}(s,t,\tilde{u}_{ij}(s,t)),$$

which yields

$$\|g_{ij}^* - \tilde{g}_{ij}\|_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)} \le \|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}.$$

Using again (77), we conclude

$$\begin{aligned}
&\left(\mathbb{E}\,\left\|A^{1,r}_{\mu(n_j),i,\omega}g_{ij}^* - A^{1,r}_{\mu(n_j),i,\omega}\tilde{g}_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\
&= \left(\mathbb{E}\,\mathbb{E}\,\left(\left\|A^{1,r}_{\mu(n_j),i,\omega}g_{ij}^* - A^{1,r}_{\mu(n_j),i,\omega}\tilde{g}_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\,\Big|\tilde{u}_{ij}\right)\right)^{1/2}\\
&\le (c(1)+1)m^{-1}\left(\mathbb{E}\,\|g_{ij}^* - \tilde{g}_{ij}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\
&\le (c(1)+1)m^{-1}\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}.
\end{aligned} \tag{94}$$

Combining (78), and (92)–(94), we obtain

$$\begin{aligned}
&\left(\mathbb{E}\,\|u_{i,j+1} - \tilde{u}_{i,j+1}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\
&\le \left(\mathbb{E}\,\|u_{i,0} - \tilde{u}_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} + (c(1)+1)m^{-1}\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}\\
&\quad +\theta\left(\mathbb{E}\,\left\|u_{ij} - \tilde{u}_{ij}\right\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}.
\end{aligned}$$

Recursion over j together with (88) gives

$$\begin{aligned}
&\left(\mathbb{E}\,\|u_{ik} - \tilde{u}_{ik}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\\
&\le \left(\left(\mathbb{E}\,\|u_{i,0} - \tilde{u}_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2} + (c(1)+1)m^{-1}\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}\right)\\
&\quad \times(1+\theta+\cdots+\theta^{k-1}) + \theta^k\left(\mathbb{E}\,\|u_{i,0} - \tilde{u}_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}\\
&\le (1+2\theta)(c(1)+1)m^{-1}\|f - \tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}\\
&\quad +(1+2\theta)\left(\mathbb{E}\,\|u_{i,0} - \tilde{u}_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}.
\end{aligned} \tag{95}$$

Consequently, recalling (64) and using recursion over i, we obtain

$$\begin{aligned}
&\left(\mathbb{E}\ \|u_{i+1,0}-\tilde{u}_{i+1,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}\\
&\le (1+2\theta)(c(1)+1)m^{-1}\|f-\tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}\\
&\quad +(1+2\theta)\left(\mathbb{E}\ \|u_{i,0}-\tilde{u}_{i,0}\|^2_{C(Q,\mathbb{R}^q)}\right)^{1/2}\\
&\le (1+2\theta+(1+2\theta)^2+\cdots+(1+2\theta)^{i+1})(c(1)+1)m^{-1}\\
&\quad \times\|f-\tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}+(1+2\theta)^{i+1}\|u_0-\tilde{u}_0\|_{C(Q,\mathbb{R}^q)}\\
&\le c(\|f-\tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}+\|u_0-\tilde{u}_0\|_{C(Q,\mathbb{R}^q)}).
\end{aligned}$$

Combining this with (95) yields

$$\left(\mathbb{E}\,\|u_{ik}-\tilde{u}_{ik}\|^2_{C(Q\times[t_i,t_{i+1}],\mathbb{R}^q)}\right)^{1/2}\le c(\|f-\tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}+\|u_0-\tilde{u}_0\|_{C(Q,\mathbb{R}^q)}),$$

and finally

$$\begin{aligned}
&\left(\mathbb{E}\,\|A^{2,r}_{\nu(n),\omega}(f,u_0)-A^{2,r}_{\nu(n),\omega}(\tilde{f},\tilde{u}_0)\|^2_{C(Q\times[0,1],\mathbb{R}^q)}\right)^{1/2}\\
&\le c(\|f-\tilde{f}\|_{C(Q\times[0,1]\times\mathbb{R}^q,\mathbb{R}^q)}+\|u_0-\tilde{u}_0\|_{C(Q,\mathbb{R}^q)}).
\end{aligned}$$

□

Now we will work in the setting of information-based complexity theory (IBC), see [23, 25]. For the precise notions used here we also refer to [10, 11]. An abstract numerical problem is described by a tuple (F, G, S, K, Λ), with F an arbitrary set—the set of input data, G a normed linear space and $S : F \to G$ an arbitrary mapping, the solution operator, which maps the input $f \in F$ to the exact solution Sf. Furthermore, K is an arbitrary set and Λ is a set of mappings from F to K—the class of admissible information functionals.

The cardinality of an algorithm A, denoted by card(A), is the number of information functionals used in A. Let $e_n^{\mathrm{det}}(S, F, G)$, respectively $e_n^{\mathrm{ran}}(S, F, G)$, denote the n-th minimal error in the deterministic, respectively randomized setting, that is, the minimal possible error among all deterministic, respectively randomized algorithms of cardinality at most n. The cardinality of an algorithm A is closely related to the arithmetic cost, that is, the number of arithmetic operations needed to carry out A. For many concrete algorithms, including all those considered here, the arithmetic cost is within a constant or a logarithmic factor of card(A).

To put the parametric ODE problem into the setting above, let

$$S = S_2,\quad F = F_2^r(\kappa),\quad G = C(Q\times[0,1],\mathbb{R}^q),\quad K = \mathbb{R}^q,$$

and let Λ_2 be the following class of function values

$$\Lambda_2 = \{\delta_{s,t,z} : \ s \in Q, t \in [0,1], z \in \mathbb{R}^q\} \cup \{\delta_s : \ s \in Q\},$$

where $\delta_{s,t,z}(f, u_0) = f(s,t,z)$ and $\delta_s(f, u_0) = u_0(s)$.

The following theorem extends a result on the complexity of parametric ODEs from [4]. There the Lipschitz condition was imposed on f and on certain derivatives of f up to order r, here the Lipschitz condition is required for f alone. This is also of importance for the applications to PDEs in the next section.

Theorem 1 *Let $r \in \mathbb{N}_0$, $d, q \in \mathbb{N}$, $\kappa > 0$. Then the deterministic n-th minimal errors satisfy*

$$e_n^{\mathrm{det}}(S_2, F_2^r(\kappa), C(Q \times [0,1], \mathbb{R}^q)) \asymp n^{-\frac{r}{d+1}}.$$

For the randomized n-th minimal errors we have the following: If $r/d > 1/2$ then

$$e_n^{\mathrm{ran}}(S_2, F_2^r(\kappa), C(Q \times [0,1], \mathbb{R}^q)) \asymp n^{-\frac{r+1/2}{d+1}} (\log n)^{\frac{1}{2}},$$

if $r/d = 1/2$ then

$$n^{-\frac{1}{2}} (\log n)^{\frac{1}{2}} \preceq e_n^{\mathrm{ran}}(S_2, F_2^r(\kappa), C(Q \times [0,1], \mathbb{R}^q)) \preceq n^{-\frac{1}{2}} (\log n)^2,$$

and if $r/d < 1/2$ then

$$e_n^{\mathrm{ran}}(S_2, F_2^r(\kappa), C(Q \times [0,1], \mathbb{R}^q)) \asymp n^{-\frac{r}{d}} (\log n)^{\frac{r}{d}}.$$

Proof Proposition 4 gives the upper bounds. To prove the lower bounds, we let $u_0 \equiv 0$ and consider functions $f = f(s,t)$ not depending on z. In this sense we have $\kappa B_{C^r(Q \times [0,1], \mathbb{R}^q)} \subset F_2^r(\kappa)$ and for $f \in \kappa B_{C^r(Q \times [0,1], \mathbb{R}^q)}$

$$(S_2(0,f))(s,1) = \int_0^1 f(s,t)dt \quad (s \in Q).$$

This means that parametric definite integration of $C^r(Q \times [0,1], \mathbb{R}^q)$ functions reduces to S_2, so that the required lower bounds for parametric ODEs follow from [15]. □

6 Almost Linear First Order PDEs

Let $d, r \in \mathbb{N}$ (note that throughout this section we assume $r \geq 1$), $Q = [0,1]^d$, and $\kappa > 0$. Given

$$(f, g, u_0) \in F_3^r(\kappa) := \kappa B_{C^r([0,1] \times \mathbb{R}^d, \mathbb{R}^d)} \times \kappa B_{C^r([0,1] \times \mathbb{R}^d \times \mathbb{R})} \times \kappa B_{C^r(\mathbb{R}^d)}, \qquad (96)$$

$f=(f_1,\dots,f_d)$, we consider the following scalar first order almost linear PDE

$$\frac{\partial u(t,x)}{\partial t}+\sum_{i=1}^{d} f_i(t,x)\frac{\partial u(t,x)}{\partial x_i}=g(t,x,u(t,x)) \quad (x\in\mathbb{R}^d, t\in[0,1]), \tag{97}$$

$$u(0,x)=u_0(x). \tag{98}$$

A solution is a continuously differentiable function $u:[0,1]\times\mathbb{R}^d\to\mathbb{R}$ satisfying (97)–(98). Due to the definition of $F_3^r(\kappa)$, the solution exists and is unique, see, e.g., [24], as well as the discussion of the relations to ODEs below. We seek to determine the solution at time $t=1$ on Q, thus, we set $G_3=C(Q)$ and define the solution operator by

$$S_3: F_3^r(\kappa)\to C(Q), \quad (S_3(f,g,u_0))(x)=u(1,x) \quad (x\in Q).$$

Furthermore, we put $K=\mathbb{R}^d\cup\mathbb{R}$ and let Λ_3 be the following class of function values

$$\Lambda_3=\{\delta_{t,x}:\ t\in[0,1], x\in\mathbb{R}^d\}\cup\{\delta_{t,x,z}:\ t\in[0,1], x\in\mathbb{R}^d, z\in\mathbb{R}\}\cup\{\delta_x:\ x\in Q\},$$

where

$$\delta_{t,x}(f,g,u_0)=f(t,x), \quad \delta_{t,x,z}(f,g,u_0)=g(t,x,z), \quad \delta_x(f,g,u_0)=u_0(x).$$

We use the method of characteristics. We want to find $\xi: Q\times[0,1]\to\mathbb{R}^d$ such that for $s\in Q$, $t\in[0,1]$,

$$\frac{\partial\xi(s,t)}{\partial t}=f(t,\xi(s,t)) \tag{99}$$

$$\xi(s,1)=s. \tag{100}$$

Observe that, due to (96) and the assumption $r\ge 1$,

$$\|f\|_{C^r_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^d,\mathbb{R}^d)}\le\sqrt{d}\kappa \tag{101}$$

(in the sense that $f=f(t,z)$ is considered as a function not depending on $s\in Q$). Thus, the solution of (99)–(100) exists and is unique. Denote

$$\xi_0: Q\to\mathbb{R}^d, \quad \xi_0(s)=s \quad (s\in Q).$$

Let

$$\tilde{S}_2: C_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R}^d,\mathbb{R}^d)\times C(Q,\mathbb{R}^d)\to C(Q\times[0,1],\mathbb{R}^d)$$

be the solution operator of parametric ODEs studied in Sect. 5, with the difference that the starting time is $t = 1$ and the ODE is considered backward in time (clearly, this does not affect the error estimates of Proposition 5, provided the algorithms are modified in the corresponding way). So we have

$$\xi = \tilde{S}_2(f, \xi_0).$$

Furthermore, $\|\xi_0\|_{C^r(Q,\mathbb{R}^d)} = \sqrt{d}$, and consequently, by (55) and (101), there is a $\kappa_1 > 0$ depending only on r, d and κ such that

$$\|\xi\|_{C^r(Q\times[0,1],\mathbb{R}^d)} \le \kappa_1. \tag{102}$$

We define $h \in C_{\mathrm{Lip}}(Q \times [0, 1] \times \mathbb{R})$ and $w_0 \in C(Q)$ by setting

$$h(s, t, z) = g(t, \xi(s, t), z) \tag{103}$$

$$w_0(s) = u_0(\xi(s, 0)) \tag{104}$$

for $s \in Q$, $t \in [0, 1]$, $z \in \mathbb{R}$. By (96) and (102), there is a $\kappa_2 > 0$ also depending only on r, d and κ such that

$$(h, w_0) \in F_2^r(\kappa_2) \subseteq F_2^0(\kappa_2). \tag{105}$$

Next we seek to find $w : Q \times [0, 1] \to \mathbb{R}$ with

$$\frac{\partial w(s,t)}{\partial t} = h(s, t, w(s, t)) \quad (s \in Q, t \in [0, 1])$$
$$w(s, 0) = w_0(s) \quad (s \in Q).$$

Then we have

$$w = S_2(h, w_0), \tag{106}$$

where

$$S_2 : C_{\mathrm{Lip}}(Q \times [0, 1] \times \mathbb{R}) \times C(Q) \to C(Q \times [0, 1])$$

is the respective solution operator of parametric ODEs, here with $q = 1$ and starting time $t = 0$. The following is well-known (see again, e.g., [24]).

Lemma 1 *If $u(t, x)$ is the solution of (97)–(98), then*

$$u(t, \xi(s, t)) = w(s, t) \quad (s \in Q, t \in [0, 1]). \tag{107}$$

It follows from (100) and (107) that

$$u(1, s) = w(s, 1) \quad (s \in Q),$$

hence by (106)

$$S_3(f,g,u_0) = (S_2(h,w_0))\,(\,\cdot\,,1). \tag{108}$$

Now let $\sigma = (\tilde{\nu},\nu) \in \mathscr{N}^2$, where $\mathscr{N}$ was defined in (60), and let $\tilde{A}^{2,r}_{\tilde{\nu},\omega}$ be the algorithm (62)–(66) for $\tilde{S}_2$. Similarly, let $A^{2,r}_{\nu,\omega}$ be the respective algorithm for S_2. We assume that the random variables $\tilde{A}^{2,r}_{\tilde{\nu},\omega}$ and $A^{2,r}_{\nu,\omega}$ are independent. Define $\xi_{\tilde{\nu}} \in C(Q\times[0,1],\mathbb{R}^d)$ by

$$\xi_{\tilde{\nu}} = \tilde{A}^{2,r}_{\tilde{\nu},\omega}(f,\xi_0) \tag{109}$$

and $h_{\tilde{\nu}} \in C_{\mathrm{Lip}}(Q\times[0,1]\times\mathbb{R})$, $w_{0,\tilde{\nu}} \in C(Q)$ by setting for $s\in Q$, $t\in[0,1]$, $z\in\mathbb{R}$

$$h_{\tilde{\nu}}(s,t,z) = g(t,\xi_{\tilde{\nu}}(s,t),z) \tag{110}$$

$$w_{0,\tilde{\nu}}(s) = u_0(\xi_{\tilde{\nu}}(s,0)). \tag{111}$$

It follows from (96) that

$$(h_{\tilde{\nu}},w_{0,\tilde{\nu}}) \in F_2^0(\kappa). \tag{112}$$

We define algorithm $A^{3,r}_{\sigma,\omega}$ for S_3 by setting for $s\in Q$

$$\left(A^{3,r}_{\sigma,\omega}(f,g,u_0)\right)(s) = \left(A^{2,r}_{\nu,\omega}(h_{\tilde{\nu}},w_{0,\tilde{\nu}})\right)(s,1). \tag{113}$$

We have $A^{3,r}_{\sigma,\omega}(f,g,u_0) \in C(Q)$. The following result provides error estimates for $A^{3,r}_{\sigma,\omega}$ (recall also the definition (61) of $\mathscr{N}_0$).

Proposition 5 *Let $r,d\in\mathbb{N}$, $\kappa>0$. There are constants $c_1,\dots,c_4>0$ such that the following holds. For each $n\in\mathbb{N}$ there is a $\sigma_0(n)\in\mathscr{N}_0^2$ such that for all $\omega\in\Omega$*

$$\operatorname{card}\left(A^{3,r}_{\sigma_0(n),\omega}\right) \le c_1 n \tag{114}$$

$$\sup_{(f,g,u_0)\in F_3^r(\kappa)} \|S_3(f,g,u_0) - A^{3,r}_{\sigma_0(n),\omega}(f,g,u_0)\|_{C(Q)} \le c_2 n^{-\frac{r}{d+1}}. \tag{115}$$

Moreover, for each $n\in\mathbb{N}$ there is a $\sigma(n)\in\mathscr{N}^2$ such that

$$\sup_{\omega\in\Omega}\operatorname{card}\left(A^{3,r}_{\sigma(n),\omega}\right) \le c_3 n \tag{116}$$

$$\sup_{(f,g,u_0)\in F_3^r(\kappa)} \left(\mathbb{E}\,\|S_3(f,g,u_0) - A^{3,r}_{\sigma(n),\omega}(f,g,u_0)\|^2_{C(Q)}\right)^{1/2}$$
$$\le c_4 n^{-\gamma_1}(\log(n+1))^{\gamma_2}, \tag{117}$$

with γ_1,γ_2 given by (34).

Proof Again we only prove the stochastic case (116)–(117), the deterministic case being analogous. Let $(\tilde{\nu}(n))_{n=1}^{\infty}$ be such that (70)–(72) of Proposition 4 hold for $\tilde{S}_2$. Similarly, let $(\nu(n))_{n=1}^{\infty}$ be a respective sequence for S_2. We put $\sigma(n) = (\tilde{\nu}(n), \nu(n))$. Now let $n \in \mathbb{N}$, $(f, g, u_0) \in F_3^r(\kappa)$. By (71)–(72) of Proposition 4, (105), (112), (113), and (108)

$$\begin{aligned}
&\left(\mathbb{E}\left\|S_3(f,g,u_0) - A^{3,r}_{\sigma(n),\omega}(f,g,u_0)\right\|^2_{C(Q)}\right)^{1/2} \\
&= \left(\mathbb{E}\left\|(S_2(h,w_0))(\,\cdot\,,1) - \left(A^{2,r}_{\nu(n),\omega}(h_{\tilde{\nu}(n)}, w_{0,\tilde{\nu}(n)})\right)(\,\cdot\,,1)\right\|^2_{C(Q)}\right)^{1/2} \\
&\le \left(\mathbb{E}\left\|S_2(h,w_0) - A^{2,r}_{\nu(n),\omega}(h_{\tilde{\nu}(n)}, w_{0,\tilde{\nu}(n)})\right\|^2_{C(Q\times[0,1])}\right)^{1/2} \\
&\le \left(\mathbb{E}\left\|S_2(h,w_0) - A^{2,r}_{\nu(n),\omega}(h, w_0)\right\|^2_{C(Q\times[0,1])}\right)^{1/2} \\
&\quad + \left(\mathbb{E}\,\mathbb{E}\left\|A^{2,r}_{\nu(n),\omega}(h,w_0) - A^{2,r}_{\nu(n),\omega}(h_{\tilde{\nu}(n)}, w_{0,\tilde{\nu}(n)})\right\|^2_{C(Q\times[0,1])}\,\Big|\,\xi_{\tilde{\nu}(n)}\right)^{1/2} \\
&\le cn^{-\gamma_1}(\log(n+1))^{\gamma_2} + c\left(\mathbb{E}\,\|h - h_{\tilde{\nu}(n)}\|^2_{C(Q\times[0,1]\times\mathbb{R})}\right)^{1/2} \\
&\quad + c\left(\mathbb{E}\,\|w_0 - w_{0,\tilde{\nu}(n)}\|^2_{C(Q)}\right)^{1/2}. \qquad (118)
\end{aligned}$$

By (103), (110), and (96), for $s \in Q$, $t \in [0,1]$, $z \in \mathbb{R}$

$$\begin{aligned}
|h(s,t,z) - h_{\tilde{\nu}(n)}(s,t,z)| &= |g(t,\xi(s,t),z) - g(t,\xi_{\tilde{\nu}(n)}(s,t),z)| \\
&\le \sqrt{d}\kappa\|\xi(s,t) - \xi_{\tilde{\nu}(n)}(s,t)\|_{\mathbb{R}^d}, \qquad (119)
\end{aligned}$$

and similarly, by (104), (111), and (96),

$$\begin{aligned}
|w_0(s) - w_{0,\tilde{\nu}(n)}(s)| &= |u_0(\xi(s,0)) - u_0(\xi_{\tilde{\nu}(n)}(s,0))| \\
&\le \sqrt{d}\kappa\|\xi(s,0) - \xi_{\tilde{\nu}(n)}(s,0)\|_{\mathbb{R}^d}. \qquad (120)
\end{aligned}$$

Furthermore, using (101) and (71) of Proposition 4, we obtain

$$\begin{aligned}
&\left(\mathbb{E}\,\|\xi - \xi_{\tilde{\nu}(n)}\|^2_{C(Q\times[0,1],\mathbb{R}^d)}\right)^{1/2} \\
&= \left(\mathbb{E}\left\|\tilde{S}_2(f,\xi_0) - \tilde{A}^{2,r}_{\tilde{\nu}(n),\omega}(f,\xi_0)\right\|^2_{C(Q\times[0,1],\mathbb{R}^d)}\right)^{1/2} \\
&\le cn^{-\gamma_1}(\log(n+1))^{\gamma_2}. \qquad (121)
\end{aligned}$$

From (119)–(121) we conclude

$$\left(\mathbb{E}\,\|h - h_{\tilde{\nu}(n)}\|^2_{C(Q\times[0,1]\times\mathbb{R})}\right)^{1/2} \le cn^{-\gamma_1}(\log(n+1))^{\gamma_2} \tag{122}$$

$$\left(\mathbb{E}\,\|w_0 - w_{0,\tilde{\nu}(n)}\|^2_{C(Q)}\right)^{1/2} \le cn^{-\gamma_1}(\log(n+1))^{\gamma_2}. \tag{123}$$

Combining (118) and (122)–(123), we obtain the desired result (117). Relation (116) follows from the definition of $A^{3,r}_{\sigma(n),\omega}$ and (70) of Proposition 4. □

The following theorem gives the deterministic and randomized minimal errors of the first order almost linear PDE problem.

Theorem 2 *Let* $r, d \in \mathbb{N}$ *and* $\kappa > 0$. *Then in the deterministic setting,*

$$e_n^{\mathrm{det}}(S_3, F_3^r(\kappa), C(Q)) \asymp n^{-\frac{r}{d+1}}.$$

In the randomized setting, if $r/d > 1/2$

$$e_n^{\mathrm{ran}}(S_3, F_3^r(\kappa), C(Q)) \asymp n^{-\frac{r+1/2}{d+1}}(\log n)^{\frac{1}{2}},$$

if $r/d = 1/2$, *then*

$$n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}} \preceq e_n^{\mathrm{ran}}(S_3, F_3^r(\kappa), C(Q)) \preceq n^{-\frac{1}{2}}(\log n)^2,$$

and if $r/d < 1/2$, *then*

$$e_n^{\mathrm{ran}}(S_3, F_3^r(\kappa), C(Q)) \asymp n^{-\frac{r}{d}}(\log n)^{\frac{r}{d}}.$$

Proof The upper bounds follow from Proposition 5 above. To show the lower bounds, we set $f \equiv 0$, $u_0 \equiv 0$, and consider $g = g(t, x)$ not depending on z. Let $C_Q^r([0,1]\times\mathbb{R}^d)$ be the subspace of $C^r([0,1]\times\mathbb{R}^d)$ consisting of all functions g satisfying

$$\operatorname{supp} g(t,\cdot) \subseteq Q \quad (t \in [0,1]).$$

Then $g \in \kappa B_{C_Q^r([0,1]\times\mathbb{R}^d)}$ implies $(0, g, 0) \in F_3^r(\kappa)$. Moreover,

$$(S_3(0,g,0))(x) = \int_0^1 g(t,x)dt \quad (x \in Q),$$

thus parametric definite integration of $C_Q^r([0,1]\times\mathbb{R}^d)$ functions reduces to S_3, and the lower bounds follow from [15] (it is readily seen from the proof in [15] that the lower bound also holds for the subclass of functions with support in Q). □

Note that to obtain this result it was crucial to have Proposition 4 and Theorem 1 for parametric ODEs under the Lipschitz condition as imposed in definitions (50), (51), and (54). If we wanted to apply the results of [4] to get the upper bounds as stated in Theorem 2, we would have to ensure the stronger Lipschitz condition from [4] (involving derivatives up to order r). This would mean to require $(f, g, u_0) \in F_3^{r+1}(\kappa)$, which, in turn, would lead to gaps between the upper and lower bounds in Theorem 2 (in the lower bounds r would have to be replaced by $r + 1$).

References

1. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. Anal. Appl. (Singap.) **9**(1), 11–47 (2011)
2. Daun, Th.: On the randomized solution of initial value problems. J. Complex. **27**, 300–311 (2011)
3. Daun, Th., Heinrich, S.: Complexity of Banach space valued and parametric integration. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012. Proceedings in Mathematics and Statistics, vol. 65, pp. 297–316. Springer, Berlin/Heidelberg (2013)
4. Daun, Th., Heinrich, S.: Complexity of parametric initial value problems in Banach spaces. J. Complex. **30**, 392–429 (2014)
5. Daun, Th., Heinrich, S.: Complexity of parametric initial value problems for systems of ODEs. Math. Comput. Simul. **135**, 72–85 (2017)
6. Defant, A., Floret, K.: Tensor Norms and Operator Ideals. North Holland, Amsterdam (1993)
7. Dick, J., Kuo, F., Le Gia, Q., Schwab, C.: Multilevel higher order QMC Petrov-Galerkin discretization for affine parametric operator equations. SIAM J. Numer. Anal. **54**(4), 2541–2568 (2016)
8. Frolov, A.S., Chentsov, N.N.: On the calculation of definite integrals dependent on a parameter by the Monte Carlo method (in Russian). Zh. Vychisl. Mat. Fiz. **2**, 714–717 (1962)
9. Hansen, M., Schwab, C.: Sparse adaptive approximation of high dimensional parametric initial value problems. Vietnam J. Math. **41**, 181–215 (2013)
10. Heinrich, S.: Monte Carlo approximation of weakly singular integral operators. J. Complex. **22**, 192–219 (2006)
11. Heinrich, S.: The randomized information complexity of elliptic PDE. J. Complex. **22**, 220–249 (2006)
12. Heinrich, S.: Complexity of initial value problems in Banach spaces. J. Math. Phys. Anal. Geom. **9**, 73–101 (2013)
13. Heinrich, S., Milla, B.: The randomized complexity of initial value problems. J. Complex. **24**, 77–88 (2008)
14. Heinrich, S., Milla, B.: The randomized complexity of indefinite integration. J. Complex. **27**, 352–382 (2011)
15. Heinrich, S., Sindambiwe, E.: Monte Carlo complexity of parametric integration. J. Complex. **15**, 317–341 (1999)
16. Kacewicz, B.: How to increase the order to get minimal-error algorithms for systems of ODE. Numer. Math. **45**, 93–104 (1984)
17. Kacewicz, B.: Randomized and quantum algorithms yield a speed-up for initial-value problems. J. Complex. **20**, 821–834 (2004)
18. Kacewicz, B.: Almost optimal solution of initial-value problems by randomized and quantum algorithms. J. Complex. **22**, 676–690 (2006)

19. Kuo, F., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**(6), 3351–3374 (2012)
20. Kuo, F., Schwab, C., Sloan, I.H.: Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. Found. Comput. Math. **15**(2), 411–449 (2015)
21. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin (1991)
22. Light, W.A., Cheney, W.: Approximation Theory in Tensor Product Spaces. Lecture Notes in Mathematics, vol. 1169, Springer, Berlin (1985)
23. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, vol. 1349, Springer, Berlin (1988)
24. Petrovski, I.G.: Ordinary Differential Equations. Dover Publications, New York (1973)
25. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic, New York (1988)

Adaptive Quasi-Monte Carlo Methods for Cubature

Fred J. Hickernell, Lluís Antoni Jiménez Rugama, and Da Li

Dedicated to Ian H. Sloan on the occasion of his 80th birthday with many thanks for his friendship and leadership.

Abstract High dimensional integrals can be approximated well by quasi-Monte Carlo methods. However, determining the number of function values needed to obtain the desired accuracy is difficult without some upper bound on an appropriate semi-norm of the integrand. This challenge has motivated our recent development of theoretically justified, adaptive cubatures based on digital sequences and lattice nodeset sequences. Our adaptive cubatures are based on error bounds that depend on the discrete Fourier transforms of the integrands. These cubatures are guaranteed for integrands belonging to cones of functions whose true Fourier coefficients decay steadily, a notion that is made mathematically precise. Here we describe these new cubature rules and extend them in two directions. First, we generalize the error criterion to allow both absolute and relative error tolerances. We also demonstrate how to estimate a function of several integrals to a given tolerance. This situation arises in the computation of Sobol' indices. Second, we describe how to use control variates in adaptive quasi-Monte cubature while appropriately estimating the control variate coefficient.

1 Introduction

An important problem studied by Ian Sloan is evaluating multivariate integrals by quasi-Monte Carlo methods. After perhaps a change of variable, one may pose the problem as constructing an accurate approximation to

$$\mu = \int_{[0,1)^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

F. J. Hickernell (✉) · L. A. Jiménez Rugama · D. Li
Illinois Institute of Technology, Chicago, IL, USA
e-mail: hickernell@iit.edu; ljimene1@hawk.iit.edu; dli37@hawk.iit.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_27

given a black-box function f that provides $f(\boldsymbol{x})$ for any $\boldsymbol{x} \in [0,1)^d$. Multivariate integrals arise in applications such as evaluating financial risk, computing multivariate probabilities, statistical physics, and uncertainty quantification.

We have developed and implemented quasi-Monte Carlo (qMC) cubature algorithms that adaptively determine the sample size needed to guarantee that an error tolerance is met, provided that the integrand belongs to a cone $\mathscr{C}$ of well-behaved functions [2, 10, 13, 15]. That is, given a low discrepancy sequence $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ and function data $f(\boldsymbol{x}_0), f(\boldsymbol{x}_1), \ldots$, we have a stopping rule based on the function data obtained so far that chooses n for which

$$|\mu - \widehat{\mu}_n| \le \varepsilon, \qquad \text{where } \widehat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i), \quad f \in \mathscr{C}. \tag{1}$$

Here, $\widehat{\mu}_n$ is the sample average of function values taken at well-chosen points whose empirical distribution mimics the uniform distribution. The cone $\mathscr{C}$ contains integrands whose Fourier coefficients decay in a reasonable manner, thus allowing the stopping rule to succeed. Specifically, the size of the high wavenumber components of an integrand in $\mathscr{C}$ cannot be large in comparison to the size of the low wavenumber components. Rather than choosing the $\boldsymbol{x}_i$ to be independent and identically distributed (IID) $\mathscr{U}[0,1)^d$ points, we use shifted digital sequences [3, 20] and sequences of nodesets of shifted rank-1 lattices [11, 17, 18, 24]. Sequences that are more evenly distributed than IID points are the hallmark of qMC algorithms.

Traditional qMC error analysis produces error bounds of the form [4, 8]

$$|\mu - \widehat{\mu}_n| \le D(\{\boldsymbol{x}_i\}_{i=0}^{n-1})\, \|f\|,$$

where the integrand, f, is assumed to lie in some Banach space with (semi-)norm $\|\cdot\|$, and $\|f\|$ is often called the *variation* of f. Moreover, the *discrepancy* $D(\cdot)$ is a measure of quality of the sample, $\{\boldsymbol{x}_i\}_{i=0}^{n-1}$. For integrands lying in the ball $\mathscr{B} := \{f : \|f\| \le \sigma\}$ one may construct a *non-adaptive* algorithm guaranteeing $|\mu - \widehat{\mu}_n| \le \varepsilon$ by choosing $n = \min\bigl\{n' \in \mathbb{N} : D(\{\boldsymbol{x}_i\}_{i=0}^{n'-1}) \le \varepsilon/\sigma\bigr\}$.

Our interest is in *adaptive* qMC algorithms, where n depends on the function data observed, not on a priori bounds on the variation of the integrand. A few methods have been proposed for choosing n:

Independent and identically distributed (IID) replications. [22] Compute

$$\widehat{\mu}_{n,R} = \frac{1}{R}\sum_{r=1}^{R} \widehat{\mu}_n^{(r)}, \qquad \widehat{\mu}_n^{(r)} = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i^{(r)}), \quad r = 1, \ldots, R,$$

where $\bigl\{\boldsymbol{x}_i^{(1)}\bigr\}_{i=0}^{\infty}, \ldots, \bigl\{\boldsymbol{x}_i^{(R)}\bigr\}_{i=0}^{\infty}$ are IID randomizations of a low discrepancy sequence, and $\mathbb{E}\bigl(\widehat{\mu}_n^{(r)}\bigr) = \mu$. A multiple of the standard deviation of these $\widehat{\mu}_n^{(r)}$ is proposed as an upper bound for $|\mu - \widehat{\mu}_{n,R}|$. To justify this approach one must make assumptions about the higher order moments of the $\widehat{\mu}_n^{(r)}$ to confidently bound their variance and then employ a Berry-Esseen inequality, or some similar inequality, to construct a confidence interval for μ. This may require rather strict assumptions on the higher order moments or an unattractively large number of replications.

Internal replications. [22] Compute

$$\widehat{\mu}_{nR} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mu}_n^{(r)} = \frac{1}{nR}\sum_{r=1}^{nR} f(\boldsymbol{x}_i), \qquad \widehat{\mu}_n^{(r)} = \frac{1}{n}\sum_{i=(r-1)n}^{rn-1} f(\boldsymbol{x}_i), \quad r = 1,\ldots,R.$$

A multiple of the standard deviation of these $\widehat{\mu}_n^{(r)}$ is proposed as an upper bound for $|\mu - \widehat{\mu}_{nR}|$. This method does not have a supporting theory.

Quasi-standard error. [7] Compute

$$\widehat{\mu}_{n,R} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mu}_n^{(r)}, \qquad \widehat{\mu}_n^{(r)} = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_{i,(r-1)d+1:rd}), \quad r = 1,\ldots,R,$$

where $\{\boldsymbol{x}_i\}_{i=0}^{\infty}$ is now an Rd dimensional sequence, and $\boldsymbol{x}_{i,(r-1)d+1:rd}$ denotes the $(r-1)d+1$st through rdth components of the ith point in the sequence. A multiple of the standard deviation of these $\widehat{\mu}_n^{(r)}$ is proposed as an upper bound for $|\mu - \widehat{\mu}_{n,R}|$. However, see [23] for cautions regarding this method.

Since the error bounds proposed above are homogeneous, the sets of integrands for which these error bounds are correct are *cones*. That is, if one of the above error bounds above is correct for integrand f, it is also correct for integrand cf, where c is an arbitrary constant. In this article we review our recent work developing adaptive qMC algorithms satisfying (1). We describe the cones $\mathscr{C}$ for which our algorithms succeed. We also extend our earlier algorithms in two directions:

- Meeting more general error criteria than simply absolute error, and
- Using control variates to improve efficiency.

Our data-based cubature error bounds are described in Sect. 2. This section also emphasizes the similar algebraic structures of our two families of qMC sequences. In Sect. 3, we describe how our error bounds can be used to satisfy error criteria that are more general than that in (1). Section 4 describes the implementation of our new adaptive qMC algorithms and provides numerical examples. Control variates with adaptive qMC cubature is described in Sect. 5. We conclude with a discussion that identifies problems for further research.

2 Error Estimation for Digital Net and Lattice Cubature

Here we summarize some of the key properties of cubature based on digital sequences and rank-1 lattice node sequences. We use a common notation for both cases to highlight the similarities in analysis. We focus on the base 2 setting for simplicity and because it is most common in practice. Always, $n = 2^m$ for non-negative integer m. See [10] and [15] for more details.

Let $\{\mathbf{0} = z_0, z_1, \ldots\}$ be a sequence of distinct points that is either a digital sequence or a rank-1 lattice node sequence. Let $\oplus : [0,1)^d \times [0,1)^d \to [0,1)^d$ denote an addition operator under which the sequence is a group and the first 2^m points form a subgroup. For some shift, $\boldsymbol{\Delta} \in [0,1)^d$, the data sites used for cubature in (1) are given by $\boldsymbol{x}_i = z_i \oplus \boldsymbol{\Delta}$ for all $i \in \mathbb{N}_0$. Typical examples of a digital sequence and a rank-1 lattice node sequence are given in Fig. 1.

There is a set of integer vector wavenumbers, $\mathbb{K}$, which is a group under its own addition operator, also denoted $\oplus$. There is also a bilinear functional, $\langle \cdot, \cdot \rangle : \mathbb{K} \times [0,1)^d \to [0,1)$, which is used to define a Fourier basis for $L^2[0,1)^d$, given by $\left\{ \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{k}, \cdot \rangle} \right\}_{\boldsymbol{k} \in \mathbb{K}}$. The integrand is expressed as a Fourier series,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{K}} \hat{f}(\boldsymbol{k}) \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{k}, \boldsymbol{x} \rangle} \quad \forall \boldsymbol{x} \in [0,1)^d,\ f \in L^2[0,1)^d,$$

$$\text{where } \hat{f}(\boldsymbol{k}) := \int_{[0,1)^d} f(\boldsymbol{x}) \mathrm{e}^{-2\pi\sqrt{-1}\langle \boldsymbol{k}, \boldsymbol{x} \rangle} \, \mathrm{d}\boldsymbol{x}.$$

Cubature requires function values so we assume throughout that this Fourier series is absolutely convergent, i.e., $\sum_{\boldsymbol{k} \in \mathbb{K}} |\hat{f}(\boldsymbol{k})| < \infty$.

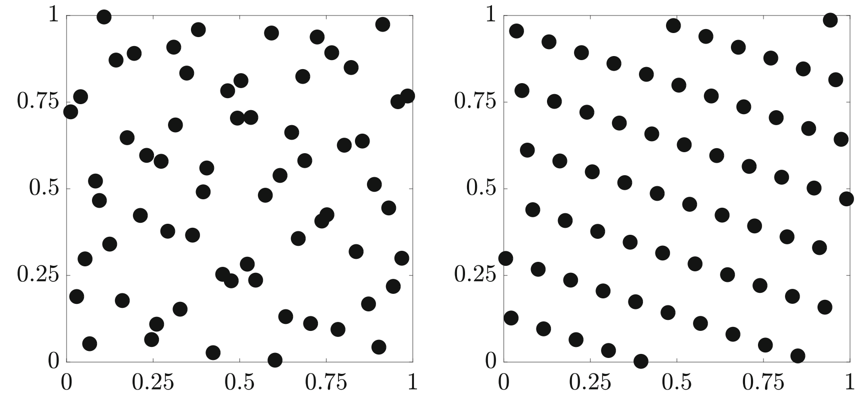


Fig. 1 Two dimensional projections of a digitally shifted and Matoušek [19] scrambled digital sequence (left) and a shifted rank-1 lattice node set (right)

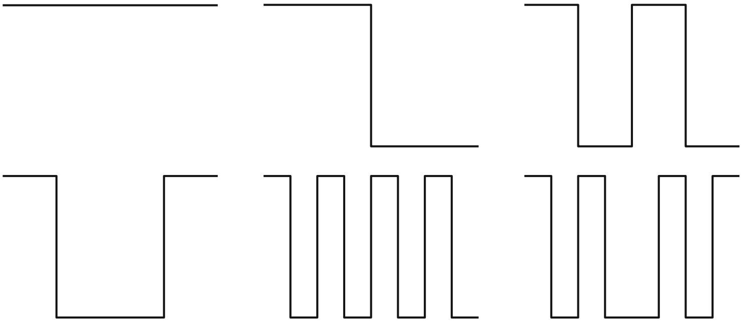


Fig. 2 One-dimensional Walsh functions corresponding to $k = 0, 1, 2, 3, 4$, and 5

In the case of digital sequences, $\oplus$ denotes digit-wise addition modulo 2 for points in $[0,1)^d$ and wavenumbers in $\mathbb{K} = \mathbb{N}_0^d$. The digits of $z_1, z_2, z_4, z_8, \ldots$ correspond to elements in the generator matrices for the usual method for constructing digital sequences [3, Sec. 4.4]. Also, $\langle \boldsymbol{k}, \boldsymbol{x} \rangle$ is one half of an ℓ^2 inner product of the digits of $\boldsymbol{k}$ and $\boldsymbol{x}$ modulo 2, implying that $\langle \boldsymbol{k}, \boldsymbol{x} \rangle \in \{0, 1/2\}$. The $\mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{k}, \cdot \rangle} = (-1)^{2\langle \boldsymbol{k}, \cdot \rangle}$ are multivariate Walsh functions (see Fig. 2).

In the case of rank-1 lattice node sequences, $\oplus$ denotes addition modulo **1** for points in $[0,1)^d$ and ordinary addition for wavenumbers in $\mathbb{K} = \mathbb{Z}^d$. Moreover, $\langle \boldsymbol{k}, \boldsymbol{x} \rangle = \boldsymbol{k}^T \boldsymbol{x} \bmod 1$. The $\mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{k}, \cdot \rangle}$ are multivariate complex exponential functions.

The *dual set* corresponding to the first $n = 2^m$ unshifted points, $\{z_0, \ldots, z_{2^m-1}\}$, is denoted $\mathbb{K}_m$ and defined as

$$\mathbb{K}_0 := \mathbb{K}, \qquad \mathbb{K}_m := \{\boldsymbol{k} \in \mathbb{K} : \langle \boldsymbol{k}, z_{2^\ell} \rangle = 0 \text{ for all } \ell = 0, \ldots, m-1\}, \quad m \in \mathbb{N}.$$

The dual set satisfies

$$\frac{1}{2^m} \sum_{i=0}^{2^m-1} \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{k}, z_i \rangle} = \begin{cases} 1, & \boldsymbol{k} \in \mathbb{K}_m, \\ 0, & \text{otherwise.} \end{cases}$$

The *discrete Fourier transform* of a function f using $n = 2^m$ data is denoted $\tilde{f}_m$ and defined as

$$\begin{aligned} \tilde{f}_m(\boldsymbol{k}) &:= \frac{1}{2^m} \sum_{i=0}^{2^m-1} f(\boldsymbol{x}_i) \mathrm{e}^{-2\pi\sqrt{-1}\langle \boldsymbol{k}, \boldsymbol{x}_i \rangle} \\ &= \hat{f}(\boldsymbol{k}) + \sum_{\boldsymbol{l} \in \mathbb{K}_m \setminus \{\boldsymbol{0}\}} \hat{f}(\boldsymbol{k} \oplus \boldsymbol{l}) \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{l}, \boldsymbol{\Delta} \rangle}, \end{aligned} \tag{2}$$

after applying some of the properties alluded to above. This last expression illustrates how the discrete Fourier coefficient $\tilde{f}_m(\boldsymbol{k})$ differs from its true counterpart, $\hat{f}(\boldsymbol{k})$, by the aliasing terms, which involve the other wavenumbers in the coset $\boldsymbol{k} \oplus \mathbb{K}_m$. As m increases, wavenumbers leave $\mathbb{K}_m$, and so the aliasing decreases.

The sample mean of the function data is the $\boldsymbol{k} = \boldsymbol{0}$ discrete Fourier coefficient:

$$\widehat{\mu}_n = \frac{1}{2^m} \sum_{i=0}^{2^m-1} f(\boldsymbol{x}_i) = \tilde{f}_m(\boldsymbol{0}) = \sum_{\boldsymbol{l} \in \mathbb{K}_m} \hat{f}(\boldsymbol{l}) \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{l}, \boldsymbol{\Delta} \rangle}.$$

Hence, an error bound for the sample mean may be expressed in terms of those Fourier coefficients corresponding to wavenumbers in the dual set:

$$|\mu - \widehat{\mu}_n| = \left| \hat{f}(\boldsymbol{0}) - \tilde{f}_m(\boldsymbol{0}) \right| = \left| \sum_{\boldsymbol{l} \in \mathbb{K}_m \setminus \{\boldsymbol{0}\}} \hat{f}(\boldsymbol{l}) \mathrm{e}^{2\pi\sqrt{-1}\langle \boldsymbol{l}, \boldsymbol{\Delta} \rangle} \right| \le \sum_{\boldsymbol{l} \in \mathbb{K}_m \setminus \{\boldsymbol{0}\}} \left| \hat{f}(\boldsymbol{l}) \right|. \tag{3}$$

Our aim is to bound the right hand side of this cubature error bound in terms of function data, more specifically, in terms of the discrete Fourier transform. However, this requires that the true Fourier coefficients of the integrand do not decay too erratically. This motivates our definition of $\mathscr{C}$, the cone of integrands for which our adaptive algorithms succeed.

To facilitate the definition of $\mathscr{C}$ we construct a bijective ordering of the wavenumbers, $\tilde{\boldsymbol{k}} : \mathbb{N}_0 \to \mathbb{K}$ satisfying $\tilde{\boldsymbol{k}}(0) = \boldsymbol{0}$ and $\{\tilde{\boldsymbol{k}}(\kappa + \lambda 2^m)\}_{\lambda=0}^{\infty} = \tilde{\boldsymbol{k}}(\kappa) \oplus \mathbb{K}_m$ for $\kappa = 0, \ldots, 2^m - 1$ and $m \in \mathbb{N}_0$, as described in [10, 15]. This condition implies the crucial fact that $\left| \tilde{f}_m(\tilde{\boldsymbol{k}}(\kappa + \lambda 2^m)) \right|$ is the same for all $\lambda \in \mathbb{N}_0$. Although there is some arbitrariness in this ordering, it is understood that $\tilde{\boldsymbol{k}}(\kappa)$ generally increases in magnitude as κ tends to infinity. We adopt the shorthand notation $\hat{f}_\kappa := \hat{f}(\tilde{\boldsymbol{k}}(\kappa))$ and $\tilde{f}_{m,\kappa} := \tilde{f}_m(\tilde{\boldsymbol{k}}(\kappa))$. Then, the error bound in (3) may be written as

$$|\mu - \widehat{\mu}_n| \le \sum_{\lambda=1}^{\infty} \left| \hat{f}_{\lambda 2^m} \right|. \tag{4}$$

The cone of functions whose Fourier series are absolutely convergent and whose true Fourier coefficients, $\hat{f}_\kappa$, decay steadily as κ tends to infinity is

$$\begin{aligned} \mathscr{C} = \{f \in AC([0,1)^d) : \widehat{S}_{\ell,m}(f) \le \widehat{\omega}(m-\ell)\mathring{S}_m(f), \ \ \ell \le m, \\ \mathring{S}_m(f) \le \mathring{\omega}(m-\ell) S_\ell(f), \ \ \ell_* \le \ell \le m\}, \end{aligned} \tag{5a}$$

where

$$S_m(f) := \sum_{\kappa=\lfloor 2^{m-1} \rfloor}^{2^m-1} \left| \hat{f}_\kappa \right|, \qquad \widehat{S}_{\ell,m}(f) := \sum_{\kappa=\lfloor 2^{\ell-1} \rfloor}^{2^\ell-1} \sum_{\lambda=1}^{\infty} \left| \hat{f}_{\kappa+\lambda 2^m} \right|, \tag{5b}$$

$$\overset{\circ}{S}_m(f) := \widehat{S}_{0,m}(f) + \cdots + \widehat{S}_{m,m}(f) = \sum_{\kappa=2^m}^{\infty} \left| \hat{f}_\kappa \right|, \tag{5c}$$

and where $\ell, m \in \mathbb{N}_0$ and $\ell \le m$. The positive integer ℓ_* and the bounded functions $\widehat{\omega}, \overset{\circ}{\omega} : \mathbb{N}_0 \to [0, \infty)$ are parameters that determine how inclusive $\mathscr{C}$ is and how robust our algorithm is. Moreover, $\overset{\circ}{\omega}(m) \to 0$ as $m \to \infty$. The default values are provided in Sect. 4.

We now explain the definition of the cone $\mathscr{C}$ and the data driven cubature error bound that we are able to derive. The argument may be outlined as follows:

1. The absolute error is bounded by the a sum of the absolute Fourier coefficients, $\widehat{S}_{0,m}$ as given in (4), a well-known result.
2. The sum $\widehat{S}_{0,m}$ is assumed to be no larger than a multiple, $\widehat{\omega}(m)$, of the tail sum of the absolute Fourier coefficients, $\overset{\circ}{S}_m(f)$, by (5).
3. Also by (5), the tail sum $\overset{\circ}{S}_m(f)$ is in turn no larger than a multiple, $\overset{\circ}{\omega}(r)$, of a some sum of the lower wavenumber absolute Fourier coefficients, $S_{m-r}(f)$. The sum $S_{m-r}(f)$ involves the Fourier coefficients that are r blocks from the end of the first 2^m coefficients.
4. Finally, $S_{m-r}(f)$, which involves the true Fourier coefficients may be bounded by a multiple of the analogous sum based on the discrete Fourier coefficients, $\widetilde{S}_{m-r,m}(f)$, defined in (7). This is justified by definition of $\mathscr{C}$ in (5), which limits the effects of aliasing.

For illustration we use the functions depicted in Fig. 3. The one on the left lies inside $\mathscr{C}$ because its Fourier coefficients decay steadily (but not necessarily monotonically), while the one on the right lies outside $\mathscr{C}$ because its Fourier coefficients decay erratically. The function lying outside $\mathscr{C}$ resembles the one lying inside $\mathscr{C}$ but with high wavenumber noise.

The sum of the absolute values of the Fourier coefficients appearing on the right side of error bound (4) is $\widehat{S}_{0,m}(f)$ according to the definition in (5b). In Fig. 3, $m = 11$, and $\widehat{S}_{0,11}(f)$ corresponds to the sum of $|\hat{f}_\kappa|$ for $\kappa = 2048, 4096, 6144, \ldots$. Since only $n = 2^m$ function values are available, it is impossible to estimate the Fourier coefficients appearing in $\widehat{S}_{0,m}(f)$ directly by discrete Fourier coefficients.

By the definition in (5c), it follows that $\widehat{S}_{0,m}(f) \le \overset{\circ}{S}_m(f)$. In Fig. 3, $\overset{\circ}{S}_{11}(f)$ corresponds to the sum of all $|\hat{f}_\kappa|$ with $\kappa \ge 2048$. The definition of $\mathscr{C}$ assumes that $\widehat{S}_{0,m}(f) \le \widehat{\omega}(m)\overset{\circ}{S}_m(f)$, where $\widehat{\omega}(m)$ could be chosen as 1 or could decay with m. This is up to the user.

Still, $\overset{\circ}{S}_m(f)$ involves high wave number Fourier coefficients that still cannot be approximated by discrete Fourier coefficients. The definition of $\mathscr{C}$ also assumes

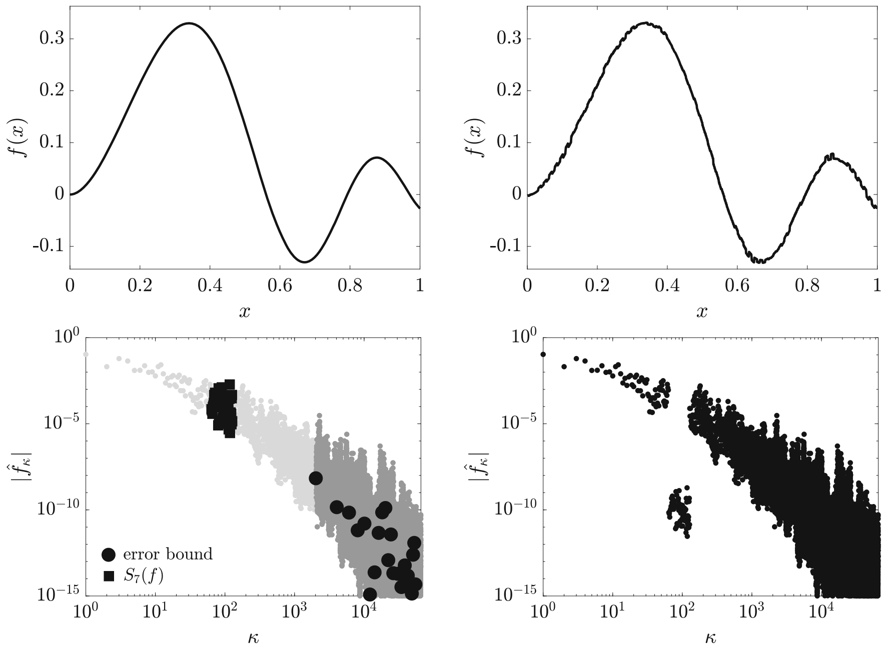


Fig. 3 A typical function lying inside $\mathscr{C}$ and its Fourier Walsh coefficients (left) in contrast to a function lying outside $\mathscr{C}$ and its Fourier Walsh coefficients (right)

that $\overset{\circ}{S}_m(f) \leq \overset{\circ}{\omega}(r) S_{m-r}(f)$ for any non-negative $r \leq m - \ell_*$. This means that the infinite sum of the high wavenumber coefficients, $\overset{\circ}{S}_m(f)$ cannot exceed some factor, $\overset{\circ}{\omega}(r)$, times the finite sum of modest wavenumber coefficients $S_{m-r}(f)$. In Fig. 3, $r = 4$, and the graph on the left shows $\overset{\circ}{S}_{11}(f) \approx 0.0003$ and $S_7(f) \approx 0.01$. Thus, $\overset{\circ}{S}_{11}(f)$ is bounded above by $\overset{\circ}{\omega}(4) S_7(f)$ for a modest value of $\overset{\circ}{\omega}(4)$. Recall from the definition in (5b) that $S_7(f)$ is the sum of the absolute values of the Fourier coefficients corresponding to $64, \ldots, 127$. In contrast, the function depicted in the right of Fig. 3 violates the assumption that $\overset{\circ}{S}_{11}(f) \leq \overset{\circ}{\omega}(4) S_7(f)$ because $S_7(f) \approx 0.00000001$ in that case. Thus, the function on the right in Fig. 3 lies outside $\mathscr{C}$.

Based on the above argument, it follows in general that for $f \in \mathscr{C}$,

$$\sum_{\lambda=1}^{\infty} \left| \hat{f}_{\lambda 2^m} \right| = \widehat{S}_{0,m}(f) \leq \widehat{\omega}(m) \overset{\circ}{S}_m(f) \leq \widehat{\omega}(m) \overset{\circ}{\omega}(r) S_{m-r}(f),$$

$$m \geq r + \ell_* \geq \ell_*. \qquad (6)$$

This implies an error bound in terms of the true Fourier coefficients with modest wavenumber. In particular (6) holds for the function depicted on the left side of Fig. 3, but not the one on the right side.

Before going on, we explain the roles of the parameters $\ell_*, r, \widehat{\omega}$, and $\mathring{\omega}$, which have not been specified. They reflect the robustness desired by the user, and they are meant to be kept constant rather than changed for every problem. The wavenumber 2_*^{ℓ} is the minimum wavenumber for which we expect steady decay to set in. A sum of absolute values of discrete Fourier coefficients in a given block is used to bound a sum of true Fourier coefficients r or more blocks away. The functions $\widehat{\omega}$ and $\mathring{\omega}$ are inflation factors for bounding one sum of Fourier coefficients in terms of another. Equation (20) in Sect. 4 provides the default choices in our algorithm implementations.

While (6) is a step forward, it involves the unknown true Fourier coefficients and not the known discrete Fourier coefficients. We next bound the infinite sum $S_{m-r}(f)$ in terms of a finite sum of discrete Fourier coefficients:

$$\widetilde{S}_{\ell,m}(f) := \sum_{\kappa=\lfloor 2^{\ell-1} \rfloor}^{2^{\ell}-1} \left|\tilde{f}_{m,\kappa}\right|. \tag{7}$$

By (2), the triangle inequality, and the definition of $\mathscr{C}$, it follows that

$$\begin{aligned}
\widetilde{S}_{m-r,m}(f) &= \sum_{\kappa=\lfloor 2^{m-r-1} \rfloor}^{2^{m-r}-1} \left|\tilde{f}_{m,\kappa}\right| \\
&\geq \sum_{\kappa=\lfloor 2^{m-r-1} \rfloor}^{2^{m-r}-1} \left[\left|\hat{f}_{\kappa}\right| - \sum_{\lambda=1}^{\infty} \left|\hat{f}_{\kappa+\lambda 2^m}\right|\right] \\
&= S_{m-r}(f) - \widehat{S}_{m-r,m}(f) \\
&\geq S_{m-r}(f)[1 - \widehat{\omega}(r)\mathring{\omega}(r)].
\end{aligned} \tag{8}$$

This provides an upper bound on $S_{m-r}(f)$ in terms of the data-based $\widetilde{S}_{m-r,m}(f)$, provided that r is large enough to satisfy $\widehat{\omega}(r)\mathring{\omega}(r) < 1$. Such a choice of r ensures that the aliasing errors are modest.

Combining (6) and (8) with (4), it is shown in [10, 15] that for any $f \in \mathscr{C}$,

$$|\mu - \widehat{\mu}_n| \leq \mathrm{err}_n := \mathfrak{C}(m,r)\widetilde{S}_{m-r,m}(f), \qquad m \geq \ell_* + r, \tag{9a}$$

$$\text{where } \mathfrak{C}(m,r) := \frac{\widehat{\omega}(m)\mathring{\omega}(r)}{1 - \widehat{\omega}(r)\mathring{\omega}(r)}, \tag{9b}$$

provided that $\widehat{\omega}(r)\mathring{\omega}(r) < 1$. Since $\widetilde{S}_{m-r,m}(f)$ depends only on the discrete Fourier coefficients, (9) is a data-based cubature error bound. One may now increment m (keeping r fixed) until err_n is small enough, where again $n = 2^m$.

If $\$(f)$ denotes the cost of one function value, then evaluating $f(\boldsymbol{x}_0), \ldots, f(\boldsymbol{x}_{2^m-1})$ requires $\$(f)n$ operations. A fast transform then computes $\tilde{f}_{m,0}, \ldots, \tilde{f}_{m,2^m-1}$ in an additional $\mathscr{O}(n \log(n)) = \mathscr{O}(m2^m)$ operations. So computing err_{2^m} for each m costs $\mathscr{O}\big([\$(f) + m]2^m\big)$ operations. For integrands that are cheap to evaluate the $\$(f)$ term is negligible compared to m, but for integrands that are expensive to integrate $\$(f)$ may be comparable to m given that m might be ten to twenty.

Using an analogous reasoning as in (8),

$$\begin{aligned} S_\ell(f) &= \sum_{\kappa=\lfloor 2^{\ell-1}\rfloor}^{2^\ell-1} \left|\hat{f}_\kappa\right| \\ &\geq \sum_{\kappa=\lfloor 2^{\ell-1}\rfloor}^{2^\ell-1} \left[|\tilde{f}_{m,\kappa}| - \sum_{\lambda=1}^{\infty} \left|\hat{f}_{\kappa+\lambda 2^m}\right| \right] \\ &= \widetilde{S}_{\ell,m}(f) - \widehat{S}_{\ell,m}(f) \\ &\geq \widetilde{S}_{\ell,m}(f)/[1 + \widehat{\omega}(m-\ell)\mathring{\omega}(m-\ell)]. \end{aligned} \tag{10}$$

Therefore, from (8) and (10), for any $\ell, m, m' \in \mathbb{N}$ such that $\ell_* \leq \ell \leq \min(m, m')$, it must be the case that

$$\frac{\widetilde{S}_{\ell,m}(f)}{1 + \widehat{\omega}(m-\ell)\mathring{\omega}(m-\ell)} \leq S_\ell(f) \leq \frac{\widetilde{S}_{\ell,m'}(f)}{1 - \widehat{\omega}(m'-\ell)\mathring{\omega}(m'-\ell)}, \tag{11}$$

provided $\widehat{\omega}(m'-\ell)\mathring{\omega}(m'-\ell) < 1$. Equation (11) is a data-based *necessary* condition for an integrand, f, to lie in $\mathscr{C}$. If it is found that the right hand side of (11) is smaller than the left hand side of (11), then f must lie outside $\mathscr{C}$. In this case the parameters defining the cone should be adjusted to expand the cone appropriately, e.g., by increasing $\widehat{\omega}$ or $\mathring{\omega}$ by a constant.

By substituting inequality (10) in the error bound (9), we get

$$\mathrm{err}_n \leq \frac{\widehat{\omega}(m)\mathring{\omega}(r)}{1 - \widehat{\omega}(r)\mathring{\omega}(r)}[1 + \widehat{\omega}(r)\mathring{\omega}(r)]S_{m-r}(f).$$

We define m^*,

$$m^* := \min\left\{ m \geq \ell_* + r : \frac{\widehat{\omega}(m)\mathring{\omega}(r)}{1 - \widehat{\omega}(r)\mathring{\omega}(r)}[1 + \widehat{\omega}(r)\mathring{\omega}(r)]S_{m-r}(f) \leq \varepsilon \right\}, \tag{12}$$

Here m^* depends on the fixed parameters of the algorithm, $\ell_*, r, \widehat{\omega}$, and $\mathring{\omega}$. Note that $\mathrm{err}_{2^{m^*}} \leq \varepsilon$.

Recall from above that at each step m in our algorithm the computational cost is $\mathscr{O}\big([\$(f) + m]2^m\big)$. Thus, the computational cost for our adaptive algorithm to satisfy the absolute error tolerance, as given in (1), is $\mathscr{O}(\Phi(m^*)2^{m^*})$, where

$\Phi(m^*) = [\$(f) + 0]2^{-m^*} + \cdots + [\$(f) + m^*]2^0$. Since

$$\Phi(m^* + 1) - \Phi(m^*) = \$(f)2^{-m^*-1} + 2^{-m^*} + \cdots + 2^0 \le \$(f)2^{-m^*-1} + 2,$$

it follows that

$$\begin{aligned}\Phi(m^*) &= [\Phi(m^*) - \Phi(m^* - 1)] + \cdots + [\Phi(1) - \Phi(0)]\\ &\le [\$(f)2^{-m^*} + 2] + \cdots + [\$(f)2^{-1} + 2]\\ &\le 2[\$(f) + m^*].\end{aligned}$$

Thus, the cost of making our data based error bound no greater than ε is bounded above by $\mathcal{O}\big([\$(f) + m^*]2^{m^*}\big)$.

The algorithm does not assume a rate of decay of the Fourier coefficients but automatically senses the rate of decay via the discrete Fourier coefficients. From (12) it is evident that the dependence of the computational cost with ε depends primarily on the unknown rate of decay of $S_{m-r}(f)$ with m, and secondarily on the specified rate of decay of $\widehat{\omega}(m)$, since all other parameters are fixed. For example, assuming $\widehat{\omega}(m) = \mathcal{O}(1)$, if $\hat{f}_\kappa = \mathcal{O}(\kappa^{-p})$, then $S_{m-r}(f) = \mathcal{O}(2^{-(p-1)m})$, and the total computational cost is $\mathcal{O}(\varepsilon^{-1/(p-1)-\delta})$ for all $\delta > 0$. If $\widehat{\omega}(m)$ decays with m, then the computational cost is less.

3 General Error Criterion

The algorithms summarized above are described in [10, 15] and implemented in the Guaranteed Automatic Integration Library (GAIL) [2] as `cubSobol_g` and `cubLattice_g`, respectively. They satisfy the absolute error criterion (1) by increasing n until err_n defined in (9) is no greater than the absolute error tolerance, ε.

There are situations requiring a more general error criterion than (1). In this section we generalize the cubature problem to involve a p-vector of integrals, $\boldsymbol{\mu}$, which are approximated by a p-vector of sample means, $\widehat{\boldsymbol{\mu}}_n$, using n samples, and for which we have a p-vector of error bounds, $\mathbf{err}_n$, given by (9). This means that $\boldsymbol{\mu} \in [\widehat{\boldsymbol{\mu}}_n - \mathbf{err}_n, \widehat{\boldsymbol{\mu}}_n + \mathbf{err}_n]$ for integrands in $\mathscr{C}$. Given some

- function $v : \Omega \subseteq \mathbb{R}^p \to \mathbb{R}$,
- positive absolute error tolerance ε_{a}, and
- relative error tolerance $\varepsilon_{\mathrm{r}} < 1$,

the goal is to construct an *optimal* approximation to $v(\boldsymbol{\mu})$, denoted $\hat{v}$, which depends on $\widehat{\boldsymbol{\mu}}_n$ and $\mathbf{err}_n$ and satisfies the error criterion

$$\sup_{\boldsymbol{\mu} \in \Omega \cap [\widehat{\boldsymbol{\mu}}_n - \mathbf{err}_n, \widehat{\boldsymbol{\mu}}_n + \mathbf{err}_n]} \mathrm{tol}(v(\boldsymbol{\mu}), \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) \le 1, \tag{13a}$$

Table 1 Examples of the tolerance function in (13) and the optimal approximation to the integral when $p = 1$ and $v(\mu) = \mu$

Kind	$\operatorname{tol}(\mu, \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}})$	Optimal $\hat{v}$	Optimal $\operatorname{tol}(\mu, \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}})$
Absolute $\varepsilon_{\mathrm{r}} = 0$	$\dfrac{(\mu - \hat{v})^2}{\varepsilon_{\mathrm{a}}^2}$	$\widehat{\mu}_n$	$\dfrac{\mathrm{err}_n^2}{\varepsilon_{\mathrm{a}}^2}$
Relative $\varepsilon_{\mathrm{a}} = 0$	$\dfrac{(\mu - \hat{v})^2}{\varepsilon_{\mathrm{r}}^2 \mu^2}$	$\dfrac{\max(\widehat{\mu}_n^2 - \mathrm{err}_n^2, 0)}{\widehat{\mu}_n}$	$\dfrac{\mathrm{err}_n^2}{\varepsilon_r^2 \max(\widehat{\mu}_n^2, \mathrm{err}_n^2)}$
Hybrid	$\dfrac{(\mu - \hat{v})^2}{\max(\varepsilon_{\mathrm{a}}^2, \epsilon_r^2 \mu^2)}$	See (18)	See (19)

$$\operatorname{tol}(v, \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) := \frac{(v - \hat{v})^2}{\max(\varepsilon_{\mathrm{a}}^2, \epsilon_r^2 \, |v|^2)}, \qquad (\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) \in [0, \infty) \times [0, 1) \setminus \{\mathbf{0}\}. \tag{13b}$$

Our hybrid error criterion is satisfied if the actual error is no greater than either the absolute error tolerance or the relative error tolerance times the absolute value of the true answer. If we want to satisfy both an absolute error criterion and a relative error criterion, then "max" in the definition of $\operatorname{tol}(\cdot)$ should be replaced by "min". This would require a somewhat different development than what is presented here. By optimal we mean that the choice of $\hat{v}$ we prescribe yields the smallest possible left hand side of (13a). This gives the greatest chance of satisfying the error criterion. The dependence of $\hat{v}$ on n is suppressed in the notation for simplicity.

The common case of estimating the integral itself, $p = 1$ and $v(\mu) = \mu$, is illustrated in Table 1. This includes (1) an absolute error criterion (see (1)), (2) a relative error criterion, and (3) a hybrid error criterion that is satisfied when either the absolute or relative error tolerances are satisfied. Note that $\hat{v}$ is not necessarily equal to $\hat{\mu}_n$. For a pure relative error criterion, $\hat{v}$ represents a shrinkage of the sample mean towards zero. Figure 4 illustrates how the optimal choice of $\hat{v}$ may satisfy (13), when $\hat{v} = \hat{\mu}$ does not.

Define $v_{\pm}$ as the extreme values of $v(\mu)$ for $\widehat{\boldsymbol{\mu}}$ satisfying the given error bound:

$$v_- = \inf_{\boldsymbol{\mu} \in \Omega \cap [\widehat{\boldsymbol{\mu}}_n - \mathbf{err}_n, \widehat{\boldsymbol{\mu}}_n + \mathbf{err}_n]} v(\boldsymbol{\mu}), \qquad v_+ = \sup_{\boldsymbol{\mu} \in \Omega \cap [\widehat{\boldsymbol{\mu}}_n - \mathbf{err}_n, \widehat{\boldsymbol{\mu}}_n + \mathbf{err}_n]} v(\boldsymbol{\mu}) \tag{14}$$

Then the following criterion is equivalent to (13):

$$\sup_{v_- \le v' \le v_+} \operatorname{tol}(v', \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) \le 1. \tag{15}$$

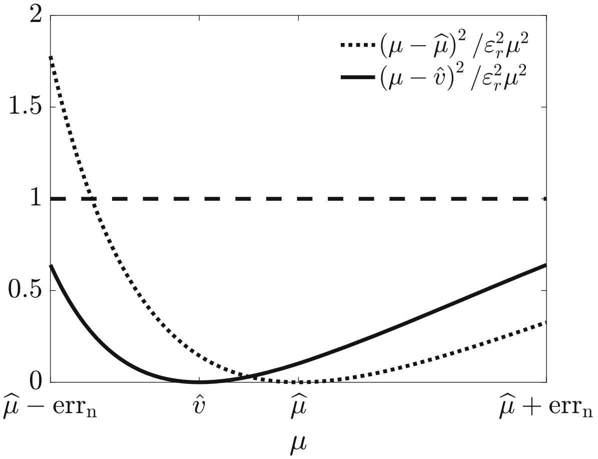


Fig. 4 Example of $v(\mu) = \mu$ with the relative error criterion, i.e. $\varepsilon_{\mathrm{a}} = 0$. For the optimal choice of $\hat{v}$, $\sup_{\mu \in [\widehat{\mu}_n - \mathrm{err}_n, \widehat{\mu}_n + \mathrm{err}_n]} \mathrm{tol}(\mu, \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) < 1 < \sup_{\mu \in [\widehat{\mu}_n - \mathrm{err}_n, \widehat{\mu}_n + \mathrm{err}_n]} \mathrm{tol}(\mu, \widehat{\mu}_n, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}})$

We claim that the optimal value of the estimated integral, i.e., the value of $\hat{v}$ satisfying (15), is

$$\hat{v} = \frac{v_- \max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}} |v_+|) + v_+ \max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}} |v_-|)}{\max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}} |v_+|) + \max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}} |v_-|)} \tag{16a}$$

$$= \begin{cases} \dfrac{v_- + v_+}{2}, & \varepsilon_{\mathrm{r}} |v_\pm| \le \varepsilon_{\mathrm{a}}, \\ \dfrac{v_s[\varepsilon_{\mathrm{a}} + v_{-s}\varepsilon_{\mathrm{r}}\mathrm{sign}(v_s)]}{\varepsilon_{\mathrm{a}} + \varepsilon_{\mathrm{r}} |v_s|}, & \varepsilon_{\mathrm{r}} |v_{-s}| \le \varepsilon_{\mathrm{a}} < \varepsilon_{\mathrm{r}} |v_s|, \ s \in \{+, -\}, \\ \dfrac{|v_+ v_-| \left[\mathrm{sign}(v_+) + \mathrm{sign}(v_-)\right]}{|v_+| + |v_-|}, & \varepsilon_{\mathrm{a}} < \varepsilon_{\mathrm{r}} |v_\pm|. \end{cases} \tag{16b}$$

From (16a) it follows that $\hat{v} \in [v_-, v_+]$. Moreover, by (16b) $\hat{v}$ is a shrinkage estimator: it is either zero or has the same sign as $(v_- + v_+)/2$, and its magnitude is no greater than $|(v_- + v_+)/2|$. Our improved GAIL algorithms `cubSobol_g` and `cubLattice_g`, which are under development, are summarized in the following theorem.

Theorem 1 *Let our goal be the computation of $v(\boldsymbol{\mu})$, as described at the beginning of this section. Let the tolerance function be defined as in* (13b)*, let the extreme possible values of $v(\boldsymbol{\mu})$ be defined as in* (14)*, and let the approximation to $v(\boldsymbol{\mu})$ be defined in terms of $\widehat{\boldsymbol{\mu}}_n$ and* $\mathbf{err}_n$ *as in* (16)*. Then, $\hat{v}$ is the optimal approximation to*

$v(\boldsymbol{\mu})$*, and the tolerance function for this optimal choice is given as follows:*

$$\inf_{\hat{v}'} \sup_{\boldsymbol{\mu}\in\Omega\cap[\widehat{\boldsymbol{\mu}}_n-\mathbf{err}_n,\widehat{\boldsymbol{\mu}}_n+\mathbf{err}_n]} \operatorname{tol}(v(\boldsymbol{\mu}), \hat{v}', \varepsilon_a, \varepsilon_r)$$

$$= \inf_{\hat{v}'} \sup_{v_-\le v'\le v_+} \operatorname{tol}(v', \hat{v}', \varepsilon_a, \varepsilon_r) \tag{17a}$$

$$= \sup_{v_-\le v'\le v_+} \operatorname{tol}(v', \hat{v}, \varepsilon_a, \varepsilon_r) \tag{17b}$$

$$= \operatorname{tol}(v_\pm, \hat{v}, \varepsilon_a, \varepsilon_r) \tag{17c}$$

$$= \frac{(v_+ - v_-)^2}{[\max(\varepsilon_a, \varepsilon_r\,|v_+|) + \max(\varepsilon_a, \varepsilon_r\,|v_-|)]^2}. \tag{17d}$$

By optimal, we mean that the infimum in (17a) *is satisfied by* $\hat{v}$ *as claimed in* (17b)*. Moreover, it is shown that the supremum in* (17b) *is obtained simultaneously at* v_+ *and* v_-*.*

Our new adaptive quasi-Monte Carlo cubature algorithms increase $n = 2^m$ *by incrementing* m *by one until the right side of* (17d) *is no larger than one. The resulting* $\hat{v}$ *then satisfies the error criterion* $\operatorname{tol}(v(\boldsymbol{\mu}), \hat{v}, \varepsilon_a, \varepsilon_r) \le 1$.

Proof The gist of the proof is to establish the equalities in (17). Equality (17d) follows from the definition of $\hat{v}$ and $v_\pm$. Equality (17c) is proven next, and (17b) is proven after that. Equality (17a) follows from definition (14).

The derivative of $\operatorname{tol}(\cdot, \hat{v}, \varepsilon_\mathrm{a}, \varepsilon_\mathrm{r})$ is

$$\frac{\partial \operatorname{tol}(v', \hat{v}, \varepsilon_\mathrm{a}, \varepsilon_\mathrm{r})}{\partial v'} = \begin{cases} \dfrac{2(v' - \hat{v})}{\varepsilon_\mathrm{a}^2}, & |v'| < \dfrac{\varepsilon_\mathrm{a}}{\varepsilon_\mathrm{r}}, \\[2ex] \dfrac{2(v' - \hat{v})\hat{v}}{\varepsilon_\mathrm{r}^2 v'^3}, & |v'| > \dfrac{\varepsilon_\mathrm{a}}{\varepsilon_\mathrm{r}}. \end{cases}$$

The sign of this derivative is shown in Fig. 5. For either $\varepsilon_\mathrm{r}\,|v_\pm| \le \varepsilon_\mathrm{a}$ or $\varepsilon_\mathrm{a} \le \varepsilon_\mathrm{r}\,|v_\pm|$, the only critical point in $[v_-, v_+]$ is $v' = \hat{v}$, where the tolerance function vanishes.

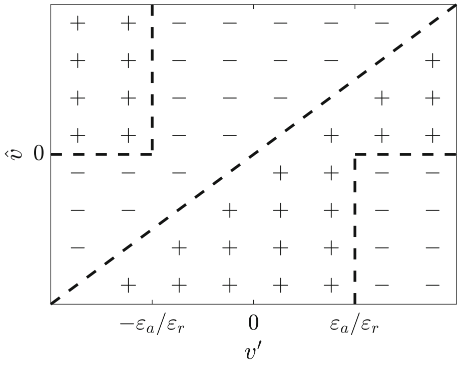


Fig. 5 The sign of $\partial \operatorname{tol}(v', \hat{v}, \varepsilon_\mathrm{a}, \varepsilon_\mathrm{r})/\partial v'$

Thus, the maximum value of the tolerance function always occurs at the boundaries of the interval. For $\varepsilon_{\mathrm{r}}\,|v_{-s}| \le \varepsilon_{\mathrm{a}} < \varepsilon_{\mathrm{r}}\,|v_s|$, $s \in \{+,-\}$, there is also a critical point at $v' = \operatorname{sign}(v_s)\varepsilon_{\mathrm{a}}/\varepsilon_{\mathrm{r}}$. However, since v_s and $\hat{v}$ have the same sign (see (16b)), the partial derivative of the tolerance function with respect to v' does not change sign at this critical point. Hence, the maximum value of the tolerance function still occurs at the boundaries of the interval, and (17c) is established.

To prove assertion (17b), consider $\hat{v}'$, some alternative to $\hat{v}$. Then

$$\begin{aligned}\operatorname{tol}(v_\pm, \hat{v}', \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) - \operatorname{tol}(v_\pm, \hat{v}, \varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}) &= \frac{(v_\pm - \hat{v}')^2 - (v_\pm - \hat{v})^2}{\max(\varepsilon_{\mathrm{a}}^2, \varepsilon_{\mathrm{r}}^2 v_\pm^2)} \\ &= \frac{(\hat{v}' - \hat{v} - 2v_\pm)(\hat{v}' - \hat{v})}{\max(\varepsilon_{\mathrm{a}}^2, \varepsilon_{\mathrm{r}}^2 v_\pm^2)}.\end{aligned}$$

This difference is positive for the + sign if $\hat{v}' \in (-\infty, \hat{v})$ and positive for the − sign if $\hat{v}' \in (\hat{v}, \infty)$. Thus, the proof of Theorem 1 is complete. □

We return to the special case of $v(\mu) = \mu$. The following corollary interprets Theorem 1 for this case, and the theorem that follows extends the computational cost upper bound in (12) for these new quasi-Monte Carlo cubature algorithms.

Corollary 1 *For $p = 1$ and $v(\mu) = \mu$, it follows that $v_\pm = \mu_n \pm \operatorname{err}_n$,*

$$\hat{v} = \frac{(\widehat{\mu}_n - \operatorname{err}_n)\max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n + \operatorname{err}_n|) + (\widehat{\mu}_n + \operatorname{err}_n)\max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n - \operatorname{err}_n|)}{\max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n + \operatorname{err}_n|) + \max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n - \operatorname{err}_n|)}, \tag{18}$$

$$\begin{aligned}&\sup_{\widehat{\mu}_n - \operatorname{err}_n \le \mu \le \widehat{\mu}_n + \operatorname{err}_n} \operatorname{tol}(\mu, \hat{v}, \varepsilon_a, \varepsilon_r) \\ &\qquad = \frac{4\operatorname{err}_n^2}{[\max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n + \operatorname{err}_n|) + \max(\varepsilon_a, \varepsilon_r\,|\widehat{\mu}_n - \operatorname{err}_n|)]^2}.\end{aligned} \tag{19}$$

Theorem 2 *For the special case described in Corollary 1, the computational cost of obtaining an approximation to the integral μ satisfying the generalized error criterion* $\operatorname{tol}(\mu, \hat{v}, \varepsilon_a, \varepsilon_r) \le 1$ *according to the adaptive quasi-Monte Carlo cubature algorithm described in Theorem 1 is $\mathscr{O}\big([\$(f) + m^*]2^{m^*}\big)$, where*

$$\begin{aligned}m^* := \min\Big\{m \ge \ell_* + r :& \\ &(1+\varepsilon_r)\frac{\widehat{\omega}(m)\mathring{\omega}(r)}{1 - \widehat{\omega}(r)\mathring{\omega}(r)}[1 + \widehat{\omega}(r)\mathring{\omega}(r)]S_{m-r}(f) \le \max(\varepsilon_a, \varepsilon_r\,|\mu|)\Big\}.\end{aligned}$$

Proof For each $n = 2^m$, we know that our algorithm produces $\widehat{\mu}_n$ and err_n satisfying $\widehat{\mu}_n - \operatorname{err}_n \le \mu \le \widehat{\mu}_n + \operatorname{err}_n$. This implies that

$$\max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}\,|\widehat{\mu}_n + \operatorname{err}_n|) + \max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}\,|\widehat{\mu}_n - \operatorname{err}_n|) \ge 2\max(\varepsilon_{\mathrm{a}}, \varepsilon_{\mathrm{r}}\,|\mu|) - 2\varepsilon_{\mathrm{r}}\operatorname{err}_n.$$

Thus, the right hand side of (19) must be no greater than one if

$$\mathrm{err}_n \le \frac{\max(\varepsilon_\mathrm{a}, \varepsilon_\mathrm{r} \, |\mu|)}{1 + \varepsilon_\mathrm{r}}.$$

Applying the logic that leads to (12) completes the proof. □

The cost upper bound depends on various parameters as one would expect. The computational cost may increase if

- ε_a decreases,
- ε_r decreases,
- $|\mu|$ decreases,
- the Fourier coefficients of the integrand increase, or
- the cone $\mathscr{C}$ expands because ℓ_*, $\widehat{\omega}$, or $\mathring{\omega}$ increase.

4 Numerical Implementation

The adaptive algorithm described here is included in the latest release of GAIL [2] as `cubSobol_g` and `cubLattice_g`, coded in MATLAB. These two functions use the Sobol' sequences provided by MATLAB [27] and the lattice generator `exod2_base2_m20.txt` from Dirk Nuyens' website [21], respectively. Our algorithm sets its default parameters as follows:

$$\ell_* = 6, \quad r = 4, \quad \widehat{\omega}(m) = \mathring{\omega}(m) = 2^{-m+3}, \quad \mathfrak{C}(m,4) = \frac{16}{3} \times 2^{-m}. \tag{20}$$

These choices are based on experience and are used in the examples below. A larger ℓ_* allows the Fourier coefficients of the integrand to behave erratically over a larger initial segment of wavenumbers. A larger r decreases the impact of aliasing in estimating the true Fourier coefficients by their discrete analogues. Increasing ℓ_* or r increases 2^{ℓ_*+r}, the minimum number of sample points used by the algorithms. The inputs to the algorithms are

- a black-box p-vector function $\boldsymbol{f}$, such that $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{f}(\boldsymbol{X})]$ for $\boldsymbol{X} \sim \mathscr{U}[0,1]^d$,
- a solution function $v : \mathbb{R}^p \to \mathbb{R}$,
- functions for computing $v_\pm$ as described in (14),
- an absolute error tolerance, ε_a, and
- a relative error tolerance ε_r.

The algorithm increases m incrementally until the right side of (17d) does not exceed one. At this point the algorithm returns $\hat{v}$ as given by (16).

Example 1 We illustrate the hybrid error criterion by estimating multivariate normal probabilities for a distribution with mean **0** and covariance matrix Σ:

$$v(\mu) = \mu = \mathbb{P}\left[\boldsymbol{a} \le \boldsymbol{X} \le \boldsymbol{b}\right] = \int_{[\boldsymbol{a},\boldsymbol{b}]} \frac{\mathrm{e}^{-\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}/2}}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\,\mathrm{d}\boldsymbol{x}. \qquad (21)$$

The transformation proposed by Genz [5] is used write this as an integral over the $d-1$ dimensional unit cube. As discussed in [5, 9], when $\boldsymbol{a} = -\infty$, $\Sigma_{ij} = \sigma$ if $i \neq j$, and $\Sigma_{ii} = 1$, the exact value of (21) reduces to a 1-dimensional integral that can be accurately estimated by a standard quadrature rule. This value is taken to be the true μ.

We perform 1000 adaptive integrations: 500 using our cubature rule based on randomly scrambled and digitally shifted Sobol' sequences (`cubSobol_g`) and 500 using our cubature rule based on randomly shifted rank-1 lattice node sequences, (`cubLattice_g`). Default parameters are used. For each case we choose $\sigma \sim \mathscr{U}[0,1]$, dimension $d = \lfloor 500^D \rfloor$ with $D \sim \mathscr{U}[0,1]$, and $\boldsymbol{b} \sim \mathscr{U}[0,\sqrt{d}]^d$. The dependence of $\boldsymbol{b}$ on the dimension of the problem ensures that the estimated probabilities are of the same order of magnitude for all d. Otherwise, the probabilities being estimated would decrease substantially as the dimension increases. The execution time and $\mathrm{tol}(\mu, \hat{v}, 0.01, 0.05)$ are shown in Fig. 6.

Satisfying the error criterion is equivalent to having $\mathrm{tol}(\mu, \hat{v}, 0.01, 0.05) \le 1$, which happens in every case. A very small value of $\mathrm{tol}(\mu, \hat{v}, 0.01, 0.05)$ means that the approximation is much more accurate than required, which may be due to coincidence or due to the minimum sample size used, $n = 2^{10}$. In Fig. 6 the error tolerances are fixed and do not affect the computation time. However, the computation time does depend on the dimension, d, since higher dimensional

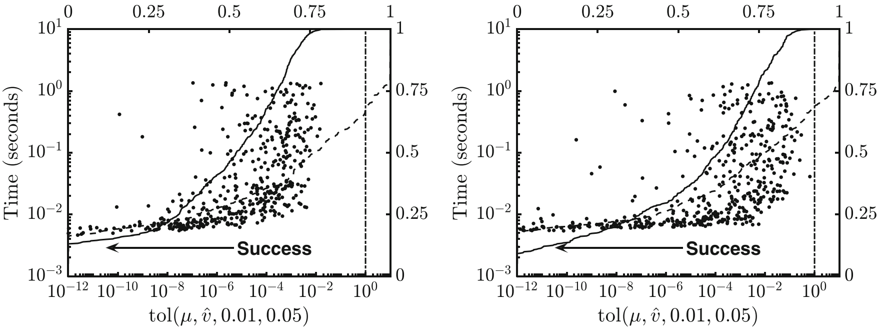


Fig. 6 On the left, 500 integration results using scrambled and digitally shifted Sobol' sequences, `cubSobol_g`. On the right, tolerance values and computation times of integrating 500 multivariate normal probabilities using randomly shifted rank-1 lattice node sequences, `cubLattice_g`. If an integrand is in $\mathscr{C}$, its dot must lie to the left of the vertical dot-dashed line denoting $\mathrm{tol}(\mu, \hat{v}, 0.01, 0.05) = 1$. The solid and dashed curves represent the empirical distributions of tolerance values and times respectively

problems tend to be harder to solve. The performances of `cubSobol_g` and `cubLattice_g` are similar.

Example 2 Sobol' indices [25, 26], which arise in uncertainty quantification, depend on more than one integral. Suppose that one is interested in how an output, $Y := g(\boldsymbol{X})$ depends on the input $\boldsymbol{X} \sim \mathscr{U}[0, 1]^d$, and g has a complicated or unknown structure. For example, g might be the output of a computer simulation. For any coordinate indexed by $j = 1, \ldots, d$, the normalized closed first-order Sobol' index for coordinate j, commonly denoted as $\underline{\tau}_j^2/\sigma^2$, involves three integrals:

$$v(\boldsymbol{\mu}) := \frac{\mu_1}{\mu_2 - \mu_3^2}, \qquad \mu_1 := \int_{[0,1)^{2d}} [g(\boldsymbol{x}_j : \boldsymbol{x}'_{-j}) - g(\boldsymbol{x}')] g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{x}', \tag{22a}$$

$$\mu_2 := \int_{[0,1)^d} g(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x}, \qquad \mu_3 := \int_{[0,1)^d} g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{22b}$$

Here, $(\boldsymbol{x}_j : \boldsymbol{x}'_{-j}) \in [0, 1)^d$ denotes a point whose rth coordinate is x_r if $r = j$, and x'_r otherwise. By definition, the values of these normalized indices must lie between 0 and 1, and both the numerator and denominator in the expression for $v(\boldsymbol{\mu})$ are non-negative. Therefore, the domain of the function v is $\Omega := \{\boldsymbol{\mu} \in [0, \infty)^2 \times \mathbb{R} : 0 \le \mu_1 \le \mu_2 - \mu_3^2\}$. Thus, given $\widehat{\boldsymbol{\mu}}_n$ and $\mathbf{err}_n$, the values of $v_\pm$ defined in (14) are

$$v_\pm = \begin{cases} 0, & \mu_{n,1} \pm \mathrm{err}_{n,1} \le 0, \\ 1, & \mu_{n,1} \pm \mathrm{err}_{n,1} > \max\bigl(0, \mu_{n,2} \mp \mathrm{err}_{n,2} - (\mu_{n,3} \pm \mathrm{err}_{n,3})^2\bigr), \\ \dfrac{\mu_{n,1} \pm \mathrm{err}_{n,1}}{\mu_{n,2} \mp \mathrm{err}_{n,2} - (\mu_{n,3} \pm \mathrm{err}_{n,3})^2}, & \text{otherwise.} \end{cases} \tag{23}$$

We estimate the first-order Sobol' indices of the test function in Bratley et al. [1] using randomly scrambled and digitally shifted Sobol' sequences and the same algorithm parameters as in Example 1 for an absolute error tolerance of 0.005:

$$g(\boldsymbol{X}) = \sum_{i=1}^{6} (-1)^i \prod_{j=1}^{i} X_j.$$

j	1	2	3	4	5	6
n	8 192	4 096	1 024	1 024	1 024	1 024
v	0.6529	0.1791	0.0370	0.0133	0.0015	0.0015
$\hat{v}$	0.6554	0.1781	0.0409	0.0111	0.0012	0.0013
$v(\widehat{\boldsymbol{\mu}}_n)$	0.6660	0.1734	0.0408	0.0111	0.0012	0.0013
$\mathrm{tol}(v, \hat{v}, 0.005, 0)$	0.2540	0.0461	0.6008	0.1962	0.0033	0.0020
$\mathrm{tol}(v, v(\widehat{\boldsymbol{\mu}}_n), 0.005, 0)$	6.9343	1.2952	0.5760	0.2006	0.0034	0.0021

The value of n chosen by our adaptive algorithm and the actual value of the tolerance function, $\mathrm{tol}(v, \hat{v}, 0.005, 0)$, are shown. Since none of those tolerance values exceed one, our algorithm correctly provides $\hat{v}$ for each coordinate j. In the fifth row above, we replaced our optimal $\hat{v}$ defined in (16) by $v(\widehat{\boldsymbol{\mu}}_n)$ for the same n as returned by our algorithm. Interestingly, this approximation to the Sobol' indices, while perhaps intuitive, does not satisfy the absolute error criterion because sometimes $\operatorname{tol}(v, v(\widehat{\boldsymbol{\mu}}_n), 0.005, 0)$ exceeds one. This reflects how $v(\widehat{\boldsymbol{\mu}}_n)$ differs from v much more than $\hat{v}$ does. An extensive study on how to estimate first-order and total effect Sobol' indices using the automatic quasi-Monte Carlo cubature is provided in [14].

5 Control Variates

The results in this section mainly follow the work of Da Li [16]. Control variates are commonly used to improve the efficiency of IID Monte Carlo integration. If one chooses a vector of functions $\boldsymbol{g} : [0,1)^d \to \mathbb{R}^q$ for which $\boldsymbol{\mu}_{\boldsymbol{g}} := \int_{[0,1)^d} \boldsymbol{g}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ is known, then

$$\mu := \int_{[0,1)^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^d} h_{\boldsymbol{\beta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \qquad \text{where } h_{\boldsymbol{\beta}}(\boldsymbol{x}) := f(\boldsymbol{x}) + \boldsymbol{\beta}^T(\boldsymbol{\mu}_{\boldsymbol{g}} - \boldsymbol{g}(\boldsymbol{x})),$$

for any choice of $\boldsymbol{\beta}$. The goal is to choose an optimal $\boldsymbol{\beta}$ to make

$$\widehat{\mu}_{\boldsymbol{\beta},n} := \frac{1}{n}\sum_{i=0}^{n-1} h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$$

sufficiently close to μ with the least expense, n, possible.

If $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ are IID $\mathscr{U}[0,1)^d$, then $\widehat{\mu}_{\boldsymbol{\beta},n}$ is an unbiased estimator for μ for any choice of $\boldsymbol{\beta}$, and the variance of the control variates estimator may be expressed as

$$\operatorname{var}(\widehat{\mu}_{\boldsymbol{\beta},n}) = \frac{\operatorname{var}(h_{\boldsymbol{\beta}}(\boldsymbol{x}_0))}{n} = \frac{1}{n}\sum_{\kappa=1}^{\infty}\left|\hat{f}_\kappa - \boldsymbol{\beta}^T\hat{\boldsymbol{g}}_\kappa\right|^2,$$

where $\hat{\boldsymbol{g}}_\kappa$ are the Fourier coefficients of $\boldsymbol{g}$. Since $\boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{g}}$ is constant, it does not enter into the calculation of the variance. The optimal choice of $\boldsymbol{\beta}$, which minimizes $\operatorname{var}(\widehat{\mu}_{\boldsymbol{\beta},n})$, is

$$\boldsymbol{\beta}_{\mathrm{MC}} = \frac{\operatorname{cov}\big(f(\boldsymbol{x}_0), \boldsymbol{g}(\boldsymbol{x}_0)\big)}{\operatorname{var}\big(\boldsymbol{g}(\boldsymbol{x}_0)\big)}. \tag{24}$$

Although $\boldsymbol{\beta}_{\mathrm{MC}}$ cannot be computed exactly, it may be well approximated in terms of sample estimates of the quantities on the right hand side.

However, if $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ are the points described in Sect. 2, then the error depends on only some of the Fourier coefficients, and (4) and (9) lead to

$$\left|\mu - \widehat{\mu}_{\boldsymbol{\beta},n}\right| \le \sum_{\lambda=1}^{\infty} \left|\hat{f}_{\lambda 2^m} - \boldsymbol{\beta}^T \hat{\boldsymbol{g}}_{\lambda 2^m}\right| \le \frac{\widehat{\omega}(m)\mathring{\omega}(r)}{1 - \widehat{\omega}(r)\mathring{\omega}(r)} \widetilde{S}_{m-r,m}(f - \boldsymbol{\beta}^T \boldsymbol{g}),$$

$$\text{provided} f - \boldsymbol{\beta}^T \boldsymbol{g} \in \mathscr{C}. \qquad (25)$$

Assuming that $f - \boldsymbol{\beta}^T\boldsymbol{g} \in \mathscr{C}$ for all $\boldsymbol{\beta}$, it makes sense to choose $\boldsymbol{\beta}$ to minimize the rightmost term. There seems to be some advantage to choose $\boldsymbol{\beta}$ based on $\widetilde{S}_{m-r,m}(f - \boldsymbol{\beta}^T\boldsymbol{g}), \ldots, \widetilde{S}_{m,m}(f - \boldsymbol{\beta}^T\boldsymbol{g})$. Our experience suggests that this strategy makes $\boldsymbol{\beta}$ less dependent on the fluctuations of the discrete Fourier coefficients over a small range of wave numbers. In summary,

$$\boldsymbol{\beta}_{\mathrm{qMC}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{t=0}^{r} \widetilde{S}_{m-t,m}(f - \boldsymbol{b}^T\boldsymbol{g}) = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{\kappa=\lfloor 2^{m-r-1} \rfloor}^{2^m-1} \left|\tilde{f}_{m,\kappa} - \boldsymbol{b}^T\tilde{\boldsymbol{g}}_{m,\kappa}\right|.$$

As already noted in [12], the optimal control variate coefficients for IID and low discrepancy sampling are generally different. Whereas $\boldsymbol{\beta}_{\mathrm{MC}}$ might be strongly influenced by low wavenumber Fourier coefficients of the integrand, $\boldsymbol{\beta}_{\mathrm{qMC}}$ depends on rather high wavenumber Fourier coefficients.

Minimizing the sum of absolute values is computationally more time consuming than minimizing the sum of squares. Thus, in practice we choose $\boldsymbol{\beta}$ to be

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{qMC}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{\kappa=\lfloor 2^{m-r-1} \rfloor}^{2^m-1} \left|\tilde{f}_{m,\kappa} - \boldsymbol{b}^T\tilde{\boldsymbol{g}}_{m,\kappa}\right|^2.$$

This choice performs well in practice. Moreover, we often find that there is little advantage to updating $\widetilde{\boldsymbol{\beta}}_{\mathrm{qMC}}$ for each m.

Example 3 Control variates may be used to expedite the pricing of an exotic option when one can identify a similar option whose price is known exactly. This often happens with geometric Brownian motion asset price models. The *geometric* mean Asian payoff is a good control variate for estimating the price of an *arithmetic* mean Asian option. The two payoffs are,

$$f(\boldsymbol{x}) = \mathrm{e}^{-rT} \max\left(\frac{1}{d}\sum_{j=1}^{d} S_{t_j}(\boldsymbol{x}) - K, 0\right) = \text{arithmetic mean Asian call},$$

$$g(\boldsymbol{x}) = \mathrm{e}^{-rT} \max\left(\left[\prod_{j=1}^{d} S_{t_j}(\boldsymbol{x})\right]^{1/d} - K, 0\right) = \text{geometric mean Asian call},$$

$$S_{t_j}(\boldsymbol{x}) = S_0 \mathrm{e}^{(r-\sigma^2/2)t_j+\sigma Z_j(\boldsymbol{x})} = \text{stock price at time } t_j,$$

$$\begin{pmatrix} Z_1(\boldsymbol{x}) \\ \vdots \\ Z_d(\boldsymbol{x}) \end{pmatrix} = \mathsf{A} \begin{pmatrix} \Phi^{-1}(x_1) \\ \vdots \\ \Phi^{-1}(x_d) \end{pmatrix}, \qquad \mathsf{A}\mathsf{A}^T = \mathsf{C} := \Big(\min(t_i, t_j)\Big)_{i,j=1}^{d}.$$

Here C is the covariance matrix of the values of a Brownian motion at the discrete times $t_1, \dots, t_d$. We choose A via a principal component analysis (singular value) decomposition of C as this tends to provide quicker convergence to the answer than other choices of A.

The option parameters for this example are $S_0 = 100$, $r = 2\%$, $\sigma = 50\%$, $K = 100$, and $T = 1$. We employ weekly monitoring, so $d = 52$, and $t_j = j/52$, where the option price is about \$11.97. Parameter $\widetilde{\beta}_{\text{qMC}}$ is estimated at the first iteration of the algorithm when $m = 10$, but not updated for each m. For $\varepsilon_{\text{a}} = 0.01$ and $\varepsilon_{\text{r}} = 0$, `cubSobol_g` without control variates requires 16,384 points while only 4096 when using control variates.

Figure 7 shows the Fourier Walsh coefficients of the original payoff, f, and the function integrated using control variates, $h_{\widetilde{\beta}_{\text{qMC}}} = f + \widetilde{\beta}_{\text{qMC}}(\mu_g - g)$, with given $\widetilde{\beta}_{\text{qMC}} = 1.06$, a typical value of β chosen by our algorithm. The squares correspond to the coefficients in the sums $\widetilde{S}_{6,10}(f)$ and $\widetilde{S}_{6,10}(h_{\widetilde{\beta}_{\text{qMC}}})$, respectively, which are used to bound the Sobol' cubature error. The circles are the first coefficients from the dual net that appear in error bound (4). From this figure we can appreciate how control variates reduces the magnitude of both the squares and the circles. For this example, the control variate coefficient estimated via (24) and sample quantities is $\beta_{\text{MC}} = 1.09$, which is quite similar to $\widetilde{\beta}_{\text{qMC}}$.

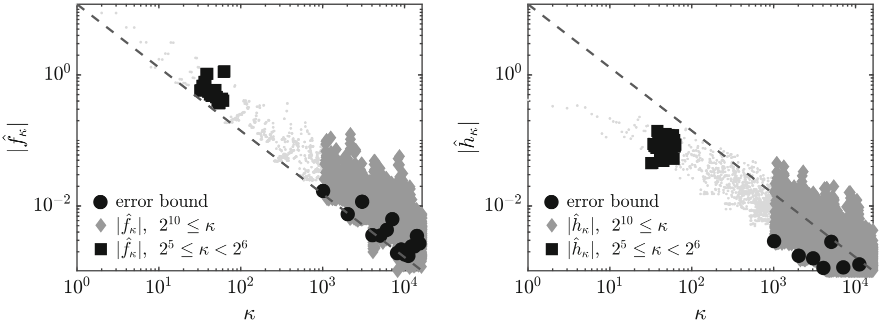


Fig. 7 Fourier Walsh coefficients for $f(\boldsymbol{x})$ on the left and $h_{1.06}(\boldsymbol{x})$ on the right. The value $\widetilde{\beta}_{\text{qMC}}$ effectively decreased the size of the coefficients involved in both the data-based error bound (9) and the error bound (4)

6 Discussion and Conclusion

Ian Sloan has made substantial contributions to the understanding and practical application of quasi-Monte Carlo cubature. One challenge is how to choose the parameters that define these cubatures in commonly encountered situations where not much is known about the integrand. These parameters include

(a) the generators of the sequences themselves,
(b) the sample size, n,
(c) the choice of importance sampling distributions,
(d) the control variate coefficients [12],
(e) the parameters defining multilevel (quasi-)Monte Carlo methods [6], and
(f) the parameters defining the multivariate decomposition method [28].

The rules for choosing these parameters should work well in practice, but not be simply heuristic as they are for some of the adaptive algorithms highlighted in the introduction. There should be a theoretical justification. Item (a) has received much attention. This article has addressed items (b) and (d). We realize that the question of choosing n is now replaced by the question of choosing the parameters defining the cone of integrands, $\mathscr{C}$. However, we have made progress because when our adaptive algorithms fail, we can pinpoint the cause. We would encourage further investigations into the best way to choose n. We also hope for more satisfying answers for the other items on the list in the future.

As demonstrated in Sect. 3, it is now possible to set relative error criteria or hybrid error criteria. We also now know how to accurately estimate a function of several means. In addition to the problem of Sobol' indices, this problem may arise in Bayesian inference, where the posterior mean of a parameter is the quotient of two integrals.

As already pointed out some years ago in [12], the choice of control variate for IID sampling is not necessarily the right choice for low discrepancy sampling. Here in Sect. 5, we have identified a natural way to determine a good control variate coefficient for digital sequence or lattice sequence sampling.

References

1. Bratley, P., Fox, B.L., Niederreiter, H.: Implementation and tests of low-discrepancy sequences. ACM Trans. Model. Comput. Simul. **2**, 195–213 (1992)
2. Choi, S.C.T., Ding, Y., Hickernell, F.J., Jiang, L., Jiménez Rugama, Ll.A., Tong, X., Zhang, Y., Zhou, X.: GAIL: Guaranteed Automatic Integration Library (versions 1.0–2.2). MATLAB software (2013–2017). http://gailgithub.github.io/GAIL_Dev/
3. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
4. Dick, J., Kuo, F., Sloan, I.H.: High dimensional integration—the Quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)

5. Genz, A.: Comparison of methods for the computation of multivariate normal probabilities. Comput. Sci. Stat. **25**, 400–405 (1993)
6. Giles, M.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)
7. Halton, J.H.: Quasi-probability: why quasi-Monte-Carlo methods are statistically valid and how their errors can be estimated statistically. Monte Carlo Methods Appl. **11**, 203–350 (2005)
8. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**, 299–322 (1998).
9. Hickernell, F.J., Hong, H.S.: Computing multivariate normal probabilities using rank-1 lattice sequences. In: Golub, G.H., Lui, S.H., Luk, F.T., Plemmons, R.J. (eds.) Proceedings of the Workshop on Scientific Computing, pp. 209–215. Springer, Singapore (1997)
10. Hickernell, F.J., Jiménez Rugama, Ll.A.: Reliable adaptive cubature using digital sequences. In: Cools, R., Nuyens, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, April 2014. Springer Proceedings in Mathematics and Statistics, vol. 163, pp. 367–383. Springer, Berlin (2016)
11. Hickernell, F.J., Hong, H.S., L'Écuyer, P., Lemieux, C.: Extensible lattice sequences for quasi-Monte Carlo quadrature. SIAM J. Sci. Comput. **22**, 1117–1138 (2000)
12. Hickernell, F.J., Lemieux, C., Owen, A.B.: Control variates for quasi-Monte Carlo. Stat. Sci. **20**, 1–31 (2005)
13. Jiménez Rugama, Ll.A.: Adaptive quasi-Monte Carlo cubature. Ph.D. thesis, Illinois Institute of Technology (2016)
14. Jiménez Rugama, Ll.A., Gilquin, L.: Reliable error estimation for Sobol' indices. Stat. Comput. (2017+, in press)
15. Jiménez Rugama, Ll.A., Hickernell, F.J.: Adaptive multidimensional integration based on rank-1 lattices. In: Cools, R., Nuyens, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, April 2014. Springer Proceedings in Mathematics and Statistics, vol. 163, pp. 407–422. Springer, Berlin (2016)
16. Li, D.: Reliable quasi-Monte Carlo with control variates. Master's thesis, Illinois Institute of Technology (2016)
17. Maize, E.: Contributions to the theory of error reduction in quasi-Monte Carlo methods. Ph.D. thesis, The Claremont Graduate School (1981)
18. Maize, E., Sepikas, J., Spanier, J.: Accelerating the convergence of lattice methods by importance sampling-based transformations. In: Plaskota, L., Woźniakowski, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2010. Springer Proceedings in Mathematics and Statistics, vol. 23, pp. 557–572. Springer, Berlin (2012)
19. Matoušek, J.: On the L_2-discrepancy for anchored boxes. J. Complexity **14**, 527–556 (1998)
20. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia (1992)
21. Nuyens, D.: https://people.cs.kuleuven.be/~dirk.nuyens/qmc-generators/genvecs/exod2_base2_m20.txt
22. Owen, A.B.: Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo. In: Niederreiter H., Spanier J. (eds.) Monte Carlo, Quasi-Monte Carlo, and Randomized Quasi-Monte Carlo, pp. 86–97. Springer, Berlin (2000)
23. Owen, A.B.: On the Warnock-Halton quasi-standard error. Monte Carlo Methods Appl. **12**, 47–54 (2006)
24. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press, Oxford (1994)
25. Sobol', I.M.: On sensitivity estimation for nonlinear mathematical models. Matem. Mod. **2**(1), 112–118 (1990)
26. Sobol', I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**(1–3), 271–280 (2001)
27. The MathWorks, Inc.: MATLAB 9.2. The MathWorks, Inc., Natick, MA (2017)
28. Wasilkowski, G.W.: On tractability of linear tensor product problems for ∞-variate classes of functions. J. Complex. **29**, 351–369 (2013)

Upwind Hybrid Spectral Difference Methods for Steady-State Navier–Stokes Equations

Youngmok Jeon and Dongwoo Sheen

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We propose an upwind hybrid spectral difference method for the steady-state Navier–Stokes equations. The (upwind) hybrid spectral difference method is based on a hybridization as follows: (1) an (upwind) spectral finite difference approximation of the Navier–Stokes equations within cells (*the cell finite difference*) and (2) an *interface finite difference* on edges of cells. The interface finite difference approximates continuity of normal stress on cell interfaces. The main advantages of this new approach are three folds: (1) they can be applied to non-uniform grids, retaining the order of convergence, (2) they are stable without using a staggered grid and (3) the schemes have an embedded static condensation property, hence, there is a big reduction in degrees of freedom in resulting discrete systems. The *inf-sup* condition is proved. Various numerical examples including the driven cavity problem with the Reynolds numbers, 5000–20,000, are presented.

1 Introduction

Although the finite difference method (FDM) is simple to implement and it can solve many physical problems efficiently [9, 22], several weak points are known compared to the finite element method and the finite volume method. For instance, some difficulties arise in the application of FDM in dealing with problems with complicated geometries, which may be overcome somehow by introducing proper geometric transforms [9, 25]. Also, uniform grids are usually necessary to have

Y. Jeon (✉)
Ajou University, Suwon, Korea
e-mail: yjeon@ajou.ac.kr

D. Sheen
Seoul National University, Seoul, Korea
e-mail: sheen@snu.ac.kr

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_28

optimal-order convergence. In particular, an application of standard FDM to the Stokes/Navier–Stokes equations using a non-staggered grid causes instability of checker-board pattern [22, 27], which may be resolved by introducing a staggered grid. Usually, using a staggered grid requires more programming efforts and treating boundary conditions is not a friendly task. In two dimension, one may avoid using a staggered grid simply by adopting the stream–vorticity formulation.

The aim of the current paper is to relax some of the above mentioned difficulties among the existing FDMs. We introduce a *hybrid spectral difference (HSD)* method and its *upwind* version. The hybrid spectral difference method is a hybridization, which is well-known in the community of domain decomposition methods, of the following two types of finite difference approximations; a *cell finite difference (cell FD)* and an *interface finite difference (interface FD).* The cell finite difference solves a local cell problem at interior nodes of each cell. The interface finite difference imposes the normal stress continuity at nodes on inter-cell boundaries. Therefore, the HSDs have a better flux conservation property, which is important to have a physically relevant solution for transport problems. When applying the cell FD for each cell the ideas in the $[Q_m]^2 \times Q_{m-2}$ spectral element [4] are employed.

The spectral difference (SD) methods are developed mainly for conservation laws, and they are very simple to implement and can be defined on both the triangular and rectangular meshes: for instance, see [2, 26, 28] and the references therein. In [10] an entropy stable finite difference method is introduced for conservation laws, where the spatial finite difference scheme within a cell is constructed to satisfy the summation-by-part (SBP) property. The SBP condition is similar to the discrete divergence theorem (Theorem 1) in this paper.

By using the two dimensional interpolation as in the SD method our HSD method may be also defined on a triangular mesh as well. However, the HSD on a rectangular mesh also can manage a complicated geometry somewhat well (see [17]), in which boundary cells are modified to contain a portion of the curved boundary. In comparison with the HDG methods the HSD can be understood as the finite difference version of the HDG method. The HSD is easier to implement than the HDG since it barely involves numerical integration. Moreover, pinpoint application of an upwind scheme is possible in the *upwind* HSD, while the upwind scheme must be applied in an integration sense in the HDG.

Our approach has the following several notable features. First of all, the HSD and *upwind* HSD are stable without resorting to a staggered grid for flow problems. Secondly, both schemes do not lose convergence order even on non-uniform grids [14]. A comparison between the HSD and its upwind version can be given as follows. The former achieves a higher order of convergence rate than the latter whenever convection does not dominate diffusion significantly. However, the upwind method is favorable if convection dominates. Numerical results will be shown to confirm these facts. Thirdly, our methods induce a natural variational formulation, which makes easy related numerical analysis. In [14] the HSD is introduced for the Poisson and Stokes equations and some elementary numerical analysis is provided. In [17] numerical analysis for HSDs of a diffusion equation is provided by introducing a discrete divergence theorem. Finally, a static condensation property is naturally embedded, hence, the degrees of freedom are reduced significantly for high-order difference methods.

The paper is organized as follows. In Sect. 2 the hybrid spectral difference method for the Oseen equations and its upwind version are presented. The steady-state Navier–Stokes equations are then to be solved by repeating the Oseen procedure. In Sect. 3 the *inf-sup* condition is proved, which is an essential part for stability of numerical methods. In Sect. 4 some numerical results illuminating convergence properties are presented and several those for a driven-cavity problem with Reynolds numbers ranging from 5000 to 20,000 on a 40 × 40 geometric grid. The streamline pictures show the primary and secondary vortices. Numerical results are also compared with well-known benchmark results in [8, 11]. Concluding remarks are briefly given in the final section.

2 The Hybrid Difference Methods

For a rectangle R denote by $Q_m(R)$ the space of polynomials in R of degree less than or equal to m in each variable. Wherever no confusion arises, the domain will be circumvented so that Q_m will mean the space of polynomials of degree less than or equal to m in each variable. The notation $\mathscr{T}(Q_m)$ represents an $(m+1)\times(m+1)$ mesh on which a polynomial $p \in Q_m$ can be uniquely determined.

One of the main objects of this paper is to introduce an inf-sup stable *upwind* hybrid spectral difference method on Q_m mesh for the Oseen equations. The velocity fields are approximated at the Q_m points while the pressure is approximated at the interior Q_{m-2} points. For this reason we will use the notation, $\mathscr{T}(Q_m) \times \mathscr{T}(Q_{m-2})$ mesh as well for the $\mathscr{T}(Q_m)$ mesh. In order to stabilize our scheme we introduce an *upwind* version by applying the upwind finite difference to the convection term.

For the sake of simplicity, let the domain Ω be a simply connected domain of which boundary is composed of lines that are parallel to axes. However, a generalization to domains with curved boundary can be treated with adequate modifications [17]. For a shape regular rectangular partition $\mathscr{T}_h$ the skeleton K_h of $\mathscr{T}_h$ is composed of all edges. In each cell of $\mathscr{T}_h$ the nodes are located as in Figs. 1 and 2. Let $\mathscr{N}_h(\Omega)$, $\mathscr{N}_h(\Gamma)$ and $\mathscr{N}_h(K_h)$ denote the set of all nodes in Ω and on Γ and K_h, respectively. Let R denote a typical element in $\mathscr{T}_h$, by h_1 and h_2 its horizontal and vertical sizes, and by η_{jk} and $\sigma_{jk} = \sigma_j\sigma_k$, $(j, k = 1, \cdots, m-1)$ the Gauss–(Legendre) points and weights in R. For each cell R, $\mathscr{N}_h(R)$ and $\mathscr{N}_h(\partial R)$ correspond to the cell versions of $\mathscr{N}_h(\Omega)$ and $\mathscr{N}_h(\Gamma)$, respectively.

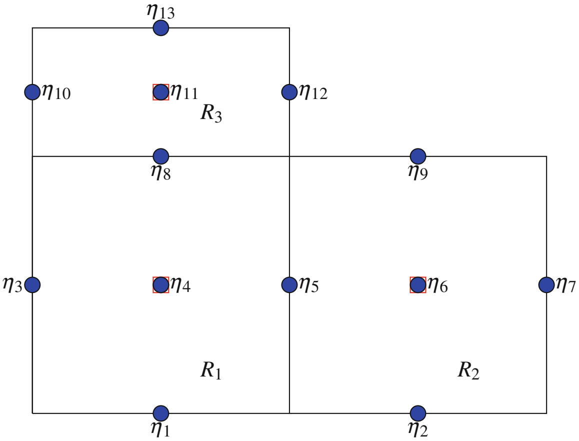


Fig. 1 $\mathscr{T}(Q_2) \times \mathscr{T}(Q_0)$ meshes: $|R_1| = h_1 \times k_1$, $|R_2| = h_2 \times k_1$, $|R_3| = h_1 \times k_2$

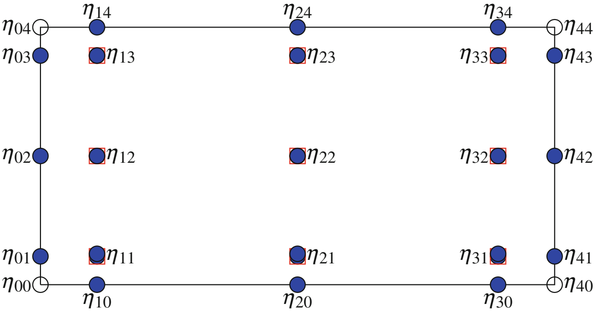


Fig. 2 A $\mathscr{T}(Q_4) \times \mathscr{T}(Q_2)$ mesh: $|R| = h_1 \times h_2$. The interior nodes are used for both $\mathbf{u}$ and p while those on the skeleton are only for $\mathbf{u}$. The point values at the four vertices $\{\eta_{jk} : i, j = 0, 4\}$ are not used in computation

2.1 Hybridization

We begin with the following steady-state Navier–Stokes equations:

$$\begin{aligned} -\frac{1}{Re}\Delta\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \nabla p = \mathbf{f} &\quad \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0 &\quad \text{in } \Omega, \\ \mathbf{u} = 0 &\quad \text{on } \Gamma, \\ \int_\Omega p\,\mathrm{d}\mathbf{x} = 0, & \end{aligned}$$

which can be solved by the following Oseen iteration:

$$-\frac{1}{Re}\Delta\mathbf{u}^{(n+1)}+\mathbf{u}^{(n)}\cdot\nabla\mathbf{u}^{(n+1)}+\nabla p^{(n+1)}=\mathbf{f}\quad\text{in }\Omega, \tag{1}$$

$$\nabla\cdot\mathbf{u}^{(n+1)}=0\quad\text{in }\Omega, \tag{2}$$

$$\mathbf{u}^{(n+1)}=0\quad\text{on }\Gamma, \tag{3}$$

$$\int_\Omega p^{(n+1)}\,\mathrm{d}\mathbf{x}=0. \tag{4}$$

Let $(\mathscr{T}_h)_{0<h<1}$ be a quasi-uniform family of rectangular triangulations such that the intersection of the closure of two rectangles is an edge or a vertex if it is not an empty set. The standard hybridization procedure is then to restrict Eqs. (1)–(4) to each rectangle with a suitable interface condition on its inter-element boundary. We thus have, for $R, R' \in \mathscr{T}_h$,

$$-\frac{1}{Re}\Delta\mathbf{u}^{(n+1)}+\mathbf{u}^{(n)}\cdot\nabla\mathbf{u}^{(n+1)}+\nabla p^{(n+1)}=\mathbf{f}\quad\text{in }R, \tag{5}$$

$$\nabla\cdot\mathbf{u}^{(n+1)}=0\quad\text{in }R, \tag{6}$$

$$\left[\!\left[-\frac{1}{Re}\partial_{\boldsymbol{\nu}}\mathbf{u}^{(n+1)}+(\boldsymbol{\nu}\cdot\mathbf{u}^{(n)})\mathbf{u}^{(n+1)}+p^{(n+1)}\boldsymbol{\nu}\right]\!\right]=0\quad\text{on }\partial R\cap\partial R', \tag{7}$$

$$\mathbf{u}^{(n+1)}=0\quad\text{on }\partial R\cap\partial\Omega, \tag{8}$$

$$\int_\Omega p^{(n+1)}\,\mathrm{d}\mathbf{x}=0. \tag{9}$$

Here, $\partial_{\boldsymbol{\nu}}\mathbf{u}=\boldsymbol{\nu}\cdot\nabla\mathbf{u}$, and $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ are the outward unit normal vectors to R and R', respectively, so that $\boldsymbol{\nu}=-\boldsymbol{\nu}'$ on $e=\partial R\cap\partial R'$. The notation $[[\cdot]]_e$ denotes the jump across e.

As the values of $\mathbf{u}$ from one side to the other side of an interface agree at the nodes on the cell interfaces, the stress continuity condition (7) can be simplified as

$$\left[\!\left[-\frac{1}{Re}\partial_{\boldsymbol{\nu}}\mathbf{u}^{(n+1)}+p^{(n+1)}\boldsymbol{\nu}\right]\!\right]=0\quad\text{on }\partial R\cap\partial R'. \tag{10}$$

The jump term in (10) will be designated by $\mathscr{J}(\mathbf{u}^{(n+1)},p^{(n+1)})$ from now on.

The hybrid spectral difference method (HSD) of the Oseen equations is based on the FD approximations of Eqs. (5)–(9).

2.2 1-d Finite Difference Formulas

We derive finite difference formulas in terms of degree of precision.

Let $m \geq 1$ be a fixed integer and consider an increasing sequence of nodes $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$ with $h = b - a$. For $\mu = 1, \cdots, m$, consider the problem to find $\left(w_{jk}^{(\mu)}\right)_{j,k=0}^{m}$ such that

$$\frac{1}{h^\mu} \sum_{j=0}^{m} (x_j)^\ell w_{jk}^{(\mu)} = \frac{d^\mu (x^\ell)}{dx^\mu} \,|_{x=x_k} \;\; (\ell, k = 0, 1, \cdots, m). \tag{11}$$

For fixed μ and k, Eq. (11) has a unique solution $\left(w_{jk}^{(\mu)}\right)_{j=0}^{m}$ as the system (11) forms a Vandermonde matrix system.

Definition 1 Based on $\left(w_{jk}^{(\mu)}\right)_{j,k=0}^{m}$ as in (11), a class of general μth-order finite difference operators $(D_x^h)^\mu$, $\mu = 1, 2, \cdots, m$, is defined as follows:

$$(D_x^h)^\mu f(x_k) = \frac{1}{h^\mu} \sum_{j=0}^{m} f(x_j) w_{jk}^{(\mu)}, \quad 0 \leq k \leq m, \quad \forall f \in C^{m+1}[a, b].$$

By the Taylor theorem, the following approximation property is immediate:

$$\left| (D_x^h)^\mu f(x_k) - \frac{d^\mu}{dx^\mu} f(x_k) \right| \lesssim h^{m-\mu+1} \| f^{(m+1)} \|_{L^\infty[a,b]} \quad \forall k = 0, 1, \cdots, m.$$

Here and in what follows, the notation "$L \lesssim R$" means that there exists a constant $c > 0$, independent of h, such that $L \leq cR$.

The upwind finite differences are obtained by using the m-nodes out of $m + 1$ nodes in the upwind direction for approximation of first derivative $\alpha D_x u$, where α is a convection coefficient.

$$D_x^{h,UP,+} f(x_k) = \frac{1}{h} \sum_{j=0}^{m-1} w_{j,k}^{(1,+)} f(x_j), \quad 1 \leq k \leq m-1 \text{ for } \alpha(x_k) > 0,$$

$$D_x^{h,UP,-} f(x_k) = \frac{1}{h} \sum_{j=1}^{m} w_{j,k}^{(1,-)} f(x_j), \quad 1 \leq k \leq m-1 \text{ for } \alpha(x_k) < 0.$$

Then,

$$|D_x f(x_k) - D_x^{h,UP,\pm} f(x_k)| \lesssim h^{m-1} \| f^{(m)} \|_{L^\infty[a,b]}, \quad 1 \leq k \leq m-1.$$

The notations $D_x^{h,UP,+}$ or $D_x^{h,UP,-}$ are simplified as $D_x^{h,UP}$ as long as there can be no misunderstanding.

Let us consider the extrapolations $\mathscr{E}_h^L$ and $\mathscr{E}_h^R$ such that $\mathscr{E}_h^L(f)(x_j) = \mathscr{E}_h^R(f)(x_j) = f(x_j)$ for $1 \leq j \leq m-1$, and

$$\mathscr{E}_h^L(f)(x_0) = \sum_{j=1}^{m-1} w_j^L f(x_j), \quad \mathscr{E}_h^R(f)(x_m) = \sum_{j=1}^{m-1} w_j^R f(x_j). \tag{12}$$

Such $w_j^L, w_j^R, j = 1, \cdots, m-1$, exist given the nodes $x_0 < x_1 < \cdots < x_m$. Designate by $\mathscr{E}_h$ either $\mathscr{E}_h^L$ or $\mathscr{E}_h^R$, depending on the direction of extension. Then,

$$|\mathscr{E}_h(f)(x_*) - f(x_*)| \lesssim h^{m-1} \| f^{(m-1)} \|_{L^\infty[a,b]}, \quad x_* = x_0,\ x_m.$$

The above one dimensional finite difference formula can be naturally extended to approximate partial derivatives in the two dimension as follows:

$$\Delta^h u = D_{xx}^h u + D_{yy}^h u, \ \ \nabla^h \cdot (u_1, u_2) = D_x^h u_1 + D_y^h u_2$$
$$\nabla^h u = (D_x^h u, D_y^h u)^T, \ \ \nabla^{h,UP} u = (D_x^{h,UP} u, D_y^{h,UP} u)^T.$$

Here, $D_{xx}^h = (D_x^h)^2$ and $D_{yy}^h = (D_y^h)^2$.

2.3 The $[Q_2]^2 \times Q_0$ HSD

To illustrate simplicity of the HSD in terms of implementation issue we consider the following case as illustrated in Fig. 1, where the pressure is defined only on nodes η_4, η_6 and η_{11}. Therefore, the extension $\mathscr{E}_h(p)$ (see (12) and the paragraph containing it, for the definition $\mathscr{E}_h$) is necessary to define the value of p at cell interface nodes. For the $[Q_2]^2 \times Q_0$ HSD, $\mathscr{E}_h(p)$ is constant on each cell, and $\nabla \mathscr{E}_h(p) = 0$ on each cell. For the sake of simplicity, denoting by $\mathbf{u} = (u_1, u_2), \mathbf{U} = (U_1, U_2)$, and p, respectively, the variables $\mathbf{u}^{(n+1)}$, $\mathbf{u}^{(n)}$, and $p^{(n+1)}$ in Eqs. (5)–(9) and choosing $R = R_1$ and the nodes as described in Fig. 1, we have the *cell finite difference* and the *interface finite difference* as follows:

$$\begin{aligned}
&-\frac{1}{Re}\Delta^h \mathbf{u}(\eta_4) + \mathbf{U} \cdot \nabla^h \mathbf{u}(\eta_4) \\
&= -\frac{1}{Re}\left[\frac{\mathbf{u}(\eta_5) - 2\mathbf{u}(\eta_4) + \mathbf{u}(\eta_3)}{(h_1/2)^2} + \frac{\mathbf{u}(\eta_8) - 2\mathbf{u}(\eta_4) + \mathbf{u}(\eta_1)}{(k_1/2)^2}\right] \\
&+ U_1(\eta_4)\frac{\mathbf{u}(\eta_5) - \mathbf{u}(\eta_3)}{h_1} + U_2(\eta_4)\frac{\mathbf{u}(\eta_8) - \mathbf{u}(\eta_1)}{k_1} = \mathbf{f}(\eta_4), \\
&\nabla^h \cdot \mathbf{u}(\eta_4) = \frac{u_1(\eta_5) - u_1(\eta_3)}{h_1} + \frac{u_2(\eta_8) - u_2(\eta_1)}{k_1} = 0,
\end{aligned}$$

and

$$\mathscr{J}(\mathbf{u},p)_{\eta_5} = \begin{cases} \frac{1}{Re}\frac{3u_1(\eta_5)-4u_1(\eta_4)+u_1(\eta_3)}{h_1} + \frac{1}{Re}\frac{3u_1(\eta_5)-4u_1(\eta_6)+u_1(\eta_7)}{h_2} \\ -p(\eta_4)+p(\eta_6) = 0, \\ \\ \frac{1}{Re}\frac{3u_2(\eta_5)-4u_2(\eta_4)+u_2(\eta_3)}{h_1} + \frac{1}{Re}\frac{3u_2(\eta_5)-4u_2(\eta_6)+u_2(\eta_7)}{h_2} = 0, \end{cases}$$

and

$$\mathscr{J}(\mathbf{u},p)_{\eta_8} = \begin{cases} \frac{1}{Re}\frac{3u_1(\eta_8)-4u_1(\eta_4)+u_1(\eta_1)}{k_1} + \frac{1}{Re}\frac{3u_1(\eta_8)-4u_1(\eta_{11})+u_1(\eta_{13})}{k_2} = 0, \\ \\ \frac{1}{Re}\frac{3u_2(\eta_8)-4u_2(\eta_4)+u_2(\eta_1)}{k_1} + \frac{1}{Re}\frac{3u_2(\eta_8)-4u_2(\eta_{11})+u_2(p_{13})}{k_2} \\ -p(\eta_4)+p(\eta_{11}) = 0. \end{cases}$$

The HSD is the finite difference version of the hybridized finite element method [7, 14, 15]. The main advantage of the HSD is that it does not involve any numerical integration except for approximation of $\int_\Omega p dx = 0$. Secondly, the finite differences are of one dimensional nature (they are not related to two or three dimensional polynomial interpolation). Therefore, we can handle the boundary data exactly even on a domain with a curved boundary by extending the line of derivative evaluation up to the given curved boundary and taking the intersection as a nodal point [17].

2.4 The $[Q_m]^2 \times Q_{m-2}$ Upwind HSD

Higher order methods can be obtained by considering the cell configuration Fig. 2, where $\mathscr{T}Q_m \times \mathscr{T}(Q_{m-2})$ nodes with $m = 4$ are illustrated for $R \in \mathscr{T}_h$, the configuration of which looks similar to the spectral element method for fluid problems in [4]. As in the previous subsection the pressure is defined only on the interior nodes. To obtain values of p on the skeleton nodes we use the extrapolation $\mathscr{E}_h(p)$. Here, the interior nodes are Gaussian points and they are projected to boundaries to obtain the skeleton nodes. Therefore, those interior and skeleton nodes does not constitute the Gauss-Lobatto nodes. The hybrid spectral difference method is to find the nodal values of $(\mathbf{u}_h, p_h)$ that satisfies the *cell FD:*

$$-\frac{1}{Re}\Delta^h \mathbf{u}_h(\eta_{jk}) + \mathbf{U}(\eta_{jk})\cdot\nabla^h \mathbf{u}_h(\eta_{jk}) + \nabla^h p_h(\eta_{jk}) = \mathbf{f}(\eta_{ij}), \qquad (13)$$

$$\nabla^h \cdot \mathbf{u}_h(\eta_{ij}) = 0$$

for each R and $1 \leq i,j \leq m-1$, and the *interface FD:*

$$\left[\!\left[-\frac{1}{Re}\partial_\nu^h \mathbf{u}_h + \boldsymbol{\nu}\mathscr{E}_h(p_h) \right]\!\right]_\eta = 0, \quad \eta \in \mathscr{N}_h(K_h), \qquad (14)$$

with $Q_h(p_h) = 0$, where $Q_h(p) \approx \int_\Omega p\,\mathrm{d}\mathbf{x}$ is a composite Gaussian quadrature approximation.

One of the main objects of this paper is to introduce an *upwind* HSD. To handle flows with a very high Reynolds number, it is necessary to introduce an upwind scheme. Our *upwind* HSD is obtained by applying the upwind finite difference to the convection term in (13) so that

$$-\frac{1}{Re}\Delta^h \mathbf{u}_h(\eta_{ij}) + \mathbf{U}(\eta_{ij}) \cdot \nabla^{h,UP}\mathbf{u}_h(\eta_{ij}) + \nabla^h p_h(\eta_{ij}) = \mathbf{f}(\eta_{ij}), \tag{15}$$

$$\nabla^h \cdot \mathbf{u}_h(\eta_{ij}) = 0 \tag{16}$$

for each R and $1 \le i,j \le m-1$, with the *interface FD* (14).

The major differences between the traditional spectral difference schemes [20, 24, 29] and ours are as follows. Most existing schemes are basically defined on staggered grids such that the solution and flux variables are approximated at the $(m-1)$ Gauss–Legendre or Chebyshev points and the $(m+1)$ Gauss–Lobatto points, respectively. The use of staggered grids is essential to guarantee the stability of their spectral difference scheme. Staggered grids based on Gauss–Chebyshev points have been also introduced [20] to approximate compressible Navier–Stokes equations. This scheme is conservative which extends the Euler method studied by Kopriva and Kolias [19, 21]. The approximation of the velocity is approximated on staggered grids componentwisely in a different setting [3]. A stable spectral collocation method has been introduced by Carpenter et al. [6]. The conservative staggered-grid method, which enforces weakly the continuity of the solution and the viscous fluxes at subdomain interfaces, has advantages over methods based on Lobatto grids, e.g., [13, 20].

However, our method adopts a single grid based on $(m-1)$ Gauss-Legendre points; the velocity fields are approximated at these $(m-1)$ Gauss-Legendre points plus the two projected points on the two cell boundaries in each direction, say -1 and 1 on the interval $[-1, 1]$, while the pressure variable is approximated at the $(m-1)$ interior Gauss-Legendre points only.

2.5 A Variational Formulation

Let $C(\Omega)$ be the space of continuous functions in Ω and introduce an equivalence relation $u \equiv v$ for $u, v \in C(\Omega)$ if $u(\eta) = v(\eta)$ for all $\eta \in \mathscr{N}_h(\Omega) \cup \mathscr{N}_h(\Gamma)$. Denote by $C^h(\Omega)$ the space of its equivalent classes. Similarly, set $\Pi_{R\in\mathscr{T}_h} C(R)$ as the space of piecewise continuous functions and introduce another equivalence relation $u \equiv v$ on $\Pi_{R\in\mathscr{T}_h} C(R)$ if $u(\eta) = v(\eta)$ for all $\eta \in \mathscr{N}_h(R) \setminus \mathscr{N}_h(\partial R)$ for all $R \in \mathscr{T}_h$, and denote this equivalent class by $L^h(\Omega)$. Next, define a discrete inner product on $C^h(\Omega) \cup L^h(\Omega)$ as follows:

$$(u, v)_h = \sum_{R\in\mathscr{T}_h} (u, v)_{R,h}, \quad (u, v)_{R,h} := |R| \sum_{1\le j,k\le m-1} \sigma_{jk} u(\eta_{jk}) v(\eta_{jk}). \tag{17}$$

where $|R|$ represents the area of R. Denote by $L^h(K_h)$ the space of piecewise continuous functions $\Pi_{e\subset\partial R, R\in\mathscr{T}_h} C(e)$ whose equivalence class is defined as $u \equiv v$ on $L^h(K_h)$ if $u(\eta) = v(\eta)$ for all $\eta \in \mathscr{N}_h(\mathscr{K}_h)$. A discrete inner product on $L^h(K_h)$ is then defined as

$$\langle u, v\rangle_h = \sum_{R\in\mathscr{T}_h} \langle u, v\rangle_{\partial R,h}, \quad \langle u, v\rangle_{\partial R,h} := \sum_{e\subset\partial R} \langle u, v\rangle_{e,h}, \tag{18}$$

$$\langle u, v\rangle_{e,h} := |e| \sum_{k=1}^{m-1} \sigma_k u(\eta_k) v(\eta_k),$$

where $\eta_j, j = 1, \cdots, m-1$, are the Gaussian nodes associated with weights $\sigma_j, j = 1, \cdots, m-1$ on e. It is worth to note that the Gaussian quadrature has the degree of precision $(2m-3)$.

The (upwind) hybrid spectral difference method ((15), (16) and (14)) can be rewritten in the following variational form: Find $(\mathbf{u}_h, p_h) \in [C^h(\Omega)]^2 \times L^h(\Omega)$ that satisfies

$$\mathscr{A}_h(\mathbf{u}_h, p_h; \mathbf{v}, q) = (\mathbf{f}, \mathbf{v})_h, \quad (\mathbf{v}, q) \in [C^h(\Omega)]^2 \times L^h(\Omega), \tag{19}$$

where

$$\begin{aligned}\mathscr{A}_h(\mathbf{u}_h, p_h; \mathbf{v}, q) = {} & (-\frac{1}{Re}\Delta^h \mathbf{u}_h + \mathbf{U}\cdot\nabla^{h,UP}\mathbf{u}_h + \nabla^h p_h, \mathbf{v})_h \\ & +(\nabla^h\cdot\mathbf{u}_h, q)_h + \langle \frac{1}{Re}\partial^h_\nu \mathbf{u}_h - \boldsymbol{\nu}\mathscr{E}_h(p_h), \mathbf{v}\rangle_h .\end{aligned} \tag{20}$$

Remark 1 With the cell configuration in Fig. 2 let us augment the boundary node set by including the four vertices of R (say, $V_R = \{\eta_{00}, \eta_{0m}, \eta_{m0}, \eta_{mm}\}$ with $m = 4$). For a rectangular subdivision $\mathscr{T}_h$ let $\mathscr{V}_h$ be the set of all vertices. Let $\widetilde{C}^h(\Omega)$ be the set of equivalence classes on $C(\Omega)$, where the equivalence relation is given as $u \equiv v$ if $u(\eta) = v(\eta)$ for $\eta \in \mathscr{N}_h(\Omega) \cup \mathscr{N}_h(\Gamma) \cup \mathscr{V}_h$. Then, $\widetilde{C}^h(\Omega)$ is an equivalent space to $Q_m(\Omega)$, and $L^h(\Omega)$ corresponds to $Q_{m-2}(\mathscr{T}_h)$, respectively. Here,

$$Q_m(\Omega) = \{u \in C(\Omega) : u|_R \in Q_m(R)\}, \quad Q_{m-2}(\mathscr{T}_h) = \{p \in L_2(\Omega) : p|_R \in Q_{m-2}(R)\}.$$

3 The *inf-sup* Condition

We begin with proving the discrete divergence theorem.

Theorem 1 (Discrete Divergence Theorem) *For* $(\mathbf{v}, p) \in [\widetilde{C}^h(R)]^2 \times L^h(R)$

$$(\nabla^h p, \mathbf{v})_{R,h} + (\nabla^h\cdot\mathbf{v}, p)_{R,h} - \langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial R,h} = 0$$

with $\widetilde{C}^h(R) = \widetilde{C}^h(\Omega)|_R$ *and* $L^h(R) = L^h(\Omega)|_R$.

Proof For $j = 1, \cdots, m-1$, set $I_j = \overline{\eta_{0j}\eta_{mj}}$ and $J_j = \overline{\eta_{j0}\eta_{jm}}$. Then, from the definition (18) it follows that

$$\langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial R,h} = h_2 \sum_{j=1}^{m-1} \sigma_j \langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial I_j} + h_1 \sum_{j=1}^{m-1} \sigma_j \langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial J_j}. \tag{21}$$

By using the equivalences of the function spaces we regard $\widetilde{C}^h(R) = Q_m(R)$ and $L^h(R) = Q_{m-2}(R)$. Set $\mathbf{v} = (v_1, v_2)$. Notice that $\mathscr{E}_h(p) = p$ as functions. By the integration by parts

$$\begin{aligned}\langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial I_j} &= \mathscr{E}_h(p)(\eta_{mj})v_1(\eta_{mj}) - \mathscr{E}_h(p)(\eta_{0j})v_1(\eta_{0j}) \\ &= \int_{I_j} \frac{\partial(v_1\mathscr{E}_h(p))}{\partial x}\,\mathrm{d}\mathbf{x} \\ &= \int_{I_j} \frac{\partial v_1}{\partial x} p\,\mathrm{d}\mathbf{x} + \int_{I_j} v_1 \frac{\partial p}{\partial x}\,\mathrm{d}\mathbf{x} \\ &= h_1 \sum_{k=1}^{m-1} \sigma_k \frac{\partial v_1}{\partial x}(\eta_{kj})p(\eta_{kj}) + h_1 \sum_{k=1}^{m-1} \sigma_k v_1(\eta_{kj}) \frac{\partial p}{\partial x}(\eta_{kj}).\end{aligned}$$

The last line follows from the fact that $\frac{\partial v_1}{\partial x}p$ and $v_1\frac{\partial p}{\partial x}$ are polynomials of degree $\leq 2m-3$ in x-variable along I_j. From (11) and (12) we have that $\frac{\partial v}{\partial x} = D_x^h v$ and $\frac{\partial p}{\partial x} = D_x^h p$ for $v \in Q_m(R)$ and $p \in Q_{m-2}(R)$, respectively. Therefore,

$$\begin{aligned} & h_2 \sum_{j=1}^{m-1} \sigma_j \langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial I_j} \\ &= |R| \sum_{j=1}^{m-1}\sum_{k=1}^{m-1} \sigma_{jk} \left\{ D_x^h v_1(\eta_{kj})p(\eta_{kj}) + v_1(\eta_{kj})D_x^h p(\eta_{kj}) \right\}. \end{aligned} \tag{22}$$

Similarly,

$$\begin{aligned} & h_1 \sum_{j=1}^{m-1} \sigma_j \langle \boldsymbol{\nu}\mathscr{E}_h(p), \mathbf{v}\rangle_{\partial J_j} \\ &= |R| \sum_{j=1}^{m-1}\sum_{k=1}^{m-1} \sigma_{jk} \left\{ D_y^h v_2(\eta_{kj})p(\eta_{kj}) + v_2(\eta_{kj})D_y^h p(\eta_{kj}) \right\}. \end{aligned} \tag{23}$$

A combination of (21)–(23) proves the theorem. □

An immediate, but important consequence of Theorem 1 is that (20) can be simplified as follows.

Corollary 1 *For any* $\mathbf{u}_h \in Q_m(R)$ *and* $p_h \in Q_{m-2}(R)$, *we have*

$$\mathscr{A}_h(\mathbf{u}_h, p_h; \mathbf{u}_h, p_h) = (-\frac{1}{Re}\Delta^h \mathbf{u}_h + \mathbf{U}\cdot\nabla^{h,UP}\mathbf{u}_h, \mathbf{u}_h)_h + \frac{1}{Re}\langle \partial_\nu^h \mathbf{u}_h, \mathbf{u}_h\rangle_h.$$

The ellipticity proof of $\mathscr{A}_h(\mathbf{u}_h, p_h; \mathbf{u}_h, p_h)$ without the advection term (the case, $\mathbf{U} = 0$) is obtained in [17]. For the Q_2 upwind method an ellipticity result for the convection diffusion equation can be found in [16]. Ellipticity of a higher order method for the nonzero advection case might be obtained by a similar manner as in [17], and it will be a subject of future research.

In order to derive the *inf-sup* condition, let us introduce a discrete L^2-norm on $L^h(\Omega)$ as follows:

$$\|u\|^2_{L_2^h(\Omega)} = \sum_{R\in\mathscr{T}_h} \|u\|^2_{L_2^h(R)}, \quad \forall u \in L^h(\Omega),$$

where

$$\|u\|^2_{L_2^h(R)} = |R| \sum_{1\le j,k\le m-1} \sigma_{jk}|u(\eta_{jk})|^2.$$

The discrete L^2-norm becomes a seminorm on $C^h(\Omega)$.

Using the finite dimensionality, scaling invariance and the fact that the partition is quasi-uniform, it holds that

$$\|u\|_{L_2^h(R)} \le c_m \|u\|_{L_2(R)} \quad \forall u \in Q_m(R). \tag{24}$$

From here on $L \lesssim R$ means that there exists a constant $c_m > 0$ such that $L \le c_m R$. We will prove the discrete inf-sup condition: there exist $\beta_m > 0$, independent of h, such that

$$\sup_{\mathbf{w}\in[\widetilde{C}^h(\Omega)]^2} \frac{(\nabla^h\cdot\mathbf{w}, q)_h}{\|\nabla^h\mathbf{w}\|_{L_2^h(\Omega)}} \ge \beta_m \|q\|_{L_2^h(\Omega)}, \quad q \in L^h(\Omega). \tag{25}$$

In order to prove (25), we introduce some useful interpolation and projection. For each $R \in \mathscr{T}_h$, define an interpolation $\Pi_R : [C(R) \cap H^1(R)]^2 \to [Q_m(R)]^2$ by $\Pi_R\mathbf{u}(\eta) = \mathbf{u}(\eta)$ for all $\eta \in \mathscr{N}_h(R) \cup \mathscr{N}_h(\partial R) \cup \mathscr{V}_h(R)$. Then, we have

$$\|\nabla \Pi_R \mathbf{u}\|_{L_2(R)} \lesssim \|\nabla \mathbf{u}\|_{L_2(R)}. \tag{26}$$

An immediate application of Theorem 1 to $\Pi_R\mathbf{u}$ leads to

$$(\nabla\cdot\Pi_R\mathbf{u},q)_{R,h}=-(\Pi_R\mathbf{u},\nabla^h q)_{R,h} \\ +\langle\boldsymbol{\nu}\cdot\mathbf{u},\mathscr{E}_h(q)\rangle_{\partial R,h},\quad \mathbf{u}\in[C(R)]^2,\quad q\in L^h(R) \tag{27}$$

since $\nabla\cdot\Pi_R\mathbf{u}=\nabla^h\cdot\Pi_R\mathbf{u}$ and $\mathbf{u}=\Pi_R\mathbf{u}$ on $\mathscr{N}_h(\partial R)\cup\mathscr{V}_h(R)$ for each $R\in\mathscr{T}_h$.

Next, for each $R\in\mathscr{T}_h$, define a projection $\widetilde{\Pi}_R:[C(\Omega)\cap H^1(R)]^2\to[Q_m(R)]^2$ such that for $\mathbf{u}\in[C(R)\cap H^1(R)]^2$

$$(\widetilde{\Pi}_R\mathbf{u},\nabla^h q)_{R,h}=-(\nabla\cdot\mathbf{u},q)_R+\langle\boldsymbol{\nu}\cdot\mathbf{u},\mathscr{E}_h(q)\rangle_{\partial R,h},\quad q\in Q_{m-2}, \tag{28}$$

and $\widetilde{\Pi}_R\mathbf{u}=\mathbf{u}$ on $\mathscr{N}_h(\partial R)\cup\mathscr{V}_h(R)$. To fulfill (28), define $\widetilde{\Pi}_R:[C(R)\cap H^1(R)]^2\to[Q_m(R)]^2$ as follows:

p1. $(\widetilde{\Pi}_R\mathbf{u}-\Pi_R\mathbf{u},\nabla^h q)_{R,h}=-(\nabla\cdot\mathbf{u},q)_R+(\nabla\cdot\Pi_R\mathbf{u},q)_{R,h}$ for $q\in Q_{m-2}(R)$,
p2. $(\widetilde{\Pi}_R\mathbf{u}-\Pi_R\mathbf{u},\mathbf{r})_{R,h}=0$ for $\mathbf{r}\in[\nabla Q_{m-2}(R)]^\perp$,
p3. $\widetilde{\Pi}_R\mathbf{u}=\Pi_R\mathbf{u}$ on ∂R

Here, $[\nabla Q_{m-2}(R)]^\perp$ is the orthogonal compliment of $\nabla Q_{m-2}(R)$ in $[Q_{m-2}(R)]^2$ with respect to the inner product $(\cdot,\cdot)_{h,R}$ on $[Q_{m-2}(R)]^2$. Notice that $\widetilde{\Pi}_R\mathbf{u}\in[Q_m(R)]^2$ is well-defined by (p1)–(p3). Also observe that (28) is obtained by subtracting (p1) from (27), and the property (p2) is considered for uniqueness of the projection.

The global projections $\Pi:[C(\Omega)\cap H^1(\Omega)]^2\to[Q_m(\Omega)]^2$ and $\widetilde{\Pi}:[C(\Omega)\cap H^1(\Omega)]^2\to[Q_m(\Omega)]^2$ are then defined elementwise such that $\Pi\ |_R=\Pi_R$ and $\widetilde{\Pi}\ |_R=\widetilde{\Pi}_R$ for each $R\in\mathscr{T}_h$.

Lemma 1 *For $\mathbf{u}\in[C(\Omega)\cap H^1(\Omega)]^2$, the projection $\widetilde{\Pi}\mathbf{u}$ satisfies the following properties:*

$$(\nabla\cdot\widetilde{\Pi}_R\mathbf{u},q)_{R,h}=(\nabla\cdot\mathbf{u},q)_R,\quad q\in Q_{m-2}(R)\quad\forall R\in\mathscr{T}_h; \tag{29}$$

$$\|\nabla\widetilde{\Pi}\mathbf{u}\|_{L_2^h(\Omega)}\lesssim\|\nabla\mathbf{u}\|_{L_2(\Omega)}. \tag{30}$$

Proof Let $\mathbf{u}\in[C(\Omega)\cap H^1(\Omega)]^2$ be arbitrary. By Theorem 1, for each $R\in\mathscr{T}_h$, $\widetilde{\Pi}_R\mathbf{u}$ satisfies

$$(\nabla\cdot\widetilde{\Pi}_R\mathbf{u},q)_{R,h}=-(\widetilde{\Pi}_R\mathbf{u},\nabla^h q)_{R,h}+\langle\boldsymbol{\nu}\cdot\mathbf{u},\mathscr{E}_h(q)\rangle_{\partial R,h},\quad q\in Q_{m-2}(R).$$

Subtracting the above equation from (28) we have (29).

Now, we proceed to prove (30). For $\mathbf{r}\in[Q_{m-2}(R)]^2$, decompose it orthogonally as $\mathbf{r}=\nabla q+(\mathbf{r}-\nabla q)$ for some $q\in Q_{m-2}(R)$ such that q vanishes at least one interior node $\eta^*\in\mathscr{N}_h(R)$. Then, a simple calculation yields that

$$q(\eta)=h\sum_{k,l=1}^{m-1}\left[c_\eta^{kl}D_x^h q(\eta_{kl})+d_\eta^{kl}D_y^h q(\eta_{kl})\right],$$

where the coefficients $c_\eta^{kl}, d_\eta^{kl}, (1 \leq k, l \leq m-1)$ are independent of h. In general, we have

$$\|q\|_{L_2^h(R)} \lesssim h\|\nabla^h q\|_{L_2^h(R)},$$
$$q \in \{q \in Q_{m-2}(R) : q(\eta^*) = 0,\ \textit{some}\ \eta^* \in \mathscr{N}_h(R)\}. \tag{31}$$

By (24) and (26) and the exactness of the Gaussian quadrature for $\int_R q^2\,\mathrm{d}\mathbf{x}$ (recalling that the degree of $q^2 \leq 2m-4 \leq 2m-3$)

$$\|\nabla \Pi_R \mathbf{u}\|_{L_2^h(R)} \lesssim \|\nabla \Pi_R \mathbf{u}\|_{L_2(R)} \lesssim \|\nabla \mathbf{u}\|_{L_2(R)}, \tag{32}$$

$$\|q\|_{L_2^h(R)} = \|q\|_{L_2(R)}. \tag{33}$$

Note that $\nabla^h q(\eta_{ij}) = \nabla q(\eta_{ij})$ $(i,j = 1,\cdots,m-1)$ for $q \in Q_{m-2}(R)$. Using (p2), (p1), (32), (33) and (31) sequentially, one has

$$\begin{aligned}
|(\widetilde{\Pi}_R\mathbf{u} - \Pi_R\mathbf{u}, \mathbf{r})_{R,h}| &= |(\widetilde{\Pi}_R\mathbf{u} - \Pi_R\mathbf{u}, \nabla^h q)_{R,h}| \\
&\leq |-(\nabla\cdot\mathbf{u}, q)_R + (\nabla\cdot\Pi_R\mathbf{u}, q)_{R,h}| \\
&\lesssim \|\nabla\mathbf{u}\|_{L_2(R)}\|q\|_{L_2(R)} + \|\nabla\Pi_R\mathbf{u}\|_{L_2^h(R)}\|q\|_{L_2^h(R)} \\
&\lesssim \|\nabla\mathbf{u}\|_{L_2(R)}\|q\|_{L_2^h(R)} \\
&\lesssim h\|\nabla\mathbf{u}\|_{L_2(R)}\|\nabla^h q\|_{L_2^h(R)} \\
&\lesssim h\|\nabla\mathbf{u}\|_{L_2(R)}\|\mathbf{r}\|_{L_2^h(R)}.
\end{aligned}$$

As a result we have

$$\|\widetilde{\Pi}_R\mathbf{u} - \Pi_R\mathbf{u}\|_{L_2^h(R)} \lesssim h\|\nabla\mathbf{u}\|_{L_2(R)}.$$

Note that $\widetilde{\Pi}_R\mathbf{u} - \Pi_R\mathbf{u} = 0$ on ∂R. Since $(\mathscr{T}_h)_{0<h<1}$ is quasi-uniform, the inverse estimate yields that

$$\|\nabla(\widetilde{\Pi}_R\mathbf{u} - \Pi_R\mathbf{u})\|_{L_2^h(R)} \lesssim \|\nabla\mathbf{u}\|_{L_2(R)}. \tag{34}$$

A combination of (32) and (34) and summation over all $R \in \mathscr{T}_h$, (30) follows:

$$\|\nabla\widetilde{\Pi}\mathbf{u}\|_{L_2^h(\Omega)} \lesssim \|\nabla\Pi\mathbf{u}\|_{L_2^h(\Omega)} + \|\nabla(\widetilde{\Pi}\mathbf{u} - \Pi\mathbf{u})\|_{L_2^h(\Omega)} \lesssim \|\nabla\mathbf{u}\|_{L_2(\Omega)}.$$

This completes the proof. □

The *inf-sup* condition (25) follows by the framework of analysis in [5, 12] as follows. Using Lemma 1

$$\sup_{\mathbf{w}\in[\widetilde{C}^h(\Omega)]^2}\frac{(\nabla\cdot\mathbf{w},q)_h}{\|\nabla\mathbf{w}\|_{L_2^h(\Omega)}}\ge\sup_{\mathbf{u}\in[C(\Omega)\cap H^1(\Omega)]^2}\frac{(\nabla\cdot\widetilde{\Pi}\mathbf{u},q)_h}{\|\nabla\widetilde{\Pi}\mathbf{u}\|_{L_2^h(\Omega)}}=\sup_{\mathbf{u}\in[C(\Omega)\cap H^1(\Omega)]^2}\frac{(\nabla\cdot\mathbf{u},q)_\Omega}{\|\nabla\widetilde{\Pi}\mathbf{u}\|_{L_2^h(\Omega)}}$$

$$\ge c_m\sup_{\mathbf{u}\in[C(\Omega)\cap H^1(\Omega)]^2}\frac{(\nabla\cdot\mathbf{u},q)_\Omega}{\|\nabla\mathbf{u}\|_{L_2(\Omega)}}\ge\widetilde{c}_m\|q\|_{L_2(\Omega)}=\widetilde{c}_m\|q\|_{L_2^h(\Omega)}$$

for $q\in L^h(\Omega)$ and $\int_\Omega q\,d\mathbf{x}=0$.

We summarize the above as in the following theorem.

Theorem 2 *The following inf-sup condition holds; for some $\beta_m>0$*

$$\sup_{\mathbf{w}\in[\widetilde{C}^h(\Omega)]^2}\frac{(\nabla^h\cdot\mathbf{w},q)_h}{\|\nabla^h\mathbf{w}\|_{L_2^h(\Omega)}}\ge\beta_m\|q\|_{L_2^h(\Omega)},\quad q\in L^h(\Omega),$$

where $(q,1)_\Omega=0$.

4 Numerical Experiments

We present three different examples. Examples 1 and 2 are to test the convergence properties of the HSD and *upwind* HSD for the Oseen equations with a smooth solution and a singular solution, respectively. Finally, Example 3 treats the driven cavity problem of the steady-state Navier–Stokes equations. For all examples the domain of computation is the unit square, [0, 1] × [0, 1], and the computational grid is the $\mathscr{T}(Q_4)\times\mathscr{T}(Q_2)$ as in Fig. 2. Since the exact solutions in Examples 2 and 3 have known singularities at the corners of the domain, it is desirable to use a graded grid (a geometric grid). Therefore, we consider the following form of graded grid for Examples 2 and 3.

$$(\tilde{x},\tilde{y})=\left(\frac{x^2}{x^2+(1-x)^2},\frac{y^2}{y^2+(1-y)^2}\right),$$

where (x,y) is a uniform grid. In Figs. 3, 4 and 5 the errors are estimated in the discrete L_2-norm for $\mathbf{u}$ and p, and the number in the slope box represents the order of convergence.

Example 1 Consider the Oseen equations:

$$-\frac{1}{Re}\Delta\mathbf{u}+\mathbf{U}\cdot\nabla\mathbf{u}+\nabla p=\mathbf{f}\quad\text{in }\Omega,$$

$$\nabla\cdot\mathbf{u}=0,\quad\text{in }\Omega,$$

$$\mathbf{u}=\mathbf{g}\quad\text{on }\Gamma,$$

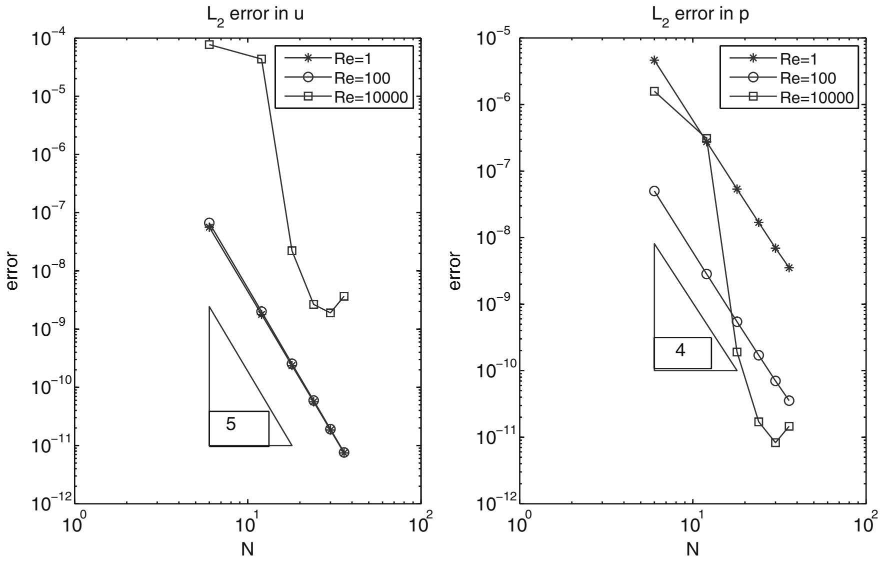


Fig. 3 Convergence history of the HSD for Example 1 with $Re = 1 \sim 10^4$

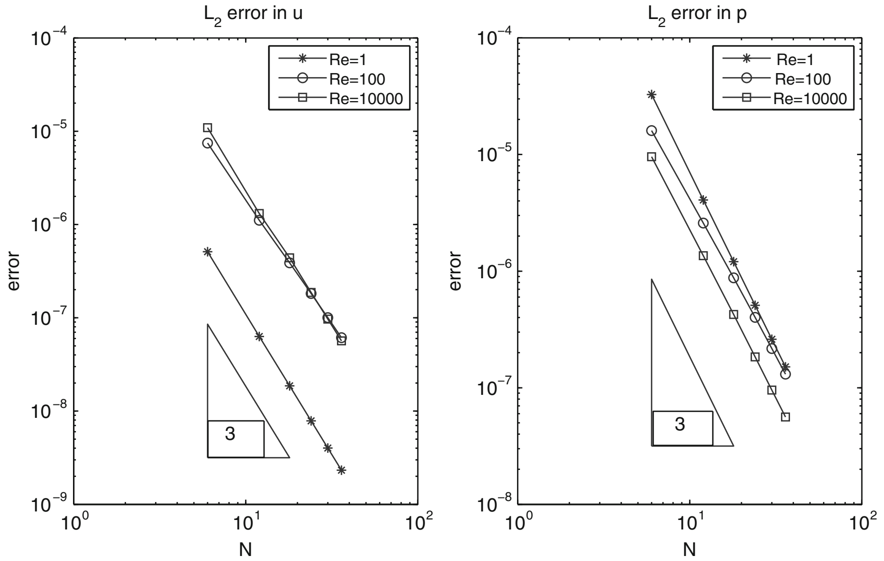


Fig. 4 Convergence history of the *upwind* HSD for Example 1 with $Re = 1 \sim 10^4$

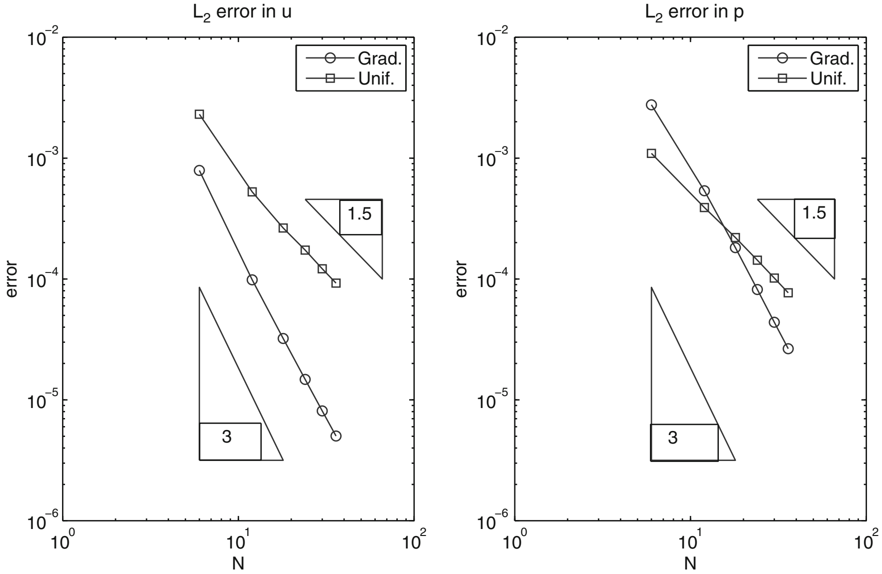


Fig. 5 Convergence history of the *upwind* HSD for Example 2 ($Re = 10^4$) with two different meshes; graded (Grad.) and uniform (Unif)

Table 1 The history of convergence for the HSD and *upwind* HSD for Example 1 with $Re = 10$, and α represents the rate of convergence

	HSD				Upwind HSD			
N	$\|u-u_h\|_{L_2^h(\Omega)}$	α	$\|p-p_h\|_{L_2^h(\Omega)}$	α	$\|u-u_h\|_{L_2^h(\Omega)}$	α	$\|p-p_h\|_{L_2^h(\Omega)}$	α
6	1.0873e−06		3.0056e−06		2.1611e−06		1.5846e−05	
12	4.1744e−08	4.70	2.2280e−07	3.75	2.4450e−07	3.14	2.2892e−06	2.79
18	5.7051e−09	4.91	4.4872e−08	3.95	7.9219e−08	2.78	7.1239e−07	2.88
24	1.3743e−09	4.95	1.4193e−08	4.00	3.5443e−08	2.80	3.0807e−07	2.91
30	4.5411e−10	4.96	5.7862e−09	4.02	1.8852e−08	2.83	1.6010e−07	2.93
36	1.8355e−10	4.97	2.7758e−09	4.03	1.1201e−08	2.86	9.3588e−08	2.94

where $\mathbf{f}$ and $\mathbf{g}$ are given to have the exact solution $\mathbf{u}(x, y) = (\exp(x)\cos(y), -\exp(x)\sin(y))$ and $p = x^2 - \frac{1}{3}$. Here, $\mathbf{U} = (x^2 + y^2 - 1, -2xy + 1)$ with $Re = 1 \sim 10^4$. From here on, Re represents the Reynolds number.

Figures 3 and 4 represent numerical results for Example 1, and Table 1 represents the corresponding numerics when $Re = 10$. For the HSD we observe the convergence, $\|\mathbf{u} - \mathbf{u}_h\|_{L_2^h(\Omega)} = O(h^5)$ and $\|p - p_h\|_{L_2^h(\Omega)} = O(h^4)$ for $Re = 1 \sim 10^2$. For $\mathbf{u}$ the above convergence is optimal, and the theoretically expected convergence is of $O(h^3)$ for p. Numerical results show a super convergence for p, which phenomenon happens often when error is measured with nodal values. However, erratic convergence is observed as Re increases up to 10^4. For the *upwind* method the apparent orders of L_2^h-convergence for the velocity and pressure variables are identical, i.e., $\|\mathbf{u}-\mathbf{u}_h\|_{L_2^h(\Omega)} = O(h^3)$ and $\|p-p_h\|_{L_2^h(\Omega)} = O(h^3)$. Over all, the HSD yields better approximation in the case of smooth solution problems with low Reynolds numbers. An advantage of using the *upwind* HSD is that the convergence order of numerical solutions is barely influenced by Reynolds numbers.

Example 2 Now consider the Oseen equations with $Re = 10^4$ as in Example 1, but with singular solutions given as $\mathbf{u}(x,y) = (y^{1.5}, x^{1.5})$ and $p = x^2 - \frac{1}{3}$. Hence, $\mathbf{u} \in [H^{2-\epsilon}(\Omega)]^2$ with arbitrary $\epsilon > 0$.

Figure 5 represents numerical results for Example 2. If one uses a graded mesh the order of convergence can be improved for the both HSD and *upwind* HSD. Here, only graphs for the *upwind* HSD are presented in this case.

Example 3 We consider the driven-cavity flow problem:

$$\begin{aligned} -\frac{1}{Re}\Delta\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \nabla p = 0 \quad &\text{in } \Omega \\ \nabla\cdot\mathbf{u} = 0 \quad &\text{in } \Omega \end{aligned}$$

with the watertight boundary condition:

$$\mathbf{u}(x_1,x_2) = \begin{cases} (1,0)^T, \; 0 < x_1 < 1, \;\; x_2 = 1, \\ (0,0)^T, \qquad\quad \mathit{otherwise}. \end{cases}$$

The solution $\mathbf{u}$ belongs to the Sobolev space $[H^{1-\epsilon}(\Omega)]^2$ for $\epsilon > 0$ with singularities at the upper two corners of the domain [1, 18]. In view of numerical experiments in Example 2 the graded mesh is used for numerical experiments of the cavity problem.

The Navier–Stokes equations is solved by iteratively solving the Oseen equations;

$$\begin{aligned} -\frac{1}{Re}\Delta\mathbf{u}^{(n+1)} + \widetilde{\mathbf{u}}^{(n)}\cdot\nabla\mathbf{u}^{(n+1)} + \nabla p^{(n+1)} = 0 \quad &\text{in } \Omega, \\ \nabla\cdot\mathbf{u}^{(n+1)} = 0 \quad &\text{in } \Omega, \end{aligned}$$

with the watertight boundary condition. A relaxation parameter is introduced for the convection term such that

$$\widetilde{\mathbf{u}}^{(n)} = \tau\mathbf{u}^{(n)} + (1-\tau)\mathbf{u}^{(n-1)}, \quad 0 < \tau \le 1.$$

Figures 6 and 7 represent convergence properties of the Oseen iteration depending on the relaxation parameter τ for the HSD and *upwind* HSD, respectively. For the HSD numerical experiments are presented up to $Re = 7000$, and it is observed that the HSD fails to converge for $Re \ge 10{,}000$ even on a relatively fine 40×40 mesh with a small τ. However, if the *upwind* HSD is adopted, the Oseen iteration performs quite stably up to $Re = 20{,}000$ if the relaxation parameter τ is properly chosen.

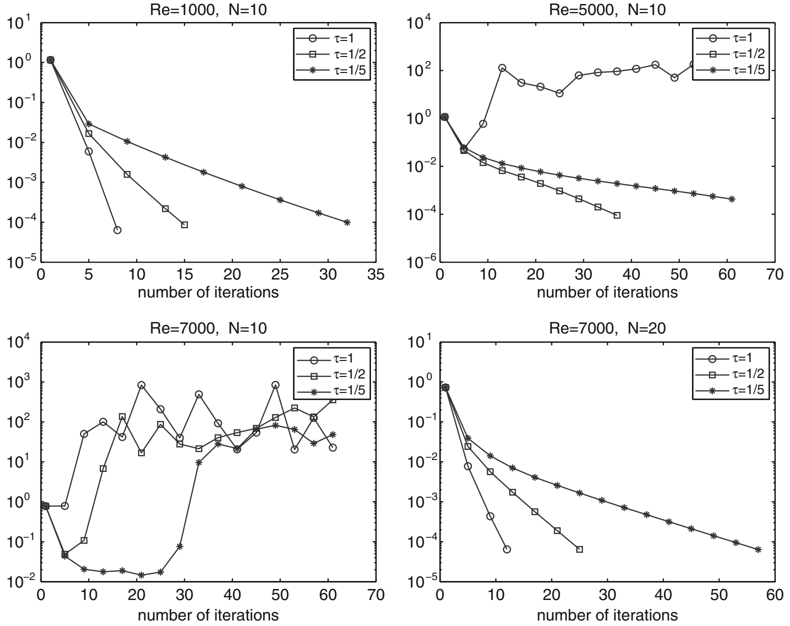


Fig. 6 Convergence history of the Oseen iteration for various choice of relaxation parameter τ with the HSD (Example 3). $Re = 1000$–7000

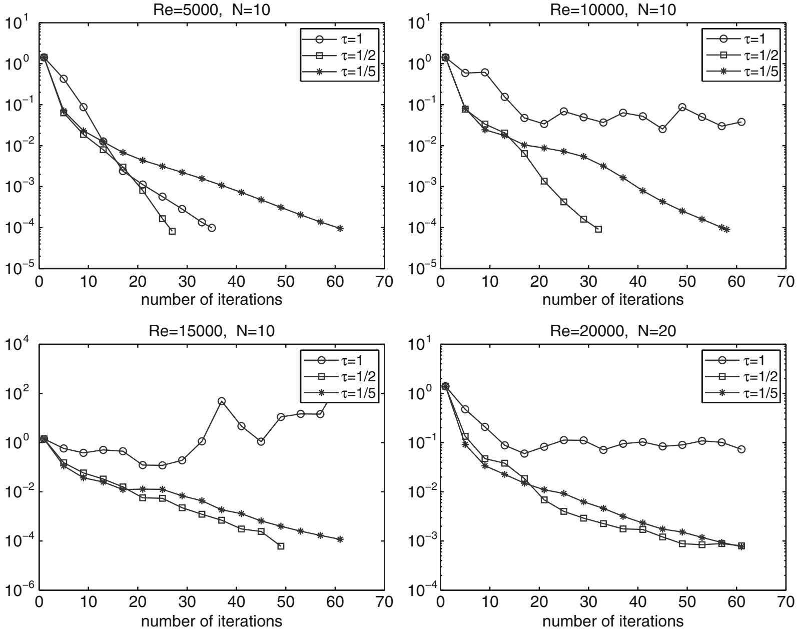


Fig. 7 Convergence history of the Oseen iteration for various choice of relaxation parameter τ for the upwind HSD (Example 3). $Re = 5000$–$20{,}000$

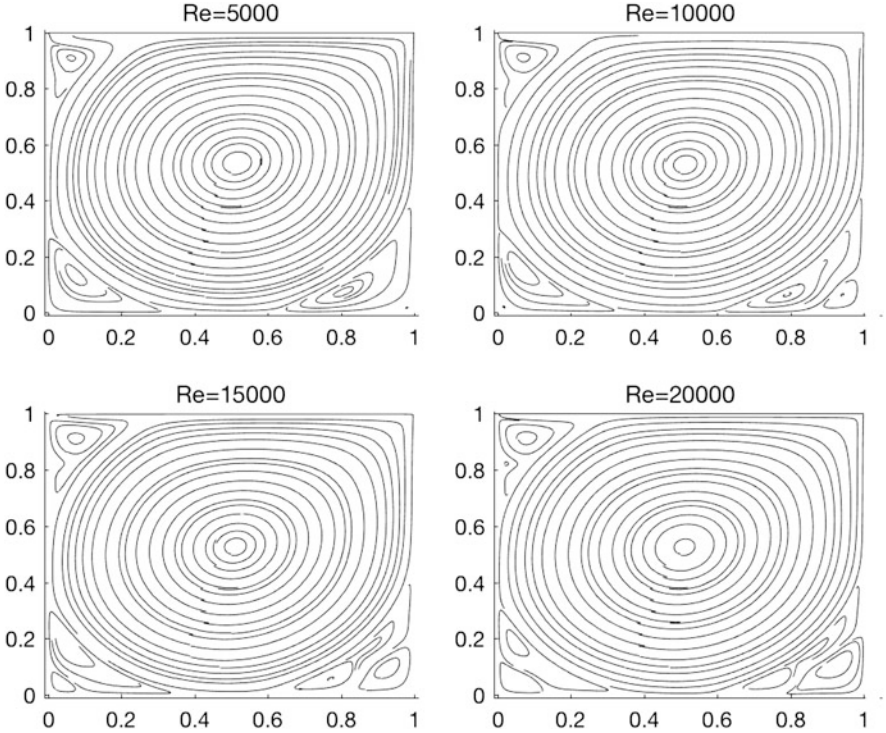


Fig. 8 Solutions of Example 3 with $Re = 5000$–20,000

If τ is small the Oseen iteration converges monotonically, but slowly. If $\tau \approx 1$ the convergence can be faster whenever the iteration is convergent; however, the iteration becomes divergent as the Reynolds number becomes larger for a fixed mesh size. As shown in the figures the *upwind* HSD performs more reliably, especially for large Reynolds number flows. In view of Fig. 7 we use the *upwind* HSD with the choices of the relaxation parameters, $\tau = \frac{1}{2}$ when $Re = 5000$–15,000 and $\tau = \frac{1}{5}$ when $Re = 20{,}000$ for our numerical experiments in Figs. 8, 9, 10 and 11. The stopping criterion is $\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\|_{L_2^h(\Omega)} < 10^{-3}$. We use the 40×40 graded mesh, which corresponds roughly to a 160×160 grid mesh for the usual FDM. As shown in Table 2 the total vorticity $\omega = \int_\Omega \nabla \times \mathbf{u}\, \mathrm{d}\mathbf{x}$ is computed almost exactly up to a machine precision. The volumetric flow rates (see [8, 23] for details) are zero for the exact solution. The approximated volumetric flow rates are computed almost exactly as well. The primary vorticity is measured and compared with the 400×400 grid case in [8].

Figure 8 shows the fluid motions with $Re = 5000$–20,000. As the Reynolds number becomes higher the primary vortex tends to shift to the center more, and more and larger secondary vortices start to form at corners. Figure 9 shows the

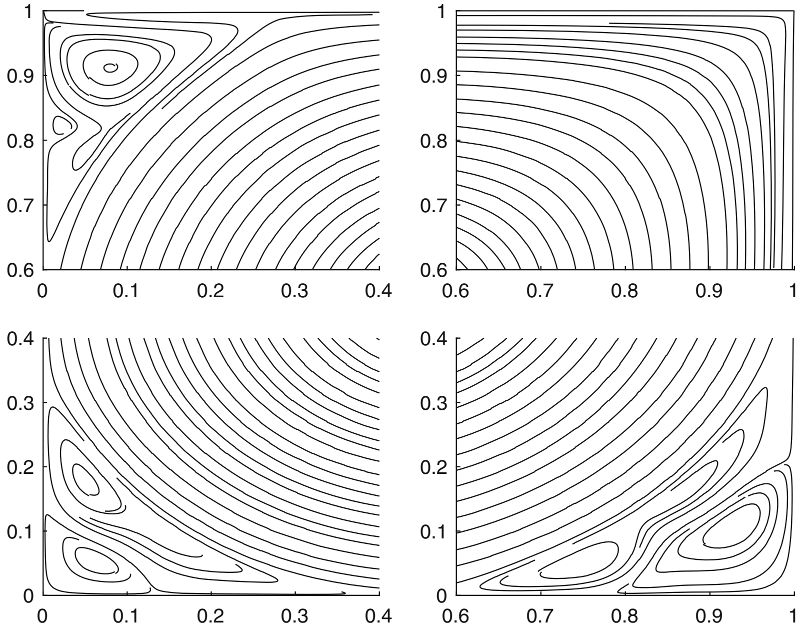


Fig. 9 Secondary vortices with $Re = 20{,}000$

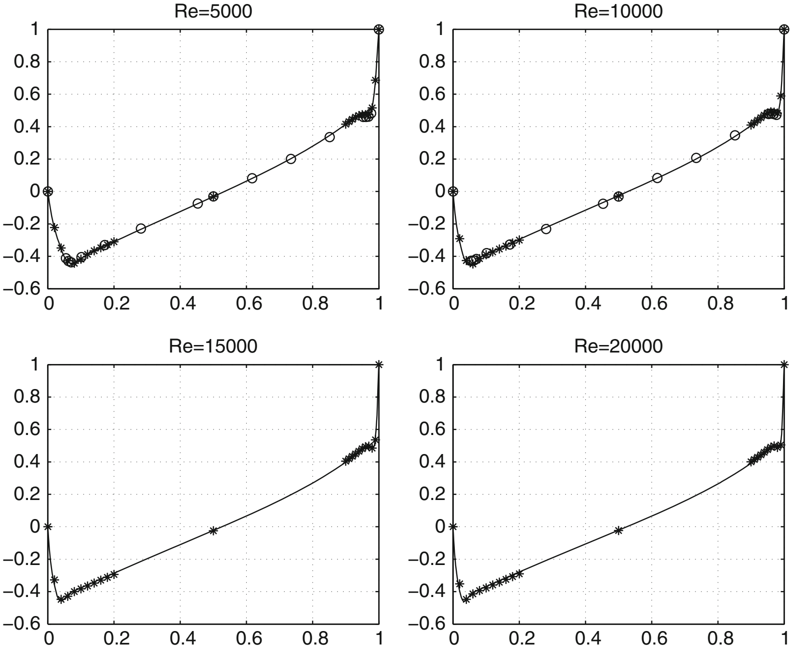


Fig. 10 Centerline velocity profiles of u with $\mathbf{u} = (u, v)$. The discrete circles represent the reference data from [11] and the *-marked data from [8] (Example 3)

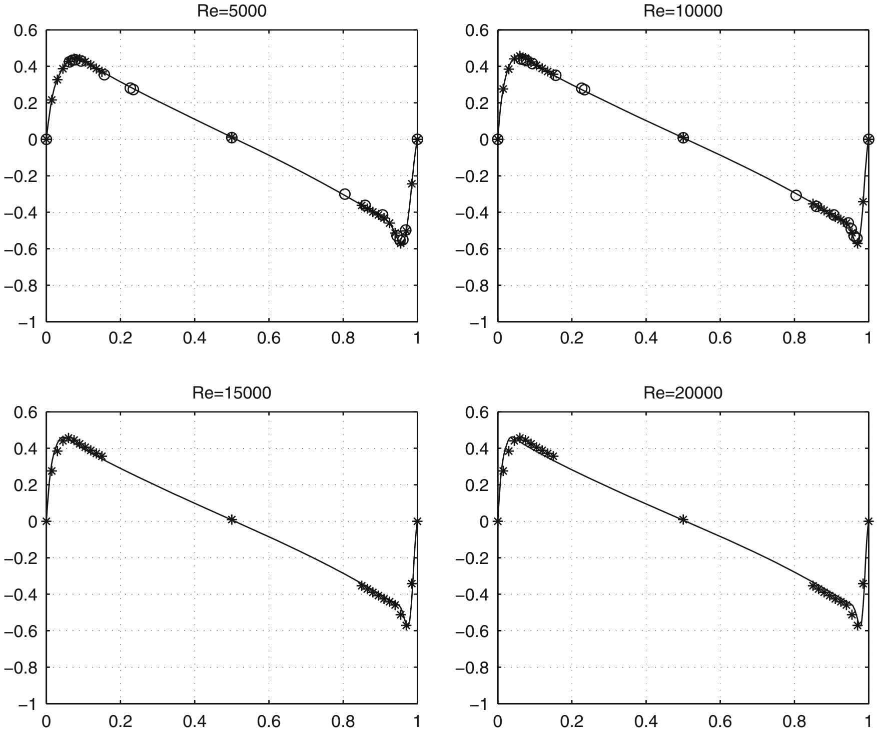


Fig. 11 Centerline velocity profiles of v with $\mathbf{u} = (u, v)$. The discrete circles represent the reference data from [11] and the *-marked data from [8] (Example 3)

Table 2 Total vorticity, volumetric flow rates and the primary vorticity for the *upwind* HSD with the graded 40 × 40 mesh for Example 3

Re	1000	5000	10,000	15,000	20,000
ω (vorticity)	1.0000	1.0000	1.0000	1.0000	1.0000
$\int u_c ds$	−8.36e−12	−3.46e−10	−1.94e−10	1.99e−10	1.12e−11
$\int v_c ds$	−1.49e−10	−1.52e−10	3.35e−10	−2.61e−10	6.07e−10
Primary vorticity: HSD	−2.057231	−1.920731	−1.847845	−1.789383	−1.741744
Primary vorticity: FDM in [8]	−2.062761	−1.909448	−1.853444	−1.809697	−1.769594

The last row is the primary vorticity with the FDM on a 400 × 400 uniform mesh in [8]

detailed fluid motions at four corners of the domain with $Re = 20{,}000$. These coincide with existing results by using other methods [8, 11].

In order to test the accuracy of our algorithms, some velocity components are compared with well-known benchmark results, and they are shown in Figs. 10 and 11, which show that the results are in good agreement. The solid line represents numerical results by our algorithm, the discrete circles represent those in [11] and the *-marked data are obtained from [8]. These graphs include computed u-velocity along the vertical center line (u_c) and v-velocity along the horizontal center line (v_c). In [11] numerical results are obtained by using the stream function-vorticity formulation and the third order FDM on the 128×128 mesh for $Re \leq 3200$ and on the 257×257 grid for $Re \geq 5000$, respectively. In [8] they use the stream function–vorticity formulation with the 2nd order central FD approximations and numerical results on the 400×400, 512×512, 600×600 uniform grids and its Richardson extrapolation are presented.

5 Concluding Remarks

In this paper an HSD and a *upwind* HSD methods are introduced, and the *inf-sup* condition is proved. For the static condensation of the methods we refer to [14].

In view of analysis in [14, 17] and the *inf-sup* condition in this paper convergence analysis can be done in discrete energy norm for $\mathbf{u}$ and discrete L_2-norm for p for Stokes flows. For a complete analysis of the HSD for the Navier-Stokes equations ellipticity of the convection term is necessary, and it remains as a future work. Some of the experimental rates of convergence in Sect. 4 are inconsistent with our intuition, and a rigorous convergence analysis will be also a subject of future research.

Numerical experiments for the steady-state Navier Stokes equations suggest that the (*upwind*) HSD is very effective for flows with a wide range of Reynolds numbers. For flow problems with very high Reynolds number and low regularity solution it is observed that the combination of the *upwind* HSD and a geometric mesh produces more reliable numerical solutions.

Acknowledgements Youngmok Jeon was supported by National Research Foundation of Korea (NRF-2015R1D1A1A09057935). Dongwoo Sheen was supported in part by National Research Foundation of Korea (NRF-2017R1A2B3012506 and NRF-2015M3C4A7065662).

References

1. Amrouche, C., Rodríguez-Bellido, M.: Stationary Stokes, Oseen and Navier–Stokes equations with singular data. Arch. Ration. Mech. Anal. **199**(2), 597–651 (2011)
2. Balan, A., May, G., Schöberl, J.: A stable high-order spectral difference method for hyperbolic conservation laws on triangular elements. J. Comput. Phys. **231**(5), 2359–2375 (2012)
3. Bernardi, C., Maday, Y.: A collocation method over staggered grids for the Stokes problem. Int. J. Numer. Methods Fluids **8**(5), 537–557 (1988)
4. Bernardi, C., Maday, Y.: Spectral methods. In: Handbook of Numerical Analysis, vol. 5, pp. 209–485. Elsevier, Amsterdam (1997)
5. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Berlin (2013)

6. Carpenter, M.H., Fisher, T.C., Nielsen, E.J., Frankel, S.H.: Entropy stable spectral collocation schemes for the Navier–Stokes equations: discontinuous interfaces. SIAM J. Sci. Comput. **36**(5), 835–867 (2014)
7. Cockburn, B., Gopalakrishnan, J., Lazarov, R.: Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal. **47**(2), 1319–1365 (2009)
8. Erturk, E., Corke, T.C., Gökçöl, C.: Numerical solutions of 2-D steady incompressible driven cavity flow at high Reynolds numbers. Int. J. Numer. Methods Fluids **48**(7), 747–774 (2005)
9. Ferziger, J.H., Peric, M.: Computational Methods for Fluid Dynamics. Springer, Berlin (2002)
10. Fisher, T.C., Carpenter, M.H.: High-order entropy stable finite difference schemes for nonlinear conservation laws: finite domains. J. Comput. Phys. **252**, 518–557 (2013)
11. Ghia, U., Ghia, K.N., Shin, C.T.: High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. J. Comput. Phys. **48**(3), 387–411 (1982)
12. Girault, V., Raviart P.-A.: Finite Element Methods for Navier-Stokes Equations, Springer. Springer Series in Computational Mathematics, vol. 5. Springer, Berlin (1986)
13. Hanley, P.: A strategy for the efficient simulation of viscous compressible flows using a multi-domain pseudospectral method. J. Comput. Phys. **108**(1), 153–158 (1993)
14. Jeon, Y.: Hybrid difference methods for PDEs. J. Sci. Comput. **64**(2), 508–521 (2015)
15. Jeon, Y., Park, E.-J.: New locally conservative finite element methods on a rectangular mesh. Numer. Math. **123**(1), 97–119 (2013)
16. Jeon, Y., Tran, M.L.: The upwind hybrid difference methods for a convection diffusion equation. Appl. Numer. Math. (2018). https://doi.org/10.1016/j.apnum.2017.12.002
17. Jeon Y., Park E.-J., Shin, D.-W.: Hybrid spectral difference methods for the Poisson equation. Comput. Methods Appl. Math. **17**, 253–267 (2017)
18. Kim, H.: Existence and regularity of very weak solutions of the stationary Navier–Stokes equations. Arch. Ration. Mech. Anal. **193**(1), 117–152 (2009)
19. Kopriva, D.A.: A conservative staggered-grid Chebyshev multidomain method for compressible flows. II. A semi-structured method. J. Comput. Phys. **128** (2), 475–488 (1996)
20. Kopriva, D.A.: A staggered-grid multidomain spectral method for the compressible Navier–Stokes equations. J. Comput. Phys. **143**(1), 125–158 (1998)
21. Kopriva, D.A., Kolias, J.H.: A conservative staggered-grid Chebyshev multidomain method for compressible flow. DTIC Document (1995)
22. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems, vol. 31. Cambridge University Press, Cambridge (2002)
23. Lim, R., Sheen, D.: Nonconforming finite element method applied to the driven cavity problem. Commun. Comput. Phys. **21**(4), 1021–1038 (2017)
24. Liu, Y., Vinokur, M., Wang, Z.J.: Spectral difference method for unstructured grids I: basic formulation. J. Comput. Phys. **216**(2), 780–801 (2006)
25. Rhie, C.M., Chow, W.L.: Numerical study of the turbulent flow past an airfoil with trailing edge separation. AIAA J. **21**(11), 1525–1532 (1983)
26. Van den Abeele, K., Lacor, C., Wang, Z.J.: On the stability and accuracy of the spectral difference method. J. Sci. Comput. **37**(2), 162–188 (2008)
27. Versteeg, H.K., Malalasekera, W.: An introduction to computational fluid dynamics: the finite volume method. Pearson Education, Harlow (2007)
28. Wang, Z.J., Liu, Y., May, G., Jameson, A.: Spectral difference method for unstructured grids II: extension to the Euler equations. J. Sci. Comput. **32**(1), 45–71 (2007)
29. Zang, T.A., Hussaini, M.Y.: Mixed spectral/finite difference approximations for slightly viscous flows. In: Seventh International Conference on Numerical Methods in Fluid Dynamics, pp. 461–466. Springer, Berlin (1981)

On Nyström and Product Integration Methods for Fredholm Integral Equations

Peter Junghanns, Giuseppe Mastroianni, and Incoronata Notarangelo

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The aim of this paper is to combine classical ideas for the theoretical investigation of the Nyström method for second kind Fredholm integral equations with recent results on polynomial approximation in weighted spaces of continuous functions on bounded and unbounded intervals, where also zeros of polynomials w.r.t. exponential weights are used.

1 Introduction

There exists a huge literature on numerical methods for Fredholm integral equations of second kind,

$$f(x) - \int_I K(x,y) f(y)\, dy = g(x)\,, \quad x \in I\,, \tag{1}$$

where I is a bounded or unbounded interval. A very famous method is the Nyström method which is based on an appropriate quadrature rule applied to the integral and on considering (1) in the space of (bounded) continuous functions on I. Such quadrature rules can be of different type. In the present paper we will focus on Gaussian rules and product integration rules based on zeros of orthogonal polynomials. The aim of this paper is to combine classical ideas for the theoretical

P. Junghanns (✉)
Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany
e-mail: peter.junghanns@mathematik.tu-chemnitz.de

G. Mastroianni · I. Notarangelo
Department of Mathematics, Computer Sciences and Economics, University of Basilicata, Potenza, Italy
e-mail: giuseppe.mastroianni@unibas.it; incoronata.notarangelo@unibas.it

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_29

investigation of the Nyström method, in particular the results of Sloan [34, 35], with recent results on polynomial approximation in weighted spaces of continuous functions on bounded and unbounded intervals, where also zeros of polynomials w.r.t. exponential weights come into the play (cf. [13, 27]). Note that the Nyström method, in general, is based on the application of a quadrature rule to the integral part of the operator. Here we focus on quadrature rules of interpolatory type, which are constructed with the help of zeros of orthogonal polynomials, i.e., which are of Gaussian type. Of course, there exists a lot of other possibilities. As an example, let us only mention the paper [12], where quasi-Monte Carlo rules are applied to the case of kernel functions of the form $K(x, y) = h(x - y)$.

Considering weighted spaces of continuous functions is motivated by the fact, that in many practical examples for the unknown function it is known that it has some kind of singularities at the endpoints of the integration interval. Moreover, the kernel function of the integral operator can have endpoint singularities in both variables. For recent attempts to combine the idea of the Nyström method with weighted polynomial approximation, we refer the reader to [11, 24, 30].

The present paper is organized as follows. In Sect. 2 we present the notion of collectively compact and strongly convergent operator sequences and the classical result on the application of this concept for proving stability and convergence of approximation methods for operator equations. After formulating the results of Sloan from the 1980s on the application of quadrature methods to Fredholm integral equations of the second kind, we show how these results can be generalized by using weighted spaces of continuous functions, where we prefer a unified approach for both bounded and unbounded integration intervals (see Definition 1 and Lemma 2). In Sect. 3 we prove a general convergence result for the classical Nyström method (see Corollary 2), where "classical" means that usual quadrature rules are used for the discretization of the integral operator, not product integration rules. In Sects. 3.1 and 3.2, this result is applied to the interval $(-1, 1)$ involving Jacobi weights and to the half line $(0, \infty)$ involving exponential weights, respectively. Finally, Sect. 4 contains the most important results of the paper and is devoted to the application of product integration rules in the Nyström method, where again the Jacobi weight case and the exponential weight case are considered separately. In particular, in both cases we show how one can use the respective $\mathbf{L}\log^+\mathbf{L}$ function classes, in order to weaken the conditions on the kernel function of the integral operator (see Propositions 5 and 6).

2 Basic Facts

In the sequel, by c we will denote real positive constants, which can assume different values at different places, and by $c \neq c(a, b, \ldots)$ we will explain, that c does not depend on $a, b, \ldots$ If α and β are positive real numbers depending on certain parameters $a, b, \ldots$, then by $\alpha \sim_{a,b,\ldots} \beta$ is meant that there is a positive constant $c \neq c(a, b, \ldots)$ such that $c^{-1}\alpha \leq \beta \leq c\alpha$.

We say, that a sequence $(\mathscr{K}_n)_{n=1}^{\infty}$ of linear operators $\mathscr{K}_n : \mathbf{X} \longrightarrow \mathbf{X}$ in the Banach space $\mathbf{X}$ is collectively compact, if the set $\{\mathscr{K}_n f : f \in \mathbf{X}, \|f\| \le 1, n \in \mathbb{N}\}$ is relatively compact in $\mathbf{X}$, i.e., the closure of this set is compact. The concept of collectively compact sets of operators goes back to Anselone and Palmer [1, 2, 4–6].

For the following proposition, see, for example, [3], or Sections 10.3 and 10.4 in [15, 16], or [17], or Section 4.1 in [7].

Proposition 1 *Let $\mathbf{X}$ be a Banach space and $\mathscr{K} : \mathbf{X} \longrightarrow \mathbf{X}$, $\mathscr{K}_n : \mathbf{X} \longrightarrow \mathbf{X}$, $n \in \mathbb{N}$ be given linear operators with $\lim\limits_{n\to\infty} \|\mathscr{K}_n f - \mathscr{K} f\| = 0$ for all $f \in \mathbf{X}$ (i.e., the operators $\mathscr{K}_n$ converge strongly to $\mathscr{K}$ in $\mathbf{X}$). For $g \in \mathbf{X}$, consider the operator equations*

$$(\mathscr{I} - \mathscr{K})f = g \tag{2}$$

where $\mathscr{I}$ is the identity operator in $\mathbf{X}$, and

$$(\mathscr{I} - \mathscr{K}_n)f_n = g\,. \tag{3}$$

If the sequence $(\mathscr{K}_n)_{n=1}^{\infty}$ is collectively compact and if $\dim\ker(\mathscr{I} - \mathscr{K}) = 0$, then, for all sufficiently large n Eq. (3) has a unique solution $f_n^* \in \mathbf{X}$, where

$$\left\| f_n^* - f^* \right\| \le c \left\| \mathscr{K}_n f^* - \mathscr{K} f^* \right\| , \quad c \ne c(n, g, f^*)\,, \tag{4}$$

and $f^* \in \mathbf{X}$ is the unique solution of (2).

Let us consider the situation that $\mathbf{X}$ is equal to the space of continuous functions $\mathbf{C}(I)$, where $I = (I, d)$ is one of the compact metric spaces $I = [-1, 1]\,, I = [0, \infty]$, or $I = [-\infty, \infty]$, the distance function of which can be given, for example, by $d(x, y) = |a(x) - a(y)|$ or $d(x, y) = \dfrac{|a(x) - a(y)|}{1 + |a(x) - a(y)|}$ with $a(x) = \arctan(x)$. As usual, the norm in $\mathbf{C}(I)$ is defined by $\|f\|_\infty := \max\{|f(x)| : x \in I\}$. As operators $\mathscr{K}$ and $\mathscr{K}_n$ we take

$$(\mathscr{K} f)(x) = \int_I K(x, y) f(y)\, dy \quad \text{as well as} \quad (\mathscr{K}_n f)(x) = \sum_{k=1}^{k_n} \Lambda_{nk}(x) f(x_{nk})\,, \tag{5}$$

where the Λ_{nk}'s are certain quadrature weights and we assume $x_{nk} \in I$ $(k = 1, \ldots, k_n)\,, x_{n1} < x_{n2} < \ldots < x_{n\,k_n}$, as well as

(K1) $\displaystyle\int_I |K(x, y)|\, dy < \infty$, i.e., $K(x, .) \in \mathbf{L}^1(I)$ for all $x \in I$,

(K2) $\displaystyle\lim_{x\to x_0} \|K(x, .) - K(x_0, .)\|_{\mathbf{L}^1(I)} = \lim_{x\to x_0} \int_I |K(x, y) - K(x_0, y)|\, dy = 0$ for all $x_0 \in I$, (K3) $\lim\limits_{n\to\infty}\sum\limits_{k=1}^{k_n}\Lambda_{nk}(x)f(x_{nk}) = \int_I K(x,y)f(y)\,dy$ for all $x \in I$ and all $f \in \mathbf{C}(I)$,

(K4) $\lim\limits_{x\to x_0}\sup\left\{\sum\limits_{k=1}^{k_n}|\Lambda_{nk}(x)-\Lambda_{nk}(x_0)| : n\in\mathbb{N}\right\} = 0$ for all $x_0 \in I$.

Note that conditions (K1) and (K2) are necessary and sufficient for the operator $\mathscr{K} : \mathbf{C}(I) \longrightarrow \mathbf{C}(I)$ being a compact one, which is a consequence of the Arzela-Ascoli Theorem characterizing the relatively compact subsets of $\mathbf{C}(I)$. Moreover, the following lemma is true and crucial for our further considerations (see [34, Section 2, Lemma] and [35, Section 3, Theorem 1]).

Lemma 1 *Suppose that conditions* (K1) *and* (K2) *are fulfilled. The operators* $\mathscr{K}_n : \mathbf{C}(I) \longrightarrow \mathbf{C}(I)$, $n \in \mathbb{N}$, *defined in* (5)*, form a collectively compact sequence, which converges strongly to* $\mathscr{K}$*, if and only if* (K3) *and* (K4) *are satisfied.*

Remark 1 For example, in case $I = [0,\infty]$, conditions (K1)–(K4) can be written equivalently as (cf. [35, (3.1)–(3.3)])

(K1') $K(x,.) \in \mathbf{L}^1(0,\infty)\ \forall\, x \in [0,\infty)$,
(K2') $\lim\limits_{x\to x_0}\|K(x,.)-K(x_0,.)\|_{\mathbf{L}^1(0,\infty)} = 0\ \forall\, x_0 \in [0,\infty)$,

(K3') $\lim\limits_{x\to\infty}\sup\left\{\int_0^\infty \left|K(x',y)-K(x,y)\right|\,dy : x' > x\right\} = 0$,

(K4') $\lim\limits_{n\to\infty}\sum\limits_{k=1}^{k_n}\Lambda_{nk}(x)f(x_{nk}) = \int_0^\infty K(x,y)f(y)\,dy\ \ \forall\, x \in [0,\infty)$ and $\forall f \in \mathbf{C}[0,\infty]$,

(K5') $\lim\limits_{x\to x_0}\sup\left\{\sum\limits_{k=1}^{k_n}|\Lambda_{nk}(x)-\Lambda_{nk}(x_0)| : n\in\mathbb{N}\right\} = 0$ for all $x_0 \in [0,\infty)$,

(K6') $\lim\limits_{x\to\infty}\sup\limits_{x'>x}\sup\left\{\sum\limits_{k=1}^{k_n}|\Lambda_{nk}(x)-\Lambda_{nk}(x_0)| : n\in\mathbb{N}\right\} = 0$.

Now, we assume that the kernel function $K(x,y)$ and the quadrature weights $\Lambda_{nk}(x)$ in (5) are represented in the form

$$K(x,y) = H(x,y)S(x,y) \quad \text{and} \quad \Lambda_{nk}(x) = \lambda_{nk}^F(H(x,.))S(x,x_{nk})\,, \tag{6}$$

respectively, and consider the conditions (H1)–(H3) below. For this, we need the following notions.

Definition 1 Let $I_0 = (-1,1)$, $I_0 = (0,\infty)$, or $I_0 = (-\infty,\infty)$, and let v be a positive weight function on I_0, where $v : I \longrightarrow [0,\infty)$ is assumed to be continuous and having the property that $p(x)v(x)$ is continuous in I for all polynomials $p(x)$. By $\widetilde{\mathbf{C}}_v = \widetilde{\mathbf{C}}_v(I_0)$ we denote the Banach space of all functions $f : I_0 \longrightarrow \mathbb{C}$, for which $vf : I_0 \longrightarrow \mathbb{C}$ can be extended to a continuous function on the whole interval I, where the norm on $\widetilde{\mathbf{C}}_v$ is given by $\|g\|_{\widetilde{\mathbf{C}}_v} = \|g\|_{v,\infty} := \max\{|v(x)g(x)| : x \in I\}$.

Moreover, let $\mathbf{C}_v \subset \widetilde{\mathbf{C}}_v$ be the closure (w.r.t. the $\widetilde{\mathbf{C}}_v$-norm) of the set $\mathbf{P}$ of all algebraic polynomials.

Now, we formulate the above mentioned conditions.

(H1) The λ_{nk}^F's, $k = 1, \ldots, k_n$, $n \in \mathbb{N}$, are linear and bounded functionals on a Banach space $\mathbf{X}_0$ continuously imbedded in $\mathbf{L}^1_{v^{-1}}(I)$, where $\mathbf{L}_{v^{-1}}(I) = \{f : v^{-1}f \in \mathbf{L}^1\}$ with $\|f\|_{\mathbf{L}^1_{v^{-1}}} = \left\|v^{-1}f\right\|_1 := \left\|v^{-1}f\right\|_{\mathbf{L}^1}$.

(H2) For all $x \in I$, $H(x, .) \in \mathbf{X}_0$ and $S(x, .) \in \mathbf{C}_v$, and for all $x_0 \in I$,

$$\lim_{x\to x_0} \|H(x, .) - H(x_0, .)\|_{\mathbf{X}_0} = 0 .$$

(H3) It holds $\lim\limits_{n\to\infty} \sum\limits_{k=1}^{k_n} \lambda_{nk}^F(f)\, g(x_{nk}) = \int_I f(y)g(y)\, dy$ for all $f \in \mathbf{X}_0$ and all $g \in \mathbf{C}_v(I_0)$.

In case of $v(x) \equiv 1$ and $I = [-1, 1]$, the following lemma is proved in [34, Section 3, Theorem 2].

Lemma 2 *Assume that $K(x,y)$ and $\Lambda_{nk}(x)$ in* (5) *are of the form* (6)*, where the conditions* (H1)–(H3) *are fulfilled and where $S(x, y)v(y)$ is continuous on I^2. Then, conditions* (K1)–(K4) *are satisfied.*

Proof Condition (K1) follows from

$$\int_I |K(x, y)|\, dy \le \|H(x, .)\|_{\mathbf{L}^1_{v^{-1}}} \|S(x, .)\|_{v,\infty} \le c\, \|H(x, .)\|_{\mathbf{X}_0} \|S(x, .)\|_{v,\infty}$$

and condition (H2). Moreover,

$$\begin{aligned} &\|K(x, .) - K(x_0, .)\|_{\mathbf{L}^1} \\ &\le \|H(x, .) - H(x_0, .)\|_{\mathbf{L}^1_{v^{-1}}} \|S(x, .)v\|_\infty + \|H(x_0, .)\|_{\mathbf{L}^1_{v^{-1}}} \|S(x, .)v - S(x_0, .)v\|_\infty \\ &\le \|H(x, .) - H(x_0, .)\|_{\mathbf{X}_0} \|S(x, .)v\|_\infty + \|H(x_0, .)\|_{\mathbf{X}_0} \|S(x, .)v - S(x_0, .)v\|_\infty \longrightarrow 0 \end{aligned}$$

if $x \to x_0 \in [-1, 1]$ because of (H2) and the (uniform) continuity of $S(x, y)v(y)$ on I^2. Hence, (K2) is also satisfied. Using (6), (H2), and (H3), we get, for $f \in \mathbf{C}(I)$,

$$\sum_{k=1}^{k_n} \Lambda_{nk}(x)f(x_{nk}) = \sum_{k=1}^{k_n} \lambda_{nk}^F(H(x, .))S(x, x_{nk})f(x_{nk}) \longrightarrow \int_I H(x, y)S(x, y)f(y)\, dy ,$$

since together with $S(x,.) \in \mathbf{C}_v$ also $S(x,.)f$ belongs to $\mathbf{C}_v$. This shows the validity of (K3). It remains to consider (K4). For this, define $\mathscr{G}_n : \mathbf{X}_0 \longrightarrow \mathbf{C}_v^*, f \mapsto \mathscr{G}_n f$ with

$$(\mathscr{G}_n f)(g) = \sum_{k=1}^{k_n} \lambda_{nk}^F(f) g(x_{nk}) \quad \text{for all} \quad g \in \mathbf{C}_v \,.$$

Indeed, $\mathscr{G}_n f \in \mathbf{C}_v^*$, since $|(\mathscr{G}_n f)(g)| \le \displaystyle\sum_{k=1}^{k_n} \frac{|\lambda_{nk}^F(f)|}{v(x_{nk})} \|g\|_{v,\infty}$. Moreover, it is easily seen that

$$\|\mathscr{G}_n f\|_{\mathbf{C}_v^*} = \sum_{k=1}^{k_n} \frac{|\lambda_{nk}^F(f)|}{v(x_{nk})} \,.$$

If we fix $f \in \mathbf{X}_0$, then $\sup\{|(\mathscr{G}_n f)(g)| : n \in \mathbb{N}\} < \infty$ for every $g \in \mathbf{C}_v$, due to (H3). Consequently, in virtue of the principle of uniform boundedness,

$$\sup\left\{\|\mathscr{G}_n f\|_{\mathbf{C}_v^*} : n \in \mathbb{N}\right\} < \infty \quad \text{for every} \quad f \in \mathbf{X}_0 \,. \tag{7}$$

Taking into account $\lambda_{nk}^F \in \mathbf{X}_0^*$ and

$$\|\mathscr{G}_n f\|_{\mathbf{C}_v^*} = \sum_{k=1}^{k_n} \frac{|\lambda_{nk}^F(f)|}{v(x_{nk})} \le \sum_{k=1}^{k_n} \frac{\|\lambda_{nk}^F\|_{\mathbf{X}_0^*}}{v(x_{nk})} \|f\|_{\mathbf{X}_0} \,, \tag{8}$$

we see that $\mathscr{G}_n$ belongs to $\mathscr{L}(\mathbf{X}_0, \mathbf{C}_v^*)$. Again by the principle of uniform boundedness and by (7), we obtain $c_0 := \sup\left\{\|\mathscr{G}_n\|_{\mathbf{X}_0 \to \mathbf{C}_v^*} : n \in \mathbb{N}\right\} < \infty$. This implies, together with (8),

$$\sum_{k=1}^{k_n} \frac{|\lambda_{nk}^F(f)|}{v(x_{nk})} \le c_0 \|f\|_{\mathbf{X}_0} \quad \forall f \in \mathbf{X}_0 \,.$$

Hence,

$$\begin{aligned}
&\sum_{k=1}^{k_n} \left|\Lambda_{nk}(x) - \Lambda_{nk}(x_0)\right| \\
&\quad = \sum_{k=1}^{k_n} \Big| \left[\lambda_{nk}^F(H(x,.)) - \lambda_{nk}^F(H(x_0,.))\right] S(x, x_{nk}) \\
&\qquad\qquad\qquad + \lambda_{nk}^F(H(x_0,.)) \left[S(x, x_{nk}) - S(x_0, x_{nk})\right] \Big|
\end{aligned}$$

$$\leq \sum_{k=1}^{k_n} \frac{\left|\lambda_{nk}^F(H(x,.)-H(x_0,.))\right|}{v(x_{nk})} \; \|S(x,.)v\|_\infty$$

$$+\sum_{k=1}^{k_n} \frac{\left|\lambda_{nk}^F(H(x_0,.))\right|}{v(x_{nk})} \; \|S(x,.)v - S(x_0,.)v\|_\infty$$

$$\leq c_0 \left[\|H(x,.)-H(x_0,.)\|_{\mathbf{X}_0} \, \|S(x,.)v\|_\infty + \|H(x_0,.)\|_{\mathbf{X}_0} \, \|S(x,.)v - S(x_0,v)\|_\infty\right] ,$$

and (K4) follows by (H2) and the continuity of $S(x,y)v(y)$ on I^2. □

3 The Classical Nyström Method

Let u be a positive weight function and w, w_1 be weight functions on I_0, where $u : I \longrightarrow [0,\infty)$ is assumed to be continuous. For example, all these three weight functions can be Jacobi weights (see Sect. 3.1) or weights of exponential type (see Sect. 3.2). Consider a Fredholm integral equation of the second kind

$$\widetilde{f}(x) - \int_I \widetilde{K}(x,y)w(y)\widetilde{f}(y)\,dy = \widetilde{g}(x)\,, \quad x \in I_0\,, \tag{9}$$

where $\widetilde{g} \in \widetilde{\mathbf{C}}_u$ and $\widetilde{K} : I^2 \longrightarrow \mathbb{C}$ are given functions and $\widetilde{f} \in \widetilde{\mathbf{C}}_u$ is looked for. Using a set of nodes $x_{nk} \in I_0$ satisfying

$$x_{n1} < x_{n2} < \ldots < x_{n,k_n} \tag{10}$$

and a quadrature rule

$$\int_I \widetilde{f}(x)w(x)\,dx \sim \sum_{k=1}^{k_n} \lambda_{nk}\widetilde{f}(x_{nk})\,, \tag{11}$$

we look for an approximate solution $\widetilde{f}_n(x)$ for Eq. (9) by solving

$$\widetilde{f}_n(x) - \sum_{k=1}^{k_n} \lambda_{nk}\widetilde{K}(x,x_{nk})\widetilde{f}_n(x_{nk}) = \widetilde{g}(x)\,. \tag{12}$$

If we define $f(x) := u(x)\widetilde{f}(x)$, $g(x) := u(x)\widetilde{g}(x)$,

$$K(x,y) = \frac{u(x)\widetilde{K}(x,y)w(y)}{u(y)}\,, \tag{13}$$

and

$$\Lambda_{nk}(x) = \frac{\lambda_{nk} u(x)\widetilde{K}(x, x_{nk})}{u(x_{nk})} =: \lambda_{nk} K_1(x, x_{nk}) \,, \tag{14}$$

then (9) considered in $\widetilde{\mathbf{C}}_u(I_0)$ together with (12) is equivalent to (2) considered in $\mathbf{C}(I)$ together with (3), where $\mathscr{K}$ and $\mathscr{K}_n$ are given by (5).

Recall, that the function (cf. (12))

$$\widetilde{f}_n(x) = \sum_{k=1}^{k_n} \lambda_{nk}\widetilde{K}(x, x_{nk})\widetilde{f}_n(x_{nk}) + \widetilde{g}(x)$$

is called Nyström interpolant at the nodes x_{nk}. For its construction, one needs the values $\xi_{nk} = \widetilde{f}_n(x_{nk})$, which can be computed by considering (12) for $x = x_{nj}$, $j = 1, \ldots, k_n$ and solving the system of linear equations

$$\xi_{nj} - \sum_{k=1}^{k_n} \lambda_{nk}\widetilde{K}(x_{nj}, x_{nk})\xi_{nk} = \widetilde{g}(x_{nj}) \,, \quad j = 1, \ldots, k_n \,.$$

Note, that the convergence of the Nyström interpolant to the solution of the original integral equation is the main feature of the Nyström method. For that reason, the natural spaces, in which the Nyström method together with the integral equation should be considered, are spaces of continuous functions. Moreover, the natural class of integral equations, to which the Nyström method together with the concept of collectively compact and strongly convergent operator sequences can be applied, is the class of second kind Fredholm integral equations, since collective compactness and strong convergence imply the compactness of the limit operator.

Nevertheless, there were developed modifications of the Nyström method applicable to integral equations with noncompact integral operators (see, for example, [9, 10, 21]).

We formulate the conditions

(A) $K_0(x, y) := u(x)\widetilde{K}(x, y)w_1(y)$ is continuous on I^2,
(B) $(w_1 u)^{-1}\, w \in \mathbf{L}^1(I)$,
(C) there exists a positive weight function $u_1 : I_0 \longrightarrow [0, \infty)$ continuous on I, such that $K_1(x, .) = u(x)\widetilde{K}(x, .)u^{-1}(.) \in \mathbf{C}_{u_1}(I_0)$ for all $x \in I$,
(D) $u_1^{-1} w \in \mathbf{L}^1(I)$,
(E) for the quadrature rule (11), we have

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} \lambda_{nk} f(x_{nk}) = \int_I f(x)w(x)\, dx$$

for all $f \in \mathbf{C}_{u_1}(I_0)$,

(F) the inequalities

$$\sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u(x_{nk})w_1(x_{nk})} \le c \tag{15}$$

hold true for all $n \in \mathbb{N}$, where $c \neq c(n)$.

The following corollary is concerned with condition (E).

Corollary 1 *Let* (D) *be satisfied. If the quadrature rule* (11) *is exact for polynomials of degree less than* $\kappa(n)$*, where* $\kappa(n)$ *tends to infinity if* $n \longrightarrow \infty$*, and if*

$$\sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u_1(x_{nk})} \le c \tag{16}$$

for all $n \in \mathbb{N}$*, where* $c \neq c(n)$*, then*

(a) $\displaystyle\lim_{n\to\infty} \sum_{k=1}^{k_n} \lambda_{nk} f(x_{nk}) = \int_I f(x)w(x)\,dx \quad \forall f \in \mathbf{C}_{u_1}(I_0)$,

(b) $\displaystyle\left| \int_I f(x)w(x)\,dx - \sum_{k=1}^{k_n} \lambda_{nk} f(x_{nk}) \right| \le c\, E_{\kappa(n)-1}(f)_{u_1,\infty}\,, \quad c \neq c(n,f)$,

where $E_m(f)_{u_1,\infty} = \inf\{\|f-p\|_{u_1,\infty} : p \in \mathbf{P}_m\}$ *is the best weighted uniform approximation of the function* f *by polynomials of degree less or equal to* m*. Moreover, if* (E) *is satisfied then* (16) *and* (b) *hold.*

Proof Define the linear functionals $\mathscr{F}_n : \mathbf{C}_{u_1}(I_0) \longrightarrow \mathbb{C}$ by

$$\mathscr{F}_n f = \sum_{k=1}^{k_n} \lambda_{nk} f(x_{nk})\,.$$

Then, in virtue of (16),

$$|\mathscr{F}_n f| \le \sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u_1(x_{nk})} \|f\|_{u_1,\infty} \le c\|f\|_{u_1,\infty} \quad \forall f \in \mathbf{C}_{u_1}\,,\ c \neq c(n,f)\,.$$

Hence, the linear functionals $\mathscr{F}_n : \mathbf{C}_{u_1}(I_0) \longrightarrow \mathbb{C}$ are uniformly bounded. Moreover, due to our assumptions,

$$\lim_{n\to\infty} \mathscr{F}_n f = \int_I f(x)w(x)\,dx \quad \forall f \in \mathbf{P}\,,$$

and the Banach-Steinhaus Theorem gives the assertion (a). For all $p \in \mathbf{P}_{\kappa(n)-1}$, we get

$$\left|\int_I f(x)w(x)\,dx - \sum_{k=1}^{k_n} \lambda_{nk} f(x_{nk})\right|$$

$$\leq \int_I |f(x)-p(x)|\,w(x)\,dx + \sum_{k=1}^{k_n} \lambda_{nk}\,|f(x_{nk})-p(x_{nk})|$$

$$\leq \left[\int_I \frac{w(x)\,dx}{u_1(x)} + \sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u_1(x_{nk})}\right] \|f-p\|_{u_1,\infty}\,.$$

It remains to take into account (D) and (16), and also (b) is proved.

Finally, we make the following observation. The norm of the functionals $\mathscr{F}_n : \mathbf{C}_{u_1}(I_0) \longrightarrow \mathbb{C}$ is equal to $\displaystyle\sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u_1(x_{nk})}$. Hence, due to the uniform boundedness principle, condition (16) is also necessary for assertion (a) to be fulfilled. □

Proposition 2 *If the conditions* (A)–(F) *are fulfilled, then the operators* $\mathscr{K}_n \in \mathscr{L}(\mathbf{C}(I))$*, defined in* (5) *and* (14)*, form a collectively compact sequence of strongly convergent to* $\mathscr{K}$ (*cf.* (5) *and* (13)) *operators in* $\mathbf{C}(I)$.

Proof We check if conditions (K1)–(K4) are fulfilled. Condition (K1) is a consequence of

$$\int_I |K(x,y)|\,dy \overset{(13)}{=} \int_I |K_1(x,y)|\,w(y)\,dy \overset{(C),(D)}{\leq} \|K_1(x,.)\|_{u_1,\infty} \left\|(u_1)^{-1}w\right\|_{\mathbf{L}^1(I)}\,.$$

Analogously, (K2) follows from

$$\int_I |K(x,y)-K(x_0,y)|\,dy = \int_I |K_0(x,y)-K_0(x_0,y)|\,\frac{w(y)}{u(y)w_1(y)}\,dy$$

by applying the continuity of $K_0(x,y)$ and condition (B). In view of (14), condition (C), and condition (E),

$$\sum_{k=1}^{k_n} \Lambda_{nk}(x)f(x_{nk}) = \sum_{k=1}^{k_n} \lambda_{nk}K_1(x,x_{nk})f(x_{nk})$$

$$\longrightarrow \int_I K_1(x,y)f(y)w(y)\,dy = \int_I K(x,y)f(y)\,dy$$

if $n \longrightarrow \infty$ for all $f \in \mathbf{C}(I)$ and all $x \in I$, i.e., $K(x,y)$ satisfies also (K3). Finally, for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|K_0(x,y) - K_0(x_0,y)| < \varepsilon$ for all $(x,y) \in U_\delta(x_0) \times I$, where $U_\delta(x_0) = \{x \in I : d(x,x_0) < \delta\}$. Consequently, according to (15),

$$\sum_{k=1}^{k_n} |\Lambda_{nk}(x) - \Lambda_{nk}(x_0)| = \sum_{k=1}^{k_n} \lambda_{nk} |K_1(x,x_{nk}) - K_1(x_0,x_{nk})|$$

$$= \sum_{k=1}^{k_n} \frac{\lambda_{nk}}{u(x_{nk})w_1(x_{nk})} |K_0(x,x_{nk}) - K_0(x_0,x_{nk})| < c\,\varepsilon$$

for all $x \in U_\delta(x_0)$, which shows the validity of (K4). The application of Lemma 1 completes the proof. □

Remark 2 In case of $u^{-1}u_1 = w_1$, for the proof of Proposition 2, one can also use Lemma 2. Indeed, if we set $v = u_1$ and define $H(x,y) = w(y)$, $S(x,y) = K_1(x,y)$, $\mathbf{X}_0 = \operatorname{span}\{w\}$ with $\|.\|_{\mathbf{X}_0} = \|.\|_{\mathbf{L}^1_{v^{-1}}(I)}$, $\lambda^F_{nk}(\gamma w) = \gamma \lambda_{nk}$ for $\gamma \in \mathbb{C}$, then, we have $\mathbf{X}_0 \subset \mathbf{L}^1_{v^{-1}}(I)$ continuously (see (D) which now coincides with (B)), $K(x,y) = H(x,y)S(x,y)$ with the continuous function $S(x,y)v(y)$ (see (A)), and $\Lambda_{nk}(x) = \lambda^F_{nk}(w)S(x,x_{nk})$ (cf. (14)). Moreover, for all $f = \gamma w \in \mathbf{X}_0$ and all $g \in \mathbf{C}_v(I_0)$,

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} \lambda^F_{nk}(f)g(x_{nk}) = \lim_{n\to\infty} \gamma \sum_{k=1}^{k_n} \lambda_{nk} g(x_{nk}) = \int_I f(y)g(y)\,dy$$

in view of condition (E). Consequently, conditions (H1)–(H3) are fulfilled and Lemma 2 can be applied.

Corollary 2 *Assume* (A)–(F). *Consider the Eqs.* (9) *and* (12) *with* $\widetilde{g} \in \widetilde{\mathbf{C}}_u(I_0)$. *Assume further, that the homogeneous equation* (9) (*i.e.,* $\widetilde{g} \equiv 0$) *has in* $\widetilde{\mathbf{C}}_u(I_0)$ *only the trivial solution. Then, for all sufficiently large n, Eq.* (12) *possesses a unique solution* $\widetilde{f}^*_n \in \widetilde{\mathbf{C}}_u(I_0)$ *converging to* $\widetilde{f}^*$, *where* $\widetilde{f}^* \in \widetilde{\mathbf{C}}_u$ *is the unique solution of* (9). *If the assumptions of Corollary 1 are satisfied, then*

$$\|\widetilde{f}^* - \widetilde{f}^*_n\|_{u,\infty} \le c \sup\left\{E_{2n-1}\left(u(x)\widetilde{K}(x,.)\widetilde{f}^*\right)_{u_1,\infty} : x \in I\right\}, \tag{17}$$

where $c \neq c(n,g)$. *(Note that, due to condition* (C), $u(x)\widetilde{K}(x,.)\widetilde{f}^* \in \mathbf{C}_{u_1}(I_0)$ *for all* $x \in I$.*)*

Proof In virtue of Proposition 2, we can apply Proposition 1 with $\mathbf{X} = \mathbf{C}(I)$ to Eqs. (2) and (3) with the above definitions (13) and (14). Estimate (4) gives

$$\|\widetilde{f}^*_n - \widetilde{f}^*\|_{u,\infty} = \|f^*_n - f^*\|_\infty \le c\,\|\mathscr{K}_n f^* - \mathscr{K} f^*\|_\infty ,$$

where $f^* \in \mathbf{C}(I)$ and $f_n^* \in \mathbf{C}(I)$ are the solutions of (2) and (3), respectively, and where

$$\begin{aligned}
&\left\| \mathscr{K}_n f^* - \mathscr{K} f^* \right\|_\infty \\
&\quad = \sup\left\{ \left| \sum_{k=1}^{k_n} \varLambda_{nk}(x) f^*(x_{nk}) - \int_I K(x,y) f^*(y)\,dy \right| : x \in I \right\} \\
&\quad = \sup\left\{ \left| \sum_{k=1}^{k_n} \lambda_{nk} u(x) \widetilde{K}(x,x_{nk}) \widetilde{f}^*(x_{nk}) - \int_I u(x) \widetilde{K}(x,y) \widetilde{f}^*(y) w(y)\,dy \right| : x \in I \right\} .
\end{aligned}$$

It remains to use $u(x)\widetilde{K}(x,.)\widetilde{f}^* \in \mathbf{C}_{u_1}(I_0)$ (cf. (C)) and Corollary 1, (b). □

3.1 The Case of Jacobi Weights

Let us apply the above described Nyström method in case of

$$\widetilde{f}(x) - \int_{-1}^{1} \widetilde{K}(x,y) v^{\alpha,\beta}(y) \widetilde{f}(y)\,dy = \widetilde{g}(x)\,, \quad -1 < x < 1\,, \tag{18}$$

where $\widetilde{g} \in \widetilde{\mathbf{C}}_u = \widetilde{\mathbf{C}}_u(-1,1)$ and $\widetilde{K} : (-1,1)^2 \longrightarrow \mathbb{C}$ are given continuous functions and where $v^{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > -1$, and $u(x) = v^{\gamma,\delta}(x)$, $\gamma, \delta \geq 0$, are Jacobi weights, and $\widetilde{\mathbf{C}}_u = \widetilde{\mathbf{C}}_{v^{\gamma,\delta}}$. We set $u_1(x) = v^{\gamma_1,\delta_1}(x)$, $w_1(x) = v^{\alpha_1,\beta_1}(x)$ and assume that

(A1) $K_0 : [-1,1]^2 \longrightarrow \mathbb{C}$ is continuous, where $K_0(x,y) = v^{\gamma,\delta}(x)\widetilde{K}(x,y)v^{\alpha_1,\beta_1}(y)$,

(B1) $\displaystyle\int_{-1}^{1} \frac{v^{\alpha,\beta}(x)\,dx}{v^{\gamma,\delta}(x) v^{\alpha_1,\beta_1}(x)} < \infty$, i.e., $\gamma + \alpha_1 < \alpha + 1$ and $\delta + \beta_1 < \beta + 1$,

(C1) $0 \leq \gamma_1$, $0 \leq \delta_1$, and $\gamma + \alpha_1 < \gamma_1 < \alpha + 1$, $\delta + \beta_1 < \delta_1 < \beta + 1$.

Setting $w(x) := v^{\alpha,\beta}(x)$, the conditions (A1) and (B1) are equivalent to (A) and (B) in the present situation, respectively. Condition (C1) leads immediately to (C) and (D), since in case $u(x) = v^{\gamma,\delta}(x)$ and $\gamma, \delta \geq 0$, the set $\mathbf{C}_u$ is equal to the set of all $f \in \widetilde{\mathbf{C}}_u$ satisfying

$$\lim_{x \to 1-0} u(x) f(x) = 0 \text{ if } \gamma > 0 \quad \text{and} \quad \lim_{x \to -1+0} u(x) f(x) = 0 \text{ if } \delta > 0\,.$$

As quadrature rule (11) we take the Gaussian rule w.r.t. the Jacobi weight $w(x) = v^{\alpha,\beta}(x)$, i.e., $k_n = n$, the $x_{nk} = x_{nk}^{\alpha,\beta}$'s are the zeros of the nth (normalized) Jacobi polynomial $p_n^{\alpha,\beta}(x)$ w.r.t. $w(x) = v^{\alpha,\beta}(x)$ and the $\lambda_{nk} = \lambda_{nk}^{\alpha,\beta}$'s are the respective Christoffel numbers. Then, for Corollary 1 we have $\kappa(n) = 2n - 1$. Moreover, condition (C1) guarantees that (15) and (16) are also fulfilled, which is due to the following lemma.

Lemma 3 ([31], Theorem 9.25) *For $v^{\alpha,\beta}(x)$ and $v^{\alpha_1,\beta_1}(x)$, assume that $\alpha + \alpha_1 > -1$ and $\beta + \beta_1 > -1$, and let $j \in \mathbb{N}$ be fixed. Then, for each polynomial $q(x)$ with $\deg q \le jn$,*

$$\sum_{k=1}^{n} \lambda_{nk}^{\alpha,\beta} \left|q(x_{nk}^{\alpha,\beta})\right| v^{\alpha_1,\beta_1}\left(x_{nk}^{\alpha,\beta}\right) \le c \int_{-1}^{1} |q(x)| v^{\alpha,\beta}(x) v^{\alpha_1,\beta_1}(x)\, dx\,,$$

where $c \neq c(n,q)$.

Hence, all conditions (A)–(F) are in force and we can apply Corollary 2 together with the estimate (b) of Corollary 1 to Eq. (18) and the Nyström method

$$\widetilde{f}_n(x) - \sum_{k=1}^{n} \lambda_{nk}^{\alpha,\beta} \widetilde{K}(x, x_{nk}^{\alpha,\beta}) \widetilde{f}_n(x_{nk}^{\alpha,\beta}) = \widetilde{g}(x)\,, \quad -1 < x < 1\,, \tag{19}$$

to get the following proposition.

Proposition 3 *Assume that (A1), (B1), and (C1) are fulfilled and that Eq. (18) has only the trivial solution in $\widetilde{\mathbf{C}}_{v^{\gamma,\delta}}$ in case of $\widetilde{g}(x) \equiv 0$. Then, for $\widetilde{g} \in \widetilde{\mathbf{C}}_{v^{\gamma,\delta}}$ and all sufficiently large n, Eq. (19) has a unique solution $\widetilde{f}_n^* \in \widetilde{\mathbf{C}}_{v^{\gamma,\delta}}$ and*

$$\|\widetilde{f}^* - \widetilde{f}_n^*\|_{\gamma,\delta,\infty} \le c \sup\left\{ E_{2n-1}\left(v^{\gamma,\delta}(x)\widetilde{K}(x,.)\widetilde{f}^*\right)_{v^{\gamma_1,\delta_1},\infty} : -1 \le x \le 1 \right\}\,,$$

where $\widetilde{f}^* \in \widetilde{\mathbf{C}}_{v^{\gamma,\delta}}$ is the unique solution of (18) and $c \neq c(n, g)$. (Again we note that the assumptions of the proposition guarantee that $v^{\gamma,\delta}(x)\widetilde{K}(x,.)\widetilde{f}^* \in \mathbf{C}_{v^{\gamma_1,\delta_1}}$ for all $x \in [-1,1]$, cf. Corollary 2.)

For checking (15) and (16), we used Lemma 3. The following Lemma will allow us to prove these assumptions also in other cases.

Lemma 4 *Let $w : I_0 \longrightarrow [0,\infty)$ and $v : I_0 \longrightarrow [0,\infty)$ be weight functions and $\lambda_{nk} > 0$, $x_{nk} \in I_0$, $k = 1,\ldots,n$, be given numbers satisfying the conditions $x_{n1} < x_{n2} < \ldots < x_{nn}$ and*

(a) *$v^{-1}w \in \mathbf{L}^1(I)$,*
(b) *$\lambda_{nk} \sim_{n,k} \Delta x_{nk} w(x_{nk})$, $k = 1,\ldots,n$, where $\Delta x_{nk} = x_{nk} - x_{n,k-1}$ and $x_{n0} < x_{n1}$ is appropriately chosen,*
(c) *$\Delta x_{nk} \sim_{n,k} \Delta x_{n,k-1}$, $k = 2,\ldots,n$,*
(d) *for each closed subinterval $[a,b] \subset I_0$, $v^{-1}w : [a,b] \longrightarrow \mathbb{R}$ is continuous and*

$$\lim_{n\to\infty} \max\left\{\Delta x_{nk} : x_{nk} \in [a,b]\right\} = 0\,, \tag{20}$$

(e) *there exists a subinterval* $[A,B] \subset I_0$ *such that* $v^{-1}w : \{x \in I_0 : x \le A\} \longrightarrow \mathbb{R}$ *and* $v^{-1}w : \{x \in I_0 : x \ge B\} \longrightarrow \mathbb{R}$ *are monotone.*

Then, there is a constant $c \ne c(n)$ *such that*

$$\sum_{k=1}^{n} \frac{\lambda_{nk}}{v(x_{nk})} \le c \int_I \frac{w(x)}{v(x)}\, dx\,. \tag{21}$$

Proof By assumption (b) we have $\sum_{k=1}^{n} \frac{\lambda_{nk}}{v(x_{nk})} \sim_n \sum_{k=1}^{n} \frac{w(x_{nk})}{v(x_{nk})}\, \Delta x_{nk}$. Moreover,

$$\lim_{n\to\infty} \sup \left\{ \left| \frac{w(x)}{v(x)} - \frac{w(x_{nk})}{v(x_{nk})} \right| : x \in [x_{n,k-1}, x_{nk}],\ x_{nk} \in [A,B] \right\} = 0\,,$$

due to assumption (d). Hence,

$$\frac{w(x_{nk})}{v(x_{nk})}\, \Delta x_{nk} \le c \int_{x_{n,k-1}}^{x_{nk}} \frac{w(x)}{v(x)}\, dx \quad \forall\, x_{nk} \in [A,B] \quad \text{with} \quad c \ne c(n,k)\,.$$

If $v^{-1}w : \{x \in I_0 : x \le A\} \longrightarrow \mathbb{R}$ is non-increasing, then

$$\frac{w(x_{nk})}{v(x_{nk})}\, \Delta x_{nk} \le \int_{x_{n,k-1}}^{x_{nk}} \frac{w(x)}{v(x)}\, dx \quad \forall\, x_{nk} < A,\ k \ge 1\,.$$

If $v^{-1}w : \{x \in I_0 : x \le A\} \longrightarrow \mathbb{R}$ is non-decreasing, then we use assumption (c) and get

$$\frac{w(x_{nk})}{v(x_{nk})}\, \Delta x_{nk} \sim_{n,k} \frac{w(x_{nk})}{v(x_{nk})}\, \Delta x_{n,k+1} \le \int_{x_{nk}}^{x_{n,k+1}} \frac{w(x)}{v(x)}\, dx \quad \forall\, x_{nk} < A,\ k \ge 1\,,$$

with (if necessary) an appropriately chosen $x_{n,n+1} > x_{nn}$. For $x_{nk} > B$ we can proceed analogously (noting that B can be chosen sufficiently large such that, for all $n \ge n_0$, $v^{-1}w$ is monotone on the interval $[x_{n,k_0-1}, x_{n,k_0})$ containing B). Summarizing we obtain (21). □

It is obvious how we have to formulate Lemma 4 in case $\lambda_{nk} > 0$ and $x_{nk} \in I_0$ are given for $k = k_1(n), \dots, k_2(n)$.

3.2 The Case of an Exponential Weight on $(0,\infty)$

Consider the integral equation

$$\widetilde{f}(x) - \int_0^\infty \widetilde{K}(x,y) w(y) \widetilde{f}(y)\, dy = \widetilde{g}(x)\,, \quad 0 < x < \infty, \tag{22}$$

where $\widetilde{g} \in \widetilde{\mathbf{C}}_u(0,\infty)$ and $\widetilde{K} : (0,\infty)^2 \longrightarrow \mathbb{C}$ are given functions and where $w(x) = w^{\alpha,\beta}(x) = e^{-x^{-\alpha}-x^{\beta}}$, $\alpha > 0$, $\beta > 1$, $u(x) = u^{a,\delta}(x) = (1+x)^{\delta}[w(x)]^{a}$, $a \ge 0$, $\delta \ge 0$. Here we use the Gaussian rule w.r.t. the weight $w(x) = w^{\alpha,\beta}(x)$ and study the Nyström method

$$\widetilde{f}_n(x) - \sum_{k=1}^{n} \lambda_{nk}^{w} \widetilde{K}(x, x_{nk}^{w}) \widetilde{f}_n(x_{nk}^{w}) = \widetilde{g}(x)\,, \quad 0 < x < \infty\,. \tag{23}$$

Let us check conditions (A)–(F), for which we choose

$$w_1(x) = u^{a_0,\delta_0}(x) := (1+x)^{\delta_0}[w(x)]^{a_0}\,, \quad \delta_0, a_0 \in \mathbb{R}\,,$$

and

$$u_1(x) = u^{a_1,\delta_1}(x) = (1+x)^{\delta_1}[w(x)]^{a_1}\,, \quad \delta_1 \ge 0\,,\ 0 < a_1 \le 1\,,$$

and assume that

(A2) $K_0(x,y) := u(x)\widetilde{K}(x,y)w_1(y)$ is continuous on $[0,\infty]^2$,
(B2) $0 < a + a_0 < 1\,, \delta + \delta_0 \ge 0$ or $a + a_0 = 1\,, \delta + \delta_0 > 1$,
(C2) $0 < a_1 < 1\,, \delta_1 \ge 0$ or $a_1 = 1\,, \delta_1 > 1$,
(D2) $a_1 > a_0 + a$.

Note that, due to Lemma 4 (cf. [22, Prop. 3.8], for checking the conditions of Lemma 4 see also [14, 19, 27])

$$\sum_{k=1}^{n} \frac{\lambda_{nk}^{w}}{u_1(x_{nk}^{w})} \le c \quad \text{with} \quad c \ne c(n) \tag{24}$$

if $u_1^{-1}w \in \mathbf{L}^1(0,\infty)$, which is equivalent to assumption (C2). We also see that (B2) implies $(w_1 u)^{-1} w \in \mathbf{L}^1(0,\infty)$. Condition (A2) together with (D2) guarantees that $u(x)\widetilde{K}(x,.)u^{-1} \in \mathbf{C}_{u_1}(0,\infty)$ for all $x \in [0,\infty]$. Hence, we see that (A2)–(D2) together with Corollary 1, (a) imply (A)–(F), and we can apply Corollary 2 together with Corollary 1, (b) to (22) and (23) to get the following.

Proposition 4 *Let* $w(x) = e^{-x^{-\alpha}-x^{\beta}}$, $\alpha > 0$, $\beta > 1$, *and* $u(x) = (1+x)^{\delta}[w(x)]^{a}$, $a \ge 0$, $\delta \ge 0$. *Assume that* (A2), (B2), (C2), *and* (D2) *are fulfilled and that Eq.* (22) *has only the trivial solution in* $\widetilde{\mathbf{C}}_u(0,\infty)$ *in case of* $\widetilde{g}(x) \equiv 0$. *Then, for* $\widetilde{g} \in \widetilde{\mathbf{C}}_u(0,\infty)$ *and all sufficiently large n, Eq.* (23) *has a unique solution* $\widetilde{f}_n^* \in \widetilde{\mathbf{C}}_u(0,\infty)$ *and*

$$\|\widetilde{f}^* - \widetilde{f}_n^*\|_{u,\infty} \le c \sup\left\{ E_{2n-1}\big(u(x)\widetilde{K}(x,.)\widetilde{f}^*\big)_{u_1,\infty} : 0 \le x \le \infty \right\}\,,$$

where $\widetilde{f}^* \in \widetilde{\mathbf{C}}_u(0,\infty)$ *is the unique solution of* (22) *and* $c \ne c(n,g)$.

4 The Nyström Method Based on Product Integration Formulas

Let again I_0 and I be equal to $(-1,1)$, $(0,\infty)$, or $(-\infty,\infty)$ and $[-1,1]$, $[0,\infty]$, or $[-\infty,\infty]$, respectively. Here we discuss the numerical solution of the Fredholm integral equation (9) by means of approximating the operator

$$\widetilde{\mathscr{K}} : \widetilde{\mathbf{C}}_u(I_0) \longrightarrow \widetilde{\mathbf{C}}_u(I_0)\,,, \quad \widetilde{f} \mapsto \int_I \widetilde{K}(.,y)w(y)\widetilde{f}(y)\,dy \tag{25}$$

by

$$\big(\widetilde{\mathscr{K}}_n\widetilde{f}\big)(x) = \int_I \frac{\widetilde{H}(x,y)}{u(y)}\left[\mathscr{L}_n\widetilde{S}(x,.)u\widetilde{f}\,\right](y)w(y)\,dy\,, \quad x \in I_0\,, \tag{26}$$

where $\widetilde{K}(x,y) = \widetilde{H}(x,y)\widetilde{S}(x,y)$ and $\mathscr{L}_n g$ is the algebraic polynomial of degree less than n with $(\mathscr{L}_n g)(x_{nk}) = g(x_{nk})$, $k = 1,\ldots,n$. Using the formula

$$(\mathscr{L}_n g)(x) = \sum_{k=1}^{n} g(x_{nk})\ell_{nk}(x) \quad \text{with} \quad \ell_{nk}(x) = \prod_{j=1, j\neq k}^{n} \frac{x - x_{nj}}{x_{nk} - x_{nj}}\,,$$

we conclude

$$\big(\widetilde{\mathscr{K}}_n\widetilde{f}\big)(x) = \sum_{k=1}^{n} \int_I \frac{\widetilde{H}(x,y)}{u(y)}\ell_{nk}(y)w(y)\,dy\,\widetilde{S}(x,x_{nk})u(x_{nk})\widetilde{f}(x_{nk})\,.$$

So, here we have $k_n = n$. Furthermore, this means that, for Eq. (2) considered in the space $\mathbf{C}(I)$, the operator $\mathscr{K} : \mathbf{C}(I) \longrightarrow \mathbf{C}(I)$ defined in (5) is approximated by $\mathscr{K}_n : \mathbf{C}(I) \longrightarrow \mathbf{C}(I)$ also given by (5), where $K(x,y)$ is defined in (13) and where (cf. (6))

$$\Lambda_{nk}(x) = \int_I H(x,y)\ell_{nk}(y)\,dy\,S(x,x_{nk}) = \lambda_{nk}^F(H(x,.))S(x,x_{nk}) \tag{27}$$

with $H(x,y) = \dfrac{u(x)\widetilde{H}(x,y)w(y)}{u(y)}$, $S(x,y) = \widetilde{S}(x,y)$, and

$$\lambda_{nk}^F(f) = \int_I f(y)\ell_{nk}(y)\,dy\,. \tag{28}$$

In order to check, under which conditions the assumption (H3) is satisfied, we should use

$$\begin{aligned}\left|\sum_{k=1}^{n}\lambda_{nk}^{F}(f)g(x_{nk})-\int_I f(y)g(y)\,dy\right| &= \left|\int_I f(y)\left[(\mathscr{L}_n g)(y)-g(y)\right]\,dy\right| \\ &\leq \left(\int_I\left|\frac{f(y)}{u(y)}\right|^p dy\right)^{\frac{1}{p}}\|(\mathscr{L}_n g-g)u\|_{\mathbf{L}^q(I)}\,,\end{aligned}\tag{29}$$

where $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, and u is an appropriate weight function.

4.1 The Case of Jacobi Weights

Consider the case where $w(x)=v^{\alpha,\beta}(x)$, $\alpha,\beta>-1$, and $v(x)=v^{\gamma,\delta}(x)$, $\gamma,\delta\geq 0$.

Lemma 5 *Let* $w=v^{\alpha,\beta}$, $\alpha,\beta>-1$, $p>1$, $\gamma_0,\delta_0\geq 0$, $\gamma_0>\frac{\alpha}{2}+\frac{1}{4}+\frac{1}{p}-1$, *and* $\delta_0>\frac{\beta}{2}+\frac{1}{4}+\frac{1}{p}-1$. *Then, condition* (H3) *is fulfilled for* $\ell_{nk}(x)=\ell_{nk}^{w}(x)=\prod_{j=1,j\neq k}^{n}\frac{x-x_{nj}^{\alpha,\beta}}{x_{nk}^{\alpha,\beta}-x_{nj}^{\alpha,\beta}}$ *in* (28) *as well as* $\mathbf{X}_0=\mathbf{L}^p_{v^{-\gamma_0,-\delta_0}}$ *and* $\mathbf{C}_v=\mathbf{C}$, *i.e.* $v\equiv 1$.

Proof First, $\mathbf{X}_0=\mathbf{L}^p_{v^{-\gamma_0,-\delta_0}}$ is continuously embedded in $\mathbf{L}^1$, since $\gamma_0,\delta_0\geq 0$. Second, we can use the fact (cf. [32, Theorems 1 and 2]) that there is a constant $c>0$ such that $\left\|(g-\mathscr{L}_n^w g)v^{\gamma_0,\delta_0}\right\|_q\leq c\,E_{n-1}(g)_\infty$ for all $g\in\mathbf{C}$ if and only if $\frac{v^{\gamma_0,\delta_0}}{\sqrt{w\varphi}}\in\mathbf{L}^q$ with $\varphi(x)=\sqrt{1-x^2}$, i.e.,

$$\gamma_0-\frac{\alpha}{2}-\frac{1}{4}>-\frac{1}{q}\quad\text{and}\quad\delta_0-\frac{\beta}{2}-\frac{1}{4}>-\frac{1}{q}\,.$$

Hence, (29) can be applied to all $f\in\mathbf{X}_0$, all $g\in\mathbf{C}$, and $u=v^{\gamma_0,\delta_0}$. □

Remark 3 We remark that Lemma 5 improves the result mentioned in [35, Section 4.5], where γ_0 and δ_0 are chosen as

$$\max\left\{\frac{\alpha}{2}+\frac{1}{4},0\right\}\quad\text{and}\quad\max\left\{\frac{\beta}{2}+\frac{1}{4},0\right\}\,,$$

respectively.

As a consequence of Lemma 5 and of Lemma 2, we have to assume that $H(x,.)$ satisfies condition (H2) for $\mathbf{X}_0=\mathbf{L}^p_{v^{-\gamma_0,-\delta_0}}$ with appropriate γ_0,δ_0 and p as in

Lemma 5. The aim of the remaining part of this subsection is to weaken this condition in a certain way.

By $\mathbf{L}\log^+\mathbf{L}(a,b)$ we denote the set of all measurable functions $f:(a,b)\longrightarrow\mathbb{C}$ for which the integral $\rho_+(f):=\displaystyle\int_a^b |f(x)|\left(1+\log^+|f(x)|\right)dx$ is finite. For $f\in\mathbf{L}^1(a,b)$, by $\mathscr{H}_a^b f$ we denote the Hilbert transform of f,

$$\left(\mathscr{H}_a^b f\right)(x):=\int_a^b \frac{f(y)\,dy}{y-x}\,,\quad a<x<b$$

(as Cauchy principal value integral). From [33, (1),(2)] we infer the following.

Lemma 6 *Let $-\infty<a<b<\infty$. If $f\in\mathbf{L}\log^+\mathbf{L}(a,b)$ and $g\in\mathbf{L}^\infty(a,b)$, then*

$$\left\|g\mathscr{H}_a^b f\right\|_1+\left\|f\mathscr{H}_a^b g\right\|_1\le c\|g\|_\infty\rho_+(f) \tag{30}$$

with $c\neq c(f,g)$ and

$$\int_a^b g(x)\left(\mathscr{H}_a^b f\right)(x)\,dx=-\int_a^b f(x)\left(\mathscr{H}_a^b g\right)(x)\,dx\,. \tag{31}$$

Let us use the abbreviations $w(x)=v^{\alpha,\beta}(x)$, $p_n(x)=p_n^{\alpha,\beta}(x)$, $x_{nk}=x_{nk}^{\alpha,\beta}$, and $\Delta x_{nk}=x_{nk}-x_{n,k-1}$, $k=1,\ldots,n$, $x_{n0}=-1$, $\mathbf{L}^p=\mathbf{L}^p(-1,1)$, and $\mathbf{L}\log^+\mathbf{L}=\mathbf{L}\log^+\mathbf{L}(-1,1)$, as well as $\mathscr{H}=\mathscr{H}_{-1}^1$. The relations

(R1) $|p_n(x)|\sqrt{w(x)\varphi(x)}\le c$ for $x\in A_n:=\left[\dfrac{x_{n1}-1}{2},\dfrac{x_{nn}+1}{2}\right]$, $c\neq c(n)$,

(R2) $\dfrac{1}{\left|p_n'(x_{nk})\right|}\sim_{n,k}\Delta x_{nk}\sqrt{w(x_{nk})\varphi(x_{nk})}$ (see [32, (14)]),

(R3) for a fixed summable function $v:[-1,1]\longrightarrow\mathbb{C}$ and a fixed $\ell\in\mathbb{N}$,

$$\sum_{k=1}^n \Delta x_{nk}\,|p(x_{nk})v(x_{nk})|\le c\int_{A_n}|p(x)v(x)|\,dx$$

for all polynomials $p\in\mathbf{P}_{\ell n}:=\{P\in\mathbf{P}:\deg P\le\ell n\}$ and with $c\neq c(n,p)$

are well-known. Note that (R1) is a consequence of the estimate (see [8, Theorem 1.1])

$$\left|p_n^{\alpha,\beta}(x)\right|\left(\sqrt{1-x}+\frac{1}{n}\right)^{\alpha+\frac12}\left(\sqrt{1+x}+\frac{1}{n}\right)^{\beta+\frac12}\le c\neq c(n,x)\,, \tag{32}$$

$-1 < x < 1$, and the relation $\theta_{n,k-1} - \theta_{nk} \sim_{n,k} \frac{1}{n}$, $k = 1, \dots, n+1$, $n \in \mathbb{N}$, where $\theta_{nk} \in [0, \pi]$ and $x_{nk} = \cos\theta_{nk}$, $\theta_{n,n+1} = 0$ (cf. [28, (5)]).

Lemma 7 *Let $w(x) = v^{\alpha,\beta}(x)$ and $v(x) = v^{\gamma,\delta}(x)$ be Jacobi weights satisfying*

$$\frac{\alpha}{2} + \frac{1}{4} > \gamma \geq 0 \quad \textit{and} \quad \frac{\beta}{2} + \frac{1}{4} > \delta \geq 0\,. \tag{33}$$

Then, there is a constant $c \neq c(n, f, g)$ such that, for all functions $f : (-1, 1) \longrightarrow \mathbb{C}$ with $fv \in \mathbf{L}^\infty$ and all g with $\frac{g}{\sqrt{w\varphi}} \in \mathbf{L}\log^+\mathbf{L}$,

$$\left\| g\mathscr{L}_n^w f \right\|_1 \leq c\,\rho_+\left(\frac{g}{\sqrt{w\varphi}}\right) \|fv\|_\infty\,.$$

Proof Write $\left\| g\mathscr{L}_n^w f \right\|_1 = J_1 + J_2 + J_3$, where

$$J_1 = \left\| g\mathscr{L}_n^w f \right\|_{\mathbf{L}^1(A_n)}\,, \quad J_2 = \left\| g\mathscr{L}_n^w f \right\|_{\mathbf{L}^1\left(-1, \frac{x_{n1}-1}{2}\right)}\,, \quad J_3 = \left\| g\mathscr{L}_n^w f \right\|_{\mathbf{L}^1\left(\frac{x_{nn}+1}{2}, 1\right)}\,.$$

Define

$$\widetilde{p}_n(y) := \begin{cases} p_n(y) : y \in A_n\,, \\ \quad 0 \quad\; : y \notin A_n\,, \end{cases} \quad \text{and} \quad \widetilde{g}_n(y) := \begin{cases} g(y) : y \in A_n\,, \\ \quad 0 \quad : y \notin A_n\,, \end{cases}$$

as well as $h_n(y) := \operatorname{sgn}\left[g(y)\left(\mathscr{L}_n^w f\right)(y)\right]$, and consider

$$\begin{aligned} J_1 &= \int_{A_n} h_n(y) g(y) \left(\mathscr{L}_n^w f\right)(y)\,dy = \sum_{k=1}^n \frac{f(x_{nk})}{p_n'(x_{nk})} \int_{A_n} \frac{p_n(y)}{y - x_{nk}}\, g(y) h_n(y)\,dy \\ &\overset{\text{(R2)}}{\leq} c\|fv\|_\infty \sum_{k=1}^n \Delta x_{nk} \frac{\sqrt{w(x_{nk})\varphi(x_{nk})}}{v(x_{nk})}\, |G_n(x_{nk})|\,, \end{aligned}$$

where

$$G_n(x) = \int_{A_n} \frac{p_n(y)Q_n(y) - p_n(x)Q_n(x)}{y - x}\, \frac{g(y)h_n(y)}{Q_n(y)}\,dy$$

for some polynomial $Q_n \in \mathbf{P}_{\ell n}$ positive on A_n ($\ell \in \mathbb{N}$ fixed). Then, due to $G_n \in \mathbf{P}_{\ell n+n-1}$ and (R3),

$$\begin{aligned}
J_1 &\le c\|fv\|_\infty \int_{A_n} |G_n(x)|\, \frac{\sqrt{w(x)\varphi(x)}}{v(x)}\, dx \\
&\le c\|fv\|_\infty \left[\int_{-1}^1 \frac{\sqrt{w(x)\varphi(x)}}{v(x)} \left(\mathscr{H}\widetilde{p}_n\widetilde{g}_n h_n\right)(x)\, k_n^1(x)\, dx \right. \\
&\qquad \left. + \int_{-1}^1 \frac{\sqrt{w(x)\varphi(x)}}{v(x)}\, |\widetilde{p}_n(x)|\, Q_n(x) \left(\mathscr{H}\frac{gh_n}{Q_n}\right)(x)\, k_n^2(x)\, dx \right] \\
&=: c\|fv\|_\infty \left[J_1' + J_1'' \right],
\end{aligned}$$

where $k_n^1(x) = \operatorname{sgn}\left[\left(\mathscr{H}\widetilde{p}_n\widetilde{g}_n h_n\right)(x)\right]$ and $k_n^2(x) = \operatorname{sgn}\left[\left(\mathscr{H}\frac{gh_n}{Q_n}\right)(x)\right]$. With the help of (31), (R1), and (30), we get

$$\begin{aligned}
J_1' &= -\int_{-1}^1 \widetilde{p}_n(x)\widetilde{g}_n(x)h_n(x) \left(\mathscr{H}\frac{\sqrt{w\varphi}}{v}\, k_n^1\right)(x)\, dx \\
&\le c \left\| \frac{g}{\sqrt{w\varphi}}\, \mathscr{H}\frac{\sqrt{w\varphi}}{v}\, k_n^1 \right\|_1 \le c \left\| \frac{\sqrt{w\varphi}}{v} \right\|_\infty \rho_+\left(\frac{g}{\sqrt{w\varphi}}\right)
\end{aligned}$$

and, by choosing $Q_n(x) \sim_{n,x} \sqrt{w(x)\varphi(x)}$ for $x \in A_n$ (see [26, Lemma 2.1]),

$$\begin{aligned}
J_1'' &\le c \int_{-1}^1 \frac{\sqrt{w(x)\varphi(x)}}{v(x)} \left(\mathscr{H}\frac{gh_n}{Q_n}\right)(x)\, k_n^2(x)\, dx \\
&= -c \int_{-1}^1 \frac{g(x)h_n(x)}{Q_n(x)} \left(\mathscr{H}\frac{\sqrt{w\varphi}}{v} k_n^2\right)(x)\, dx \\
&\le c \left\| \frac{g}{\sqrt{w\varphi}}\, \mathscr{H}\frac{\sqrt{w\varphi}}{v}\, k_n^2 \right\|_1 \le c \left\| \frac{\sqrt{w\varphi}}{v} \right\|_\infty \rho_+\left(\frac{g}{\sqrt{w\varphi}}\right).
\end{aligned}$$

Now, let us estimate J_3, the term J_2 can be handled analogously. We get

$$\begin{aligned}
J_3 &= \int_{\frac{x_{nn}+1}{2}}^1 h_n(y)g(y)\left(\mathscr{L}_n^w f\right)(y)\, dy = \sum_{k=1}^n \frac{f(x_{nk})}{p_n'(x_{nk})} \int_{\frac{x_{nn}+1}{2}}^1 \frac{p_n(y)}{y-x_{nk}}\, g(y)h_n(y)\, dy \\
&\overset{\text{(R2)}}{\le} c\|fv\|_\infty \sum_{k=1}^n \varDelta x_{nk} \frac{\sqrt{w(x_{nk})\varphi(x_{nk})}}{v(x_{nk})} \int_{\frac{x_{nn}+1}{2}}^1 \frac{|p_n(y)g(y)|}{y-x_{nk}}\, dy
\end{aligned}$$

Note that, due to the assumptions on w and u, $\alpha + \frac{1}{2} > 0$. Hence, in view of (32),

$$\frac{|p_n(y)|\sqrt{w(y)\varphi(y)}}{y - x_{nk}} \le \frac{c}{1 - x_{nk}}, \quad y \in \left[\frac{x_{nn}+1}{2}, 1\right],$$

since, for $y \in \left[\frac{x_{nn}+1}{2}, 1\right]$, we have $y - x_{nk} \ge \frac{1 - x_{nk}}{2}$. We conclude

$$J_3 \le c\|fv\|_\infty \sum_{k=1}^{n} \frac{\sqrt{w(x_{nk})\varphi(x_{nk})}}{v(x_{nk})(1 - x_{nk})} \int_{\frac{x_{nn}}{2}+1}^{1} \frac{|g(y)|\,dy}{\sqrt{w(y)\varphi(y)}}$$

$$\overset{\text{(R3)}}{\le} c\|fv\|_\infty \int_{-1}^{1} \frac{\sqrt{w(x)\varphi(x)}}{(1-x)v(x)}\,dx \left\|\frac{g}{\sqrt{w\varphi}}\right\|_1 \le c\|fv\|_\infty\, \rho_+\left(\frac{g}{\sqrt{w\varphi}}\right),$$

since $\frac{\alpha}{2} + \frac{1}{4} - \gamma - 1 > -1$. □

Lemma 8 *Let $v : I \longrightarrow [0, \infty)$ be a weight function as in Definition 1 and $R : I^2 \longrightarrow \mathbb{C}$ be a function such that $R_x \in \mathbf{C}_v$ for all $x \in I$, where $R_x(y) = R(x, y)$, and such that $R(x, y)v(y)$ is continuous on I^2. Then, for every $n \in \mathbb{N}$, there is a function $P_n(x, y)$ such that $P_{n,x}(y) = P_n(x, y)$ belongs to $\mathbf{P}_n$ for every $x \in I$ and $\lim_{n\to\infty} \sup\left\{|R(x, y) - P_n(x, y)|v(y) : (x, y) \in I^2\right\} = 0$.*

Proof Let $\varepsilon_n > 0$ and, for every $x \in I$, choose $P_{n,x} \in \mathbf{P}_n$ such that

$$\|(R_x - P_{n,x})v\|_\infty < E_n(R_x)_{v,\infty} + \varepsilon_n\,.$$

It remains to prove that $\lim_{n\to\infty} \sup\{E_n(R_x)_{v,\infty} : x \in I\} = 0$. If this is not the case, then there are an $\varepsilon > 0$ and $n_1 < n_2 < \ldots$ such that $E_{n_k}(R_{x_k})_{v,\infty} \ge 2\varepsilon$ for certain $x_k \in I$. Due to the compactness of I, we can assume that $x_k \longrightarrow x^*$ for $k \longrightarrow \infty$. In virtue of the continuity of $R(x, y)v(y)$, we can conclude that $\|(R_{x_k} - R_{x^*})v\|_\infty < \varepsilon$ for all $k \ge k_0$. Since $\|(R_{x_k} - p)v\|_\infty \ge 2\varepsilon$ for all $p \in \mathbf{P}_{n_k}$ and $k \in \mathbb{N}$, we obtain, for $p \in \mathbf{P}_{n_k}$ and $k \ge k_0$,

$$2\varepsilon \le \|(R_{x_k} - p)v\|_\infty \le \|(R_{x_k} - R_{x^*})v\|_\infty + \|(R_{x^*} - p)v\|_\infty < \varepsilon + \|(R_{x^*} - p)v\|_\infty$$

and, consequently, $\|(R_{x*} - p)v\|_\infty > \varepsilon$ for all $p \in \mathbf{P}_{n_k}$ and $k \in \mathbb{N}$, in contradiction to $R_{x*} \in \mathbf{C}_v$. □

Let us come back to the integral operator $\mathscr{K} : \mathbf{C}[-1, 1] \longrightarrow \mathbf{C}[-1, 1]$,

$$(\mathscr{K}f)(x) = \int_{-1}^{1} K(x, y)f(y)\,dy \tag{34}$$

and its product integration approximation $\mathscr{K}_n : \mathbf{C}[-1,1] \longrightarrow \mathbf{C}[-1,1]$,

$$(\mathscr{K}_n f)(x) = \sum_{k=1}^{n} \Lambda_{nk}(x) f(x_{nk}^w) = \int_{-1}^{1} H(x,y) \left(\mathscr{L}_n^w S_x f\right)(x)\, dx\,, \tag{35}$$

where $S_x(y) = S(x,y)$,

$$K(x,y) = H(x,y)S(x,y)\,, \quad \text{and} \quad \Lambda_{nk}(x) = S(x, x_{nk}^w) \int_{-1}^{1} H(x,y)\ell_{nk}^w(y)\, dy\,. \tag{36}$$

Proposition 5 *Consider* (34) *and* (35) *together with* (36) *in the Banach space* $\mathbf{C}[-1,1]$. *If the Jacobi weights* $w = w^{\alpha,\beta}$ *and* $v = v^{\gamma,\delta}$ *satisfy the conditions of Lemma* 7 *and if*

(a) $\dfrac{H_x}{\sqrt{w\varphi}} \in \mathbf{L}\log^+\mathbf{L}$ *for all* $x \in [-1,1]$, *where* $H_x(y) = H(x,y)$,

(b) $\sup\left\{\rho_+\left(\dfrac{H_x}{\sqrt{w\varphi}}\right) : -1 \le x \le 1\right\} < \infty$,

(c) $\lim\limits_{x \to x_0} \rho_+\left(\dfrac{H_x - H_{x_0}}{\sqrt{w\varphi}}\right) = 0$ *for all* $x_0 \in [-1,1]$,

(d) *the map* $[-1,1]^2 \longrightarrow \mathbb{C}$, $(x,y) \mapsto S(x,y)v(y)$ *is continuous with* $S_x \in \mathbf{C}_v$ *for all* $x \in [-1,1]$,

then the operators $\mathscr{K}_n$ *form a collectively compact sequence, which converges strongly to the operator* $\mathscr{K}$.

Proof At first we show that $\mathscr{K}_n$ converges strongly to $\mathscr{K}$. Indeed, for $f \in \mathbf{C}[-1,1]$, a function $P(x,y)$, which is a polynomial in y of degree less than n, and $P_x(y) = P(x,y)$, we have

$$\begin{aligned}
&|(\mathscr{K}_n f)(x) - (\mathscr{K} f)(x)| \\
&\le \int_{-1}^{1} \left|H(x,y)\left[\mathscr{L}_n^w(S_x f - P_x)\right](x)\right| dx + \int_{-1}^{1} |H(x,y)\left[S(x,y)f(y) - P(x,y)\right]|\, dx \\
&\le c\left[\rho_+\left(\frac{H_x}{\sqrt{w\varphi}}\right) + \left\|H_x v^{-1}\right\|_1\right] \|(S_x f - P_x)v\|_\infty\,,
\end{aligned}$$

where we took into account Lemma 7 and that condition (a) together with (33) implies $H_x v^{-1} \in \mathbf{L}^1(-1,1)$. Moreover, $\sup\left\{\left\|H_x v^{-1}\right\|_1 : -1 \le x \le 1\right\} < \infty$ due to condition (b). Thus,

$$\|\mathscr{K}_n f - \mathscr{K} f\|_\infty \le c \sup_{-1 \le x \le 1} \|(S_x f - P_x)v\|_\infty\,,$$

which proves the desired strong convergence by referring to Lemma 8. A consequence of this is that the set $\{\|\mathscr{K}_n f\|_\infty : f \in \mathbf{C}[-1,1],\ \|f\|_\infty \le 1\}$ is bounded. Furthermore, for $\|f\|_\infty \le 1$,

$$\begin{aligned}
&|(\mathscr{K}_n f)(x) - (\mathscr{K}_n f)(x_0)| \\
&\le \int_{-1}^{1} \left| H(x,y) \left[\mathscr{L}_n^w (S_x - S_{x_0}) f\right](y) \right| dy \\
&\qquad\qquad + \int_{-1}^{1} \left| [H(x,y) - H(x_0,y)] \left(\mathscr{L}_n^w S_{x_0} f\right)(y) \right| dy \\
&\overset{\text{Lemma 7}}{\le} c \left[\rho_+\left(\frac{H_x}{\sqrt{w\varphi}}\right) \|(S_x - S_{x_0})v\|_\infty + \rho_+\left(\frac{H_x - H_{x_0}}{\sqrt{w\varphi}}\right) \|S_{x_0} v\|_\infty \right].
\end{aligned}$$

Hence, due to (b)–(d), the set $\{\mathscr{K}_n f : f \in \mathbf{C}[-1,1],\ \|f\|_\infty \le 1\}$ is equicontinuous in each point $x_0 \in [-1,1]$, and so equicontinuous on $[-1,1]$. □

4.2 The Case of an Exponential Weight on $(0,\infty)$

Here, in case $w(x) = w_{\alpha,\beta}(x) = x^\alpha e^{-x^\beta}$, $0 < x < \infty$, $\alpha \ge 0$, $\beta > \frac{1}{2}$, we are going to prove results analogous to Lemma 7 and Proposition 5. Note that quadrature rules with such weights were introduced and investigated in [25]. Moreover, we mention that in [20] there are considered numerical methods and presented numerical results for Fredholm integral equations of second kind, basing on interpolation processes w.r.t. the nodes x_{nk}^w.

We again set $p_n(x) = p_n^w(x)$ and $x_{nk} = x_{nk}^w$ and, additionally, $x_{n,n+1} = a_n$, where $a_n = a_n(\sqrt{w}) \sim_n n^{\frac{1}{\beta}}$ is the Mhaskar-Rahmanov-Saff number associated with the weight $\sqrt{w(x)}$. Let us fix $\theta \in (0,1)$, set $n_\theta = \{\min k \in 1,\ldots,n : x_{nk} \ge \theta a_n\}$, and define, for a function $f : (0,\infty) \to \mathbb{C}$,

$$\mathscr{L}_n^* f = \sum_{k=1}^{n_\theta} f(x_{nk}) \ell_{nk}^*, \quad \ell_{nk}^*(x) = \frac{p_n^w(x)(a_n - x)}{p_n'(x_{nk})(x - x_{nk})(a_n - x_{nk})}. \tag{37}$$

Then, we have $\left(\mathscr{L}_n^* f\right)(x_{nk}) = f(x_{nk})$ for $k = 1,\ldots,n_\theta$ and $\left(\mathscr{L}_n^* f\right)(x_{nk}) = 0$ for $k = n_\theta + 1,\ldots,n+1$, as well as, for $\Delta x_{nk} = x_{nk} - x_{n,k-1}$, $k = 1,\ldots,n$, $x_{n0} = 0$,

(R4) $\sup\left\{ |p_n(x)| \sqrt{w(x)\sqrt{|a_n - x|x}} : 0 < x < \infty \right\} \le c < \infty$ with a constant $c \ne c(n)$ (see [14, 19]),

(R5) $\dfrac{1}{\left|p_n'(x_{nk})\right|} \sim_{n,k} \Delta x_{nk} \sqrt{w(x_{nk})\sqrt{(a_n - x_{nk})x_{nk}}}$, $k = 1,\ldots,n$ (see [14, 19]), (R6) for fixed $\ell \in \mathbb{N}$, there is a constant $c \neq c(n,p)$ such that (see [18])

$$\sum_{k=1}^{n_\theta} \Delta x_{nk}\,|p(x_{nk})| \le c\int_0^{\theta a_n} |p(x)|\,dx \quad \text{for all} \quad p \in \mathbf{P}_{\ell n}\,.$$

Remark 4 The constant on the right-hand side of (30) does not depend on the interval $[a,b]$, i.e., we have, for $-\infty < a < b < \infty$,

$$\left\|g\mathscr{H}_a^b f\right\|_1 + \left\|f\mathscr{H}_a^b g\right\|_1 \le c\|g\|_\infty \rho_+(f) \tag{38}$$

for all $g \in \mathbf{L}^\infty(a,b)$ and$f \in \mathbf{L}\log^+\mathbf{L}(a,b)$, where $c \neq c(f,g,a,b)$.

Indeed, if c_1 is the constant in (30) in case $a = 0$ and $b = 1$, then, by setting $x = \chi(t) = (b-a)t + a$ and $y = \chi(s)$,

$$\begin{aligned}
&\int_a^b \left|g(x)\int_a^b \frac{f(y)\,dy}{y-x}\right| dx + \int_a^b \left|f(x)\int_a^b \frac{g(y)\,dy}{y-x}\right| dx \\
&\quad = (b-a)\left[\int_0^1 \left|g(\chi(t))\int_0^1 \frac{f(\chi(s))\,ds}{s-t}\right| dt + \int_0^1 \left|f(\chi(t))\int_0^1 \frac{g(\chi(s))\,ds}{s-t}\right| dt\right] \\
&\quad \le c_1\|g\|_{\infty,[a,b]}\int_0^1 |f(\chi(t))|\left(1+\log^+|f(\chi(t))|\right) dt = c_1\|g\|_{\infty,[a,b]}\,\rho_{+,[a,b]}(f)\,.
\end{aligned}$$

Lemma 9 *Let $\psi(x) = \sqrt{x}$, $x \ge 0$ and $v(x) = (1+x)^\delta\sqrt{w(x)}$, $\delta \ge \frac{1}{4}$. Then, there is a constant $c \neq c(n,f,g)$ such that, for all functions $f : (0,\infty) \longrightarrow \mathbb{C}$ with $fv \in \mathbf{L}^\infty(0,\infty)$ and all g with $\dfrac{g}{\sqrt{w\psi}} \in \mathbf{L}\log^+\mathbf{L}(0,\infty)$,*

$$\left\|g\mathscr{L}_n^* f\right\|_{\mathbf{L}^1(0,\infty)} \le c\,\rho_+\left(\frac{g}{\sqrt{w\psi}}\right)\|fv\|_\infty\,.$$

Proof Write $\left\|g\mathscr{L}_n^* f\right\|_{\mathbf{L}^1(0,\infty)} = \left\|g\mathscr{L}_n^* f\right\|_{\mathbf{L}^1(0,2a_n)} + \left\|g\mathscr{L}_n^* f\right\|_{\mathbf{L}^1(2a_n,\infty)} =: J_1 + J_2$. Using (R5) we get, with $h_n(y) = \operatorname{sgn}\left[g(y)\left(\mathscr{L}_n^* f\right)(y)\right]$,

$$\begin{aligned}
J_1 &\le c\|fv\|_\infty \sum_{k=1}^{n_\theta} \Delta x_{nk}\frac{\sqrt{w(x_{nk})\psi(x_{nk})}}{v(x_{nk})(a_n-x_{nk})^{\frac{3}{4}}}\left|\int_0^{2a_n} \frac{p_n(y)(a_n-y)g(y)h_n(y)}{y-x_{nk}}\,dy\right| \\
&= c\|fv\|_\infty \sum_{k=1}^{n_\theta} \Delta x_{nk}\frac{(x_{nk})^{\frac{1}{4}}}{(1+x_{nk})^\delta(a_n-x_{nk})^{\frac{3}{4}}}\left|\int_0^{2a_n} \frac{p_n(y)(a_n-y)g(y)h_n(y)}{y-x_{nk}}\,dy\right| \\
&\le c\frac{\|fv\|_\infty}{(a_n)^{\frac{3}{4}}}\sum_{k=1}^{n_\theta} \Delta x_{nk}\,|G_n(x_{nk})|\ ,
\end{aligned}$$

where

$$G_n(t) = \int_0^{2a_n} \frac{p_n(y)(a_n-y)Q_n(y) - p_n(t)(a_n-t)Q_n(t)}{y-t}\,\frac{g(y)h_n(y)}{Q_n(y)}\,dy$$

and $Q_n \in \mathbf{P}_{\ell n}$ a polynomial positive on $(0, a_n)$ ($\ell \in \mathbb{N}$ fixed). Since $G_n \in \mathbf{P}_{(\ell+1)n}$, with the help of (R6) we can estimate

$$\begin{aligned} J_1 \le \frac{c\|fv\|_\infty}{(a_n)^{\frac{3}{4}}} \Bigg[&\int_0^{2a_n} \left|\left(\mathcal{H}_0^{2a_n} p_n(a_n-\cdot)gh_n\right)(x)\right| dx \\ &+ \int_0^{2a_n} \left| p_n(x)(a_n-x)Q_n(x)\left(\mathcal{H}_0^{2a_n}\frac{gh_n}{Q_n}\right)(x)\right| dx \Bigg] =: J_1' + J_1''\,. \end{aligned}$$

Defining $k_n^1(x) = \operatorname{sgn}\left[\left(\mathcal{H}_0^{2a_n} p_n(a_m-\cdot)gh_n\right)(x)\right]$ and using (31) and (R4), we obtain

$$\begin{aligned} J_1' &\le \frac{c\|fv\|_\infty}{(a_n)^{\frac{3}{4}}} \int_0^{2a_n} p_n(x)(a_n-x)g(x)h_n(x)\left(\mathcal{H}_0^{2a_n}k_n^1\right)(x)\,dx \\ &\le c\|fv\|_\infty \int_0^{2a_n} \frac{|g(x)|}{\sqrt{w(x)\psi(x)}}\left|\left(\mathcal{H}_0^{2a_n}k_n^1\right)(x)\right| dx \overset{(38)}{\le} c\|fv\|_\infty \rho_+\left(\frac{g}{\sqrt{w\psi}}\right). \end{aligned}$$

In order to estimate J_1'', we choose $Q_n \in \mathbf{P}_{\ell n}$ such that $Q_n(x) \sim_{n,x} \sqrt{w(x)\psi(x)}$ for $x \in (0, 2a_n)$ (see [29]). Then, due to (R4) and (31),

$$\begin{aligned} J_1'' &\le c\|fv\|_\infty k_n^2(x)\left(\mathcal{H}_0^{2a_n}\frac{gh_n}{Q_n}\right)(x)\,dx \\ &\le c\|fv\|_\infty \int_0^{2a_n}\left|\frac{g(x)}{\sqrt{w(x)\psi(x)}}\left(\mathcal{H}k_n^2\right)(x)\right| dx \overset{(38)}{\le} c\|fv\|_\infty \rho_+\left(\frac{g}{\sqrt{w\psi}}\right), \end{aligned}$$

where $k_n^2(x) = \operatorname{sgn}\left[\left(\mathcal{H}_0^{2a_n}\frac{gh_n}{Q_n}\right)(x)\right]$. Finally, let us consider J_2. Again taking into account (R5), we get

$$\begin{aligned} J_2 &\le c\|fv\|_\infty \sum_{k=1}^{n_\theta} \Delta x_{nk} \frac{\sqrt{w(x_{nk})\psi(x_{nk})}}{v(x_{nk})(a_n-x_{nk})^{\frac{3}{4}}}\left|\int_{2a_n}^{\infty}\frac{p_n(y)(a_n-y)g(y)h_n(y)}{y-x_{nk}}dy\right| \\ &= c\|fv\|_\infty \sum_{k=1}^{n_\theta} \Delta x_{nk} \frac{(x_{nk})^{\frac{1}{4}}}{(1+x_{nk})^{\delta}(a_n-x_{nk})^{\frac{3}{4}}}\left|\int_{2a_n}^{\infty}\frac{p_n(y)(a_n-y)g(y)h_n(y)}{y-x_{nk}}dy\right| \\ &\le \frac{c\|fv\|_\infty}{(a_n)^{\frac{3}{4}}} \sum_{k=1}^{n_\theta} \Delta x_{nk} \int_{2a_n}^{\infty} \frac{|p_n(y)|\sqrt{w(y)\sqrt{y(y-a_n)}}}{(a_n)^{\frac{1}{4}}}\left(\frac{y-a_n}{y-x_{nk}}\right)^{\frac{3}{4}}\frac{|g(y)|}{\sqrt{w(y)\psi(y)}}dy\,, \end{aligned}$$

where we also used that $y - x_{nk} \geq 2a_n - a_n = a_n$. Hence, in virtue of $\left(\dfrac{y - a_n}{y - x_{nk}}\right)^{\frac{3}{4}} \leq 1$ for $y > 2a_n$, $\displaystyle\sum_{k=1}^{n_\theta} \Delta x_{nk} \leq a_n$, and (R1),

$$J_2 \leq c\|fv\|_\infty \int_{2a_n}^\infty \frac{|g(y)|}{\sqrt{w(y)\psi(y)}}\, dy \leq c\|fv\|_\infty \rho_+\left(\frac{g}{\sqrt{w\psi}}\right).$$

□

Let us apply Lemma 9 to the integral operator $\mathscr{K} : \mathbf{C}[0,\infty] \longrightarrow \mathbf{C}[0,\infty]$,

$$(\mathscr{K}f)(x) = \int_0^\infty K(x,y)f(y)\, dy \tag{39}$$

and its product integration approximation $\mathscr{K}_n : \mathbf{C}[0,\infty] \longrightarrow \mathbf{C}[0,\infty]$,

$$(\mathscr{K}_n f)(x) = \sum_{k=1}^{n_\theta} \Lambda_{nk}^*(x) f(x_{nk}^w) = \int_0^\infty H(x,y)\left(\mathscr{L}_n^* S_x f\right)(x)\, dx\,, \tag{40}$$

where $w(x) = w_{\alpha,\beta}(x) = x^\alpha e^{-x^\beta}$, $\alpha > -1$, $\beta > \frac{1}{2}$, where $\mathscr{L}_n^*$ is defined in (37), and where $S_x(y) = S(x,y)$,

$$K(x,y) = H(x,y)S(x,y)\,, \quad \Lambda_{nk}^*(x) = S(x, x_{nk}^w)\int_0^\infty H(x,y)\ell_{nk}^*(y)\, dy\,. \tag{41}$$

Proposition 6 *Consider* (39) *and* (40) *together with* (41) *in the Banach space* $\mathbf{C}[0,\infty]$. *If* $v(x) = (1+x)^\delta\sqrt{w(x)}$ *with* $\delta \geq \frac{1}{4}$ *and if*

(a) $\dfrac{H_x}{\sqrt{w\psi}} \in \mathbf{L}\log^+\mathbf{L}(0,\infty)$ *for all* $x \in [0,\infty]$, *where* $H_x(y) = H(x,y)$,
(b) $\sup\left\{\rho_+\left(\dfrac{H_x}{\sqrt{w\varphi}}\right) : 0 \leq x \leq \infty\right\} < \infty$,
(c) $\displaystyle\lim_{d(x,x_0)\to 0} \rho_+\left(\frac{H_x - H_{x_0}}{\sqrt{w\psi}}\right) = 0$ *for all* $x_0 \in [0,\infty]$,
(d) *the map* $[0,\infty]^2 \longrightarrow \mathbb{C}$, $(x,y) \mapsto S(x,y)v(y)$ *is continuous with* $S_x \in \mathbf{C}_v$ *for all* $x \in [-1,1]$,

then the operators $\mathscr{K}_n$ *form a collectively compact sequence, which converges strongly to the operator* $\mathscr{K}$.

Proof We proceed in an analogous way as in the proof of Proposition 5. For $f \in \mathbf{C}[0,\infty]$ and a function $P(x,y) = P_x(y)$, which is a polynomial in y of degree less than n, we have

$$\begin{aligned} &|(\mathscr{K}_n f)(x) - (\mathscr{K} f)(x)| \\ &\le \int_{-1}^{1} \left| H(x,y) \left[\mathscr{L}_n^*(S_x f - P_x) \right](x) \right| dx + \int_0^\infty \left| H(x,y) \sum_{k=n_\theta+1}^{n+1} P_x(x_{nk}^w) \ell_{nk}^*(y) \right| dy \\ &\qquad + \int_{-1}^{1} |H(x,y) \left[S(x,y) f(y) - P(x,y) \right]| \, dx =: J_1 + J_1 + J_3 \,, \end{aligned}$$

By Lemma 9,

$$J_1 \le c\, \rho_+ \left(\frac{H_x}{\sqrt{w\varphi}} \right) \|(S_x f - P_x) v\|_\infty \; .$$

Condition (a) together with $\delta \ge \frac{1}{4}$ implies $H_x v^{-1} \in \mathbf{L}^1(-1,1)$, and hence

$$J_3 \le \left\| H_x v^{-1} \right\|_1 \|(S_x f - P_x) v\|_\infty \; .$$

Consequently, since we have also $\sup \left\{ \left\| H_x v^{-1} \right\|_1 : -1 \le x \le 1 \right\} < \infty$ by condition (b), we get

$$J_1 + J_3 \le c \sup_{-1 \le x \le 1} \|(S_x f - P_x) v\|_\infty \; . \tag{42}$$

To estimate J_2, we recall that (see [27, (2.3)])

$$\| P_n u \|_{\mathbf{L}^\infty(x_{n_\theta}, \infty)} \le c\, e^{-\widetilde{c}\, n} \| P_n u \|_\infty \quad \text{for} \quad P_n \in \mathbf{P}_{m(n)}$$

($m(n) < n$, $\lim_{n \to \infty} m(n) = \infty$) for some positive constants $c \ne c(n, P_n)$ and $\widetilde{c} \ne \widetilde{c}(n, P_n)$ and (cf. [23, pp. 362,373])

$$\sum_{k=n_\theta+1}^{n+1} \frac{v(x) \ell_{nk}^*(x)}{v(x_{nk}^w)} \le c\, n^\sigma$$

for some $\sigma > 0$ and $c \ne c(n,x)$. Thus,

$$J_2 \le c\, n^\sigma \left\| H_x v^{-1} \right\|_1 \| P_x v \|_{\mathbf{L}^\infty(x_{n\theta}, \infty)} \le c n^\sigma e^{-\widetilde{c}\, n} \left\| H_x v^{-1} \right\|_1 \| P_x v \|_\infty \; ,$$

where $P_x \in \mathbf{P}_{m(n)}$ can be chosen in such a way that $\sup\{ \| P_x v \|_\infty : x \in [0, \infty] \} < \infty$ (in view of Lemma 8). Hence, together with (42) we conclude the strong convergence of $\mathscr{K}_n$ to $\mathscr{K}$. Consequently, the set

$$\{\|\mathscr{K}_n f\|_\infty : f \in \mathbf{C}[0,\infty],\ \|f\|_\infty \le 1\}$$

is bounded. Furthermore, for $\|f\|_\infty \le 1$,

$$\begin{aligned}
&|(\mathscr{K}_n f)(x) - (\mathscr{K}_n f)(x_0)| \\
&\le \int_{-1}^{1} \left|H(x,y)\left[\mathscr{L}_n^*(S_x - S_{x_0})f\right](y)\right| dy \\
&\qquad\qquad + \int_{-1}^{1} \left|[H(x,y) - H(x_0,y)]\left(\mathscr{L}_n^* S_{x_0} f\right)(y)\right| dy \\
&\overset{\text{Lemma 9}}{\le} c\left[\rho_+\left(\frac{H_x}{\sqrt{w\varphi}}\right)\|(S_x - S_{x_0})v\|_\infty + \rho_+\left(\frac{H_x - H_{x_0}}{\sqrt{w\varphi}}\right)\|S_{x_0}v\|_\infty\right].
\end{aligned}$$

Hence, due to (b)–(d), the set $\{\mathscr{K}_n f : f \in \mathbf{C}[-1,1],\ \|f\|_\infty \le 1\}$ is equicontinuous in each point $x_0 \in [0,\infty]$, and so equicontinuous on $[0,\infty]$. □

Acknowledgements The second author was partially supported by University of Basilicata (local funds).

The third author was partially supported by University of Basilicata (local funds) and by National Group of Computing Science GNCS-INdAM.

References

1. Anselone, P.M.: Collectively compact and totally bounded sets of linear operators. J. Math. Mech. **17**, 613–621 (1967/1968)
2. Anselone, P.M.: Collectively compact approximations of integral operators with discontinuous kernels. J. Math. Anal. Appl. **22**, 582–590 (1968)
3. Anselone, P.M.: Collectively Compact Operator Approximation Theory and Applications to Integral Equations. Prentice-Hall, Englewood Cliffs, NJ (1971). With an appendix by Joel Davis, Prentice-Hall Series in Automatic Computation
4. Anselone, P.M., Palmer, T.W.: Collectively compact sets of linear operators. Pac. J. Math. **25**, 417–422 (1968)
5. Anselone, P.M., Palmer, T.W.: Spectral properties of collectively compact sets of linear operators. J. Math. Mech. **17**, 853–860 (1967/1968)
6. Anselone, P.M., Palmer, T.W.: Spectral analysis of collectively compact, strongly convergent operator sequences. Pac. J. Math. **25**, 423–431 (1968)
7. Atkinson, K.E.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge Monographs on Applied and Computational Mathematics, vol. 4. Cambridge University Press, Cambridge (1997)
8. Badkov, V.M.: Convergence in the mean and almost everywhere of Fourier series in polynomials that are orthogonal on an interval. Mat. Sb. (N.S.) **95**(137), 229–262, 327 (1974). English translation in Math. USSR Sbornik **24**(2), 223–256 (1974)

9. Chandler, G.A., Graham, I.G.: Product integration-collocation methods for noncompact integral operator equations. Math. Comput. **50**(181), 125–138 (1988)
10. Chandler, G.A., Graham, I.G.: The convergence of Nyström methods for Wiener-Hopf equations. Numer. Math. **52**(3), 345–364 (1988)
11. De Bonis, M.C., Mastroianni, G.: Nyström method for systems of integral equations on the real semiaxis. IMA J. Numer. Anal. **29**(3), 632–650 (2009)
12. Dick, J., Kritzer, P., Kuo, F.Y., Sloan, I.H.: Lattice-Nyström method for Fredholm integral equations of the second kind with convolution type kernels. J. Complex. **23**(4–6), 752–772 (2007)
13. Ditzian, Z., Totik, V.: Moduli of Smoothness. Springer Series in Computational Mathematics, vol. 9. Springer, New York (1987)
14. Kasuga, T., Sakai, R.: Orthonormal polynomials with generalized Freud-type weights. J. Approx. Theory **121**(1), 13–53 (2003)
15. Kress, R.: Linear Integral Equations. Applied Mathematical Sciences, vol. 82. Springer, Berlin (1989)
16. Kress, R.: Linear Integral Equations. Applied Mathematical Sciences, vol. 82, 2nd edn. Springer, New York (1999)
17. Kress, R.: Linear Integral Equations, Applied Mathematical Sciences, vol. 82, 3rd edn. Springer, New York (2014)
18. Laurita, C., Mastroianni, G.: L^p-convergence of Lagrange interpolation on the semiaxis. Acta Math. Hungar. **120**(3), 249–273 (2008)
19. Levin, A.L., Lubinsky, D.S.: Christoffel functions, orthogonal polynomials, and Nevai's conjecture for Freud weights. Constr. Approx. **8**(4), 463–535 (1992)
20. Mastroianni, G., Milovanović, G.V.: Some numerical methods for second-kind Fredholm integral equations on the real semiaxis. IMA J. Numer. Anal. **29**(4), 1046–1066 (2009)
21. Mastroianni, G., Monegato, G.: Nyström interpolants based on the zeros of Legendre polynomials for a noncompact integral operator equation. IMA J. Numer. Anal. **14**(1), 81–95 (1994)
22. Mastroianni, G., Notarangelo, I.: A Lagrange-type projector on the real line. Math. Comput. **79**(269), 327–352 (2010)
23. Mastroianni, G., Notarangelo, I.: Some Fourier-type operators for functions on unbounded intervals. Acta Math. Hungar. **127**(4), 347–375 (2010)
24. Mastroianni, G., Notarangelo, I.: A Nyström method for Fredholm integral equations on the real line. J. Integr. Equ. Appl. **23**(2), 253–288 (2011)
25. Mastroianni, G., Occorsio, D.: Some quadrature formulae with nonstandard weights. J. Comput. Appl. Math. **235**(3), 602–614 (2010)
26. Mastroianni, G., Russo, M.G.: Lagrange interpolation in weighted Besov spaces. Constr. Approx. **15**(2), 257–289 (1999)
27. Mastroianni, G., Szabados, J.: Polynomial approximation on the real semiaxis with generalized Laguerre weights. Stud. Univ. Babeş-Bolyai Math. **52**(4), 105–128 (2007)
28. Mastroianni, G., Totik, V.: Uniform spacing of zeros of orthogonal polynomials. Constr. Approx. **32**(2), 181–192 (2010)
29. Mastroianni, G., Notarangelo, I., Szabados, J.: Polynomial inequalities with an exponential weight on $(0, +\infty)$. Mediterr. J. Math. **10**(2), 807–821 (2013)
30. Mastroianni, G., Milovanović, G.V., Notarangelo, I.: Gaussian quadrature rules with an exponential weight on the real semiaxis. IMA J. Numer. Anal. **34**(4), 1654–1685 (2014)
31. Nevai, P.G.: Orthogonal polynomials. Mem. Am. Math. Soc. **18**(213), v+185 (1979)
32. Nevai, P.: Mean convergence of Lagrange interpolation. III. Trans. Am. Math. Soc. **282**(2), 669–698 (1984)
33. Nevai, P.: Hilbert transforms and Lagrange interpolation. Addendum to: "mean convergence of Lagrange interpolation. III" [Trans. Am. Math. Soc. **282**(2), 669–698 (1984); MR0732113 (85c:41009)]. J. Approx. Theory **60**(3), 360–363 (1990)
34. Sloan, I.H.: Analysis of general quadrature methods for integral equations of the second kind. Numer. Math. **38**(2), 263–278 (1981/1982)
35. Sloan, I.H.: Quadrature methods for integral equations of the second kind over infinite intervals. Math. Comput. **36**(154), 511–523 (1981)

Properties and Numerical Solution of an Integral Equation to Minimize Airplane Drag

Peter Junghanns, Giovanni Monegato, and Luciano Demasi

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In this paper, we consider an (open) airplane wing, not necessarily symmetric, for which the optimal circulation distribution has to be determined. This latter is the solution of a constraint minimization problem, whose (Cauchy singular integral) Euler-Lagrange equation is known. By following an approach different from a more classical one applied in previous papers, we obtain existence and uniqueness results for the solution of this equation in suitable weighted Sobolev type spaces. Then, for the collocation-quadrature method we propose to solve the equation, we prove stability and convergence and derive error estimates. Some numerical examples, which confirm the previous error estimates, are also presented. These results apply, in particular, to the Euler-Lagrange equation and the numerical method used to solve it in the case of a symmetric wing, which were considered in the above mentioned previous papers.

1 Introduction

In [3], the authors have studied the induced drag minimization problem for an open symmetric airplane wing. In particular, by applying a classical variational approach, they have derived the associated Euler-Lagrange (integral) equation (ELE) for

P. Junghanns
Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany
e-mail: peter.junghanns@mathematik.tu-chemnitz.de

G. Monegato (✉)
Dipartimento di Scienze Mathematiche, Politecnico di Torino, Turin, Italy
e-mail: giovanni.monegato@polito.it

L. Demasi
Department of Aerospace Engineering, San Diego State University, San Diego, CA, USA
e-mail: ldemasi@mail.sdsu.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_30

the unknown wing circulation distribution. In its final form, this equation is a Cauchy singular one, for which existence and uniqueness of its solution have been assumed. For the solution of this equation, the authors have proposed a discrete polynomial collocation method, based on Chebyshev polynomials and a corresponding Gaussian quadrature. Although the convergence of this method has been confirmed by an intensive numerical testing, no error estimates have been obtained.

Later, in [4], by using an alternative (weakly singular) formulation of the above ELE of Symm's type, existence and uniqueness of the optimal circulation has been proved under the assumption that the curve transfinite diameter is different from 1. The authors have however conjectured that this property should hold without this restriction.

In this paper, we consider an open wing (also called lifting curve), not necessarily symmetric, and examine the associated Euler-Lagrange equation. The main physical quantities and formulas, that are needed to describe the minimization problem, are briefly recalled in Sect. 2. Then, in Sect. 3, by following an approach different from the more classical one applied in [4], we obtain existence and uniqueness results in suitable weighted Sobolev type spaces, without requiring the above mentioned curve restriction. In Sect. 3, we derive an error estimate for the collocation-quadrature method we use to solve the ELE. In the case of a symmetric lifting line, the method naturally reduces to that proposed in [3]. Finally, in the last section, to test the efficiency of the proposed method and the error estimate previously obtained for it, we apply the method to four different open curves.

2 The Drag Minimization Problem

Following [3], we consider a wing defined by a single open lifting line ℓ in the cartesian y-z plane. This is represented by a curve ℓ, having parametric representation $\psi(t) = \begin{bmatrix} \psi_1(t) & \psi_2(t) \end{bmatrix}^T$, $|\psi'(t)| \neq 0$, $t \in [-1, 1]$. The corresponding arc length abscissa η is then defined by

$$\eta(t) = \int_0^t \left|\psi'(s)\right| \mathrm{d}s, \tag{1}$$

where, here and in the following, $|\cdot|$ denotes the Euclidean norm. This abscissa will run from $\eta(-1) = -b$ to $\eta(1) = a$ for some positive real numbers a and b. Moreover, $\eta(0) = 0$.

For simplicity, it is also assumed that the lifting line ℓ is sufficiently smooth. That is, it is assumed that $\psi_i(t)$, $i = 1, 2$, are continuous functions together with their first $m \geq 2$ derivatives on the interval $[-1, 1]$ (i.e., $\psi_i \in \mathbf{C}^m[-1, 1]$). A point on the lifting line, where the aerodynamic forces are calculated, is denoted by $\mathbf{r} = \begin{bmatrix} y \; z \end{bmatrix}^T \in \ell$, where $\mathbf{r} = \mathbf{r}(\eta) = \begin{bmatrix} y(\eta) & z(\eta) \end{bmatrix}^T = \begin{bmatrix} \psi_1(t) & \psi_2(t) \end{bmatrix}^T$ in correspondence with (1).

Using the arc length abscissa, the expressions of the wing *lift* L and *induced drag* D_{ind} are obtained in terms of the (unknown) *circulation* Γ:

$$L = L(\Gamma) = -\rho_\infty V_\infty \int_{-b}^{a} \tau_y(\eta)\Gamma(\eta)\,\mathrm{d}\eta \tag{2}$$

$$D_{\text{ind}} = D_{\text{ind}}(\Gamma) = -\rho_\infty \int_{-b}^{a} v_n(\eta)\Gamma(\eta)\mathrm{d}\eta. \tag{3}$$

The quantities ρ_∞ and V_∞ are given positive constants which indicate the density and freestream velocity, respectively. Further, $\tau_y(\eta) = y'(\eta)$ is the projection on the y-axis of the unit vector tangent to the lifting line, while v_n is the so-called *normalwash*. This latter has the representation

$$v_n(\eta) = \frac{1}{4\pi} ⨍_{-b}^{a} \Gamma'(\xi)\; Y(\eta,\xi)\;\mathrm{d}\xi, \quad -b < \eta < a, \tag{4}$$

where

$$Y(\eta,\xi) = -\frac{\mathrm{d}}{\mathrm{d}\eta}\ln|\mathbf{r}(\xi) - \mathbf{r}(\eta)|\,. \tag{5}$$

The function $Y(\eta,\xi)$, which is the *kernel* of the associated integral transform, has a singularity of order 1 when $\eta = \xi$, and the integral in (4) is a Cauchy principal value one.

The problem we need to solve is the minimization, in a suitable space, of the functional $D_{\text{ind}}(\Gamma)$, subject to the prescribed lift constraint

$$L(\Gamma) = L_{\text{pres}}. \tag{6}$$

In the next sections we will go back to the interval $[-1,1]$. For this, we use the notations

$$\Gamma_0(t) := \Gamma(\eta(t)), \quad \mathbf{r}_0(t) := \mathbf{r}(\eta(t)) = \psi(t),$$

and

$$Y_0(t,s) := -\frac{\mathrm{d}}{\mathrm{d}t}\ln|\mathbf{r}_0(s) - \mathbf{r}_0(t)| \tag{7}$$

for $t, s \in [-1,1]$, as well as the respective relations

$$\Gamma_0'(t) = \Gamma'(\eta(t))\eta'(t), \quad \psi_1'(t) = y'(\eta(t))\eta'(t), \quad Y_0(t,s) = Y(\eta(t),\eta(s))\eta'(t).$$

Condition (6) then takes the new form

$$\int_{-1}^{1} \psi_1'(t)\Gamma_0(t)\,\mathrm{d}t = \gamma_0 := -\frac{L_{\text{pres}}}{\rho_\infty V_\infty}. \tag{8}$$

Moreover, from (3) and (4) we get

$$D_{\text{ind}} = D_{\text{ind}}(\Gamma_0) = -\frac{\rho_\infty}{4\pi}\int_{-1}^{1} ⨍_{-1}^{1} Y_0(t,s)\Gamma_0'(s)\,\mathrm{d}s\,\Gamma_0(t)\,\mathrm{d}t. \tag{9}$$

3 The Euler-Lagrange Equation and Its Properties

For a Jacobi weight $\rho(t) := v^{\alpha,\beta}(t) = (1-t)^\alpha(1+t)^\beta$ with $\alpha, \beta > -1$, let us recall the definition of the Sobolev-type space (cf. [1]) $\mathbf{L}_\rho^{2,r} = \mathbf{L}_\rho^{2,r}(-1,1)$, $r \geq 0$. For this, by $\mathbf{L}_\rho^2 = \mathbf{L}_\rho^{2,0}$ we denote the real Hilbert space of all (classes of) quadratic summable (w.r.t. the weight $\rho(t)$) functions $f : (-1,1) \longrightarrow \mathbb{R}$ equipped with the inner product

$$\langle f, g\rangle_\rho := \int_{-1}^{1} f(t)g(t)\rho(t)\,\mathrm{d}t$$

and the norm $\|f\|_\rho = \sqrt{\langle f,f\rangle_\rho}$. In case $\alpha = \beta = 0$, i.e., $\rho \equiv 1$, we write $\langle f, g\rangle$ and $\|f\|$ instead of $\langle f, g\rangle_\rho$ and $\|f\|_\rho$, respectively. If $\left\{p_n^\rho : n \in \mathbb{N}_0\right\}$ denotes the system of orthonormal (w.r.t. $\rho(t)$) polynomials $p_n^\rho(t)$ of degree n with positive leading coefficient, then

$$\mathbf{L}_\rho^{2,r} := \left\{ f \in \mathbf{L}_\rho^2 : \sum_{n=0}^{\infty}(1+n)^{2r}\left|\langle f, p_n^\rho\rangle_\rho\right|^2 < \infty \right\}.$$

Equipped with the inner product

$$\langle f, g\rangle_{\rho,r} = \sum_{n=0}^{\infty}(1+n)^{2r}\langle f, p_n^\rho\rangle_\rho\langle g, p_n^\rho\rangle_\rho$$

and the norm $\|f\|_{\rho,r} := \sqrt{\langle f,f\rangle_{\rho,r}}$, the set $\mathbf{L}_\rho^{2,r}$ becomes a Hilbert space. Note that, in cases $\alpha = \beta = -\frac{1}{2}$ and $\alpha = \beta = \frac{1}{2}$, the spaces $\mathbf{L}_\rho^{2,r}$ were also introduced in [6, Section 1] with a slightly different notation. Let $\varphi(t) = \sqrt{1-t^2}$ and define

$$\mathbf{V} := \left\{ f = \varphi u : u \in \mathbf{L}_\varphi^{2,1} \right\}$$

together with $\langle f, g\rangle_{\mathbf{V}} := \langle \varphi^{-1}f, \varphi^{-1}g\rangle_{\varphi,1}$ and $\|f\|_{\mathbf{V}} := \left\|\varphi^{-1}f\right\|_{\varphi,1}$.

In the following, we denote by $\mathscr{D}$ the operator of generalized differentiation. An important property of this operator with respect to the $\mathbf{L}^{2,r}_\rho$ spaces is recalled in the next lemma, where we have set $\rho^{(1)}(t) = (1-t)^{1+\alpha}(1+t)^{1+\beta} = \rho(t)(1-t^2)$.

Lemma 1 ([2], Lemma 2.7; cf. also [1], Theorem 2.17) *For $r \geq 0$, the operator $\mathscr{D} : \mathbf{L}^{2,r+1}_\rho \longrightarrow \mathbf{L}^{2,r}_{\rho^{(1)}}$ is continuous.*

Lemma 2 *For $f \in \mathbf{V}$, we have $f \in \mathbf{C}[-1,1]$ with $f(\pm 1) = 0$.*

Proof Let $f = \varphi g$ with $g \in \mathbf{L}^{2,1}_\varphi$. Due to Lemma 1, $\mathscr{D}g \in \mathbf{L}^2_{\varphi^3}$. Hence, for $0 < t < 1$,

$$|g(t)| = \left| g(0) + \int_0^t (\mathscr{D}g)(s)\,\mathrm{d}s \right| \leq |g(0)| + \sqrt{\int_0^t (1-s^2)^{-\frac{3}{2}}\mathrm{d}s}\;\|\mathscr{D}g\|_{\varphi^3}$$

and

$$\int_0^t (1-s^2)^{-\frac{3}{2}}\mathrm{d}s \leq \int_0^t (1-s)^{-\frac{3}{2}}\mathrm{d}s = 2\left(\frac{1}{\sqrt{1-t}} - 1\right).$$

This implies $f(1) = \lim_{t\to 1-0}\varphi(t)g(t) = 0$. Analogously, one can show that also $f(-1) = 0$ holds for $f \in \mathbf{V}$. □

Now, the problem we aim to solve (cf. [3]) is the following:

(P) *Find a function $\Gamma_0 \in \mathbf{V}$, which minimizes the functional* (cf. (9))

$$F(\Gamma_0) := -\int_{-1}^1 \fint_{-1}^1 Y_0(t,s)\Gamma_0'(s)\,\mathrm{d}s\,\Gamma_0(t)\,\mathrm{d}t$$

subject to the condition (cf. (8)) $\langle \psi_1', \Gamma_0\rangle = \gamma_0$.

If we define the linear operator

$$(\mathscr{A}f)(t) = -\frac{1}{\pi}\fint_{-1}^1 Y_0(t,s)f'(s)\,\mathrm{d}s, \quad -1 < s < 1, \tag{10}$$

then the problem can be reformulated as follows:

(P) *Find a function $\Gamma_0 \in \mathbf{V}$, which minimizes the functional $F(\Gamma_0) := \langle \mathscr{A}\Gamma_0, \Gamma_0\rangle$* on $\mathbf{V}$ *subject to the condition $\langle \psi_1', \Gamma_0\rangle = \gamma_0$.*

The formulation of this problem is correct, which can be seen from the following lemma.

Lemma 3 *If $\psi_j \in \mathbf{C}^m[-1,1]$ for some integer $m \geq 2$ and $|\psi'(t)| \neq 0$ for $t \in [-1,1]$, then the function $Y_0(t,s)$ has the representation*

$$Y_0(t,s) = \frac{1}{s-t} + K(t,s), \tag{11}$$

where the function $K : [-1,1]^2 \longrightarrow \mathbb{R}$ is continuous together with its partial derivatives $\dfrac{\partial^{j+k}K(t,s)}{\partial t^j \partial s^k}$*,* $k, j \in \mathbb{N}_0$*,* $j+k \le m-2$*.*

Proof Note that, by definition,

$$Y_0(t,s) = \frac{[\psi_1(s)-\psi_1(t)]\psi_1'(t) + [\psi_2(s)-\psi_2(t)]\psi_2'(t)}{[\psi_1(s)-\psi_1(t)]^2 + [\psi_2(s)-\psi_2(t)]^2}.$$

Hence,

$$K(t,s) = Y_0(t,s) - \frac{1}{s-t} = \frac{\Omega(t,s)}{\Psi(t,s)},$$

where

$$\Omega(t,s) = G_1(t,s)g_1(t,s) + G_2(t,s)g_2(t,s), \quad \Psi(t,s) = [g_1(t,s)]^2 + [g_2(t,s)]^2,$$

$$g_j(t,s) = \frac{\psi_j(s)-\psi_j(t)}{s-t} = \int_0^1 \psi_j'\big(sv + t(1-v)\big)\,\mathrm{d}v$$

$$G_j(t,s) = \frac{\psi_j'(t) - g_j(t,s)}{s-t} = \int_0^1 \psi_j''\big(sv + t(1-v)\big)(1-v)\,\mathrm{d}v,$$

and the assertion of the lemma follows by taking into account $\Psi(t,s) \ne 0$ for all $(t,s) \in [-1,1]^2$. □

Lemma 4 *The operator $\mathscr{A} : \mathbf{V} \longrightarrow \mathbf{L}^2_\varphi$ is a linear and bounded one and, consequently, $\langle \mathscr{A}f, f\rangle$ is well defined for all $f \in \mathbf{V}$.*

Proof Let $U_n = p_n^\varphi$ and $T_n = p_n^{\varphi^{-1}}$. Then, for $f \in \mathbf{V}$,

$$\begin{aligned}\|\mathscr{D}f\|_\varphi^2 &= \sum_{n=0}^\infty \left|\langle \mathscr{D}f, \varphi^{-1}T_n\rangle_\varphi\right|^2 = \sum_{n=0}^\infty |\langle \mathscr{D}f, T_n\rangle|^2 = \sum_{n=1}^\infty |\langle f, n\,U_{n-1}\rangle|^2 \\ &= \sum_{n=0}^\infty (1+n)^2 \left|\langle \varphi^{-1}f, U_n\rangle_\varphi\right|^2 = \left\|\varphi^{-1}f\right\|_{\varphi,1}^2 = \|f\|_{\mathbf{V}}^2,\end{aligned}$$

i.e., $\mathscr{D}f \in \mathbf{L}^2_\varphi$. By relation (11), the operator $\mathscr{A}$ defined in (10) can be written in the form $\mathscr{A} = -(\mathscr{S} + \mathscr{K})\mathscr{D}$ with

$$(\mathscr{S}f)(t) := \frac{1}{\pi}\,⨍_{-1}^1 \frac{f(s)\,\mathrm{d}s}{s-t}, \quad (\mathscr{K}f)(t) = \frac{1}{\pi}\int_{-1}^1 K(t,s)f(s)\,\mathrm{d}s, \quad -1 < t < 1.$$

It is well known that the Cauchy singular integral operator $\mathscr{S} : \mathbf{L}^2_\varphi \longrightarrow \mathbf{L}^2_\varphi$ is bounded [5, Theorem 4.1] and that $\mathscr{K} : \mathbf{L}^2_\varphi \longrightarrow \mathbf{L}^2_\varphi$ is compact. Consequently, for $f = \varphi u \in \mathbf{V}$ we have that $\langle \mathscr{A} f, f \rangle = \langle \mathscr{A} f, u \rangle_\varphi$ is a finite number, since both $\mathscr{A} f$ and u belong to $\mathbf{L}^2_\varphi$. □

In the following lemma we give a representation of the operator $\mathscr{A}$ defined in (10), which is crucial for our further investigations. From this representation, it is seen that the operator $\mathscr{A}$ is an example of a hypersingular integral operator in the sense of Hadamard (cf., for example, the representation of Prandtl's integro-differential operator in [2, Section 1], where $\mathbf{r}_0(t) = t$ and $\mathscr{B}$ is equal to the Cauchy singular integral operator).

Lemma 5 *For all $f \in \mathbf{V}$, the relation*

$$\mathscr{A} f = \mathscr{D} \mathscr{B} f \tag{12}$$

holds true, where

$$(\mathscr{B} f)(t) = \frac{1}{\pi} \int_{-1}^{1} \ln |\mathbf{r}_0(s) - \mathbf{r}_0(t)| f'(s) \, \mathrm{d}s \tag{13}$$

and where $\mathscr{D}$ is the operator of generalized differentiation already used in the proof of Lemma 2.

Proof Since $\mathscr{D} : \mathbf{V} \longrightarrow \mathbf{L}^2_\varphi$ is an isometrical mapping (cf. the proof of Lemma 4), it suffices to show that $-(\mathscr{S} + \mathscr{K})g = \mathscr{D} \mathscr{B}_0 g$ is valid for all $g \in \mathbf{L}^2_\varphi$, where

$$(\mathscr{B}_0 g)(t) = \frac{1}{\pi} \int_{-1}^{1} \ln |\mathbf{r}_0(s) - \mathbf{r}_0(t)| \, g(s) \, \mathrm{d}s.$$

Since

$$Z_0(t,s) := \ln |\mathbf{r}_0(s) - \mathbf{r}_0(t)| = \ln |s - t| + K_0(t,s) \tag{14}$$

with a function $K_0 : [-1,1]^2 \longrightarrow \mathbb{R}$ which is continuous together with $\dfrac{\partial K_0(t,s)}{\partial t} = -K(t,s)$, the operator $\mathscr{B}_0 : \mathbf{L}^2_\varphi \longrightarrow \mathbf{L}^{2,1}_{\varphi^{-1}}$ is bounded (see [7, Section 5] and [1, Lemma 4.2]). Moreover, $\mathscr{D} : \mathbf{L}^{2,1}_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_\varphi$ is continuous [2, Lemma 2.7], such that on the one hand, the operator $\mathscr{D} \mathscr{B}_0 : \mathbf{L}^2_\varphi \longrightarrow \mathbf{L}^2_\varphi$ is linear and bounded. On the other hand, the operator $\mathscr{S} + \mathscr{K} : \mathbf{L}^2_\varphi \longrightarrow \mathbf{L}^2_\varphi$ is also linear and bounded. Thus, it remains to prove that

$$⨍_{-1}^{1} Y_0(t,s) g(s) \, \mathrm{d}s = -\frac{d}{dt} \int_{-1}^{1} Z_0(t,s) g(s) \, \mathrm{d}s, \quad -1 < t < 1 \tag{15}$$

for all g from a linear and dense subset of $\mathbf{L}^2_\varphi$. For this, let $g : [-1,1] \longrightarrow \mathbb{R}$ be a continuously differentiable function and consider

$$\psi_0(t) := \int_{-1}^{1} Z_0(t,s)g(s)\,\mathrm{d}s = \lim_{\varepsilon\to+0} \psi_\varepsilon(t)$$

with $\psi_\varepsilon(t) := \left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) Z_0(t,s)g(s)\,\mathrm{d}s$. For every $t \in (-1,1)$, it follows

$$\begin{aligned}
\psi_\varepsilon'(t) &= -\left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) Y_0(t,s)g(s)\,\mathrm{d}s + Z_0(t,t-\varepsilon)g(t-\varepsilon) - Z_0(t,t+\varepsilon)g(t+\varepsilon)\\
&= -\left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) Y_0(t,s)g(s)\,\mathrm{d}s + \ln\varepsilon[g(t-\varepsilon) - g(t+\varepsilon)]\\
&\qquad + K_0(t,t-\varepsilon)g(t-\varepsilon) - K_0(t,t+\varepsilon)g(t+\varepsilon)\\
&\longrightarrow -\,⨍_{-1}^{1} Y_0(t,s)g(s)\,\mathrm{d}s \quad \text{if} \quad \varepsilon \longrightarrow +0,
\end{aligned}$$

where, as before, the last integral is defined in the Cauchy principal value sense. For every $\delta \in (0,1)$, this convergence is uniform w.r.t. $t \in [-1+\delta, 1-\delta]$. Indeed, for $0 < \varepsilon_1 < \varepsilon_2 < \delta$ and for

$$g_\varepsilon(t) := \left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) Y_0(t,s)g(s)\,\mathrm{d}s,$$

we have

$$\begin{aligned}
g_{\varepsilon_1}(t) - g_{\varepsilon_2}(t) &= \int_{t-\varepsilon_2}^{t-\varepsilon_1} Y_0(t,s)g(s)\,\mathrm{d}s + \int_{t+\varepsilon_1}^{t+\varepsilon_2} Y_0(t,s)g(s)\,\mathrm{d}s\\
&= \int_{t-\varepsilon_2}^{t-\varepsilon_1} Y_0(t,s)[g(s)-g(t)]\,\mathrm{d}s + \int_{t+\varepsilon_1}^{t+\varepsilon_2} Y_0(t,s)[g(s)-g(t)]\,\mathrm{d}s\\
&\qquad + g(t)\left[\int_{t-\varepsilon_2}^{t-\varepsilon_1} Y_0(t,s)\,\mathrm{d}s + \int_{t+\varepsilon_1}^{t+\varepsilon_2} Y_0(t,s)\,\mathrm{d}s\right]\\
&= \left(\int_{t-\varepsilon_2}^{t-\varepsilon_1} + \int_{t+\varepsilon_1}^{t+\varepsilon_2}\right)[1+(s-t)K(t,s)]\,\frac{g(s)-g(t)}{s-t}\,\mathrm{d}s\\
&\qquad + g(t)\left(\int_{t-\varepsilon_2}^{t-\varepsilon_1} + \int_{t+\varepsilon_1}^{t+\varepsilon_2}\right) K(t,s)\,\mathrm{d}s.
\end{aligned}$$

Consequently,

$$|g_{\varepsilon_1}(t) - g_{\varepsilon_2}(t)| \leq M(\varepsilon_2 - \varepsilon_1)$$

with $M = 2\,(M_1\,\|g'\|_\infty + M_2\|g\|_\infty)$, $M_1 = 1 + \max\{|(s-t)K(t,s)| : -1 \leq s,t \leq 1\}$, and $M_2 = \max\{|K(t,s)| : -1 \leq s,t \leq 1\}$. This uniform convergence implies that $\psi_0(t)$ is differentiable for all $t \in (-1,1)$, where

$$\psi_0'(t) = \frac{d}{dt}\left[\lim_{\varepsilon\to+0}\psi_\varepsilon(t)\right] = \lim_{\varepsilon\to+0}\psi_\varepsilon'(t) = -\fint_{-1}^{1} Y_0(t,s)g(s)\,\mathrm{d}s$$

and $\psi_0'(t) = \dfrac{d}{dt}\displaystyle\int_{-1}^{1} Z_0(t,s)g(s)\,\mathrm{d}s$, and (15) is proved. □

Lemma 6 *The operator* $\mathscr{A} : \mathbf{V} \longrightarrow \mathbf{L}^2_\varphi$ *is symmetric and positive, i.e.* $\forall f,g \in \mathbf{V}$, $\langle \mathscr{A}f, g\rangle = \langle f, \mathscr{A}g\rangle$ *and,* $\forall f \in \mathbf{V}\setminus\{0\}$, $\langle \mathscr{A}f,f\rangle > 0$.

Proof Using relation (12), Lemma 2, partial integration, and Fubini's theorem, we get, for all $f,g \in \mathbf{V}$,

$$\langle \mathscr{A}f, g\rangle = -\frac{1}{\pi}\int_{-1}^{1}\int_{-1}^{1} f'(s)\ln|\mathbf{r}_0(s) - \mathbf{r}_0(t)|\,\mathrm{d}s\,g'(t)\,\mathrm{d}t = \langle f, \mathscr{A}g\rangle. \tag{16}$$

Hence,

$$\langle \mathscr{A}f,f\rangle = \frac{1}{\pi}\int_{-1}^{1}\int_{-1}^{1}\ln\frac{1}{|\mathbf{r}_0(s) - \mathbf{r}_0(t)|}f'(s)f'(t)\,\mathrm{d}s\,\mathrm{d}t$$

corresponds to the logarithmic energy of the function f', where $\displaystyle\int_{-1}^{1} f'(t)\,\mathrm{d}t = 0$ due to Lemma 2. Consequently (see [8], Section I.1, and in particular Lemma 1.8), $\langle \mathscr{A}f,f\rangle$ is positive if $f' \neq 0$ a.e. Hence, $\langle \mathscr{A}f,f\rangle = 0$ implies $f'(t) = 0$ for almost all $t \in (-1,1)$ and, due to $f(\pm 1) = 0$, also $f(t) = 0$ for all $t \in [-1,1]$. □

For $\gamma \in \mathbb{R}$, define the (affine) manifold $\mathbf{V}_\gamma := \left\{f \in \mathbf{V} : \langle f, \psi_1'\rangle = \gamma\right\}$. The following result then holds.

Proposition 1 *The element* $\Gamma_0^* \in \mathbf{V}_{\gamma_0}$ *is a solution of Problem* (P) *if and only if there is a number* $\beta \in \mathbb{R}$ *such that*

$$\mathscr{A}\Gamma_0^* = \beta\psi_1'. \tag{17}$$

This solution is unique, if it exists.

Proof Assume that $\varGamma_0^* \in \mathbf{V}_{\gamma_0}$ and $F(\varGamma_0^*) = \min\left\{F(\varGamma_0) : \varGamma_0 \in \mathbf{V}_{\gamma_0}\right\}$. This implies $G'(0) = 0$ for $G(\alpha) = F(\varGamma_0^* + \alpha f)$ and for all $f \in \mathbf{V}_0 \setminus \{0\}$. Since

$$G(\alpha) = F(\varGamma_0^*) + 2\alpha\langle \mathscr{A}\varGamma_0^*, f\rangle + \alpha^2\langle f, f\rangle \tag{18}$$

and

$$G'(\alpha) = 2\langle \mathscr{A}\varGamma_0^*, f\rangle + 2\alpha\langle f, f\rangle,$$

this condition gives $\langle \mathscr{A}\varGamma_0^*, g\rangle_\varphi = 0$ for all $g \in \mathbf{L}_\varphi^{2,1}$ with $\langle g, \psi_1'\rangle_\varphi = 0$, which is equivalent to (17). On the other hand, if $\varGamma_0^* \in \mathbf{V}_{\gamma_0}$ and $\beta \in \mathbb{R}$ fulfil (17) and if $f \in \mathbf{V}_0 \setminus \{0\}$, then we get from (18) for $\alpha = 1$

$$\begin{aligned} F(\varGamma_0^* + f) &= F(\varGamma_0^*) + 2\langle \mathscr{A}\varGamma_0^*, f\rangle + \langle f, f\rangle \\ &= F(\varGamma_0^*) + 2\langle \mathscr{A}\varGamma_0^* - \beta\psi_1', f\rangle + \langle f, f\rangle \\ &= F(\varGamma_0^*) + \langle f, f\rangle > F(\varGamma_0^*), \end{aligned}$$

which shows the uniqueness of the solution (if it exists). □

Remark 1 Using relation (12), Eq. (17) can be written equivalently as

$$\mathscr{B}\varGamma_0^* = \beta\psi_1 + \gamma, \quad \varGamma_0^* \in \mathbf{V}_{\gamma_0},\ \beta, \gamma \in \mathbb{R}. \tag{19}$$

Moreover, by applying partial integration to the integral in (13) and taking into account $f(\pm 1) = 0$ for $f \in \mathbf{V}$ (see Lemma 2), we get

$$\begin{aligned} (\mathscr{B}f)(t) &= \lim_{\varepsilon \to +0} \frac{1}{\pi}\left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) \ln|\mathbf{r}_0(s) - \mathbf{r}_0(t)|\, f'(s)\, \mathrm{d}s \\ &= \lim_{\varepsilon \to +0} \frac{1}{\pi}\big[f(t-\varepsilon)\ln|\mathbf{r}_0(t-\varepsilon) - \mathbf{r}_0(t)| - f(t+\varepsilon)\ln|r_0(t+\varepsilon) - \mathbf{r}_0(t)|\big] \\ &\quad - \lim_{\varepsilon \to +0} \frac{1}{\pi}\left(\int_{-1}^{t-\varepsilon} + \int_{t+\varepsilon}^{1}\right) f(s)\,\frac{\mathrm{d}}{\mathrm{d}s}\ln|\mathbf{r}_0(s) - \mathbf{r}_0(t)|\, \mathrm{d}s \\ &\overset{(7)}{=} \frac{1}{\pi} \fint_{-1}^{1} Y_0(s,t) f(s)\, \mathrm{d}s \end{aligned}$$

Hence, we obtain the identity

$$\mathscr{B}f = \mathscr{A}_0 f \quad \forall f \in \mathbf{V}, \tag{20}$$

where (cf. (11))

$$(\mathscr{A}_0 f)(t) = \frac{1}{\pi} ⨍_{-1}^{1} Y_0(s,t) f(s)\,\mathrm{d}s = -(\mathscr{S} f)(t) + (\mathscr{K}_0 f)(t)$$

with

$$(\mathscr{K}_0 f)(t) = \frac{1}{\pi} \int_{-1}^{1} K(s,t) f(s)\,\mathrm{d}s. \tag{21}$$

Note that Eq. (17), hence its equivalent representation one obtains from (19) and equality (20), defines the Euler-Lagrange equation for the drag minimization problem.

The following Lemma is a consequence of the well-known relation

$$\mathscr{S} \varphi p_n^{\varphi} = -p_{n+1}^{\varphi^{-1}}, \quad n \in \mathbb{N}_0. \tag{22}$$

Lemma 7 *The operators* $\mathscr{S} : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1},0}$ *and* $\mathscr{S} : \varphi \mathbf{L}^{2,r}_{\varphi} \longrightarrow \mathbf{L}^{2,r}_{\varphi^{-1},0}$, $r > 0$, *are invertible, where* $\mathbf{L}^2_{\rho,0} = \left\{ f \in \mathbf{L}^2_{\rho} : \langle f, 1 \rangle_{\rho} = 0 \right\}$ *and* $\mathbf{L}^{2,r}_{\rho,0} = \mathbf{L}^{2,r}_{\rho} \cap \mathbf{L}^2_{\rho,0}$.

In the following proposition we discuss the solvability of (19).

Proposition 2 *Assume that* $\psi_j \in \mathbf{C}^3[-1,1]$. *Then,*

(a) *the operator* $\mathscr{A}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ *has a trivial null space, i.e.,*

$$N(\mathscr{A}_0) = \left\{ f \in \mathbf{L}^2_{\varphi^{-1}} : \mathscr{A}_0 f = 0 \right\} = \{0\};$$

(b) *if* $\psi_1(t)$ *is not a constant function, then Eq.* (19) *possesses a unique solution* $(\Gamma_0^*, \beta, \gamma) \in \mathbf{V}_{\gamma_0} \times \mathbb{R}^2$*;*

(c) *if* $\psi_1(t)$ *is not a constant function, then Problem* (P) *is uniquely solvable.*

Proof Let $f_0 \in \mathbf{L}^2_{\varphi^{-1}}$ and $\mathscr{A}_0 f_0 = 0$. Hence, $\mathscr{S} f_0 = \mathscr{K}_0 f_0 \in \mathbf{C}^1[-1,1] \subset \mathbf{L}^{2,1}_{\varphi^{-1}}$, due to Lemma 3. By Lemma 7, we get $\mathscr{K}_0 f_0 \in \mathbf{L}^{2,1}_{\varphi^{-1},0}$ and, consequently, $f_0 \in \varphi \mathbf{L}^{2,1}_{\varphi} = \mathbf{V}$. On the other hand, due to (12) and (20) (cf. also the proof of Lemma 6), we have

$$0 < \langle \mathscr{A} f, f \rangle = -\langle \mathscr{A}_0 f, \mathscr{D} f \rangle \quad \forall f \in \mathbf{V} \setminus \{0\}.$$

This implies $f_0 = 0$, and (a) is proved.

Since, by Lemma 7, the operator $\mathscr{S} : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ is Fredholm with index -1 and since, due to the continuity of the function $K(s,t)$ (cf. Lemma 3), the operator $\mathscr{K}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ is compact, also the operator $\mathscr{A}_0 = -\mathscr{S} + \mathscr{K}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ is Fredholm with index -1. Hence, we conclude that the codimension of the image

$$R(\mathscr{A}_0) = \left\{ \mathscr{A}_0 f : f \in \mathbf{L}^2_{\varphi^{-1}} \right\}$$

is equal to 1. Hence, the intersection $\mathbf{W}_1 := R(\mathscr{A}_0) \cap \{\beta\psi_1 + \gamma : \beta, \gamma \in \mathbb{R}\}$ is at least one-dimensional. If this dimension is equal to 1 and if $\mathbf{W}_1 = \operatorname{span}\{\psi_0\}$, then there is a unique $\Gamma_0 \in \mathbf{L}^2_{\varphi^{-1}}$, such that $\mathscr{A}_0\Gamma_0 = \psi_0$. Again using Lemma 3, we get $\mathscr{S}\Gamma_0 = \mathscr{K}_0\Gamma_0 - \psi_0 \in \mathbf{C}^1[-1, 1]$ and, consequently, $\Gamma_0 \in \mathbf{V}$. We show that $\langle \Gamma_0, \psi_1' \rangle \neq 0$. If this is not the case, then, because of Proposition 1 and Remark 1, Problem (P) has only a solution for $\gamma_0 = 0$. But, this (unique) solution is identically zero. This implies $\Gamma_0 = 0$ in contradiction to $\psi_0 \neq 0$. Hence, $\Gamma_0^* = \dfrac{\gamma_0}{\langle \Gamma_0, \psi_1' \rangle}\,\Gamma_0$ is the solution of (19) with $\beta\psi_1 + \gamma = \dfrac{\gamma_0}{\langle \Gamma_0, \psi_1' \rangle}\,\psi_0$.

To complete the proof of (b), finally we show that $\dim \mathbf{W}_1 = 2$ is not possible. Indeed, in that case $\mathbf{W}_1 = \operatorname{span}\{\psi_1, \psi_0\}$ with $\psi_0(t) = 1$, and we have (cf. the previous considerations) two linearly independent solutions $\Gamma_0^j \in \mathbf{V}$ of $\mathscr{A}\Gamma_0^j = \psi_j$ with $\langle \Gamma_0^j, \psi_1' \rangle \neq 0, j = 0, 1$. Hence, $\Gamma_0^{j,*} = \dfrac{\gamma_0}{\langle \Gamma_0^j, \psi_1' \rangle}\,\Gamma_0^j, j = 1, 2$, are two linearly independent solutions of (19) and, in virtue of Proposition 1, also of Problem (P) in contradiction to the uniqueness of a solution of (P).

Assertion (c) is an immediate consequence of (b), together with Proposition 1 and Remark 1. □

In the case of the wing problem examined in [3] and [4], where the line ℓ is symmetric in the $x - z$ plane with respect to the z-axis and where it cannot be a vertical segment, we note that the problem formulation (19) can be simplified significantly. In particular, for it, the following result holds (see also the associated numerical method, described at the end of the next section).

Corollary 1 *If the lifting line ℓ is symmetric w.r.t. the z-axis, i.e., $\psi_1(-t) = -\psi_1(t)$ and $\psi_2(-t) = \psi_2(t)$, and if $\psi_1(t)$ is not constant, then the unique solution $\Gamma_0^* \in \mathbf{V}$ of Problem (P) is an even function. Moreover, in (19) we have $\gamma = 0$.*

Proof In virtue of Proposition 1, Remark 1, and Proposition 2, there exist unique $\beta, \gamma \in \mathbb{R}$ such that $(\Gamma_0^*, \beta, \gamma) \in \mathbf{V}_{\gamma_0} \times \mathbb{R}^2$ is the unique solution of (19). By assumption, $Z_0(-t, -s) = Z_0(t, s)$ (cf. (14)). Set $g(t) = \Gamma_0^*(-t)$. From (19) it follows

$$\beta\psi_1(t) - \gamma = -\beta\psi_1(-t) - \gamma = -\frac{1}{\pi}\int_{-1}^{1} Z_0(-t, s)(\Gamma_0^*)'(s)\,\mathrm{d}s$$

$$= -\frac{1}{\pi}\int_{-1}^{1} Z_0(-t, -s)(\Gamma_0^*)'(-s)\,\mathrm{d}s = \frac{1}{\pi}\int_{-1}^{1} Z_0(t, s)g'(s)\,\mathrm{d}s,$$

which means that $(g, \beta, -\gamma) = (\Gamma_0^*, \beta, \gamma)$. □

Remark 2 Note that $\gamma = 0$ implies that the problem defined by (19) can be reformulated as follows: *Find* $\Gamma_0^* \in \mathbf{V}$ *and* $\beta \in \mathbb{R}$ *such that*

$$\mathscr{B}\Gamma_0^* = \beta\psi_1, \quad \langle \psi_1', \Gamma_0^* \rangle = \gamma_0. \tag{23}$$

This is only apparently a system of two equations. Indeed, since we must necessarily have $\beta \neq 0$, by introducing the new unknown $\overline{\Gamma}_0^* = \Gamma_0^*/\beta$ we obtain

$$\mathscr{B}\overline{\Gamma}_0^* = \psi_1, \quad \beta\langle \psi_1', \overline{\Gamma}_0^* \rangle = \gamma_0. \tag{24}$$

which is exactly the decoupled system that has been derived in [3]. Solving the first equation we obtain $\overline{\Gamma}_0^*$, from the second equation we get the value of β, and finally we find the solution Γ_0^*.

4 A Collocation-Quadrature Method

Here, we describe a numerical procedure for the approximate solution of Eq. (19). For this, we write this equation in the form (cf. (20), (21))

$$\mathscr{A}_0 f = \beta\psi_1 + \gamma, \quad (f, \beta, \gamma) \in \mathbf{V}_{\gamma_0} \times \mathbb{R}^2 \tag{25}$$

with $\mathscr{A}_0 = -\mathscr{S} + \mathscr{K}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ and

$$(\mathscr{S}f)(t) = \frac{1}{\pi} ⨍_{-1}^{1} \frac{f(s)\,\mathrm{d}s}{s-t} \quad \text{and} \quad (\mathscr{K}_0 f)(t) = \frac{1}{\pi}\int_{-1}^{1} K(s,t) f(s)\,\mathrm{d}s.$$

For any integer $n \geq 1$ we are looking for an approximate solution $(f_n, \beta_n, \gamma_n) \in R(\mathscr{P}_n) \times \mathbb{R}^2$ of (25), where $R(\mathscr{P}_n)$ is the image space of the orthoprojection $\mathscr{P}_n : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ defined by

$$\mathscr{P}_n f = \sum_{k=0}^{n-1} \langle f, U_k \rangle\, \varphi U_k,$$

where $U_k = p_k^{\varphi}$ denotes the normalized second kind Chebyshev polynomial of degree k, by solving the collocation equations

$$-(\mathscr{S}f_n)(t_{jn}) + (\mathscr{K}_n^0 f_n)(t_{jn}) = \beta_n\psi_1(t_{jn}) + \gamma_n, \quad j = 1, \ldots, n+1, \tag{26}$$

together with

$$\frac{\pi}{n+1}\sum_{i=1}^{n} \varphi(s_{in})\psi_1'(s_{in}) f_n(s_{in}) = \gamma_0, \tag{27}$$

where $t_{jn} = \cos\frac{(2j-1)\pi}{2n+2}$ and $s_{in} = \cos\frac{i\pi}{n+1}$ are Chebyshev nodes of first and second kind, respectively, and where

$$(\mathscr{K}_n^0 f_n)(t) = \frac{1}{n+1}\sum_{i=1}^{n} \varphi(s_{in})K(s_{in}, t)f_n(s_{in}). \tag{28}$$

Note that $f_n(t)$ can be written, with the help of the weighted Lagrange interpolation polynomials

$$\widetilde{\ell}_{kn}^{\varphi}(t) = \frac{\varphi(t)\ell_{kn}^{\varphi}(t)}{\varphi(s_{kn})} \quad \text{with} \quad \ell_{kn}^{\varphi}(t) = \frac{U_n(t)}{(t-s_{kn})U_n'(s_{kn})}, \quad k = 1, \dots, n,$$

in the form

$$f_n(t) = \sum_{k=1}^{n} \xi_{kn}\widetilde{\ell}_{kn}^{\varphi}(t), \quad \xi_{kn} = f_n(s_{kn}). \tag{29}$$

Let $\mathscr{L}_n^j$, $j = 1, 2$, denote the interpolation operators which associate to a function $g : (-1, 1) \longrightarrow \mathbb{R}$ the polynomials

$$(\mathscr{L}_n^1 g)(t) = \sum_{j=1}^{n+1} \frac{g(t_{jn})T_{n+1}(t)}{(t-t_{jn})T_{n+1}'(t_{jn})} \quad \text{and} \quad (\mathscr{L}_n^2 g)(t) = \sum_{i=1}^{n} \frac{g(s_{in})U_n(t)}{(t-s_{in})U_n'(s_{in})},$$

where $T_n = p_n^{\varphi^{-1}}$. Now, the system (26), (27) can be written as operator equation

$$\mathscr{A}_n f_n = \beta_n \mathscr{L}_n^1 \psi_1 + \gamma_n, \quad f_n \in R(\mathscr{P}_n) \tag{30}$$

together with

$$\langle \mathscr{L}_n^2 \psi_1', f_n\rangle = \gamma_0, \tag{31}$$

where $\mathscr{A}_n = -\mathscr{S}_n + \mathscr{K}_n$, $\mathscr{S}_n = \mathscr{L}_n^1 \mathscr{S}\mathscr{P}_n$, and $\mathscr{K}_n = \mathscr{L}_n^1 \mathscr{K}_n^0 \mathscr{P}_n$. The equivalence of (27) and (31) follows from the algebraic accuracy of the Gaussian rule w.r.t. the Chebyshev nodes of second kind. The assertion of the following lemma is well-known (see [9, Theorem 14.3.1]).

Lemma 8 *For all* $f \in \mathbf{C}[-1, 1]$, $\lim_{n\to\infty} \left\| f - \mathscr{L}_n^1 f \right\|_{\varphi^{-1}} = 0$ *and* $\lim_{n\to\infty} \left\| f - \mathscr{L}_n^2 f \right\|_{\varphi} = 0$.

The next lemma provides convergence rates for the interpolating polynomials and will be used in the proof of Proposition 3.

Lemma 9 ([1], Theorem 3.4) *If $r > \frac{1}{2}$, then there exists a constant $c > 0$ such that, for any real p, $0 \le p \le r$ and all $n \ge 1$,*

(a) $\left\| f - \mathscr{L}_n^1 f \right\|_{\varphi^{-1},p} \le c\, n^{p-r} \|f\|_{\varphi^{-1},r}$ *for all* $f \in \mathbf{L}^{2,r}_{\varphi^{-1}}$,

(b) $\left\| f - \mathscr{L}_n^2 f \right\|_{\varphi,p} \le c\, n^{p-r} \|f\|_{\varphi,r}$ *for all* $f \in \mathbf{L}^{2,r}_{\varphi}$.

Lemma 10 *Let $\psi_j \in \mathbf{C}^2[-1,1]$, $j = 1, 2$. Then,*

(a) $\lim\limits_{n\to\infty} \|\mathscr{K}_n - \mathscr{K}_0\|_{\mathbf{L}^2_{\varphi^{-1}} \to \mathbf{L}^2_{\varphi^{-1}}} = 0$,

(b) *there exist constants $\eta > 0$ and $n_0 \in \mathbb{N}$ such that*

$$\|\mathscr{A}_n f_n\|_{\varphi^{-1}} \ge \eta\, \|f_n\|_{\varphi^{-1}} \quad \forall f_n \in R(\mathscr{P}_n),\ \forall\, n \ge n_0.$$

Proof At first, recall that the operator $\mathscr{A}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ is Fredholm with index -1 (cf. the proof of Proposition 2). By Banach's theorem, the operator $\mathscr{A}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \left(R(\mathscr{A}_0), \|.\|_{\varphi^{-1}}\right)$ has a bounded inverse. Hence, there is a constant $\eta_0 > 0$ with

$$\|\mathscr{A}_0 f\|_{\varphi^{-1}} \ge \eta_0 \|f\|_{\varphi^{-1}} \quad \forall f \in \mathbf{L}^2_{\varphi^{-1}}. \tag{32}$$

By definition of $\mathscr{K}_n^0$ and in virtue of the algebraic accuracy of the Gaussian rule, for $f_n \in R(\mathscr{P}_n)$ we have

$$(\mathscr{K}_n^0 f_n)(t) = \frac{1}{\pi} \int_{-1}^{1} \mathscr{L}_n^2 \left[K(.,t)\varphi^{-1} f_n\right](s)\varphi(s)\,\mathrm{d}s = \frac{1}{\pi} \int_{-1}^{1} \mathscr{L}_n^2 \left[K(.,t)\right](s) f_n(s)\,\mathrm{d}s,$$

which implies

$$\left\| \left(\mathscr{K}_n^0 - \mathscr{K}_0\right) \mathscr{P}_n f \right\|_\infty \le \frac{1}{\pi} \sup \left\{ \left\| \mathscr{L}_n^2 \left[K(.,t)\right] - K(.,t) \right\|_\varphi : -1 \le t \le 1 \right\} \|f\|_{\varphi^{-1}},$$

where $\|.\|_\infty$ is the norm in $\mathbf{C}[-1,1]$, i.e., $\|f\|_\infty = \max\{|f(t)| : -1 \le t \le 1\}$. Since, due to Lemma 8 and the principle of uniform boundedness, the operator sequence $\mathscr{L}_n^1 : \mathbf{C}[-1,1] \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ is uniformly bounded, the last estimate together with Lemma 8 (applied to $\mathscr{L}_n^2$) leads to

$$\lim_{n\to\infty} \left\| \mathscr{K}_n - \mathscr{L}_n^1 \mathscr{K}_0 \mathscr{P}_n \right\|_{\mathbf{L}^2_{\varphi^{-1}} \to \mathbf{L}^2_{\varphi^{-1}}} = 0.$$

Again Lemma 8, the strong convergence of $\mathscr{P}_n = \mathscr{P}_n^* \longrightarrow \mathscr{I}$ (the identity operator), and the compactness of the operator $\mathscr{K}_0 : \mathscr{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{C}[-1,1]$ give us $\lim\limits_{n\to\infty} \left\| \mathscr{L}_n^1 \mathscr{K}_0 \mathscr{P}_n - \mathscr{K}_0 \right\|_{\mathbf{L}^2_{\varphi^{-1}} \to \mathbf{L}^2_{\varphi^{-1}}} = 0$, and (a) is proved.

Formula (22) implies the relation $\mathscr{S}_n = \mathscr{S}\mathscr{P}_n$, from which, together with (a), we conclude

$$\|(\mathscr{A}_n - \mathscr{A}_0)f_n\|_{\varphi^{-1}} \le \alpha_n \|f_n\|_{\varphi^{-1}} \quad \forall f_n \in R(\mathscr{P}_n), \tag{33}$$

where $\alpha_n \longrightarrow 0$. Together with (32), this leads to (b). □

Proposition 3 *Assume* $\psi_j \in \mathbf{C}^3[-1,1]$, $j = 1,2$, $\gamma_0 \neq 0$, *and let* $\psi_1(t)$ *be not constant. Then, for all sufficiently large* n *(say* $n \ge n_0$*), there exists a unique solution* $(f_n^*, \beta_n^*, \gamma_n^*) \in R(\mathscr{P}_n) \times \mathbb{R}^2$ *of* (30), (31). *Moreover,*

$$\lim_{n\to\infty} \sqrt{\left\| f_n^* - f^* \right\|_{\varphi^{-1}}^2 + |\beta_n^* - \beta^*|^2 + |\gamma_n^* - \gamma^*|^2} = 0, \tag{34}$$

where (f^*, β^*, γ^*) *is the unique solution of* (25). *If* $\psi_j \in \mathbf{C}^m[-1,1]$, $j = 1,2$, *for some integer* $m > 2$, *then*

$$\sqrt{\left\| f_n^* - f^* \right\|_{\varphi^{-1}}^2 + |\beta_n^* - \beta^*|^2 + |\gamma_n^* - \gamma^*|^2} \le c\, n^{2-m} \tag{35}$$

with a constant $c > 0$ *independent of* n.

Proof Due to the Fredholmness of $\mathscr{A}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ with index -1 and due to $N(\mathscr{A}_0) = \{0\}$ (Proposition 2, (a)), we have $\mathbf{L}^2_{\varphi^{-1}} = R(\mathscr{A}_0) \oplus \operatorname{span}\{g_0\}$ (direct orthogonal sum w.r.t. $\langle .,. \rangle_{\varphi^{-1}}$) for some $g_0 \in \mathbf{L}^2_{\varphi^{-1}}$ with $\|g_0\|_{\varphi^{-1}} = 1$.

By $\mathbf{H}$ we denote the Hilbert space of all pairs $(f, \delta) \in \mathbf{L}^2_{\varphi^{-1}} \times \mathbb{R}$ equipped with the inner product $\langle (f_1, \delta_1), (f_2, \delta_2) \rangle_{\mathbf{H}} = \langle f_1, f_2 \rangle_{\varphi^{-1}} + \delta_1 \delta_2$. For a continuous function $g_1 \in \mathbf{C}[-1,1]$ with $\langle g_1, g_0 \rangle_{\varphi^{-1}} \neq 0$ and for $n \in \mathbb{N}$, define the linear and bounded operators

$$\mathscr{B}_0 : \mathbf{H} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}, \quad (f, \delta) \mapsto \mathscr{A}_0 f - \delta g_1$$

and

$$\mathscr{B}_n : \mathbf{H} \longrightarrow \mathbf{L}^2_{\varphi^{-1}}, \quad (f, \delta) \mapsto \mathscr{A}_n f - \delta \mathscr{L}_n^1 g_1.$$

Let us consider the auxiliary problems

$$\mathscr{B}_0(f, \delta) = g \in \mathbf{C}[-1,1], \quad (f, \delta) \in \mathbf{L}^2_{\varphi^{-1}} \times \mathbb{R} \tag{36}$$

and

$$\mathscr{B}_n(f_n, \delta_n) = \mathscr{L}_n^1 g, \quad (f_n, \delta_n) \in R(\mathscr{P}_n) \times \mathbb{R}. \tag{37}$$

An immediate consequence of (33) is the relation

$$\begin{aligned}\|\mathscr{B}_n(f_n,\delta)-\mathscr{B}_0(f_n,\delta)\|_{\varphi^{-1}} &\le \alpha_n\,\|f_n\|_{\varphi^{-1}}+|\delta|\,\left\|\mathscr{L}_n^1 g_1-g_1\right\|_{\varphi^{-1}}\\ &\le \beta_n\,\|(f_n,\delta)\|_{\mathbf{H}}\quad \forall\,(f_n,\delta)\in R(\mathscr{P}_n)\times\mathbb{R},\end{aligned}\tag{38}$$

where $\beta_n=\sqrt{\alpha_n^2+\left\|\mathscr{L}_n^1 g_1-g_1\right\|_{\varphi^{-1}}^2}\longrightarrow 0$. Equation (36) is uniquely solvable, since $R(\mathscr{A}_0)=\left\{f\in\mathbf{L}^2_{\varphi^{-1}}:\langle f,g_0\rangle_{\varphi^{-1}}=0\right\}$ and, consequently, the part $\delta^*\in\mathbb{R}$ of the solution (f^*,δ^*) of (36) is uniquely determined by the condition

$$\langle g+\delta g_1,g_0\rangle_{\varphi^{-1}}=0,\quad \text{i.e.,}\quad \delta^*=-\frac{\langle g,g_0\rangle_{\varphi^{-1}}}{\langle g_1,g_0\rangle_{\varphi^{-1}}},$$

and $f^*\in\mathbf{L}^2_{\varphi^{-1}}$ is the unique solution (cf. Proposition 1, (a)) of $\mathscr{A}_0 f=g+\delta^* g_1$. Hence, in virtue of Banach's theorem, the operator $\mathscr{B}_0:\mathbf{H}\longrightarrow\mathbf{L}^2_{\varphi^{-1}}$ is boundedly invertible, which implies that there is a constant $\eta_1>0$, such that

$$\|\mathscr{B}_0(f,\delta)\|_{\varphi^{-1}}\ge\eta_1\,\|(f,\delta)\|_{\mathbf{H}}\quad\forall\,(f,\delta)\in\mathbf{H}.\tag{39}$$

Putting this together with (38), we can state that there is a number $n_0\in\mathbb{N}$, such that

$$\|\mathscr{B}_n(f_n,\delta)\|_{\varphi^{-1}}\ge\frac{\eta_1}{2}\,\|(f_n,\delta)\|_{\mathbf{H}}\quad\forall\,(f_n,\delta)\in R(\mathscr{P}_n)\times\mathbb{R},\ \forall\,n\ge n_0.\tag{40}$$

This implies that, for $n\ge n_0$, the map $\mathscr{B}_n:R(\mathscr{P}_n)\times\mathbb{R}\longrightarrow\operatorname{span}\left\{T_j:j=0,1,\dots,n\right\}$ is a bijection, such that (37) is uniquely solvable for all $n\ge n_0$. Moreover, if (f_n^*,δ_n^*) is the solution of (37), then

$$\begin{aligned}&\left\|(f_n^*,\delta_n^*)-(\mathscr{P}_n f^*,\delta^*)\right\|_{\mathbf{H}}\\ &\le\frac{2}{\eta_1}\,\left\|\mathscr{L}_n^1 g-\mathscr{B}_n(\mathscr{P}_n f^*,\delta^*)\right\|_{\varphi^{-1}}\\ &\le\frac{2}{\eta_1}\left(\left\|\mathscr{L}_n^1 g-\mathscr{B}_0(\mathscr{P}_n f^*,\delta^*)\right\|_{\varphi^{-1}}+\|(\mathscr{B}_0-\mathscr{B}_n)(\mathscr{P}_n f^*,\delta^*)\|_{\varphi^{-1}}\right)\\ &\le\frac{2}{\eta_1}\left(\left\|\mathscr{L}_n^1 g-\mathscr{B}_0(\mathscr{P}_n f^*,\delta^*)\right\|_{\varphi^{-1}}+\beta_n\,\|(\mathscr{P}_n f^*,\delta^*)\|_{\mathbf{H}}\right),\end{aligned}\tag{41}$$

which implies

$$\lim_{n\to\infty}\left\|(f_n^*,\delta_n^*)-(f^*,\delta^*)\right\|_{\mathbf{H}}=0.\tag{42}$$

Now, let $(f^*, \beta^*, \gamma^*) \in \mathbf{V}_{\gamma_0} \times \mathbb{R}^2$ be the unique solution of (25) (cf. Proposition 2, (b)). There exist $\beta_1^*, \gamma_1^* \in \mathbb{R}$ such that $\langle g_1, g\rangle_{\varphi^{-1}} = 0$ and $\|g_1\|_{\varphi^{-1}} = 1$, where $g = \beta^*\psi_1 + \gamma^*$ and $g_1 = \beta_1^*\psi_1 + \gamma_1^*$. Because of

$$\dim\left(R(\mathscr{A}_0) \cap \{\beta\psi_1 + \gamma : \beta, \gamma \in \mathbb{R}\}\right) = 1$$

(cf. the proof of Proposition 2), we have $g_1 \notin R(\mathscr{A}_0)$, i.e., $\langle g_1, g_0\rangle_{\varphi^{-1}} \neq 0$. With these notations, $(f^*, 0) \in \mathbf{L}^2_{\varphi^{-1}} \times \mathbb{R}$ is the unique solution of (36). Taking into account the previous considerations, we conclude that, for all sufficiently large n, there is a unique $(f_n^1, \delta_n^1) \in R(\mathscr{P}_n) \times \mathbb{R}$ satisfying

$$\mathscr{A}_n f_n^1 - \delta_n^1 \mathscr{L}_n^1(\beta_1^*\psi_1 + \gamma_1^*) = \mathscr{L}_n^1(\beta^*\psi_1 + \gamma^*)$$

or equivalently

$$\mathscr{A}_n f_n^1 = (\beta^* + \delta_n^1\beta_1^*)\mathscr{L}_n^1\psi_1 + \gamma^* + \delta_n^1\gamma_1^*,$$

where, due to (42), $\left\|f_n^1 - f^*\right\|_{\varphi^{-1}} \longrightarrow 0$ and $\delta_n^1 \longrightarrow 0$. It follows

$$\langle \mathscr{L}_n^2\psi_1', f_n^1\rangle = \langle \mathscr{L}_n^2\psi_1', \varphi^{-1}f_n^1\rangle_\varphi \longrightarrow \langle \psi_1', \varphi^{-1}f^*\rangle_\varphi = \langle \psi_1', f^*\rangle = \gamma_0.$$

Consequently, for all sufficiently large n, $\langle \mathscr{L}_n^2\psi_1', f_n^1\rangle \neq 0$ and $(f_n^*, \beta_n^*, \gamma_n^*)$ with

$$f_n^* = \frac{\gamma_0 f_n^1}{\langle \mathscr{L}_n^2\psi_1', f_n^1}\rangle, \quad \beta_n^* = \frac{\gamma_0(\beta^* + \delta_n^1\beta_1^*)}{\langle \mathscr{L}_n^2\psi_1', f_n^1\rangle}, \quad \gamma_n^* = \frac{\gamma_0(\gamma^* + \delta_n^1\gamma_1^*)}{\langle \mathscr{L}_n^2\psi_1', f_n^1\rangle}$$

is a solution of (30), (31). This solution is unique, since (f_n^1, δ_n^1) was uniquely determined. Furthermore,

$$f_n^* \longrightarrow f^* \text{ in } \mathbf{L}^2_{\varphi^{-1}} \quad \text{and} \quad \beta_n^* \longrightarrow \beta^*,\ \gamma_n^* \longrightarrow \gamma^*,$$

and (34) follows. To prove the error estimate (35), first we recall that $\psi_j \in \mathbf{C}^m[-1,1], j = 1,2$, for some $m > 2$ implies, due to Lemma 3, the continuity of the partial derivatives $\dfrac{\partial^k K(s,t)}{\partial t^k}$, $k = 1, \ldots, m-2$, for $(s,t) \in [-1,1]^2$. Consequently,

$$-\mathscr{S}f^* = \beta^*\psi_1 + \gamma^* - \mathscr{K}_0 f^* \in \mathbf{C}^{m-2}[-1,1] \subset \mathbf{L}^{2,m-2}_{\varphi^{-1}},$$

i.e., in virtue of Lemma 7, $f^* \in \varphi\mathbf{L}^{2,m-2}_{\varphi}$. Taking into account the uniform boundedness of $\mathscr{L}_n^1 : \mathbf{C}[-1,1] \longrightarrow \mathbf{L}^2_{\varphi^{-1}}$ (see Lemma 8) and Lemma 9, we get, for all $f_n \in R(\mathscr{P}_n)$,

$$\begin{aligned}&\|(\mathscr{K}_n - \mathscr{K}_0)f_n\|_{\varphi^{-1}}\\&\quad\leq \left\|\mathscr{L}_n^1(\mathscr{K}_n^0 - \mathscr{K}_0)f_n\right\|_{\varphi^{-1}} + \left\|(\mathscr{L}_n^1\mathscr{K}_0 - \mathscr{K}_0)f_n\right\|_{\varphi^{-1}}\end{aligned}$$

$$\leq \sup\left\{\left\|\mathscr{L}_n^2 K(.,t) - K(.,t)\right\|_\varphi : -1 \leq t \leq 1\right\} \|f_n\|_{\varphi^{-1}} + \left\|(\mathscr{L}_n^1 \mathscr{K}_0 - \mathscr{K}_0) f_n\right\|_{\varphi^{-1}}$$

$$\leq c\, n^{1-m} \left(\sup\left\{\|K(.,t)\|_{\varphi,m-1} : -1 \leq t \leq 1\right\} \|f_n\|_{\varphi^{-1}} + \|\mathscr{K}_0 f_n\|_{\varphi^{-1},m-1}\right)$$

$$\leq c n^{1-m} \|f_n\|_{\varphi^{-1}},$$

where we have also used that $\mathscr{K}_0 : \mathbf{L}^2_{\varphi^{-1}} \longrightarrow \mathbf{C}^{m-2}[-1,1] \subset \mathbf{L}^{2,m-2}_{\varphi^{-1}}$ is bounded (cf. [1, Lemma 4.2]). Hence, in (33) and (38) we have $\alpha_n = \mathscr{O}(n^{2-m})$ and, since $g_1 = \beta_1^* \psi_1 + \gamma^* \in \mathbf{L}^{2,m}_{\varphi^{-1}}$, also $\beta_n = \mathscr{O}(n^{2-m})$. From (41) and $g = \beta^* \psi_1 + \gamma^* \in \mathbf{L}^{2,m}_{\varphi^{-1}}$ we obtain the bound

$$\left\|(f_n^1, \delta_n^1) - (\mathscr{P}_n f^*, 0)\right\|_{\mathbf{H}}$$

$$\leq \frac{2}{\eta_1}\left(\left\|\mathscr{L}_n^1 g - g\right\|_{\varphi^{-1}} + \|\mathscr{A}_0\|_{\mathbf{L}^2_{\varphi^{-1}} \to \mathbf{L}^2_{\varphi^{-1}}} \|f^* - \mathscr{P}_n f^*\|_{\varphi^{-1}} + \beta_n \|f^*\|_{\varphi^{-1}}\right)$$

$$\leq c\, n^{2-m}.$$

Now, (35) easily follows. □

Remark 3 From the proof of Proposition 3 it is seen that the first assertion including (34) remains true if the assumption $\psi_j \in \mathbf{C}^3[-1,1]$ is replaced by $\psi_j \in \mathbf{C}^2[-1,1]$ together with $\dim N(\mathscr{A}_0) = 0$ (cf. Proposition 2).

5 Implementation Features

Let us discuss some computational aspects. Because of $f_n \in R(\mathscr{P}_n)$ we have, taking into account (22) and $T_{n+1}(t_{jn}) = 0$,

$$(\mathscr{S} f_n)(t_{jn}) = \sum_{k=1}^n \frac{f_n(s_{kn})}{\varphi(s_{kn}) U_n'(s_{kn})} \frac{1}{\pi} \int_{-1}^1 \frac{\varphi(s) U_n(s)}{(s - s_{kn})(s - t)}\, \mathrm{d}s$$

$$= \sum_{k=1}^n \frac{f_n(s_{kn})}{\varphi(s_{kn}) U_n'(s_{kn})} \frac{1}{\pi} \int_{-1}^1 \left(\frac{1}{s - s_{kn}} - \frac{1}{s - t_{jn}}\right) \varphi(s) U_n(s)\, \mathrm{d}s \, \frac{1}{s_{kn} - t_{jn}}$$

$$= -\sum_{k=1}^n \frac{T_{n+1}(s_{kn})}{\varphi(s_{kn}) U_n'(s_{kn})} \frac{f_n(s_{kn})}{s_{kn} - t_{jn}} = \sum_{k=1}^n \frac{\varphi(s_{kn})}{n+1} \frac{f_n(s_{kn})}{s_{kn} - t_{jn}}.$$

From this, the following expression is obtained:

$$-(\mathscr{S}f_n)(t_{jn}) + (\mathscr{K}_n^0 f_n)(t_{jn}) = \frac{1}{n+1}\sum_{k=1}^{n} \varphi(s_{kn})Y_0(s_{kn}, t_{jn})f_n(s_{kn}), \quad j = 1, \ldots, n+1$$

(cf. (11), (28), and (26)). Thus, to find the solution (f_n, β_n, γ_n) of (26), (27), we have to solve the algebraic linear system of equations

$$\mathbb{A}_n \xi_n = \eta_n, \tag{43}$$

where $\eta_n = \left[\, \eta_{jn} \,\right]_{j=1}^{n+2} = \left[\, 0 \ldots 0 \ \gamma_0 \,\right]^T \in \mathbb{R}^{n+2}$ is given and $\xi_n = \left[\, \xi_{kn} \,\right]_{k=1}^{n+2} = \left[f_n(s_{1n}) \ldots f_n(s_{nn}) \ \beta_n \ \gamma_n \,\right]^T \in \mathbb{R}^{n+2}$ is the vector we are looking for, and where the matrix $\mathbb{A}_n = \left[\, a_{jk} \,\right]_{j,k=1}^{n+2}$ is defined by

$$a_{jk} = \frac{\varphi(s_{kn})Y_0(s_{kn}, t_{jn})}{n+1}, \quad j = 1, \ldots, n+1, \ k = 1, \ldots, n,$$

$$a_{j,n+1} = -\psi_1(t_{jn}), \ \ a_{j,n+2} = -1, \quad j = 1, \ldots, n+1,$$

$$a_{n+2,k} = \frac{\pi\varphi(s_{kn})\psi_1'(s_{kn})}{n+1}, \quad k = 1, \ldots, n, \qquad a_{n+2,n+1} = a_{n+2,n+2} = 0.$$

In the case of a symmetric wing, the above numerical method can be significantly simplified. Indeed, it turns out that the ideas used in the proof of Corollary 1, which have led to (24), also work for the discrete system (43). This can be shown as follows.

First, note that in this case we have

$$Y_0(-t, s) = -Y_0(t, -s) \quad \text{and} \quad Y_0(-t, -s) = -Y_0(t, s). \tag{44}$$

Then, let $n = 2m$ be sufficiently large, let

$$\xi_n^* = \left[f_n^*(s_{1n}) \ldots f_n^*(s_{nn}) \ \beta_n^* \ \gamma_n^* \,\right]^T \in \mathbb{R}^{n+2}$$

be the unique solution of (43), and define $\widetilde{\xi}_n = \left[f_n^*(s_{nn}) \ldots f_n^*(s_{1n}) \ \beta_n^* \ -\gamma_n^* \,\right]^T$. Then, the jth entry of $\mathbb{A}_n\widetilde{\xi}_n$ equals

$$\begin{aligned}
&\left(\mathbb{A}_n\widetilde{\xi}_n\right)_j \\
&= \sum_{k=1}^{n} \frac{\varphi(s_{kn})Y_0(s_{kn}, t_{jn})}{n+1} f_n^*(s_{n+1-k,n}) + \beta_n^*\psi_1(t_{jn}) - \gamma_n^* \\
&= -\sum_{k=1}^{n} \frac{\varphi(s_{n+1-k,n})Y_0(s_{n+1-k,n}, t_{n+2-j,n})}{n+1} f_n^*(s_{n+1-k,n}) + \beta_n^*\psi_1(t_{n+2-j,n}) + \gamma_n^* = 0
\end{aligned}$$

for $j = 1, \ldots, n+1$, and

$$\left(\mathbb{A}_n \widetilde{\xi}_n\right)_{n+2} = \sum_{k=1}^{n} \frac{\varphi(s_{kn})\psi_1'(s_{kn})}{n+1} f_n^*(s_{n+1-k,n})$$

$$= \sum_{k=1}^{n} \frac{\varphi(s_{n+1-k,n})\psi_1'(s_{n+1-k,n})}{n+1} f_n^*(s_{n+1-k,n}) = \gamma_0$$

for $j = n+2$, where we have taken into account (44) and the identities $s_{n+1-k,n} = s_{kn}$ and $t_{n+2-j,n} = t_{jn}$. This means that $\widetilde{\xi}_n$ also solves (43) and, due to the solution uniqueness, we have only to compute the $m+1 = n/2+1$ values $\xi_{kn}^* = f_n^*(s_{kn}) = f_n^*(s_{n+1-k,n})$, $k = 1, \ldots, m$, and $\xi_{n+1,n}^* = \beta_n^*$, while $\xi_{n+1,n+1}^* = \gamma_n^* = 0$. The system we have to solve can now be written in the form

$$\sum_{k=1}^{m} b_{jk}\xi_{kn} = \beta_n\psi_1(t_{jn}), \quad j = 1, \ldots, m, \quad \sum_{k=1}^{m} \frac{2\varphi(s_{kn})\psi_1'(s_{kn})}{n+1}\xi_{kn} = \gamma_0, \tag{45}$$

where $b_{jk} = a_{jk} + a_{j,n+1-k}$, and where we have used the properties that, for $j = 1, \ldots, m$, the $(n+2-j)$th equation in (43) is identical to the jth equation and that the $(m+1)$th equation is automatically fulfilled ($b_{m+1,k} = 0$ and $\psi_1(t_{m+1,n}) = 0$, since $t_{m+1,n} = 0$), in virtue of the assumed symmetries. Of course, (45) is, with $\bar{\xi}_n = \xi_{kn}/\beta_n$, equivalent to

$$\sum_{k=1}^{m} b_{jk}\bar{\xi}_{kn} = \psi(t_{jn}), \quad j = 1, \ldots, m, \qquad \beta_n \sum_{k=1}^{m} \frac{2\varphi(s_{kn})\psi_1'(s_{kn})}{n+1}\bar{\xi}_{kn} = \gamma_0,$$

since $\beta^* \neq 0$ and $\beta_n^* \longrightarrow \beta^*$ (cf. Remark 2 with (23) and (24)), and thus for all sufficiently large n we have $\beta_n^* \neq 0$. This latter system is precisely the method used in [3], for which we have now proved its convergence and given an error estimate. A similar simplification can be obtained also for $n = 2m+1$.

Finally, we discuss the question if, under the assumptions of Proposition 3, the condition numbers of the matrices $\mathbb{A}_n$ are uniformly bounded or if it is necessary to apply a preconditioning to $\mathbb{A}_n$. Note that, under the assumptions of Proposition 3, the operator sequence $\mathscr{B}_n : R(\mathscr{P}_n) \times \mathbb{R}^2 \longrightarrow \mathbf{P}_n \times \mathbb{R}$ ($\mathbf{P}_n$ being the set of all real algebraic polynomials of degree less than or equal to n) defined by

$$\mathscr{B}_n(f_n, \beta, \gamma) = \left(\mathscr{A}_n f_n - \beta\mathscr{L}_n^1\psi_1 - \gamma, \langle\mathscr{L}_n^2\psi_1', f_n\rangle\right)$$

(cf. (30) and (31)) is a bounded and stable one, i.e., the norms of $\mathscr{B}_n$ and of $\mathscr{B}_n^{-1}$ (which exist for all sufficiently large n) are uniformly bounded (as a consequence of Proposition 3 together with Lemma 10). Hereby, the norms in $\mathbf{H}_n^1 := R(\mathscr{P}_n) \times \mathbb{R}^2$

and $\mathbf{H}_n^2 := \mathbf{P}_n \times \mathbb{R}$ are given by

$$\|(f_n,\beta,\gamma)\|_{\mathbf{H}_n^1} = \sqrt{\|f_n\|_{\varphi^{-1}}^2 + |\beta|^2 + |\gamma|^2} \quad \text{and} \quad \|(p_n,\delta)\|_{\mathbf{H}_n^2} = \sqrt{\|p_n\|_{\varphi^{-1}}^2 + |\delta|^2},$$

respectively. Set $\omega_n = \sqrt{\dfrac{\pi}{n+1}}$ and define the operators

$$\mathscr{E}_n : \mathbf{H}_n^1 \longrightarrow \mathbb{R}^{n+2}, \quad (f_n,\beta,\gamma) \mapsto \big(\omega_n f_n(s_{1n}),\ldots,\omega_n f_n(s_{nn}),\beta,\gamma\big)$$

and

$$\mathscr{F}_n : \mathbf{H}_n^2 \longrightarrow \mathbb{R}^{n+2}, \quad (p_n,\delta) \mapsto \big(\omega_n p_n(t_{1n}),\ldots,\omega_n p_n(t_{n+1,n}),\delta\big),$$

where the space $\mathbb{R}^{n+2}$ is equipped with the usual Euclidean inner product. These operators are unitary ones. To prove this, we recall the representation (29) of $f_n(t)$, in order to see that, for all $(f_n,\beta,\gamma) \in \mathbf{H}_n^1$ and $(\xi_1,\ldots,\xi_{n+2}) \in \mathbb{R}^{n+2}$,

$$\begin{aligned}\langle \mathscr{E}_n(f_n,\beta,\gamma),(\xi_1,\ldots,\xi_{n+2})\rangle &= \omega_n \sum_{k=1}^n f_n(s_{kn})\xi_k + \beta\xi_{n+1} + \gamma\xi_{n+2} \\ &= \int_{-1}^1 f_n(s)\frac{1}{\omega_n}\sum_{k=1}^n \xi_k \widetilde{\ell}_{kn}^{\varphi}(s)\,\mathrm{d}s + \beta\xi_{n+1} + \gamma\xi_{n+2} \\ &= \langle (f_n,\beta,\gamma), \mathscr{E}_n^{-1}(\xi_1,\ldots,\xi_{n+2})\rangle_{\mathbf{H}_n^1}.\end{aligned}$$

Analogously, one can show that

$$\langle \mathscr{F}_n(p_n,\delta),(\eta_1,\ldots,\eta_{n+2})\rangle = \langle (p_n,\delta), \mathscr{F}_n^{-1}(\eta_1,\ldots,\eta_{n+2})\rangle_{\mathbf{H}_n^2}$$

holds true for all $(p_n,\delta) \in \mathbf{H}_n^2$ and $(\eta_1,\ldots,\eta_{n+2}) \in \mathbb{R}^{n+2}$. As a consequence we get, that an appropriate matrix $\mathbb{B}_n = \left[\, b_{jk} \,\right]_{j,k=1}^{n+2}$ can be defined by

$$\mathscr{B}_n(\xi_1,\ldots,\xi_{n+2}) = \mathscr{F}_n\mathscr{B}_n\mathscr{E}_n^{-1}(\xi_1,\ldots,\xi_{n+2}) = \mathscr{E}_n\mathscr{B}_n\left(\omega_n^{-1}\sum_{k=1}^n \xi_k\widetilde{\ell}_{kn}^{\varphi},\xi_{n+1},\xi_{n+2}\right)$$

$$= \mathscr{E}_n\left(\omega_n^{-1}\sum_{k=1}^n \xi_k\mathscr{A}_n\widetilde{\ell}_{kn}^{\varphi} - \xi_{n+1}\mathscr{L}_n^1\psi_1 - \xi_{n+2}, \omega_n\sum_{k=1}^n \varphi(s_{kn})\psi_1'(s_{kn})\xi_k\right)$$

$$= \left(\left[\sum_{k=1}^{n} \left(\mathscr{A}_n \widetilde{\ell^{\varphi}_{kn}} \right) (t_{jn}) \xi_k - \omega_n \psi_1 (t_{jn}) \xi_{n+1} - \omega_n \xi_{n+2} \right]_{j=1}^{n+1}, \omega_n \sum_{k=1}^{n} \varphi(s_{kn}) \psi_1'(s_{kn}) \xi_k \right)$$

$$= \left(\left[\sum_{k=1}^{n} a_{jk} \xi_k + \omega_n a_{j,n+1} \xi_{n+1} + \omega_n a_{j,n+2} \xi_{n+2} \right]_{j=1}^{n+1}, \sum_{k=1}^{n} \omega_n^{-1} a_{n+2,k} \xi_k \right),$$

i.e., $b_{jk} = a_{jk}$ for $j = 1, \dots, n+1$, $k = 1, \dots, n$, $b_{jk} = \omega_n a_{jk}$ for $j = 1, \dots, n+1$, $k = n+1, n+2$, and $b_{n+2,k} = \omega_n^{-1} a_{n+2,k}$, $k = 1, \dots, n$. This means that

$$\mathbb{B}_n = \mathbb{F}_n \mathbb{A}_n \mathbb{E}_n^{-1} \quad \text{with} \quad \mathbb{E}_n = \operatorname{diag}\left[1 \, \dots \, 1 \ \omega_n^{-1} \ \omega_n^{-1} \right], \ \mathbb{F}_n = \operatorname{diag}\left[1 \, \dots \, 1 \ \omega_n^{-1} \right],$$

and we can solve the system $\mathbb{B}_n \widetilde{\xi} = \widetilde{\eta}$ instead of $\mathbb{A}_n \xi = \eta$, where $\widetilde{\eta} = \mathbb{F}_n \eta$ and $\widetilde{\xi} = \mathbb{E}_n \xi$.

Therefore, in the following numerical examples we can check the stability of the method by computing the condition number of the matrix $\mathbb{B}_n$ w.r.t. the Euclidean norm, which is equal to the quotient $\frac{s_{\max}(\mathbb{B}_n)}{s_{\min}(\mathbb{B}_n)}$ of its biggest and its smallest singular values. Moreover, the left hand side in (35) can be approximated by the following discretization of it

$$\text{err} = \sqrt{\frac{\pi}{N+1} \sum_{k=1}^{N} \left[f_n^*(s_{kN}) - f_N^*(s_{kN}) \right]^2 + \left| \beta_n^* - \beta_N^* \right|^2 + \left| \gamma_n^* - \gamma_N^* \right|^2} \qquad (46)$$

with $N >> n$.

6 Numerical Examples

To test the numerical method, we have proposed, and the associated convergence estimate (35), we have considered four simple curves. The first one is the following non symmetric part of the unit circle:

$$\psi_1(t) = \cos\left(\frac{\pi}{8}(3t + 13)\right), \ \ \psi_2(t) = \sin\left(\frac{\pi}{8}(3t + 13)\right), \ \ -1 \le t \le 1.$$

The second one is a symmetric part of the ellipse having semi-axis $a = 1, b = 0.2$ and centered at the point $(0, b)$, given by:

$$\psi_1(t) = a \cos\left((\frac{\pi}{2} + 0.01)t + 3\frac{\pi}{2} \right), \ \ \psi_2(t) = b \sin\left((\frac{\pi}{2} + 0.01)t + 3\frac{\pi}{2} \right),$$
$$-1 \le t \le 1.$$

The third one is the non symmetric C^3-continuous curve

$$\psi_1(t) = t,\ -1 \le t \le 1, \quad \psi_2(t) = \begin{cases} \frac{t^4}{4}\ , -1 \le t \le 0, \\ \frac{t^4}{2}\ ,\ \ 0 < t \le 1, \end{cases}$$

while the last one is the (non symmetric and C^2) open curve defined by the following natural (smooth) cubic spline:

$$\psi_1(t) = t,\ -1 \le t \le 1,$$

$$\psi_2(t) = \begin{cases} \frac{a+b}{4}(1+t)^3 - \left(a + \frac{a+b}{4}\right)(1+t) + a\ , -1 \le t \le 0, \\ \frac{a+b}{4}(1-t)^3 + \left(b + \frac{a+b}{4}\right)t - \frac{a+b}{4} \quad ,\ \ 0 < t \le 1, \end{cases}$$

where we have chosen $a = 0.1, b = 0.25$.

In Tables 1, 2, 3, and 4 we report the (global) error, defined by (46), and the errors $|\beta_N^* - \beta_n^*|$ and $|\gamma_N^* - \gamma_n^*|$, we have obtained for some values of n and N. In all examples, we take $\gamma_0 = -1$ (cf. (8)). Moreover, in the last two tables we also present some values $n^r *$ err for an appropriate r in order to determine the convergence rate, where err is given by (46). We can see that the convergence rate is higher than that forecasted by Proposition 3 (remember that ψ is C^3 in Example 3 and C^2 in Example 4.

Table 1 Example 1: non symmetric circular arc, $N = 256$

n	(46)	β_n^*	γ_n^*	$\lvert\beta_N^* - \beta_n^*\rvert$	$\lvert\gamma_N^* - \gamma_n^*\rvert$	cond($\mathbb{B}_n$)	cond($\mathbb{A}_n$)
4	8.99e-04	−0.6926674	0.1832556	6.68e−06	1.77e−06	2.5770	3.3097
8	3.68e-08	−0.6926607	0.1832538	1.18e−14	9.74e−15	2.5771	5.2093
16	9.98e-14	−0.6926607	0.1832538	1.22e−14	3.97e−15	2.5771	9.1997
256		−0.6926607	0.1832538			2.5771	130.1899

Table 2 Example 2: symmetric ellipse arc, $N = 256$

n	(46)	β_n^*	γ_n^*	$\lvert\beta_N^* - \beta_n^*\rvert$	$\lvert\gamma_N^* - \gamma_n^*\rvert$	cond($\mathbb{B}_n$)	cond($\mathbb{A}_n$)
4	6.05e-03	−0.5984153	0.0000000	1.88e−04	6.40e−16	2.6788	2.8100
8	1.44e-04	−0.5982318	0.0000000	4.92e−06	1.21e−15	2.6919	4.0774
16	7.15e-07	−0.5982269	0.0000000	9.14e−10	1.03e−15	2.6918	7.0137
32	1.25e-10	−0.5982269	0.0000000	2.22e−16	8.98e−16	2.6918	13.0283
256		−0.5982269	0.0000000			2.6918	97.6167

Table 3 Example 3: non symmetric C^3 arc, $N = 256$

n	(46)	β_n^*	γ_n^*	$\|\beta_N^* - \beta_n^*\|$	$\|\gamma_N^* - \gamma_n^*\|$	cond($\mathbb{B}_n$)	cond($\mathbb{A}_n$)	n^4·err
4	1.42e-03	−0.5671039	0.0227110	1.69e−04	8.82e−06	2.0341	2.7058	0.3633411
8	2.65e-05	−0.5669336	0.0227055	1.70e−06	3.22e−06	2.0339	4.4381	0.1084367
16	5.27e-07	−0.5669351	0.0227025	1.46e−07	2.87e−07	2.0339	8.0212	0.0345319
32	3.58e-08	−0.5669353	0.0227022	1.06e−08	2.13e−08	2.0339	29.6341	0.0375489
64	2.34e-09	−0.5669353	0.0227022	7.09e−10	1.44e−09	2.0339	15.2228	0.0394176
128	1.41e-10	−0.5669353	0.0227022	4.31e−11	8.75e−11	2.0339	58.4577	0.0403184
256		−0.5669353	0.0227022			2.0339	116.1042	

Table 4 Example 4: non symmetric C^2 arc, $N = 512$

n	(46)	β_n^*	γ_n^*	$\lvert\beta_N^* - \beta_n^*\rvert$	$\lvert\gamma_N^* - \gamma_n^*\rvert$	cond($\mathbb{B}_n$)	cond($\mathbb{A}_n$)	$n^{\frac{5}{2}}$·err
4	2.10e-04	−0.6297931	0.0036251	9.66e−05	3.02e−05	2.1603	2.7822	0.0067258
8	2.32e-05	−0.6297061	0.0036502	9.64e−06	5.09e−06	2.1601	4.5135	0.0042078
16	3.50e-06	−0.6296973	0.0036545	8.24e−07	7.73e−07	2.1601	8.0943	0.0035876
32	6.28e-07	−0.6296965	0.0036552	6.22e−08	1.08e−07	2.1601	15.2931	0.0036382
64	1.16e-07	−0.6296965	0.0036553	4.34e−09	1.43e−08	2.1601	29.6985	0.0037854
128	2.07e-08	−0.6296965	0.0036553	2.90e−10	1.82e−09	2.1601	58.5096	0.0038293
512		−0.6296965	0.0036553			2.1601	231.3716	

References

1. Berthold, D., Hoppe, W., Silbermann, B.: A fast algorithm for solving the generalized airfoil equation. J. Comput. Appl. Math. **43**(1–2), 185–219 (1992). In: Monegato, G. (ed.) Orthogonal Polynomials and Numerical Methods
2. Capobianco, M.R., Criscuolo, G., Junghanns, P.: A fast algorithm for Prandtl's integro-differential equation. ROLLS Symposium (Leipzig, 1996). J. Comput. Appl. Math. **77**(1–2), 103–128 (1997)
3. Demasi, L., Dipace, A., Monegato, G., Cavallaro, R.: Invariant formulation for the minimum induced drag conditions of non-planar wing systems. AIAA J. **52**(10), 2223–2240 (2014)
4. Demasi, L., Monegato, G., Dipace, A., Cavallaro, R.: Minimum induced drag theorems for joined wings, closed systems, and generic biwings: theory. J. Optim. Theory Appl. **169**(1), 200–235 (2016)
5. Gohberg, I., Krupnik, N.: Introduction. In: One-Dimensional Linear Singular Integral Equations, vol. I. Birkhäuser, Basel, Boston, Berlin (1992)
6. Hartmann, T., Stephan, E.P.: Rates of convergence for collocation with Jacobi polynomials for the airfoil equation. J. Comput. Appl. Math. **51**(2), 179–191 (1994)
7. Junghanns, P., Monegato, G., Strozzi, A.: On the integral equation formulations of some 2D contact problems. J. Comput. Appl. Math. **234**(9), 2808–2825 (2010)
8. Saff, E.B., Totik, V.: Logarithmic Potentials with External Fields. Grundlehren der Mathematischen Wissenschaften, vol. 316. Springer, Berlin/Heidelberg (1997)
9. Szegö, G.: Orthogonal Polynomials. American Mathematical Society, Providence, RI (1939)

Hyperbolic Conservation Laws and L^2

Barbara Lee Keyfitz and Hao Ying

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Taking as background the fact that conservation laws in a single space variable are well-posed in the space of functions of bounded variation, while multidimensional systems enjoy short-time well-posedness in Sobolev spaces H^s, we attempt to resolve the discrepancies between these two theories by exploring what can be said about stability of one-dimensional systems in L^2. We summarize some positive results for special cases, and also show by a conterexample that there is no straightforward way to resolve the difficulty.

1 Motivation

The topic of this paper is well-posedness for systems of hyperbolic conservation laws of the form

$$\frac{\partial F_0(\mathbf{u})}{\partial t} + \sum_1^d \frac{\partial F_i(\mathbf{u})}{\partial x_i} = \mathbf{0}\,, \quad \mathbf{u} \in \mathbb{R}^n\,, \quad \mathbf{x} \in \mathbb{R}^d\,. \tag{1}$$

Here $F_0 \in \mathbb{R}^n$ is a vector of conserved quantities, functions of the state variables $\mathbf{u}$, and the vectors F_i, $1 \le i \le d$, are flux vectors. Examples of such systems abound in mechanics, fluid dynamics, aerodynamics, and similar contexts. The systems that arise in such applications have a number of features in common, which we will draw

B. L. Keyfitz (✉)
Mathematics Department, The Ohio State University, Columbus, OH, USA
e-mail: keyfitz.2@osu.edu

H. Ying
Bank of America, Charlotte, NC, USA
e-mail: ying.32@osu.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_31

upon to motivate some constraints on the nonlinear functions F_i. One important constraint is that the system be symmetrizable hyperbolic. The meaning of this is as follows. Upon carrying out the differentiations in (1), one obtains a system where the vectors of derivatives are multiplied by $n \times n$ Jacobian matrices. This results in the so-called *quasilinear form* of the system:

$$A_0(\mathbf{u})\frac{\partial \mathbf{u}}{\partial t} + \sum_1^d A_i(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x_i} = \mathbf{0}\,, \tag{2}$$

with each $A_i = dF_i$. The original system is *symmetric hyperbolic* if all the A_j are symmetric and A_0 is in addition positive definite. It is often the case that although the Jacobian derivatives A_i are not symmetric, the system can be multiplied by a positive definite matrix that puts all the coefficient matrices simultaneously in symmetric form, with the matrix multiplying the first vector still positive definite. In that case, the system is said to be *symmetrizable hyperbolic*. Typically, it is also the case that the hyperbolicity conditions hold only for $\mathbf{u}$ in some open set $\mathscr{G} \subset \mathbb{R}^n$. For example, the equations of compressible gas dynamics, comprising the system (13) described in Sect. 2, give rise to a positive definite matrix A_0 only when the density ρ is positive.

We consider the Cauchy problem for Eq. (2) with the initial condition

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})\,.$$

For sufficiently smooth initial data, $\mathbf{u}_0 \in H^s(\Omega)$, where $\Omega \subset \mathbb{R}^d$ and $s > d/2 + 1$, there is an equally smooth solution in some time interval $[0, T]$, with $T > 0$. Specifically, we have the existence, uniqueness and continuous dependence on data of a solution

$$\mathbf{u} \in \mathscr{C}([0,T], H^s(\Omega)) \cap \mathscr{C}^1([0,T], H^{s-1}(\Omega))\,.$$

Various forms of this result were proved by Leray and Ohya [16], Lax [13], and Kato [10]. For a modern exposition of this theory, see Majda [18], Serre [21] or Taylor [23]. Two common choices for Ω are $\mathbb{R}^d$ itself and $\mathbb{T}^d$, corresponding to periodic data. To avoid some non-trivial considerations associated with boundary conditions, we will focus on the case $\Omega = \mathbb{R}^d$ in this paper.

The fact that well-posedness theory for system (1) or (2) has been developed in the Sobolev space H^s is significant. We expand on this in Sect. 1.1. Solutions in H^s, for sufficiently large s, are known as *classical* solutions; indeed, from the Sobolev Embedding Theorem, if $s > d/2 + 1$ the solutions are continuously differentiable (that is $\mathscr{C}^1$) in $\mathbf{x}$, and hence also in t.

Systems (1) and (2) are clearly reversible in time, and classical solutions to the Cauchy problem with $\mathbf{u}_0 \in H^s$, $s > d/2 + 1$, exist on an open interval around $t = 0$. In [18, 21, 23] cited above it is proved that for almost all initial data the lifespan of classical solutions is finite, and when there is a maximal time of existence, then either $\|\nabla \mathbf{u}\|_\infty$ becomes infinite as t tends to a critical value T^* or $\mathbf{u}$ leaves every compact subset of $\mathscr{G}$ as $t \to T^*$. Beyond that time, if solutions to (1) exist, it can only be that they take the form of *weak* solutions. Finding a proof of existence of weak solutions for multidimensional conservation laws is an open problem in the analysis of quasilinear hyperbolic partial differential equations.

An important point here is that the conservation form of (1) is required in the definition of a weak solution for hyperbolic systems. This follows from the distribution theory definition of a weak solution. In the case of (1), multiplying by a smooth test function and integrating by parts removes all derivatives of $\mathbf{u}$. Then we say that $\mathbf{u}$ is a *weak solution* in the sense of distributions if $\mathbf{u}$ and $F_j(\mathbf{u})$ are locally integrable functions that satisfy the integrated equation for all smooth test functions. In the case of a general quasilinear system (2), if the matrices A_i are not derivatives of functions F_i then after integration by parts the system still contains derivatives of $\mathbf{u}$, multiplying functions of $\mathbf{u}$. The difficulty which this poses for nonlinear systems is that while derivatives of integrable functions are well-defined as distributions, multiplication of integrable functions by distributions is not well-defined in general. Hence, one does not expect to find a theory for weak solutions of any quasilinear systems other than systems that arise from conservation laws of the form (1). We emphasize that there is at present no existence theory for weak solutions of systems of conservation laws in dimensions $d > 1$. Not only is there no general theory, but to our knowledge there are not even any examples of systems where existence of nontrivial solutions to the Cauchy problem has been proved.

In practice, the criterion that defines a weak solution of (1) as a distribution satisfying the integrated form of the equation is not sufficient to define a weak solution uniquely. A definition of weak solution that guarantees well-posedness was developed by Kružkov [12] for scalar equations ($n = 1$). For systems in a single space variable ($d = 1$), the original existence theorem for weak solutions is due to Glimm [6]. Much more recently it has been extended by Bressan and his colleagues (see [3] and the references therein) to well-posedness results, though, as is also the case for Glimm's result [6], only for data close to a constant. Dafermos's monograph [5] provides a good exposition of the one-dimensional theory.

In one space dimension, the theory is based on constructing approximate solutions, and the construction incorporates some constraints, usually known as "entropy conditions", that guarantee uniqueness. Entropy conditions also, as the name suggests, destroy the time-reversibility of the system. Both Glimm and Bressan considered data and solutions with $\mathbf{u}(\cdot, t)$ in the function space $\mathscr{BV} \cap L^1$. (By $\mathscr{BV}$ we mean the space of functions of bounded variation; L^1 can generally be replaced by L^1_{loc}, to handle data that approach different constants at $x = -\infty$ and $x = \infty$.) The usual statement is that "conservation laws in one space dimension are well-posed in $\mathscr{BV}$". The spaces $\mathscr{BV}$ do not fit easily into the classification of Sobolev spaces. In particular, $\mathscr{BV} \cap L^1$ contains as a proper subset the Sobolev space $W^{1,1}$ consisting of absolutely continuous functions of a single variable. However, functions in $\mathscr{BV}$ may have countably many discontinuities, and $\mathscr{BV}$ is not a separable Banach space.

A simple example shows that solutions to conservation laws have better properties in L^1 than in L^2. The prototype for a scalar, one-dimensional equation is Burgers' equation,

$$u_t + uu_x = 0\,.$$

A typical feature, both for Burgers' equation and more complex systems, is a *centered rarefaction wave*. An example of such a wave is

$$u(x,t) = \begin{cases} 0\,, \; x < 0\,, \\ \frac{x}{t}\,, \; 0 < x < t\,, \\ 1\,, \; x > t\,, \end{cases} \tag{3}$$

a solution of Burgers' equation. This has first derivative $u_x = 1/t$ for $0 < x < t$, and so

$$\int_0^t |u_x(x,t)|^p \, dx = t^{1-p}$$

is unbounded for $p > 1$. In this example, the $W^{1,p}$ norm blows up for $p > 1$ as t decreases to zero. This solution is time-reversible and we see that for $u = x/(1-t)$, corresponding to a compression wave, the $W^{1,p}$ norm, for $p > 1$, increases without bound until $t = 1$ when a shock forms.

Shocks, the other typical feature of Burgers' and all conservation law systems, are jump discontinuities. The indicator function of an interval is in $W^{s,p}$ for $sp < 1$, which means that in one dimension a function with a jump discontinuity is in $W^{s,1}$ only for $s < 1$ and in $W^{s,2}$ only for $s < 1/2$.

Linear hyperbolic equations in any number of space dimensions are well-posed in $H^s = W^{s,2}$, for any s, positive or negative, while, as we noted above, quasilinear hyperbolic systems are locally well-posed (that is, for short times) in H^s for $s > d/2 + 1$, where d is the space dimension. Specifically, this rules out times after shocks have formed. The monograph of Taylor [23] has complete statements on well-posedness for both linear and quasilinear systems. The definitions of Sobolev spaces $W^{s,p}$ for $p \neq 2$ and s not an integer are somewhat intricate. As we will not use them in this paper, we omit the definitions, but refer to [25] for details.

If the space dimension is greater than one, then most hyperbolic equations are well-posed *only* in Sobolev spaces based on L^2, and not in L^p for any $p \neq 2$. We expand on this in Sect. 1.1. This important consideration leads to the current quest for a function space other than $\mathscr{BV}$ in which to seek well-posedness, even for systems in a single space dimension.

Now we can explain the point of this paper. A linear system with constant or bounded coefficients is not sensitive to scalings of the form $u \mapsto \alpha u$ where α is a constant scale factor. In particular, this means that, no matter what norm is used, bounds of the form

$$\|\mathbf{u}(\cdot,t)\| \leq C(t)\|\mathbf{u}(\cdot,0)\| \tag{4}$$

are precisely the types of bounds one might seek in order to confirm wellposedness for linear equations. Expectations for continuity of the data-to-solution map for quasilinear equations might be quite different. First, of course, any candidates for useful bounds will be of the form

$$\|\mathbf{v}(\cdot,t) - \mathbf{w}(\cdot,t)\| \le C(t)\|\mathbf{v}(\cdot,0) - \mathbf{w}(\cdot,0)\|\,, \tag{5}$$

in the norm under consideration (this property is often called "stability"), and in addition might require pointwise restrictions of the form $\mathbf{u}(x,t) \in \mathscr{G}$, where $\mathscr{G} \subset \mathbb{R}^n$ was defined in the discussion following (2) as the subset of state space in which the system is symmetric hyperbolic. Pointwise constraints are consistent with the existence theorems for classical solutions, which show that solutions remain in a compact subset of $\mathscr{G}$ for some time.

Another difficulty in attempting to establish well-posedness for weak solutions of quasilinear systems (1) is that we expect it may be necessary to impose more restrictive size constraints of the form $\|\mathbf{u}(\mathbf{x},t) - \text{const}\| < \varepsilon$, since such constraints are required for well-posedness of systems in one space dimension. The value of ε depends on the geometry (nonlinearity) of the flux vectors. In one space dimension, when the data $\mathbf{u}_0$ are bounded in $\mathscr{BV}$ then the solution $\mathbf{u}(\cdot,t)$ is uniformly bounded in $\mathscr{BV}$ for all t, and this implies that the solution satisfies uniform pointwise bounds. However, it may be the case that in $d > 1$ space dimensions we should not expect pointwise boundedness. For example, in the case of the linear wave equation, there is a generic phenomenon of focusing, which produces solutions that become unbounded on lower dimensional subsets. In the case of nonlinear systems the compressible gas dynamics equations are an example where lower bounds for the density may not exist because a vacuum state at which hyperbolicity fails may form spontaneously.

1.1 Expectations for Well-Posedness

The fact, mentioned earlier, that hyperbolic systems are not well-posed in any Sobolev space $W^{s,p}$ for $p \neq 2$, was originally proved for the wave equation by Littman [17], and later for general first order linear hyperbolic systems by Brenner [2]. The observation that Brenner's argument extends to quasilinear systems is due to Rauch [20]. Littman's argument is easy to articulate. Looking at the case $d = 2$ (from which all higher-dimensional cases follow), one can check that

$$u(x,y,t) \equiv u(r,t) = \frac{1}{2\pi}\int_r^t \frac{g(t-\tau)}{\sqrt{\tau^2 - r^2}}\,d\tau\,, \tag{6}$$

is a rotationally invariant solution of the wave equation

$$u_{xx} + u_{yy} - u_{tt} \equiv u_{rr} + \frac{1}{r}u_r - u_{tt} = 0 \tag{7}$$

whenever $t \notin \operatorname{supp} g$. Littman defines the L_p energy of a solution,

$$E_p(u;t) \equiv E_p(t) \equiv \int_{\mathbb{R}^2} \left(|u_x|^p + |u_y|^p + |u_t|^p\right) dx\, dy\,, \tag{8}$$

for any p, and then for a suitable choice of g, with support in $[0, t_0]$, shows that for a fixed $t_1 > t_0$, the ratio $E_p(t_1)/E_p(t_0)$ tends to zero or to infinity as $t_0 \to 0$. This follows from the form of the solution in (6). In fact, if g were constant then we would have, in polar coordinates $(x, y) = (r\cos\theta, r\sin\theta)$,

$$u(r,t) = \log\left(\frac{1+\sqrt{1-\xi^2}}{\xi}\right) \equiv U(\xi)\,, \quad \xi = \frac{r}{t}\,, \quad 0 < \xi \le 1\,. \tag{9}$$

That is, u depends on r and t through the ratio r/t alone, and

$$u_t = -\frac{r}{t^2}U'(\xi) = -\frac{1}{t}\xi U'(\xi) \quad \text{and} \quad u_r = \frac{1}{t}U'(\xi)\,. \tag{10}$$

We may ignore the difference between $|u_r|^p$ and $|u_x|^p + |u_y|^p$. We find that the L_p energy defined by (8) at time t is

$$\begin{aligned} E_p(u;t) &= \int_0^t \left(|u_t|^p + |u_r|^p\right) r\, dr = t^2 \int_0^1 \left(|u_t|^p + |u_r|^p\right)\xi\, d\xi \\ &= t^{2-p}\int_0^1 (\xi^p + 1)|U'(\xi)|^p\, d\xi\,. \end{aligned} \tag{11}$$

This equation makes it clear that $E_p(t_1)/E_p(t_0)$ is bounded, above and below, only if $p = 2$. The simple argument presented here is merely heuristic, since u does not satisfy the homogeneous wave equation with this choice of g. To produce a solution of (7) for all r and t, one needs a better choice for g in (6), and Littman's paper [17] achieves this and turns the simple estimate in (11) into a rigorous proof. The expression in (9), which is closely related to the fundamental solution of the wave equation, gives the basic reason that we expect E_p to behave badly for any $p \neq 2$.

The theorem of Brenner [2] is much more general; it applies to any system of the form (2) where the A_j are constant. Brenner's result is that unless either all the A_j commute or $d = 1$, then no bound in $W^{s,p}$ of the form (4) can exist for $p \neq 2$. The somewhat startling condition of commuting matrices can be explained as follows. The fundamental solution of the wave equation, which appears in Eq. (6), comes directly from the equation for the characteristic normals, $\tau^2 = \xi^2 + \eta^2$. A homogeneous quadratic equation for the characteristic normals appears because the characteristic determinant of (the first-order system derived from) the wave equation cannot be factored into linear factors. Now observe that

$$\det\left(A_0\tau + \sum A_i\xi_i\right)$$

factors into linear factors if and only if the matrices have a common basis of eigenvectors, which is the case precisely when the matrices commute. It is easy to verify that if the two-dimensional wave equation is written as a first-order system, the matrices do not commute. This makes intuitively reasonable Brenner's result that any system where the solution, or any part of it, behaves like a solution of the wave equation will be well-posed only when $p = 2$. In the case of commuting eigenvectors, all components of $\mathbf{u}$ (in a coordinate system where the matrices are simultaneously diagonalized) propagate in time like solutions $u = f(x - at)$ of a linear transport equation, $u_t + au_x = 0$. Thus the norm of solutions is well-controlled in time, in any norm. Brenner's masterful proof uses Hörmander's theory of Fourier multipliers in L^p, but the result, when you see the connection to Littman's proof, is not unexpected.

To complete this argument, Rauch [20] showed that, at least as long as classical solutions exist, injecting some nonlinearity into the system cannot cure the difficulty. Again, intuitively, one can see the reason for this in Littman's original observation that for the mode of wave propagation associated with the wave equation the L^2 norm, alone, is well-controlled in time.

Dafermos [4] has proved that one can actually show L^2 stability—that is, inequality (5) using the L^2 norm—for conservation law systems in any number of space variables when only one of the functions $\mathbf{v}$ and $\mathbf{w}$ is a classical solution, provided the other is an admissible weak solution. His proof requires only that the system have a convex entropy function (which is equivalent to the system being symmetrizable hyperbolic). This is not a significant restriction. However, the proof relies in an essential way on the hypothesis that one solution is classical. Entropies and an explanation of Dafermos' result [4] are the subject of Sect. 2 of this paper.

1.2 Solution Bounds and Derivative Bounds

The derivatives of a linear system satisfy the original system of equations up to lower-order terms; and so if the data and solution are in H^s, then first derivatives are in H^{s-1} and so on. Among other things, this makes it reasonable to talk about distribution solutions of linear systems, and to expect well-posedness even for negative values of s. The situation is different for quasilinear systems. A curious situation obtains here. The first derivatives satisfy a system whose structure is different from that of the original system in an important way, but the form of the equations for higher derivatives reverts to that of the original system. A simple calculation using Burgers' equation as an example should suffice to make the point. Differentiating $u_t + uu_x = 0$ with respect to x and letting $v = u_x$ we obtain

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right) v + v^2 = 0\,, \tag{12}$$

from which one discovers that v becomes infinite in finite time along the characteristic from $(x_0, 0)$ with speed u if $v(x_0, 0) = u_x(x_0, 0) < 0$. This is the well-known gradient catastrophe leading to shock formation, and Eq. (12) is just another way to see it. This observation is, by the way, the basis of Fritz John's famous proof in [9] that for genuinely nonlinear hyperbolic systems in one space dimension data of compact support always lead to the formation of shocks.

However, if we differentiate Burgers' equation a second time with respect to x and let $u_{xx} = w$ then we have

$$\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right) w + 3vw = 0\,.$$

That is, the higher-order derivatives of the solution u satisfy linear equations with coefficients that depend on the lower-order derivatives. These do not generate further singularities. This fact was used in the proof of nonuniform dependence of solutions on initial data for the equations of compressible hydrodynamics in [7] and [11]. Proofs that it holds in the general case can be found in Majda's book [18, Theorem 2.2, Corollary 1], and in Taylor's [23, Chapter 16, Corollary 1.6].

2 Entropies in Conservation Laws

This short section gives some background intended to help understand the role of entropy functions in proving stability. It is not needed for the main result in Sect. 4.

We illustrate the notion of entropy with an example. The system of two equations representing isentropic gas dynamics in one space dimension is

$$\begin{aligned} \rho_t + m_x &= 0\,, \\ m_t + \left(\frac{m^2}{\rho} + \frac{1}{\gamma}\rho^\gamma\right)_x &= 0\,. \end{aligned} \tag{13}$$

Here ρ is the density and m the momentum of the gas at a point x and time t, suitably non-dimensionalized. The constant γ, the so-called ratio of specific heats, is typically between 1 and 3; for air it is approximately 1.4. The two equations represent conservation of mass and momentum. That they form a closed system results from an assumption that the non-dimensionalized pressure is a function of the density, $p(\rho) = \rho^\gamma/\gamma$. A brief calculation verifies that if ρ and m are differentiable solutions to these two equations then a third equation also holds:

$$\eta_t + q_x \equiv \left(\frac{m^2}{\rho} + \frac{2}{\gamma(\gamma-1)}\rho^\gamma\right)_t + \left(\frac{m^3}{\rho^2} + \frac{2}{\gamma-1}m\rho^{\gamma-1}\right)_x = 0\,. \tag{14}$$

The quantity η is a strictly convex function of ρ and m; it is the specific energy (kinetic plus potential), again suitably scaled; q is called the energy flux. Now, weak solutions of (13) will not in general satisfy (14). For example, at a shock discontinuity with speed $dx/dt = s$ the weak form of a conservation law system $\mathbf{u}_t + F(\mathbf{u})_x = 0$ is $s[\mathbf{u}] = [F(\mathbf{u})]$, where the brackets indicate jumps in quantities across the discontinuity, and the weak form of (13) is inconsistent with the weak form of (14). In short, although the energy is also a conserved quantity for classical solutions, energy is not conserved in the presence of discontinuities.

One method of identifying *admissible* weak solutions is to require that for any weak solution the quantity

$$\int_{-\infty}^{\infty} \eta(x,t)\, dx \equiv \int_{-\infty}^{\infty} \left(\frac{m^2}{\rho} + \frac{2}{\gamma(\gamma-1)}\rho^{\gamma} \right) dx \tag{15}$$

not increase in time. This has the reasonable physical interpretation that energy, conserved for classical solutions, will decrease rather than increase in time for weak solutions. If one considers, instead of (13), a more complete formulation of gas dynamics with an additional equation for conservation of energy and a third state variable, either the internal energy or the pressure, and closes the system with a modeling assumption about the energy flux, then there is a fourth equation that follows from the first three for classical solutions. In this case, the fourth equation represents conservation of entropy. In this more complete formulation, entropy is the quantity that fails to be conserved when classical solutions break down. True thermodynamic entropy is a concave function of density, momentum and pressure, and is an increasing function of time for weak solutions. Mathematical terminology changes the sign to obtain a convex function and in conservation law parlance we always speak of "entropy decrease". It can be proved that a conservation law system that can be put in symmetric hyperbolic form always possesses a convex entropy, and conversely [4].

There are several ways that the existence of a convex entropy like (15) has been used to establish some L^2 stability results.

First, a smooth strictly convex function can be approximated locally by a quadratic function, so $\eta(\rho, m)$ above is equivalent to a homogeneous quadratic function, and hence bounds the L^2 norm. The same is true for any strictly convex entropy for a symmetrizable hyperbolic system in the form (1), in any number of space variables.

Second, one can look at the difference $\eta(\mathbf{v}) - \eta(\tilde{\mathbf{w}})$, where one of $\mathbf{v}$ and $\tilde{\mathbf{w}}$ is a classical solution and the other is a weak solution, admissible in the sense that the convex entropy function decreases in time. The L^2 norm of $\eta(\mathbf{v}) - \eta(\tilde{\mathbf{w}})$ will satisfy an inequality which can be used to bound the L^2 difference of $\mathbf{v} - \tilde{\mathbf{w}}$ as a function of t. This nice result is due to Dafermos [4]. The relative entropy method used to obtain the results in Sect. 3.1 is a variant of this approach.

3 Current Progress on L^2 Stability in One Space Dimension

Given the background of Sect. 1, it appears that a promising direction in which to resolve the gap between results for weak solutions in one dimension and results for classical solutions in multidimensional systems is to establish results in L^2 for one-dimensional systems. In this section, we summarize some interesting results. However, we have found a significant obstruction to this program, which we present in Sect. 4.

For the remainder of this paper, we assume $d = 1$.

3.1 Stability in L^2 of a Single Shock

Recent interest in L^2 results for conservation laws in one space dimension has been stimulated by work of Vasseur and his student Leger. We mention the papers [14, 15] and [22], though this is not a complete list. The technique exploited in this research is the *relative entropy method*. In Sect. 2 we noted that "entropy" is one of several admissibility criteria, and that convex entropies also provide a way of bounding L^p norms of solutions for $p > 1$. Relative entropy is a way of measuring the difference between two solutions, or between a solution and a constant (solution). For scalar equations, this idea is the basis of Kružkov's definition [12] of a weak solution for scalar equations in any number of space dimensions. Specifically, in [12] Kružkov gave a complete characterization of admissible weak solutions for a scalar equation by comparing a candidate for a weak solution with all constant solutions. No generalization of this characterization to systems is known, even in a single space dimension.

The research of Leger, Vasseur and others looks at a specific type of stability: the behavior of a shock wave under perturbation. For a one-dimensional system,

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}F(\mathbf{u}) = 0\,,$$

they consider a self-similar shock discontinuity of the form

$$\mathbf{w}(x,t) = \begin{cases} \mathbf{w}_L\,, & x < \sigma t\,, \\ \mathbf{w}_R\,, & x \geq \sigma t\,, \end{cases}$$

where $\mathbf{w}_L$, $\mathbf{w}_R$ and σ satisfy the Rankine-Hugoniot relation

$$\sigma(\mathbf{w}_R - \mathbf{w}_L) = F(\mathbf{w}_R) - F(\mathbf{w}_L)\,,$$

which is necessary for a discontinuous function to be a weak solution. They compare this solution to a weak solution $\mathbf{v}$ corresponding to a nearby initial condition $\mathbf{v}_0$.

The result in [15] applies to *extremal* shocks. Extremal shocks have the property that on one side ($x < 0$ for the slowest, $x > 0$ for the fastest waves) there are no outgoing characteristics. Stability is proved under the condition that the perturbation is smaller on the upstream side of the shock ($x < 0$ for the slowest shocks), where all the characteristics are incoming, than downstream (the half-line $x > 0$ in the case of slow shocks), where there are $n - 1$ outgoing characteristics in a system of n equations. Their result is stated in terms of the integral of the squared distance between **w** and **v** on each side of the discontinuity, with different conditions on the two sides. One might think of these as "one-sided" L^2 norms. Specializing to a slow shock for concreteness, we summarize their result by saying that if the difference between $\mathbf{v}_0$ and $\mathbf{w}_L$ in this one-sided L^2 norm (integrating from $-\infty$ to 0) is of order ε^2, and the L^2 norm of the perturbation $\mathbf{v}_0 - \mathbf{w}_R$ on the right is of order $\sqrt{\varepsilon}$, Leger and Vasseur [15] prove that the path of the perturbed shock, $s(t)$, will differ from the unperturbed path by an amount also of order $\sqrt{\varepsilon}$:

$$|s(t) - \sigma t| \leq C\sqrt{\varepsilon t(1+t)}\,, \tag{16}$$

and the difference between the perturbed solution and its unperturbed shadow, up to that translation, is of the same order as the perturbation in the initial data; that is,

$$\int_{-\infty}^{0} |\mathbf{v}(x+s(t),t) - \mathbf{w}_L|^2\,dx \leq \varepsilon^4\,, \quad \int_{0}^{\infty} |\mathbf{v}(x+s(t),t) - \mathbf{w}_R|^2\,dx \leq C\varepsilon(1+t)\,.$$

There is a similar result, with the bounds on the left and right sides of the shock reversed, for a fast shock. It is noted by Texier and Zumbrun [24] that the relative entropy technique allows a strengthening of earlier results using other, weaker entropy criteria.

In an earlier paper on scalar equations, Leger [14] finds a consistent result. For a scalar equation with convex flux, any shock is extremal from both sides and there are no outgoing characteristics, so the tighter restriction of the perturbation to order ε^2 on one side is not needed. In addition, it is not necessary to assume that the perturbation is small. Leger's result in the L^2 norm is of the form

$$\|v(\cdot + s(t) + \sigma t, t) - w(\cdot)\|_2 \leq \|v_0(\cdot) - w(\cdot)\|_2\,,$$

where w is the shock solution to the Riemann problem with states w_L and w_R and speed σ, and $s(t)$ now satisfies

$$|s(t)| \leq C\|v_0 - w\|_2\sqrt{t}\,,$$

where C may depend on $\|u_0\|_\infty$, on w, and on the flux function.

An interesting paper by Adimurthi et al. [1] generalizes Leger's result to obtain stability of a single shock in L^p, for any $p > 1$, for a convex scalar equation. These authors use a completely different method, the well-known Lax-Oleĭnik formula

[13] which relates the conservation law to a Hamilton-Jacobi equation. The Lax-Oleĭnik formula is valid only for a scalar, one-dimensional equation, but the proof in [1] could possibly be made more general, since the argument relies principally on tracing characteristics and on generalized characteristics.

3.2 An Example of Stability

For a scalar conservation law, it has been known for a long time, see [12], that the stability bound (5) holds in L^1 with $C(t) \equiv 1$. The following elementary example shows that one cannot expect an L^2 estimate of this nature, even for a scalar equation. Consider two solutions of Burgers equation with square-wave data

$$w(x,0) = \begin{cases} 0\,, & x < 0\,,\ x > \ell \\ A\,, & x \in [0,\ell] \end{cases}\,, \quad v(x,0) = \begin{cases} 0\,, & x < 0\,,\ x > \ell \\ A+\varepsilon\,, & x \in [0,\ell] \end{cases}\,.$$

We have $\|v - w\|_2(0) = \varepsilon\sqrt{\ell}$. For $t > 0$, the position of the shock in w is at $x_w(t) = \ell + At/2$ and the shock in v is at $x_v(t) = \ell + (A+\varepsilon)t/2$ so a calculation gives

$$\|v - w\|_2^2 = \frac{A^2 t}{2}\varepsilon + \left(\ell + \frac{At}{2}\right)\varepsilon^2 + \frac{\varepsilon^3 t}{6}\,,$$

for small t ($t < 2\ell/A$). Thus, $\|v - w\|_2 \geq A\sqrt{\varepsilon t/2}$ for small positive times. Continuing the calculation, for $t > 2\ell/A$, the shock and the rarefaction generated by the initial data interact, and the shock strength and position change. For w, the position of the shock is $x_w(t) = \sqrt{2\ell At}$; the shock in v is at $x_v(t) = \sqrt{2\ell(A+\varepsilon)t}$, and a calculation of the L^2 norm yields

$$\|v - w\|_2^2 = \sqrt{2\ell^3 t}\left(\sqrt{A}\varepsilon + c\varepsilon^2 + \dots\right),$$

so the difference in the L^2 norms is of order $\sqrt{\varepsilon}$ for large t as well as for small t. This suggests the possibility of a stability property like

$$\|v(\cdot,t) - w(\cdot,t)\|_2 \leq C(t)\|v(\cdot,0) - w(\cdot,0)\|_2^{1/2}\,,$$

rather than the dependence suggested in (5).

In this piecewise constant example, the "mismatch" in shock speeds created by an L^2 difference of ε in the data causes a difference in the L^2 norm of solutions that is of size $\sqrt{\varepsilon}$. Leger and Vasseur's estimate (16) in [15] of the mismatch in shock speeds is optimal, and holds for an arbitrary small L^2 perturbation. The simple estimate in this example is also consistent with Leger's result [14] for a scalar equation, where a shock is simultaneously slow and fast. The occurrence of square roots in the example suggests that it might be possible to find an estimate of the form

$$\|\mathbf{v}(\cdot,t)-\mathbf{w}(\cdot,t)\|_2 \le C(t)\|\mathbf{v}(\cdot,0)-\mathbf{w}(\cdot,0)\|_2^{\alpha}\,, \tag{17}$$

in the L^2 norm, for some $\alpha < 1$. However, we show in Sect. 4 that this conjecture fails.

4 A Counterexample

Although the results of Leger, Vasseur and Serre [14, 15, 22] suggest that some form of L^2 stability may be possible, no general result of the form (17) can exist. We give a proof by constructing a counterexample for Burgers' equation. This immediately implies the result for any convex scalar equation, and for any system in a single space dimension where there is at least one genuinely nonlinear family (even locally). In more than one space dimension, our counterexample also suggests that an L^2 stability result will not be possible without further constraints.

The counterexample builds on the idea, demonstrated in the example in Sect. 3.2, that if two choices of step function data differ by an amount ε_n, then the L^2 difference in the corresponding solutions is $\sqrt{\varepsilon_n}$. Suppose that $\sum \varepsilon_n$ converges but $\sum \sqrt{\varepsilon_n}$ does not, and that the data and solutions are such that those sums bound their L^2 norms above and below. Then we have a demonstration of instability. To carry out this construction, we let v_0 and w_0, with $v_0 \ge w_0$, be two non-negative step functions of compact support, containing a large step upward and then decreasing in steps of unit length, as illustrated in Fig. 1. To be specific, for $N \ge 2$ and a positive parameter a to be determined, we define w_0 by

$$w_0(x) \equiv w_j = a(N-j+1)\,, \quad \text{for} \quad j-1 < x \le j\,, \quad \text{for} \quad 1 \le j \le N\,,$$

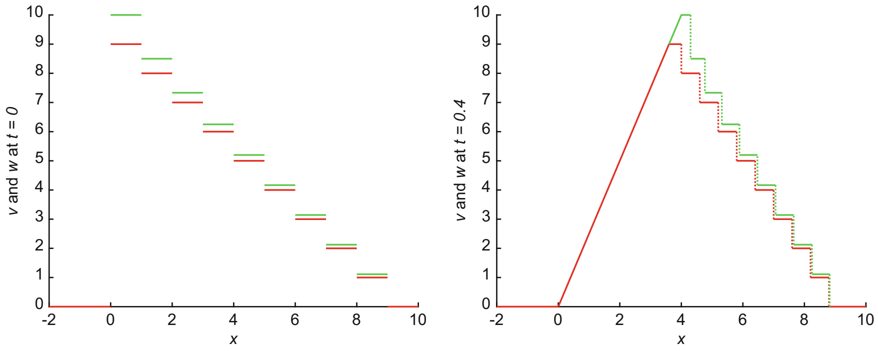


Fig. 1 Illustration of the example with $N = 9$, $a = b = 1$

and $w_0(x) = 0$ otherwise. Then take v_0, with support the same as w_0, and

$$v_0(x) = w_j + b_j\,, \quad \text{for} \quad j-1 < x \le j\,, \quad \text{and} \quad b_j = \frac{b}{j}\,, \quad \text{for} \quad 1 \le j \le N\,.$$

Here b is another positive parameter. For use in the summations below, define $w_{N+1} = 0$ and $b_{N+1} = 0$.

We have

$$u_0 \equiv v_0 - w_0 = \{b_j\}$$

(with obvious notation) and

$$\|u_0\|_2^2 = \sum_1^N b_j^2 = \sum_1^N \frac{b^2}{j^2} = b^2 S_N\,; \quad \|u_0\|_2 = b\sqrt{S_N}\,.$$

Now, $S_N \to \pi^2/6$ as $N \to \infty$. The L_1 norm of u_0 is $b\sum 1/j$, which is finite but unbounded as $N \to \infty$. Later in the argument we will see that we can adjust a and b as functions of N so that the L^1 and L^∞ norms of the data remain bounded. Thus we cannot destroy the counterexample by a introducing a constraint such as restricting consideration to small data.

Since v and w are solutions of Burgers' equation, $u_t + uu_x = 0$, in both cases the solution begins (reading from the left) with a rarefaction followed by N shocks. In the interior of the rarefaction, $u(x,t) = x/t$ as in the example (3). The rarefaction extends from $x = 0$ to $x = aNt$ (for w) and to $x = a(N + b_1)t$ (for v). For all j, $1 \le j \le N$, the jth shock in w (emanating from $x = j$) separates states w_j and w_{j+1} and has speed

$$s_{w,j} = \frac{w_j + w_{j+1}}{2} = \frac{a(N-j+1+N-j)}{2} = a\left(N-j+\frac{1}{2}\right)\,,$$

and position

$$j + s_{w,j}t = j + a\left(N-j+\frac{1}{2}\right)t\,.$$

The jth shock in v has speed

$$s_{v,j} = \frac{v_j + v_{j+1}}{2} = \frac{w_j + w_{j+1}}{2} + \frac{b_j + b_{j+1}}{2}$$
$$= \begin{cases} a\left(N-j+\frac{1}{2}\right) + \frac{b}{2}\left(\frac{1}{j} + \frac{1}{j+1}\right)\,, & 1 \le j < N\,, \\ \frac{1}{2}\left(a + \frac{b}{N}\right)\,, & j = N\,, \end{cases}$$

and position

$$\begin{cases} j + s_{w,j}t + \frac{b}{2}\left(\frac{1}{j} + \frac{1}{j+1}\right)t\,, & 1 \le j < N\,, \\ N + s_{w,N}t + \frac{bt}{2N}\,, & j = N\,, \end{cases}$$

so the corresponding shocks in the two solutions separate at the rate

$$\Delta_j t \equiv \begin{cases} \frac{b}{2}\left(\frac{1}{j} + \frac{1}{j+1}\right)t\,, & 1 \le j < N\,, \\ \frac{bt}{2N}\,, & j = N\,. \end{cases}$$

In the interval between two corresponding shocks, as in Sect. 3.2 example, the difference u between v and w is

$$u_j \equiv v_j - w_{j+1} = w_j + b_j - w_{j+1} > a\,,$$

since $b_j > 0$. In the data we have $v_0 \ge w_0$ and the same is true of the solutions. The L^2 difference between v and w is bounded below by the difference between v and w in these "gaps", and we can estimate

$$\|v(\cdot,t) - w(\cdot,t)\|_2^2 > \sum_1^N a^2 \Delta_j t = \frac{a^2 bt}{2}\left(\sum_1^{N-1}\left(\frac{1}{j} + \frac{1}{j+1}\right) + \frac{1}{N}\right)$$
$$= a^2 bt\left(\sum_1^N \frac{1}{j} - \frac{1}{2}\right) > a^2 bt\left(\int_1^{N+1} \frac{1}{x}\,dx - \frac{1}{2}\right) = a^2 bt\left(\log(N+1) - \frac{1}{2}\right)\,.$$

As a result, we get a lower bound for the difference:

$$\|u(\cdot,t)\|_2 = \|v(\cdot,t) - w(\cdot,t)\|_2 \ge a\sqrt{bt(\log(N+1) - 1/2)}\,,$$

as desired. To complete the argument, we note that we can choose t to be a fixed number, independent of a, b and N, as long as a and b are bounded above. For example, if $a \le 1$ and $b \le 1$, we find that the shocks do not intersect each other and do not intersect the rarefactions until some time after $t = 1/2$. Then for any $t \le 1/2$, we compute the ratio

$$\frac{\|v(\cdot,t) - w(\cdot,t)\|_2}{\|v_0 - w_0\|_2} \ge at^{1/2}\frac{\sqrt{\log(N+1) - 1/2}}{\sqrt{bS_N}}\,. \tag{18}$$

For fixed a and b, this grows without bound as $N \to \infty$.

To obtain two sequences $\left\{w_0^{(N)}\right\}$ and $\left\{v_0^{(N)}\right\}$ that violate L^2 stability even more dramatically, we note that

$$\|w_0\|_2 < \|v_0\|_2 = \left(\sum_1^N \left(a(N-j+1) + \frac{b}{j}\right)^2\right)^{1/2} . \tag{19}$$

We can achieve an arbitrarily large growth rate for data that are arbitrarily small. To do so, let us fix the ratio $a/\sqrt{b} = 1$, say, and fix an arbitrary constant M. Then to have $\|u(\cdot,t)\|_2/\|u_0\|_2 > M\sqrt{t}$, for all $t \leq 1/2$, it suffices to take

$$\log(N+1) > \frac{\pi^2}{6}M^2 , \quad \text{or} \quad N > e^{cM^2} ,$$

and once N has been so chosen, we can choose $a = a^{(N)}$ and $b = b^{(N)}$, using (19) to achieve a bound on the data in L^2. For any $\varepsilon > 0$, if

$$a \sim \frac{\varepsilon}{N^{3/2}} , \quad \text{and} \quad b = a^2 = \frac{\varepsilon}{N^3}$$

then we have (18) for data with L^2 norm initially of order ε. In fact, a similar calculation applied to the L^1 norms, where

$$\|w_0\|_1 < \|v_0\|_1 = \sum_1^N \left(a(N-j+1) + \frac{b}{j}\right) ,$$

shows that we can also achieve an arbitrarily large ratio in (18) with data that are small in L^1, and in $\mathscr{BV} \cap L^1$.

5 Conclusions

The construction in Sect. 4 used data of compact support, but involved a sequence of initial conditions whose support grew with N. This is an essential feature of the construction. At least for scalar equations, growth in the support of the data appears to be a necessary condition to obtain ill-posedness in L^2. For if $\operatorname{supp} u \subset \Omega$ then applying Hölder's inequality we have

$$\|u\|_1 = \int_\Omega |u|\,dx = \int_\Omega 1 \cdot |u|\,dx \leq \left(\int_\Omega 1\,dx \int_\Omega |u|^2\,dx\right)^{1/2} = (\operatorname{diam} \Omega)^{1/2}\|u\|_2 ,$$

while on the other hand we have the standard Hölder inequality

$$\|u\|_2 \leq \|u\|_\infty \|u\|_1 \,.$$

For scalar conservation laws, we can take advantage of both a maximum principle [8, 19] and L^1 contraction [12]. Thus, if v and w are both supported in Ω, we have

$$\begin{aligned}\|u(\cdot,t)\|_2 = \|v(\cdot,t) - w(\cdot,t)\|_2 &\leq \|v(\cdot,t) - w(\cdot,t)\|_\infty \|v(\cdot,t) - w(\cdot,t)\|_1 \\ &\leq \left(\|v(\cdot,t)\|_\infty + \|w(\cdot,t)\|_\infty\right) \|v_0 - w_0\|_1 \\ &\leq (\operatorname{diam}\Omega)^{1/2} \left(\|v_0\|_\infty + \|w_0\|_\infty\right) \\ &\quad \times \|v_0 - w_0\|_2 \,. \end{aligned} \tag{20}$$

Thus, a rather trivial calculation gives a form of stability in L^2, for a scalar equation. However, the condition of uniformly bounded supports is not a natural condition for hyperbolic problems. In addition, although systems of conservation laws are stable in L^1, the L^1 contraction property that holds for scalar equations does not generalize to systems, so the estimate (20) appears to be valid only for a scalar equation.

The analysis presented in this paper represents another step in a continuing search for a definitive answer to a central conundrum of multidimensional conservation laws. Based on the reasoning behind theorems, summarized in Sect. 1.1, that show that L^2-based function spaces are the only spaces in which classical solutions to quasilinear systems are well-posed in space dimensions greater than one, we have examined whether weak solutions in a single space dimension are stable in L^2. On the evidence presented in Sect. 4, this question has a negative answer for the simple reason that the speed of a discontinuity in a solution depends strongly on its amplitude.

Two possible ways to resolve the problem emerge. One might consider restricting the class of data, for example to initial conditions of uniformly bounded support. Alternatively, one might explore function spaces that are equivalent to L^2 for sufficiently smooth functions but are more forgiving of mismatched discontinuities.

Acknowledgements In a different context, Feride Tığlay suggested the idea of looking for bounds of the form (17). We are indebted to her and to John Holmes for helpful conversations about this problem.

References

1. Adimurthi, Ghoshal, S.S., Gowda, G.D.V.: L^p stability for entropy solutions of scalar conservation laws with strict convex flux. J. Differ. Equ. **256**, 3395–3416 (2014)
2. Brenner, P.: The Cauchy problem for symmetric hyperbolic systems in L_p. Math. Scand. **19**, 27–37 (1966)

3. Bressan, A.: Hyperbolic Systems of Conservation Laws: The One-Dimensional Cauchy Problem. Oxford University Press, Oxford (2000)
4. Dafermos, C.M.: Entropy and the stability of classical solutions of hyperbolic systems of conservation laws. In: Ruggeri, T. (ed.) Recent Mathematical Methods in Nonlinear Wave Propagation (Montecatini Terme, 1994), pp. 48–69. Springer, Berlin (1996)
5. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics. Springer, Berlin (2000)
6. Glimm, J.: Solutions in the large for nonlinear hyperbolic systems of equations. Commun. Pure Appl. Math. **18**, 95–105 (1965)
7. Holmes, J., Keyfitz, B.L., Tığlay, F.: Nonuniform dependence on initial data for compressible gas dynamics: the Cauchy problem on $\mathbb{R}^2$. SIAM J. Math. An. (to appear)
8. Hopf, E.: The partial differential equation $u_t + uu_x = \mu u_{xx}$. Commun. Pure Appl. Math. **III**, 201–230 (1950)
9. John, F.: Formation of singularities in one-dimensional nonlinear wave propagation. Commun. Pure Appl. Math. **XXVII**, 377–405 (1974)
10. Kato, T.: The Cauchy problem for quasi-linear symmetric hyperbolic systems. Arch. Ration. Mech. Anal. **58**, 181–205 (1975)
11. Keyfitz, B.L., Tığlay, F.: Nonuniform dependence on initial data for compressible gas dynamics: the periodic Cauchy problem. J. Differ. Equ. **263**, 6494–6511 (2017)
12. Kružkov, S.N.: First-order quasilinear equations in several independent variables. Math. USSR-Sb. **10**, 217–243 (1970)
13. Lax, P.D.: Hyperbolic systems of conservation laws, II. Commun. Pure Appl. Math. **X**, 537–566 (1957)
14. Leger, N.: L^2 estimates for shock solutions of scalar conservation laws using the relative entropy method. Arch. Ration. Mech. Anal. **199**, 761–778 (2011)
15. Leger, N., Vasseur, A.: Relative entropy and the stability of shocks and contact discontinuities for systems of conservation laws with non-bv perturbations. Arch. Ration. Mech. Anal. **201**, 271–302 (2011)
16. Leray, J., Ohya, Y.: Equations et systémes non-linéaires, hyperboliques non-stricts. Math. Ann. **170**, 167–205 (1967)
17. Littman, W.: The wave operator and L_p norms. J. Math. Mech. **12**, 55–68 (1963)
18. Majda, A.: Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables. Springer, New York (1984)
19. Oleĭnik, O.A.: Discontinuous solutions of nonlinear differential equations. Am. Math. Soc. Transl. (2) **26**, 95–172 (1963)
20. Rauch, J.: BV estimates fail in most quasilinear hyperbolic systems in dimensions greater than one. Commun. Math. Phys. **106**, 481–484 (1986)
21. Serre, D.: Systems of Conservation Laws. 1: Hyperbolicity, Entropies, Shock Waves. Cambridge University Press, Cambridge (1999). Translated from the French original by I. N. Sneddon
22. Serre, D., Vasseur, A.F.: L_2-type contraction for systems of conservation laws. J. Éc. Polytech. Math. **1**, 1–28 (2014)
23. Taylor, M.E.: Partial Differential Equations III: Nonlinear Equations. Springer, New York (1996)
24. Texier, B., Zumbrun, K.: Entropy criteria and stability of extreme shocks: a remark on a paper of Leger and Vasseur. Proc. Am. Math. Soc. **143**, 749–754 (2015)
25. Triebel, H.: Interpolation Theory, Function Spaces, Differential Operators. Barth, Heidelberg (1995)

Integral Equation Methods in Inverse Obstacle Scattering with a Generalized Impedance Boundary Condition

Rainer Kress

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The inverse problem under consideration is to reconstruct the shape of an impenetrable two-dimensional obstacle with a generalized impedance boundary condition from the far field pattern for scattering of time-harmonic acoustic or E-polarized electromagnetic plane waves. We propose an inverse algorithm that extends the approach suggested by Johansson and Sleeman (IMA J. Appl. Math. 72(1):96–112, 2007) for the case of the inverse problem for a sound-soft or perfectly conducting scatterer. It is based on a system of nonlinear boundary integral equations associated with a single-layer potential approach to solve the forward scattering problem which extends the integral equation method proposed by Cakoni and Kress (Inverse Prob. 29(1):015005, 2013) for a related boundary value problem for the Laplace equation. In addition, we also present an algorithm for reconstructing the impedance function when the shape of the scatterer is known. We present the mathematical foundations of the methods and exhibit their feasibility by numerical examples.

1 Introduction

The use of generalized impedance boundary conditions (GIBC) in the mathematical modeling of wave propagation has gained considerable attention in the literature over the last decades. This type of boundary conditions is applied to scattering problems for penetrable obstacles to model them approximately by scattering problems for impenetrable obstacles in order to reduce the cost of numerical computations. In this paper, we will consider boundary conditions that generalize

R. Kress (✉)
Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Göttingen, Germany
e-mail: kress@math.uni-goettingen.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_32

the classical impedance boundary condition, which is also known as Leontovich boundary condition, by adding a term with a second order differential operator. As compared with the Leontovich condition, this wider class of impedance conditions provides more accurate models, for example, for imperfectly conducting obstacles (see [7, 8, 16]).

To formulate the generalized impedance condition and the corresponding scattering problem, let D be a simply connected bounded domain in $\mathbb{R}^2$ with boundary ∂D of Hölder class $C^{4,\alpha}$ and denote by ν the unit normal vector to ∂D oriented towards the complement $\mathbb{R}^2 \setminus \overline{D}$. We consider the scattering problem to find the total wave $u = u^i + u^s \in H^2_{\text{loc}}(\mathbb{R}^2 \setminus \overline{D})$ satisfying the Helmholtz equation

$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^2 \setminus \overline{D} \tag{1}$$

with positive wave number k and the generalized impedance boundary condition

$$\frac{\partial u}{\partial \nu} + \mathrm{i}k\left(\lambda u - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}u}{\mathrm{d}s}\right) = 0 \quad \text{on } \partial D \tag{2}$$

where $\mathrm{d}/\mathrm{d}s$ is the tangential derivative and $\mu \in C^2(\partial D)$ and $\lambda \in C^1(\partial D)$ are complex valued functions. We note that the classical Leontovich condition is contained in (2) as the special case where $\mu = 0$. The incident wave u^i is assumed to be a plane wave $u^i(x) = e^{\mathrm{i}k\,x\cdot d}$ with a unit vector d describing the direction of propagation, but we also can allow other incident waves such as point sources. The scattered wave u^s has to satisfy the Sommerfeld radiation condition

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^s}{\partial r} - \mathrm{i}ku^s\right) = 0, \quad r = |r|, \tag{3}$$

uniformly with respect to all directions. The derivative for $u|_{\partial D} \in H^{\frac{3}{2}}(\partial D)$ with respect to arc length s in (2) has to be understood in the weak sense, that is, u has to satisfy

$$\int_{\partial D} \left(\eta\,\frac{\partial u}{\partial \nu} + \mathrm{i}k\lambda\eta u + \mathrm{i}k\mu\,\frac{\mathrm{d}\eta}{\mathrm{d}s}\,\frac{\mathrm{d}u}{\mathrm{d}s}\right)\mathrm{d}s = 0 \tag{4}$$

for all $\eta \in H^{\frac{3}{2}}(\partial D)$.

The Sommerfeld radiation condition is equivalent to the asymptotic behavior of an outgoing cylindrical wave of the form

$$u^s(x) = \frac{e^{\mathrm{i}k\,|x|}}{\sqrt{|x|}}\left\{u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right)\right\}, \quad |x| \to \infty, \tag{5}$$

uniformly for all directions $\hat{x} = x/|x|$ where the function u_∞ defined on the unit circle $\mathbb{S}^1$ is known as the far field pattern of u^s. Besides the direct scattering problem to determine the scattered wave u^s for a given incident wave u^i the two inverse scattering problems that we will consider are to determine the boundary ∂D, for given impedance functions, or the impedance coefficients μ and λ, for a given boundary, from a knowledge of the far field pattern u_∞ on $\mathbb{S}^1$ for one or several incident plane waves. The first problem we will call the inverse shape problem and the second the inverse impedance problem.

For further interpretation of the generalized impedance boundary condition we refer to [1–3] where the direct and the inverse scattering problem are analyzed by variational methods. For the solution of a related boundary value problem for the Laplace equation with the generalized impedance boundary condition of the form (2), Cakoni and Kress [4] have proposed a single-layer potential approach that leads to a boundary integral equation or more precisely a boundary integro-differential equation governed by a pseudo-differential operator of order one. In Sect. 2 we will extend this approach to the direct scattering problem (1)–(3). As to be expected, the single-layer approach fails when k^2 is an interior Dirichlet eigenvalue for the negative Laplacian in D and to remedy this deficiency we describe a modified approach by a combined single- and double-layer approach that leads to a pseudo-differential operator of order two. For simplicity, confining ourselves to the single-layer potential approach, we then proceed in Sect. 3 with describing the numerical solution of the integro-differential equation via trigonometric interpolation quadratures and differentiation that lead to spectral convergence.

Our analysis of the two inverse problems is based on a nonlinear boundary integral equation method in the spirit of Johansson and Sleeman [10] (see also [6, Section 5.4]) and follows the approach for the Laplace equation as developed by Cakoni and Kress [4]. We begin in Sect. 4 with a review on uniqueness and then proceed in Sect. 5 with the solution of the inverse shape problem followed by the solution of the inverse impedance problem in Sect. 6. In both cases we present the theoretical basis for the inverse algorithms and illustrate them by a couple of numerical examples.

2 The Boundary Integral Equation

In this section we describe a boundary integral equation method for solving the direct obstacle scattering problem and begin by establishing uniqueness of the solution. Throughout our analysis we will assume that

$$\operatorname{Re}\lambda \geq 0, \quad \operatorname{Re}\mu \geq 0, \quad |\mu| > 0, \tag{6}$$

where the first two conditions ensure uniqueness and the third condition is required for our existence analysis.

Theorem 1 *Any solution $u \in H^2_{\mathrm{loc}}(\mathbb{R}^2 \setminus \overline{D})$ to (1)–(2) satisfying the Sommerfeld radiation condition vanishes identically.*

Proof Inserting $\eta = \bar{u}|_{\partial D}$ in the weak form (4) of the boundary condition we obtain that

$$\int_{\partial D} \bar{u}\, \frac{\partial u}{\partial \nu}\, \mathrm{d}s = -\mathrm{i}k \int_{\partial D} \left\{ \lambda |u|^2 + \mu \left| \frac{\mathrm{d}u}{\mathrm{d}s} \right|^2 \right\} \mathrm{d}s.$$

Hence in view of our assumption (6) we can conclude that

$$\operatorname{Im} \int_{\partial D} \bar{u}\, \frac{\partial u}{\partial \nu}\, \mathrm{d}s \le 0$$

and from this and the radiation condition the statement of the theorem follows from Rellich's lemma, see Theorem 2.13 in [6]. □

Corollary 1 *The scattering problem (1)–(3) has at most one solution.*

We recall the fundamental solution of the Helmholtz equation

$$\Phi(x,y) = \frac{\mathrm{i}}{4}\, H_0^{(1)}(k|x-y|), \quad x \neq y,$$

in $\mathbb{R}^2$ in terms of the Hankel function $H_0^{(1)}$ of the first kind of order zero. Further, following [6, Section 3.1] we introduce the classical boundary integral operators in scattering theory given by the single- and double-layer operators

$$(S\varphi)(x) := 2 \int_{\partial D} \Phi(x,y)\varphi(y)\, \mathrm{d}s(y), \quad x \in \partial D, \tag{7}$$

$$(K\varphi)(x) := 2 \int_{\partial D} \frac{\partial \Phi(x,y)}{\partial \nu(y)}\, \varphi(y)\, \mathrm{d}s(y), \quad x \in \partial D, \tag{8}$$

and the corresponding normal derivative operators

$$(K'\varphi)(x) := 2 \int_{\partial D} \frac{\partial \Phi(x,y)}{\partial \nu(x)}\, \varphi(y)\, \mathrm{d}s(y), \quad x \in \partial D, \tag{9}$$

$$(T\varphi)(x) := 2\, \frac{\partial}{\partial \nu(x)} \int_{\partial D} \frac{\partial \Phi(x,y)}{\partial \nu(y)}\, \varphi(y)\, \mathrm{d}s(y), \quad x \in \partial D. \tag{10}$$

For the subsequent analysis in contemporary Sobolev spaces, we note that for $\partial D \in C^{4,\alpha}$ the operators $S : H^{\frac{1}{2}}(\partial D) \to H^{\frac{3}{2}}(\partial D)$, $S, K : H^{\frac{3}{2}}(\partial D) \to H^{\frac{5}{2}}(\partial D)$, $T : H^{\frac{3}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ and $K' : H^{\frac{1}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ are all bounded (see [11, 15]).

In a first attempt, extending the approach proposed in [4] for the Laplace equation, we try to find the solution of (1)–(3) in the form of a single-layer potential for the scattered wave

$$u^s(x) = \int_{\partial D} \Phi(x,y)\varphi(y)\, \mathrm{d}s(y), \quad x \in \mathbb{R}^2 \setminus \overline{D}, \tag{11}$$

with density $\varphi \in H^{\frac{1}{2}}(\partial D)$ and note that the regularity $\varphi \in H^{\frac{1}{2}}(\partial D)$ guarantees that $u \in H^2_{\mathrm{loc}}(\mathbb{R}^2 \setminus \overline{D})$ (see [15]). From the asymptotics for the Hankel function $H_0^{(1)}(t)$ as $t \to \infty$, it can be deduced that the far field pattern of u^s is given by

$$u_\infty(\hat{x}) = \gamma \int_{\partial D} e^{-\mathrm{i}k\,\hat{x}\cdot y}\varphi(y)\,\mathrm{d}s(y), \quad \hat{x} \in \mathbb{S}^1, \tag{12}$$

where

$$\gamma = \frac{e^{\mathrm{i}\frac{\pi}{4}}}{\sqrt{8\pi k}}\,. \tag{13}$$

Letting x approach the boundary ∂D from inside $\mathbb{R}^2 \setminus \overline{D}$, from the jump relations for single-layer potentials (see [6, Theorem 3.1]) we observe that the boundary condition (2) is satisfied provided φ solves the integro-differential equation

$$\varphi - K'\varphi - \mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)S\varphi = g \tag{14}$$

where we set

$$g := 2\,\frac{\partial u^i}{\partial \nu}\bigg|_{\partial D} + 2\mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)u^i|_{\partial D} \tag{15}$$

in terms of the incident wave u^i. After defining a bounded linear operator $A : H^{\frac{1}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ by

$$A\varphi := \varphi - K'\varphi - \mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)S\varphi \tag{16}$$

we can summarize the above into the following theorem.

Theorem 2 *The single-layer potential (11) solves the scattering problem (1)–(3) provided the density φ satisfies the equation*

$$A\varphi = g. \tag{17}$$

Lemma 1 *The operator $M : H^{\frac{3}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ given by*

$$M\varphi := \frac{\mathrm{d}^2\varphi}{\mathrm{d}s^2} + \int_{\partial D} \varphi\,\mathrm{d}s \tag{18}$$

is bounded and has a bounded inverse.

Proof We parametrize the boundary ∂D with the arc length s as parameter and identify $H^p(\partial D)$ with $H^p_{\mathrm{per}}[0, L]$ where L is the length of ∂D and $H^p_{\mathrm{per}}[0, L] \subset H^p[0, L]$

is the subspace of L periodic functions (or more precisely bounded linear functionals if $p < 0$) (see [12, Section 8.5]). Using the Fourier series representation of $H^r_{\rm per}[0, L]$ it can be seen that indeed $M : H^{\frac{3}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ is an isomorphism. □

Lemma 2 *The operator* $A - {\rm i}k\mu MS : H^{\frac{1}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ *is compact.*

Proof The boundedness of the operators $S, K' : H^{\frac{1}{2}}(\partial D) \to H^{\frac{3}{2}}(\partial D)$ and $K' : H^{\frac{1}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ and our assumption $\eta \in C^1(\partial D)$ implies that all terms in the sum (16) defining the operator A are bounded from $H^{\frac{1}{2}}(\partial D)$ into $H^{\frac{1}{2}}(\partial D)$ except the term

$$\varphi \mapsto {\rm i}k \, \frac{\rm d}{{\rm d}s} \, \mu \, \frac{\rm d}{{\rm d}s} \, S\varphi.$$

Therefore, after splitting

$$\frac{\rm d}{{\rm d}s} \, \mu \, \frac{{\rm d}\, S\varphi}{{\rm d}s} = \mu \, \frac{{\rm d}^2 \, S\varphi}{{\rm d}s^2} + \frac{{\rm d}\mu}{{\rm d}s} \, \frac{{\rm d}\, S\varphi}{{\rm d}s}$$

and using our assumption $\mu \in C^1(\partial D)$ we observe that the operator $A - {\rm i}k\mu MS : H^{\frac{1}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ is bounded. Hence the statement of the theorem follows from the compact embedding of $H^{\frac{1}{2}}(\partial D)$ into $H^{-\frac{1}{2}}(\partial D)$. □

Theorem 3 *Provided k^2 is not a Dirichlet eigenvalue for the negative Laplacian in D, for each $g \in H^{-\frac{1}{2}}(\partial D)$ Eq. (17) has a unique solution $\varphi \in H^{\frac{1}{2}}(\partial D)$ and this solution depends continuously on g.*

Proof Since under our assumption on k the operator $S : H^{\frac{1}{2}}(\partial D) \to H^{\frac{3}{2}}(\partial D)$ is an isomorphism, by Lemma 1 and our assumptions on μ the operator ${\rm i}k\mu MS : H^{\frac{1}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ also is an isomorphism. Therefore, in view of Lemma 2, by the Riesz theory it suffices to show that the operator A is injective. Assume that $\varphi \in H^{\frac{1}{2}}(\partial D)$ satisfies $A\varphi = 0$. Then, by Theorem 2 the single-layer potential u defined by (11) solves the scattering problem for the incident wave $u^i = 0$. Hence, by the uniqueness Theorem 1 we have $u = 0$ in $\mathbb{R}^2 \setminus \overline{D}$. Taking the boundary trace of u it follows that $S\varphi = 0$ and consequently $\varphi = 0$. □

To remedy the failure of the single-layer potential approach at the interior Dirichlet eigenvalues, as in the case of the classical impedance condition, we modify it into the form of a combined single- and double-layer potential for the scattered wave

$$u^s(x) = \int_{\partial D} \left\{ \Phi(x, y) + {\rm i} \, \frac{\partial \Phi(x, y)}{\partial \nu(y)} \right\} \varphi(y) {\rm d}s(y), \quad x \in \mathbb{R}^2 \setminus \overline{D}, \tag{19}$$

with density $\varphi \in H^{\frac{3}{2}}(\partial D)$. The regularity $\varphi \in H^{\frac{3}{2}}(\partial D)$ implies $u \in H^2_{\text{loc}}(\mathbb{R}^2 \setminus \overline{D})$. Letting x approach the boundary ∂D from inside $\mathbb{R}^2 \setminus \overline{D}$, we observe that the boundary condition (2) is satisfied provided φ solves the integro-differential equation

$$\varphi - K'\varphi - \mathrm{i}T\varphi - \mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)(S\varphi + \mathrm{i}\varphi + \mathrm{i}K\varphi) = g \tag{20}$$

with g given by (15). We define a bounded linear operator $B : H^{\frac{3}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ by

$$B\varphi := \varphi - K'\varphi - \mathrm{i}T\varphi - \mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)(S\varphi + \mathrm{i}\varphi + \mathrm{i}K\varphi) \tag{21}$$

and then have the following theorem.

Theorem 4 *The combined single- and double-layer potential (19) solves the scattering problem (1)–(3) provided the density φ satisfies the equation*

$$B\varphi = g. \tag{22}$$

Lemma 3 *The operator $B - k\mu M : H^{\frac{3}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ is compact.*

Proof The boundedness of $S, K : H^{\frac{3}{2}}(\partial D) \to H^{\frac{5}{2}}(\partial D)$, $K' : H^{\frac{1}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ and $T : H^{\frac{3}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ mentioned above implies that all terms in the sum (21) defining the operator A are bounded from $H^{\frac{3}{2}}(\partial D)$ into $H^{\frac{1}{2}}(\partial D)$ except the term

$$\varphi \mapsto k\,\frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{d\varphi}{ds}\,.$$

Therefore, as in the proof of Lemma 2 we can deduce that the operator $B - k\mu M$: $H^{\frac{3}{2}}(\partial D) \to H^{\frac{1}{2}}(\partial D)$ is bounded and the statement follows from the compact embedding of $H^{\frac{1}{2}}(\partial D)$ into $H^{-\frac{1}{2}}(\partial D)$. □

Theorem 5 *For each $g \in H^{-\frac{1}{2}}(\partial D)$ the integral equation (22) has a unique solution $\varphi \in H^{\frac{3}{2}}(\partial D)$ and this solution depends continuously on g.*

Proof By our assumption on μ we have that $k\mu M : H^{\frac{3}{2}}(\partial D) \to H^{-\frac{1}{2}}(\partial D)$ is an isomorphism. Therefore, in view of Theorem 4 and Lemma 3 by the Riesz theory it suffices to show that the operator B is injective. Assume that $\varphi \in H^{\frac{3}{2}}(\partial D)$ satisfies $B\varphi = 0$. Then, by Theorem 4 the combined single- and double-layer potential u defined by (19) solves the scattering problem for the incident wave $u^i = 0$. Hence, by the uniqueness Theorem 1 we have $u = 0$ in $\mathbb{R}^2 \setminus \overline{D}$. Taking the boundary trace of u it follows that $S\varphi + \mathrm{i}\varphi + \mathrm{i}K\varphi = 0$. From this, proceeding as in the corresponding existence proof for the scattering problem with Dirichlet boundary condition (see Theorem 3.11 in [6]) we can conclude that $\varphi = 0$. □

Summarizing, we finally have our main result of this section.

Theorem 6 *The direct scattering problem (1)–(3) has a unique solution.*

In addition to the potential approach for setting up the boundary integral equations, of course, following the so-called direct approach one can also derive integral equations based on Green's representation formula. Passing to the boundary ∂D in Huygens' principle (see Theorem 3.14 in [6]) and incorporating the boundary condition (2) we obtain the equation

$$\eta - K\eta - \mathrm{i}kS\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)\eta = 2u^i|_{\partial D} \tag{23}$$

for the boundary trace $\eta := u|_{\partial D}$ of the total field. Obviously, the operator on the left-hand side of (23) is the adjoint of A with respect to the L^2 bilinear form and therefore, by the Fredholm alternative, Eq. (23) also is uniquely solvable, provided k^2 is not a Dirichlet eigenvalue for the negative Laplacian in D.

3 Numerical Solution

For the numerical solution, for simplicity we confine ourselves to Eq. (14). We employ a collocation method based on numerical quadratures using trigonometric polynomial approximations as the most efficient method for solving boundary integral equations for scattering problems in planar domains with smooth boundaries (see [6, Section 3.5]). Here, additionally we need to be concerned with presenting an approximation for the operator $\varphi \mapsto \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{d\varphi}{ds}$ as the new feature in the integro-differential equations for the generalized impedance boundary condition. For this, we apply trigonometric differentiation.

Both for the numerical solution and later on for the presentation of our inverse algorithm we assume that the boundary curve ∂D is given by a regular 2π periodic counter clockwise parameterization

$$\partial D = \{z(t) : 0 \le t \le 2\pi\}. \tag{24}$$

Then, via $\psi = \varphi \circ z$ we introduce the parameterized single-layer operator by

$$(\widetilde{S}\psi)(t) := \frac{\mathrm{i}}{2}\int_0^{2\pi} H_0^{(1)}(k|z(t) - z(\tau)|)\,|z'(\tau)|\,\psi(\tau)\,\mathrm{d}\tau$$

and the parameterized normal derivative operator by

$$(\widetilde{K}'\psi)(t) := \frac{\mathrm{i}k}{2}\int_0^{2\pi} \frac{[z'(t)]^{\perp}\cdot[z(\tau) - z(t)]}{|z'(t)|\,|z(t) - z(\tau)|}\,H_1^{(1)}(k|z(t) - z(\tau)|)\,|z'(\tau)|\,\psi(\tau)\,\mathrm{d}\tau$$

for $t \in [0, 2\pi]$. Here we made use of $H_0^{(1)\prime} = -H_1^{(1)}$ with the Hankel function $H_1^{(1)}$ of order zero and of the first kind. Furthermore, we denote $a^\perp := (a_2, -a_1)$ for any vector $a = (a_1, a_2)$, that is, $a^\perp$ is obtained by rotating a clockwise by 90°. Then the parameterized form of (14) is given by

$$\psi - \widetilde{K}'\psi - \mathrm{i}k\lambda \circ z\widetilde{S}\psi + \frac{1}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mu \circ z}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{S}\psi = g \circ z \tag{25}$$

We construct approximations via trigonometric interpolation quadratures and trigonometric differentiation based on equidistant interpolation points $t_j = j\pi/n$ for $j = 1, \dots, 2n$ with $n \in \mathbb{N}$. For the operators $\widetilde{S}$ and $\widetilde{K}'$ we make use of approximation $\widetilde{S}_n$ and $\widetilde{K}'_n$ via trigonometric interpolation quadratures that take care of the logarithmic singularities of the Hankel functions as described in Section 3.5 of [6] or in [14]. We refrain from repeating the details.

To approximate the operator $\varphi \mapsto \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}\varphi}{\mathrm{d}s}$ we simply use numerical differentiation via trigonometric interpolation, i.e., we approximate the derivative ψ' of a 2π periodic function ψ by the derivative $(P_n\psi)'$ of the unique trigonometric polynomial $P_n\psi$ of degree n (without the term $\sin nt$) that interpolates $(P_n\psi)(t_j) = \psi(t_j)$ for $j = 1, \dots, 2n$. For the resulting weights we refer to [12, Section 13.5]. We set $P'_n\psi := (P_n\psi)'$ and approximate

$$\frac{1}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mu \circ z}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{S}\psi \approx \frac{1}{|z'|}P'_n\frac{\mu \circ z}{|z'|}P'_n\widetilde{S}_n\psi.$$

Summarizing, our numerical solution method approximates the integro-differential equation (25) by

$$\psi_n - \widetilde{K}'_n\psi_n - \mathrm{i}k\lambda \circ z\widetilde{S}_n\psi_n + \frac{1}{|z'|}P'_n\frac{\mu \circ z}{|z'|}P'_n\widetilde{S}_n\psi_n = g \circ z \tag{26}$$

which is solved for the trigonometric polynomial ψ_n by collocation at the nodal points t_j for $j = 1, \dots, 2n$.

Since the operators

$$\varphi \mapsto \frac{\mathrm{d}^2}{\mathrm{d}s^2}S\varphi \quad \text{and} \quad \varphi \mapsto \frac{\mathrm{d}}{\mathrm{d}s}S\frac{\mathrm{d}\varphi}{\mathrm{d}s}$$

have the same principal part, the error and convergence analysis for numerically solving the hypersingular equation of the first kind with the operator T, defined in (10), via Maue's formula and trigonometric differentiation as carried out in [13] and based on Theorem 13.12 and Corollary 13.13 in [12], can be transferred to the approximation (26) with only minor modifications. In particular, such an analysis would predict spectral convergence in the case of analytic μ, λ and z. However, since our main emphasis is on the inverse scattering problem we refrain from carrying out the details. Instead of this we will conclude with a numerical example exhibiting the

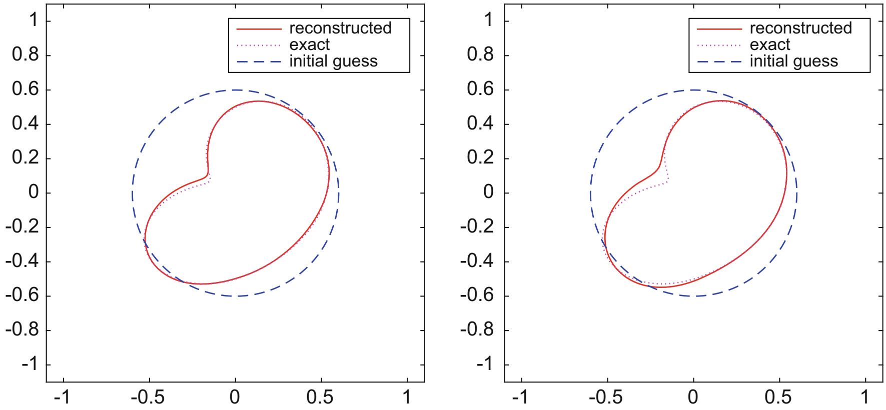


Fig. 1 Reconstruction of the apple (27) for exact data after 30 iterations (left) and for 5% noise after 10 iterations (right)

spectral convergence. Before doing so we note, that an approximate solution of (20) including an error analysis can be obtained analogously using the approximations for S, K, K' from [6, Section 3.5] and the approximation of T via Maue's formula that we just mentioned (see [13]).

For numerical examples we consider scattering by an apple-shaped obstacle with parametric representation

$$z(t) = \frac{0.5 + 0.4\cos t + 0.1\sin 2t}{1 + 0.7\cos t}\,(\cos t, \sin t), \quad 0 \le t \le 2\pi, \tag{27}$$

(see Fig. 1) and by a peanut-shaped obstacle with parametric representation

$$z(t) = \sqrt{\cos^2 t + 0.25\sin^2 t}\,(\cos t, \sin t), \quad 0 \le t \le 2\pi, \tag{28}$$

(see Fig. 2). As impedance functions we choose

$$\lambda(z(t)) = \frac{1}{1 - 0.1\sin 2t} \quad \text{and} \quad \mu(z(t)) = \frac{1}{1 + 0.3\cos t} \tag{29}$$

for $t \in [0, 2\pi]$ and note that for both examples we can interpret the impedance functions as given in a neighborhood of ∂D depending only on the polar angle.

After approximately solving the integro-differential equation for the density φ the far field pattern is obtained from (12) by the composite trapezoidal rule. Rather than presenting tables with the far field pattern for plane wave incidence we find it more convenient to just illustrate the spectral convergence by Table 1 which shows the maximum norm (over the collocation points) of the error $E_n := \|u_\infty - u_{\infty,n}\|_\infty$ between the exact and the approximate far field pattern for a point

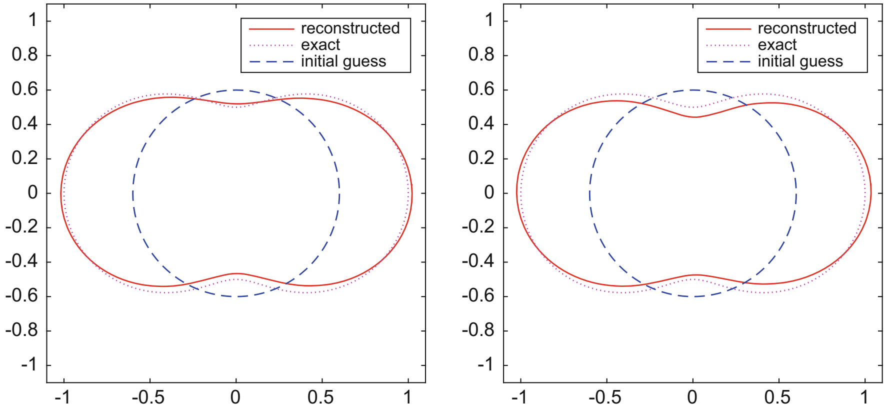


Fig. 2 Reconstruction of the peanut (28) for exact data after 30 iterations (left) and for 5% noise after 10 iterations (right)

Table 1 Error decay for apple-shaped and peanut-shaped scatterer

	$2n$	$E_{n,\text{apple}}$	$E_{n,\text{peanut}}$
$k = 2$	16	5.02e-04	5.32e-05
	32	3.55e-05	1.33e-07
	64	5.52e-08	8.19e-14
	128	1.16e-13	1.45e-14
$k = 8$	16	1.00e-02	1.00e-01
	32	2.43e-05	1.95e-05
	64	1.38e-08	3.71e-14
	128	5.86e-14	8.94e-15

source $u_s = \frac{i}{4} H_0^{(1)}(k|x - x_0|)$ located at some $x_0 \in D$ which has far field pattern $u_\infty(\hat{x}) = \gamma\, e^{-ik\,\hat{x}\cdot x_0}$. In the examples we chose $x_0 = (0.1, 0.2)$.

4 Inverse Scattering: Uniqueness

We now turn our attention to the inverse scattering problems. The most general inverse scattering problem is the *inverse shape and impedance problem* to determine ∂D, μ and λ from a knowledge of one (or finitely many) far field patterns u_∞ of solutions u to (1)–(3). In this paper we will be only concerned with two less general cases, namely the *inverse shape problem* and the *inverse impedance problem*. The inverse shape problem consists in determining ∂D from one (or finitely many) far field patterns knowing the impedance coefficients μ and λ. With the roles reversed, the inverse impedance problem requires to determine the impedance functions μ and λ from one (or finitely many) far field patterns for a known shape ∂D.

The first question to ask is what is the minimum amount of data, i.e., the minimal number of far field patterns, to guaranty the uniqueness of the solution for the inverse impedance problem or the inverse shape problem. The following theorem shows that three far field patterns uniquely determine both impedance functions λ and μ provided that ∂D is known.

Theorem 7 *For a given shape ∂D, three far field patterns corresponding to the scattering of three plane waves with different incident directions uniquely determine the impedance functions μ and λ.*

Proof Plane waves with different directions clearly are linearly independent. Consequently the corresponding total waves u_1, u_2, u_3 are also linearly independent. Therefore, the proof of Theorem 3.1 in [4] for the case of the Laplace equation can be carried over without any changes to the Helmholtz equation since it only uses the differential equation on the boundary as given by the generalized impedance boundary condition. □

Extending the counter example given in [4] for the Laplace case, the following example illustrates non-uniqueness issues for the inverse impedance problem using two far field patterns. Let D be a disc of radius R centered at the origin, let μ and λ be constants satisfying (6), and consider the two incident waves given in polar coordinates by $u^i(r,\theta) = J_n(kr)\, e^{\pm \mathrm{i} n\theta}$ in terms of the Bessel function J_n of order $n \in \mathbb{N}$. Then the corresponding total wave is given by

$$u(r,\theta) = \big(J_n(kr) - a_n H_n^{(1)}(kr)\big)\, e^{\pm \mathrm{i} n\theta}$$

with the Hankel function $H_n^{(1)}$ of the first kind of order n and

$$a_n = \frac{kR^2 J_n'(kR) + \mathrm{i}k(n^2\mu + \lambda R^2) J_n(kR)}{kR^2 H_n^{(1)\prime}(kR) + \mathrm{i}k(n^2\mu + \lambda R^2) H_n^{(1)}(kR)}\,. \tag{30}$$

We note that the uniqueness Theorem 1 ensures that the denominator in (30) is different from zero. Clearly, there are infinitely many combinations of positive real numbers μ and λ giving the same value for a_n, that is, the same two linearly independent total fields.

The following uniqueness result for the full inverse shape and impedance problem was obtained by Bourgeois et al. [3].

Theorem 8 *Both the shape and the impedance functions of a scattering obstacle with generalized impedance condition are uniquely determined by the far field patterns for an infinite number of incident waves with different incident directions and one fixed wave number.*

The main idea of the proof in [6, Theorem 5.6] for the case $\mu = 0$ remains valid. We only need to convince ourselves that the mixed reciprocity relation for scattering of point sources and plane waves, see [6, Theorem 3.16] extends from the case $\mu = 0$ to the general case as consequence of the weak form (4) of the generalized impedance condition.

We conclude this short section on uniqueness for the inverse problem with outlining the proof for the identifiability of a disc and its constant impedance coefficients from the far field pattern for one incident plane wave.

Theorem 9 *A disc with constant impedance coefficients is uniquely determined by the far field pattern for one incident plane wave.*

Proof Using polar coordinates the Jacobi–Anger expansion (see [6, p. 75]) reads

$$e^{\mathrm{i}k\,x\cdot d} = \sum_{n=-\infty}^{\infty} \mathrm{i}^n J_n(kr)\, e^{\mathrm{i}n\theta}, \quad x \in \mathbb{R}^2, \tag{31}$$

where θ is the angle between x and d. From this it can be seen that the scattered wave u^s for scattering from a disc of radius R centered at the origin has the form

$$u^s(x) = \sum_{n=-\infty}^{\infty} a_n\, \mathrm{i}^n H_n^{(1)}(kr)\, e^{\mathrm{i}n\theta}, \quad r > R, \tag{32}$$

with the coefficients a_n from (30). Using the asymptotics of the Bessel and Hankel functions for large n (see [6, Section 3.4]) uniform convergence can be established for the series (32) in compact subsets of $\mathbb{R}^2 \setminus \{0\}$. In particular, this implies that the scattered wave u^s has an extension as solution to the Helmholtz equation across the boundary into the interior of the disc with the exception of the center.

Now assume that two discs D_1 and D_2 with centers z_1 and z_2 have the same far field pattern $u_{\infty,1} = u_{\infty,2}$ for scattering of one incident plane wave. Then by Rellich's lemma (see [6]) the scattered waves coincide $u_1^s = u_2^s$ in $\mathbb{R}^2 \setminus (D_1 \cup D_2)$ and we can identify $u^s = u_1^s = u_2^s$ in $\mathbb{R}^2 \setminus (D_1 \cup D_2)$. Now assume that $z_1 \neq z_2$. Then u_1^s has an extension into $\mathbb{R}^2 \setminus \{z_1\}$ and u_2^s an extension into $\mathbb{R}^2 \setminus \{z_2\}$. Therefore, u^s can be extended from $\mathbb{R}^2 \setminus (D_1 \cup D_2)$ into all of $\mathbb{R}^2$, that is, u^s is an entire solution to the Helmholtz equation. Consequently, since u^s also satisfies the radiation condition it must vanish identically $u^s = 0$ in all of $\mathbb{R}^2$. Therefore the incident field $u^i(x) = e^{\mathrm{i}k\,x\cdot d}$ must satisfy the generalized impedance condition on D_1 with radius R_1. Parameterizing $x \cdot d = R_1 \cos\theta$ the boundary condition then implies

$$\left\{\cos\theta + \lambda + k^2\mu \sin^2\theta + \frac{1}{R_1}\,\mathrm{i}k\mu\cos\theta\right\} e^{ikR_1\cos\theta} = 0$$

for all $\theta \in [0, 2\pi]$. However this is a contradiction and therefore $z_1 = z_2$.

In order to show that D_1 and D_2 have the same radius and the same impedance coefficients, we observe that by symmetry, or by inspection of the explicit solution given above, the far field pattern for scattering of plane waves from a disc with constant impedance coefficients depends only on the angle between the observation direction and the incident direction. Hence, knowledge of the far field pattern for one incident direction implies knowledge of the far field pattern for all incident directions. Now the statement follows from the above Theorem 8. □

5 Solution of the Inverse Shape Problem

We now proceed describing an iterative algorithm for approximately solving the inverse shape problem by extending the method proposed by Johansson and Sleeman [10] for sound-soft or perfectly conducting obstacles. After introducing the far field operator

$$S_\infty : H^{\frac{1}{2}}(\partial D) \to L^2(\mathbb{S}^1)$$

by

$$(S_\infty\varphi)(\hat{x}) := \gamma \int_{\partial D} e^{-\mathrm{i}k\,\hat{x}\cdot y}\varphi(y)\,\mathrm{d}s(y), \quad \hat{x} \in \mathbb{S}^1, \tag{33}$$

from (11) and (12) we observe that the far field pattern for the solution to the scattering problem (1)–(3) is given by

$$u_\infty = S_\infty\varphi \tag{34}$$

in terms of the solution to (14). We note that S_∞ is compact and state the following theorem as theoretical basis of our inverse algorithm. For this we note that the operators and the right-hand side g depend on the boundary curve ∂D.

Theorem 10 *For a given incident field u^i and a given far field pattern u_∞, assume that ∂D and the density φ satisfy the system*

$$\varphi - K'\varphi - \mathrm{i}k\left(\lambda - \frac{\mathrm{d}}{\mathrm{d}s}\,\mu\,\frac{\mathrm{d}}{\mathrm{d}s}\right)S\varphi = g \tag{35}$$

and

$$S_\infty\varphi = u_\infty \tag{36}$$

where g is given in terms of the incident field by (15). Then ∂D solves the inverse shape problem.

The ill-posedness of the inverse shape problem is reflected through the ill-posedness of the second equation (36), the far field equation that we denote as the *data equation*. Note that the system (35)–(36) is linear with respect to the density φ and nonlinear with respect to the boundary ∂D. This opens up a variety of approaches to solve (35)–(36) by linearization and iteration. In this paper, we are going to proceed as follows. Given an approximation for the unknown ∂D we solve Eq. (35) that we denote as the *field equation* for the unknown density φ. Then, keeping φ fixed we linearize the data equation (36) with respect to the boundary ∂D to update the approximation.

To describe this in more detail, we also need the parameterized version

$$\widetilde{S}_\infty : H^{\frac{1}{2}}[0, 2\pi] \to L^2(\mathbb{S}^1)$$

of the far field operator given by

$$(\widetilde{S}_\infty \psi)(\hat{x}) := \gamma \int_0^{2\pi} e^{-\mathrm{i}k\,\hat{x}\cdot z(\tau)}\,|z'(\tau)|\,\psi(\tau)\,\mathrm{d}\tau, \quad \hat{x} \in \mathbb{S}^1. \tag{37}$$

Then the parameterized form of (35)–(36) is given by

$$\psi - \widetilde{K}'\psi - \mathrm{i}k\lambda \circ z\widetilde{S}\psi + \frac{1}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mu \circ z}{|z'|}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{S}\psi = g \circ z \tag{38}$$

and

$$\widetilde{S}_\infty(\psi, z) = u_\infty \tag{39}$$

where $\psi = \varphi \circ z$.

The Fréchet derivative $\widetilde{S}'_\infty$ of the operator $\widetilde{S}_\infty$ with respect to the boundary curve z in the direction ζ is given by

$$\widetilde{S}'_\infty(\psi;\zeta)(\hat{x}) := \gamma \int_0^{2\pi} e^{-\mathrm{i}k\,\hat{x}\cdot z(\tau)}\left[-\mathrm{i}k\,\hat{x}\cdot\zeta(\tau)\,|z'(\tau)| + \frac{z'(\tau)\cdot\zeta'(\tau)}{|z'(\tau)|}\right]\psi(\tau)\,\mathrm{d}\tau$$

for $\hat{x} \in \mathbb{S}^1$. Then the linearization of (39) at z with respect to the direction ζ becomes

$$\widetilde{S}_\infty\psi + \widetilde{S}'_\infty(\psi;\zeta) = u_\infty \tag{40}$$

and is a linear equation for the update ζ.

Now, given an approximation for the boundary curve ∂D with parameterization z, each iteration step of the proposed inverse algorithm consists of two parts.

1. We solve the well-posed field equation (38) for ψ. This can be done through the numerical method described in Sect. 3.
2. Then we solve the ill-posed linearized equation (40) for ζ and obtain an updated approximation for ∂D with the parameterization $z + \zeta$. Since the kernels of the integral operators in (40) are smooth, for its numerical approximation the composite trapezoidal rule can be employed. Because of the ill-posedness, the solution of (40) requires stabilization, for example, by Tikhonov regularization.

These two steps are now iterated until some stopping criterion is satisfied. In our numerical examples the iterations were stopped when the decrease in the residual of the data equation became smaller than a given threshold.

In principle, the parameterization of the update is not unique. To cope with this ambiguity, one possibility, which we pursued in our numerical examples, is to allow only parameterizations of the form

$$z(t) = r(t)(\cos t, \sin t), \quad 0 \leq t \leq 2\pi, \tag{41}$$

with a non-negative function r representing the radial distance of ∂D from the origin. Consequently, the perturbations are of the form

$$\zeta(t) = q(t)(\cos t, \sin t), \quad 0 \leq t \leq 2\pi, \tag{42}$$

with a real function q. In the approximations we assume r and its update q to have the form of a trigonometric polynomial of degree J, in particular,

$$q(t) = \sum_{j=0}^{J} a_j \cos jt + \sum_{j=1}^{J} b_j \sin jt. \tag{43}$$

Then the update equation (40) is solved in the least squares sense, penalized via Tikhonov regularization, for the unknown coefficients $a_0, \ldots, a_J$ and $b_1, \ldots, b_J$ of the trigonometric polynomial representing the update q. Since (43) requires the coefficients to be real valued we turn the linear system (40) with a complex matrix and complex right hand side into an equivalent real system by taking the real and imaginary parts of (40). It is advantageous to use an H^p Sobolev penalty term rather than an L^2 penalty in the Tikhonov regularization, i.e., to interpret $\widetilde{S}'_\infty$ as an ill-posed linear operator $\widetilde{S}'_\infty : H^p[0, 2\pi] \to L^2(\mathbb{S}^1)$ for some small $p \in \mathbb{N}$.

As a theoretical basis for the application of Tikhonov regularization from [9] we cite that, after the restriction to star-like boundaries of the form (42), the operator $\widetilde{S}'_\infty$ is injective if k_0^2 is not a Neumann eigenvalue for the negative Laplacian in D.

The above algorithm has a straightforward extension for the case of more than one incident wave. Assume that $u_1^i, \ldots, u_M^i$ are M incident waves with different incident directions and $u_{\infty,1}, \ldots, u_{\infty,M}$ the corresponding far field patterns for scattering from ∂D. Given an approximation z for the boundary we first solve the field equations (38) for the M different incident fields to obtain M densities $\psi_1, \ldots, \psi_M$. Then we solve the linearized equations

$$\widetilde{S}_\infty \psi_m + \widetilde{S}'_\infty(\psi_m; \zeta) = u_{\infty,m}, \quad m = 1, \ldots, M, \tag{44}$$

for the update ζ by interpreting them as one ill-posed equation with an operator from $H^p[0, 2\pi]$ into $(L^2(\mathbb{S}^1))^M$ and applying Tikhonov regularization.

The numerical examples are intended as proof of concept and not as indications of an already fully developed method. In particular, the regularization parameters and the number of iterations are chosen by trial and error instead of, for example, a discrepancy principle. In all examples, to avoid committing an inverse crime the synthetic far field data are obtained by solving the integral equation (20) for the combined single- and double-layer approach whereas the inverse solver is based on the single-layer approach via the integral equation (14).

For both examples the impedance functions are given by (29). The number of quadrature points is $2n = 64$ both on the boundary curve and on the circle for the far field pattern. The wave number is $k = 2$. The degree of the polynomials (43) is chosen as $J = 4$ and the regularization parameter for an H^2 regularization of the linearized data equation (40) is $\alpha = 0.05 \times 0.9^m$ for the m-th iteration step. For the perturbed data, random noise is added point wise with relative error in the L^2 norm. The iterations are started with an initial guess given by a circle of radius 0.6 centered at the origin. In both examples we used two incident waves, for the apple shape the incident directions are $d = (\pm 1, 0)$ and for the peanut shape $d = (0, \pm 1)$. In the figures the exact ∂D is given as dotted (magenta), the reconstruction as full (red) and the initial guess as dashed (blue) curve.

6 Solution of the Inverse Impedance Problem

Turning to the solution of the inverse impedance problem, we note that we can understand the data equation (36) as its main basis. Knowing the boundary ∂D, assuming again that k^2 is not a Dirichlet eigenvalue for the negative Laplacian in D we can represent u^s from u_∞ as a single-layer potential with density φ on ∂D. In order to attain the given far field pattern the density has to satisfy

$$S_\infty \varphi = u_\infty. \tag{45}$$

Once the density φ is known, the values of u and $\partial_\nu u$, i.e., the Cauchy data of u on the boundary can be obtained through the jump relations

$$u|_{\partial D} = u^i|_{\partial D} + \frac{1}{2}\, S\varphi \tag{46}$$

and

$$\left.\frac{\partial u}{\partial \nu}\right|_{\partial D} = \left.\frac{\partial u^i}{\partial \nu}\right|_{\partial D} + \frac{1}{2}\, K'\varphi - \frac{1}{2}\,\varphi. \tag{47}$$

For the numerical solution of (45) and the evaluation of (46) and (47) the approximations of the integral operators described in Sect. 3 are available. The derivative of $u|_{\partial D}$ with respect to s can be obtained by trigonometric differentiation. Knowing the Cauchy data on ∂D we now can recover the impedance functions μ and λ from the boundary condition (2).

The uniqueness result of Theorem 7 suggests that we need three incident plane waves with different directions leading to three far field patterns $u_{\infty,1}, u_{\infty,2}, u_{\infty,3}$ to reconstruct λ and μ. Solving the corresponding data equations (45) by Tikhonov regularization and using (46) and (47), we obtain three Cauchy pairs $u_1, \partial_\nu u_1$ and $u_2, \partial_\nu u_2$ and $u_3, \partial_\nu u_3$ for which we can exploit the boundary condition to construct μ and λ. For this we proceed somewhat differently than in [4] and mimic the idea of the proof of Theorem 7.

Multiplying the impedance condition (2) for u_1 by u_2 and the impedance condition for u_2 by u_1 and subtract we obtain

$$\mathrm{i}k\,\frac{\mathrm{d}}{\mathrm{d}s}\,\mu\left(u_1\,\frac{\mathrm{d}u_2}{\mathrm{d}s} - u_2\,\frac{\mathrm{d}u_1}{\mathrm{d}s}\right) = u_1\,\frac{\partial u_2}{\partial\nu} - u_2\,\frac{\partial u_1}{\partial\nu} \quad \text{on } \partial D.$$

From this it follows that

$$\mathrm{i}k\mu\left(u_1\,\frac{\mathrm{d}u_2}{\mathrm{d}s} - u_2\,\frac{\mathrm{d}u_1}{\mathrm{d}s}\right) = \mathscr{I}\left\{u_1\,\frac{\partial u_2}{\partial\nu} - u_2\,\frac{\partial u_1}{\partial\nu}\right\} + C_{12} \quad \text{on } \partial D \tag{48}$$

where C_{12} is a complex constant and $\mathscr{I}$ denotes integration over ∂D from a fixed $x_0 \in \partial D$ to $x \in \partial D$. Proceeding the same way with the two other possible combinations of u_2 and u_3 and of u_3 and u_1 we obtain two analogous equations with two more constants C_{23} and C_{31}. We approximate the unknown (parameterized) impedance function μ by a trigonometric polynomial of degree J and collocate the parametrized three equations of the form (48) at the $2n$ collocation points $t_j = j\pi/n$, $j = 1, \dots, 2n$. The resulting linear system of $6n$ equations for the $(2J + 1)$ Fourier coefficients of μ_{approx} and the three integration constants C_{12}, C_{23}, C_{31} then is solved in the least squares sense.

Having reconstructed μ, the remaining coefficient λ can be obtained from the impedance condition for any of the three functions u_1, u_2, or u_3. For symmetry, approximating the unknown function λ also by a trigonometric polynomial of degree J we collocate the boundary condition (2) for all three solutions u_1, u_2, and u_3 and solve the resulting linear system of $6n$ equations for the $(2J + 1)$ Fourier coefficients of λ_{approx} in the least squares sense.

For both our numerical examples the impedance functions are given by (29). The wave number is $k = 1$ and the three incident directions are $d = (1, 0)$ and $d = (\cos 2\pi/3, \pm\sin 2\pi/3)$. As in the examples of Sect. 4 the number of quadrature points is $2n = 64$ on each curve. The integration $\mathscr{I}$ is approximated by trigonometric interpolation quadrature. The degree of the polynomials for the approximation of the impedance functions is chosen as $J = 2$. We approximate the density φ via H^2 Tikhonov regularization of (45) by a trigonometric polynomial of degree $J_\varphi = 12$. The regularization parameter α is chosen by trial and error as $\alpha_{\text{exact}} = 10^{-10}$ and $\alpha_{\text{noise}} = 10^{-5}$.

Figures 3 and 4 show the reconstruction of the impedances for an ellipse with parametrization

$$z(t) = (\cos t, 0.7\sin t), \quad 0 \le t \le 2\pi, \tag{49}$$

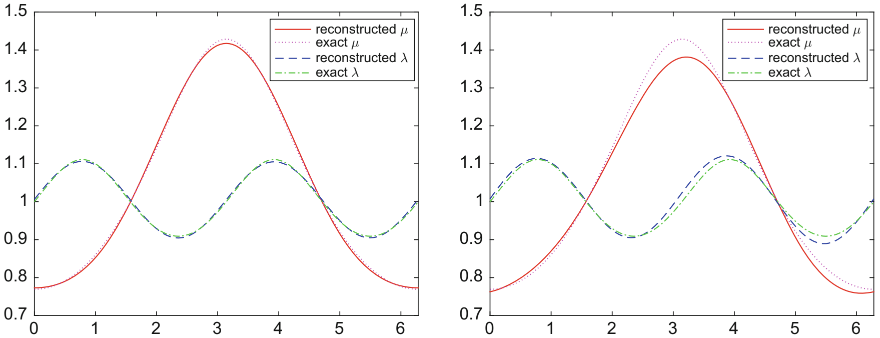


Fig. 3 Reconstruction of the impedance functions for the ellipse (49) for exact data (left) and 1% noise (right)

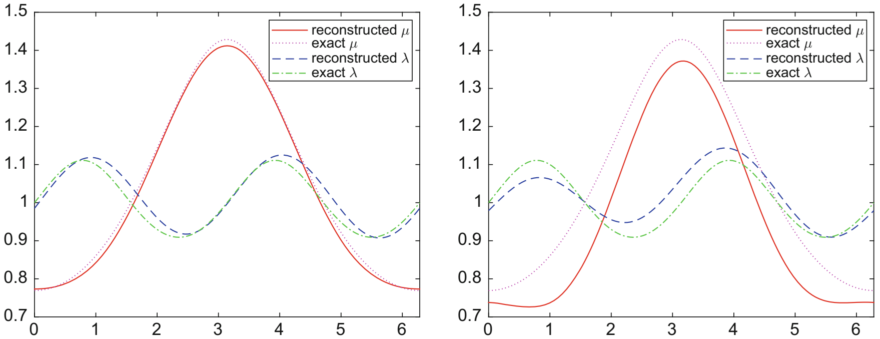


Fig. 4 Reconstruction of the impedance functions for the peanut (28) for exact data (left) and 1% noise (right)

and for the peanut (28). The exact μ is given as dotted (magenta) curve and the reconstruction as full (red) curve, the exact λ is dashed-dotted (green) and the reconstruction dashed (blue). In general, the examples and our further numerical experiments indicate that the simultaneous reconstruction of both impedance functions is very sensitive to noise.

In conclusion, we note that we have presented a method for the reconstruction of the shape (for known impedance functions) and a method for the reconstruction of the impedance functions (for known shape) with numerical examples as proof of concept. Further research is required for the solution of the full inverse problem by simultaneous linearization of the system (35) and (36) with respect to both the shape and the impedance analogous to [5].

References

1. Bourgeois, L., Haddar, H.: Identification of generalized impedance boundary conditions in inverse scattering problems. Inverse Prob. Imaging **4**(1), 19–38 (2010)
2. Bourgeois, L., Chaulet, N., Haddar, H.: Stable reconstruction of generalized impedance boundary conditions. Inverse Prob. **27**(9), 095002 (2011)
3. Bourgeois, L., Chaulet, N., Haddar, H.: On simultaneous identification of a scatterer and its generalized impedance boundary condition. SIAM J. Sci. Comput. **34**(3), A1824–A1848 (2012)
4. Cakoni, F., Kress, R.: Integral equation methods for the inverse obstacle problem with generalized impedance boundary condition. Inverse Prob. **29**(1), 015005 (2013)
5. Cakoni, F., Hu, Y., Kress, R.: Simultaneous reconstruction of shape and generalized impedance functions in electrostatic imaging. Inverse Prob. **30**(10), 105009 (2014)
6. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory. Applied Mathematical Sciences, vol. 93, 3rd edn. Springer, New York (2012)
7. Duruflé, M., Haddar, H., Joly, P.: High order generalized impedance boundary conditions in electromagnetic scattering problems. C. R. Phys. **7**(5), 533–542 (2006)
8. Haddar, H., Joly, P., Nguyen, H.M.: Generalized impedance boundary conditions for scattering by strongly absorbing obstacles: the scalar case. Math. Models Methods Appl. Sci. **15**(8), 1273–1300 (2005)
9. Ivanyshyn, O., Johansson, T.: Boundary integral equations for acoustical inverse sound-soft scattering. J. Inverse Ill-Posed Probl. **16**(1), 65–78 (2008)
10. Johansson, T., Sleeman, B.: Reconstruction of an acoustically sound-soft obstacle from one incident field and the far field pattern. IMA J. Appl. Math. **72**(1), 96–112 (2007)
11. Kirsch, A.: Surface gradient and continuity properties of some integral operators in classical scattering theory. Math. Meth. Appl. Sci. **11**(6), 789–804 (1989)
12. Kress, R.: Linear Integral Equations. Applied Mathematical Sciences, vol. 82, 3rd edn. Springer, New York (2013)
13. Kress, R.: A collocation method for a hypersingular boundary integral equation via trigonometric differentiation. J. Integral Equ. Appl. **26**(2), 197–213 (2014)
14. Kress, R., Sloan, I.H.: On the numerical solution of a logarithmic integral equation of the first kind for the Helmholtz equation. Numer. Math. **66**(1), 199–214 (1993)
15. McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)
16. Senior, T.B.A., Volakis, J.L.: Approximate Boundary Conditions in Electromagnetics. IEEE Electromagnetic Waves Series, vol. 41. The Institution of Electrical Engineers, London (1995)

Ian Sloan and Lattice Rules

Peter Kritzer, Harald Niederreiter, and Friedrich Pillichshammer

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Lattice rules are a powerful and popular form of quasi-Monte Carlo rules that are based on integration lattices. The study of the theory and application of lattice rules is intimately connected with the name Ian H. Sloan. We take the opportunity of Ian's 80th birthday to give an overview of his wide-ranging and fruitful contributions to this topic.

1 Introduction and Background

Ian Sloan is a major force in the area of numerical integration, and one of his main contributions to this subject is the theory of lattice rules. Therefore we find it very appropriate to devote an article appreciating his work on lattice rules to this anniversary volume.

Lattice rules belong to the family of quasi-Monte Carlo methods for numerical integration. The emphasis of these methods is on the multidimensional case. It is well known that quasi-Monte Carlo methods are effective even for integration problems of very high dimensions as they appear, for instance, in computational finance. We refer to the books of Dick and Pillichshammer [5] and Leobacher and Pillichshammer [43] as well as to the extensive survey article by Dick et al. [10]

P. Kritzer (✉) · H. Niederreiter
Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria
e-mail: peter.kritzer@oeaw.ac.at

F. Pillichshammer
Department of Financial Mathematics and Applied Number Theory, Johannes Kepler University Linz, Linz, Austria
e-mail: friedrich.pillichshammer@jku.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_33

for a general background on quasi-Monte Carlo methods. An older monograph on quasi-Monte Carlo methods is [47].

Quasi-Monte Carlo methods are, in a nutshell, deterministic versions of Monte Carlo methods. We normalize the integration domain to be the s-dimensional unit cube $I^s := [0, 1]^s$ for some $s \geq 1$. For a Lebesgue-integrable function f on I^s, the Monte Carlo method uses the estimate

$$\int_{I^s} f(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}_n), \tag{1}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are independent and uniformly distributed random samples from I^s. If $f \in L^2(I^s)$, then the expected error in (1) has the order of magnitude $N^{-1/2}$ (note that the convergence rate is independent of the dimension s). The quasi-Monte Carlo method for the same integral uses again the approximation (1), but now $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in I^s$ are deterministic points that are chosen to obtain a smaller error bound than the Monte Carlo error bound. Since $O\left(N^{-1/2}\right)$ is the mean-square error in (1) averaged over all samples of N points in I^s, such deterministic points must exist. One of the main issues in the theory of quasi-Monte Carlo methods is the explicit construction of deterministic point sets that yield better numerical integration schemes than the Monte Carlo method for large families of integrands. There are three principal approaches: (1) via Halton sequences and their variants (see [5, Section 3.4]); (2) via the theory of nets (see most parts of [5]); (3) via the theory of lattice rules. In this article, we focus on the third approach since it is here where Ian has made most of his contributions.

A lattice rule is a generalization of the classical one-dimensional numerical integration rule

$$\int_0^1 f(u)\mathrm{d}u \approx \frac{1}{N}\sum_{n=1}^{N} f\left(\frac{n-1}{N}\right) \tag{2}$$

called a rectangle rule which, in the case of an integrand f with $f(0) = f(1)$, agrees with the N-point trapezoidal rule for the interval [0, 1]. Lattice rules can be introduced from a group-theoretic and a geometric perspective.

We start with the first point of view. For a given dimension $s \geq 1$, the Euclidean space $\mathbb{R}^s$ is an abelian group under addition which has $\mathbb{Z}^s$ as a subgroup. Thus, we can form the factor group $\mathbb{R}^s/\mathbb{Z}^s$, also called the s-dimensional *torus group*. Now we consider, for the moment, the nodes $0, \frac{1}{N}, \ldots, \frac{N-1}{N}$ in (2) and their corresponding cosets $0+\mathbb{Z}, \frac{1}{N}+\mathbb{Z}, \ldots, \frac{N-1}{N}+\mathbb{Z}$ in the one-dimensional torus group $\mathbb{R}/\mathbb{Z}$. Clearly, these cosets form a finite subgroup of $\mathbb{R}/\mathbb{Z}$; in fact, it is the cyclic group generated by $\frac{1}{N}+\mathbb{Z}$. The generalization is now obvious. For an arbitrary s, let $L/\mathbb{Z}^s$ be any finite subgroup of $\mathbb{R}^s/\mathbb{Z}^s$ and let $\boldsymbol{y}_n + \mathbb{Z}^s$ with $\boldsymbol{y}_n \in [0, 1)^s$ for $n = 1, \ldots, N$ be the distinct cosets making up the group $L/\mathbb{Z}^s$. The point set consisting of the points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$

is called a *lattice point set* and the corresponding quasi-Monte Carlo approximation

$$\int_{I^s} f(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{y}_n) \tag{3}$$

is called a *lattice rule*.

Why do we speak of a "lattice rule" and not, for instance, of a "finite-group rule"? The reason is a nice geometric interpretation of lattice rules. Recall that an s-dimensional *lattice* is defined to be a discrete subgroup of $\mathbb{R}^s$ that is not contained in any proper linear subspace of $\mathbb{R}^s$. Equivalently, an s-dimensional lattice is obtained by taking a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_s$ of the vector space $\mathbb{R}^s$ and forming the set

$$L = \left\{ \sum_{i=1}^{s} k_i \boldsymbol{b}_i : k_i \in \mathbb{Z} \text{ for } 1 \le i \le s \right\}$$

of all linear combinations of $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_s$ with integer coefficients. The lattices corresponding to lattice rules must have an additional property expressed in the following definition.

Definition 1 An s-dimensional lattice is called an s-dimensional *integration lattice* if it contains $\mathbb{Z}^s$ as a subset.

If we take an s-dimensional integration lattice L as the starting point, then the intersection $L \cap [0, 1)^s$ is a finite set since L is discrete, and this finite set of points in $[0, 1)^s$ forms a lattice point set. Furthermore, all lattice point sets can be obtained in this way.

An important concept for the analysis of lattice rules is that of the dual lattice of a given integration lattice.

Definition 2 The *dual lattice* $L^\perp$ of the s-dimensional integration lattice L is defined by

$$L^\perp = \{\boldsymbol{h} \in \mathbb{R}^s : \boldsymbol{h} \cdot \boldsymbol{y} \in \mathbb{Z} \text{ for all } \boldsymbol{y} \in L\},$$

where $\cdot$ denotes the standard inner product on $\mathbb{R}^s$.

It is easy to see that the dual lattice of an s-dimensional integration lattice is always a subgroup of $\mathbb{Z}^s$.

An interesting special case arises if the finite subgroup $L/\mathbb{Z}^s$ of $\mathbb{R}^s/\mathbb{Z}^s$ is cyclic. Let N be the order of the group $L/\mathbb{Z}^s$ and let $\boldsymbol{y} + \mathbb{Z}^s$ be a generator of $L/\mathbb{Z}^s$. Then $N\boldsymbol{y} \in \mathbb{Z}^s$, and so $\boldsymbol{y} = \frac{1}{N}\boldsymbol{g}$ for some $\boldsymbol{g} \in \mathbb{Z}^s$. The lattice rule (3) then attains the form

$$\int_{I^s} f(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \approx \frac{1}{N}\sum_{n=1}^{N} f\left(\left\{\frac{n-1}{N}\boldsymbol{g}\right\}\right), \tag{4}$$

where the curly brackets denote fractional parts, that is, the points $\frac{n-1}{N}\boldsymbol{g}$, $n = 1, \ldots, N$, are considered modulo 1 in each coordinate. The numerical integration schemes in (4) were historically the first lattice rules and they are collectively known as the *method of good lattice points*. In this context, we call $\boldsymbol{g}$ the *generating vector* of the lattice rule.

We conclude this section with some comments on the history of lattice rules. The special case of the method of good lattice points was introduced by Korobov [28] in 1959 and a few years later independently by Hlawka [23]. The Soviet school quickly developed a quite satisfactory theory of the method of good lattice points, which is summarized in the book of Korobov [30] from 1963.

The first steps in the direction of general lattice rules were taken by Frolov [14] in 1977, but it was Ian who developed a systematic approach to the subject. His general perspective was first presented at a workshop in Canberra in December 1983 (see [68]). His fundamental paper [69] with Kachoyan on general lattice rules was submitted in August 1984, but unfortunately it took until 1987 to get published. In the meantime, his paper [63] had also advertised general lattice rules. Specialists in numerical integration got a first-hand exposure to general lattice rules through Ian's talk at an Oberwolfach workshop in November 1987 (see [73]). The second author first met Ian on a local train from Freudenstadt to Wolfach on their way to this workshop, but it was only after some curiosity-driven small talk that they were able to identify each other. Since then, Ian has had very close contacts with the Austrian community in the area of quasi-Monte Carlo methods, and we take this opportunity to thank him wholeheartedly for all his help and support.

An excellent summary of Ian's contributions to the theory of lattice rules as of 1994 can be found in his book [67] written jointly with Stephen Joe.

2 The Structure of Lattice Rules

In Sect. 1, we have defined lattice rules as quasi-Monte Carlo methods using the points of an integration lattice in I^s as the integration nodes. An obvious question is how such lattice rules can be represented in general. From (4), it can be seen that at least some lattice rules $Q_{N,s}$ for approximating the integral of a function f defined on I^s by N lattice points can be written as

$$Q_{N,s}(f) = \frac{1}{N}\sum_{j=1}^{N} f\Big(\Big\{\frac{j-1}{N}\boldsymbol{g}\Big\}\Big).$$

One may observe that the above representation is not unique. Indeed, choosing an integer $m \geq 2$ and replacing N and $\boldsymbol{g}$ by mN and $m\boldsymbol{g}$, respectively, yields an integration rule with mN points, each of them occurring with multiplicity m, such that the newly obtained rule is effectively equivalent to the one we started with. Furthermore, it is easy to construct examples where for the same N two distinct generating vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ yield rules with identical sets of integration nodes. Another question is of course whether there are maybe other lattice rules that can be represented in a different way than the one above. Hence, it is natural to ask for a canonical way of denoting lattice rules that makes it easier to structure and classify these, and to detect equivalences. Regarding this problem, it was Ian Sloan who, together with James Lyness, made substantial progress in providing a systematic way of representing lattice rules. We state a crucial theorem from [70] here.

Theorem 1 (Sloan and Lyness) *Let $Q_{N,s}$ be an s-dimensional lattice rule with $N \geq 2$ points. Then there exist uniquely determined integers*

- *r, with $r \in \{1, \dots, s\}$,*
- *and $n_1, \dots, n_r > 1$ with $n_{k+1} | n_k$ for $1 \leq k \leq r-1$, and $N = n_1 \cdots n_r$,*

such that $Q_{N,s}$ applied to a function f can be represented as

$$Q_{N,s}(f) = \frac{1}{N} \sum_{j_r=1}^{n_r} \cdots \sum_{j_1=1}^{n_1} f\left(\left\{\frac{j_1-1}{n_1} z_1 + \cdots + \frac{j_r-1}{n_r} z_r\right\}\right),$$

where $z_1, \dots, z_r$ are linearly independent (over $\mathbb{Q}$) integer vectors in $\mathbb{Z}^s$.

The proof of Theorem 1 is based on the group structure of lattice points, and can be found in [70], see also [67] and [51, Section 4.3.2].

Remark 1 The integer r in Theorem 1 is called *rank* of the lattice rule $Q_{N,s}$, and $n_1, \dots, n_r$ are the *invariants*.

Theorem 1 implies that the rank and the invariants of any given lattice rule are determined uniquely. What about the generating vectors $z_1, \dots, z_r$ of a rank-r lattice rule? In general, these cannot be determined uniquely. However, Sloan and Lyness [71] identified a class of lattice rules for which even the generating vectors can, in a certain sense, be identified unambiguously. This class is known as projection-regular lattice rules, which shall be described briefly here (we follow [67] and [71] in our terminology). Given an s-dimensional lattice rule $Q_{N,s}$ and $d \in \{1, \dots, s\}$, we speak of the d-dimensional principal projection of $Q_{N,s}$ when we consider the d-dimensional lattice rule obtained by omitting the last $s-d$ components of the integration nodes.

Note now that we can modify the representation of lattice rules outlined in Theorem 1 to a so-called extended canonical form by setting $n_{r+1} = \cdots = n_s = 1$, and by choosing arbitrary integer vectors $z_{r+1}, \dots, z_s$. Then we can represent a lattice rule $Q_{N,s}$ as

$$Q_{N,s}(f) = \frac{1}{N} \sum_{j_s=1}^{n_s} \cdots \sum_{j_1=1}^{n_1} f\left(\left\{\frac{j_1-1}{n_1} z_1 + \cdots + \frac{j_s-1}{n_s} z_s\right\}\right),$$

where we now trivially have $N = n_1 \cdots n_s$. Furthermore the rank r is in this case the maximal index such that $n_r > 1$. Using the latter representation of lattice rules, projection regularity is defined as follows (cf. [71]).

Definition 3 Let $Q_{N,s}$ be an s-dimensional lattice rule with invariants $n_1, \ldots, n_s$ in its extended canonical form. The rule $Q_{N,s}$ is called *projection regular* if for any choice of $d \in \{1, \ldots, s\}$ the d-dimensional principal projection of $Q_{N,s}$ has invariants $n_1, \ldots, n_d$.

The benefit of projection-regular lattice rules is that, by demanding some structure of the generating vectors, their choice is unique. Indeed, note that, given a lattice rule $Q_{N,s}$ in its extended canonical form, we can define an $s \times s$-matrix Z with the generating vectors $z_1, \ldots, z_s$ as its rows, i.e. $Z = (z_1, \ldots, z_s)^\top$. The matrix Z corresponding to a lattice rule is simply called Z-matrix in [71]. We say that Z is unit upper triangular if Z is upper triangular and has only 1s as the entries on the main diagonal. Using this terminology, we can state the following theorem, which is the main result of [71].

Theorem 2 (Sloan and Lyness) *A lattice rule $Q_{N,s}$ is projection-regular if and only if the rule $Q_{N,s}$ can be represented in an extended canonical form such that the corresponding Z-matrix is unit upper triangular.*

A question that remains is whether a projection-regular lattice rule can be represented in two different ways in an extended canonical form with unit upper triangular Z-matrices? The answer to this question is, as shown in [71], no, so it makes sense to speak of *the* standard form of a projection-regular lattice with unit upper triangular Z-matrix. For further details, we refer to the monograph [67] and the original paper [71]. We also remark that, for the special case of rank-2 lattices, Ian, again together with James Lyness, showed alternative representations, which subsequently made computer searches more effective (see, e.g., [44]). An overview of results related to the ones outlined in this section can also be found in [64].

A further important topic regarding the structure of lattice rules that was studied extensively by Ian and his collaborators is that of copy rules. Even though, due to their simple structure, rank-1 lattice rules have many convenient properties and have frequently been studied in many papers, Ian has pointed out on various occasions that also lattice rules of higher rank should be considered. To be more precise, as stated, e.g., in [13], lattice rules of higher or even maximal rank may perform better than rules of rank 1 with respect to certain criteria, such as the quantity P_α used in the classical literature on lattice rules. A prominent way of obtaining higher-rank lattice rules are copy rules. The basic idea of copy rules is elegant and simple: given a lattice rule $Q_{N,s}$ based on an integration lattice L, we obtain an integration rule consisting of m^s scaled "copies" of $Q_{N,s}$ by considering the lattice rule based on the integration lattice $m^{-1}L$ for some positive integer m. In other words, a copy rule is obtained by scaling the original rule $Q_{N,s}$ and copying it to each of the cubes obtained by partitioning I^s into cubes of volume m^{-s}. Using the canonical representation form, let us suppose we start with a rank-1 rule $Q_{N,s}$ with generating vector z, i.e.,

$$Q_{N,s}(f) = \frac{1}{N}\sum_{j=1}^{N} f\left(\left\{\frac{j-1}{N}z\right\}\right).$$

The m^s copy rule $Q_{m,N,s}$ is then given by

$$Q_{m,N,s}(f) = \frac{1}{m^s N}\sum_{k_s=1}^{m}\cdots\sum_{k_1=1}^{m}\sum_{j=1}^{N} f\left(\left\{\frac{j-1}{mN}z + \frac{(k_1-1,\dots,k_s-1)}{m}\right\}\right).$$

It was shown by Sloan and Lyness in [70] that an m^s copy rule has rank s; to be more precise, an s-dimensional lattice rule is of rank s if and only if it is an m^s copy rule obtained from a rule of lower rank. Furthermore, in the paper [13], Disney and Sloan gave numerical results indicating that copy rules perform better with respect to P_α than rank-1 rules of comparable size. We refer to [13] and [67] for further details.

The rank of lattice rules plays a role in yet another variant of lattice rules, namely so-called embedded (sometimes also imbedded) lattice rules. The crucial feature of embedded lattice rules is that, at least up to a certain point, these can be extended by adding additional points to the rule without having to discard previous ones. This property can be a desirable advantage in practical implementations. In this context one then more precisely speaks of sequences of embedded integration rules, a concept that not only occurs with lattice rules but also other types of quadratures. Sequences of embedded lattice rules, as they shall be presented here, were originally described by Ian together with Stephen Joe (see [25] and [67]); they have the additional property that the more points we add, the higher the rank of the lattice rule gets. In [67, p. 164], the following is remarked:

> Embedded sequences of quadrature rules open up the possibility of obtaining an error estimate with little extra cost. The hope is that such an error estimate, along with the full set of approximate integrals obtained from the sequence of embedded rules, can give valuable information about the reliability of the final (and hopefully best) approximation in the sequence.

Using a representation form similar to the canonical representation form introduced above, a sequence of embedded lattice rules is defined as follows. For a fixed positive integer m that is relatively prime to N, let for $r \in \{0,1,\dots,s\}$ the rule $Q_{r,m,N,s}$ be defined by

$$Q_{r,m,N,s}(f) = \frac{1}{m^r N}\sum_{k_r=1}^{m}\cdots\sum_{k_1=1}^{m}\sum_{j=1}^{N} f\left(\left\{\frac{j-1}{N}z + \frac{(k_1-1,\dots,k_r-1,0,\dots,0)}{m}\right\}\right),$$

where z is a generating vector with components that are relatively prime to m. As pointed out in [25], the $Q_{r,m,N,s}$ have the property that the integration nodes in $Q_{r,m,N,s}$ also occur in $Q_{r+1,m,N,s}$ for $0 \le r \le s-1$, and that $Q_{r,m,N,s}$ is a lattice rule with $m^r N$ points and rank r for $0 \leq r \leq s$, the only exception being the rank of $Q_{0,m,N,s}$, which is 1. Hence, we see that $Q_{s,m,N,s}$, which is the rule based on the largest number of integration nodes and, thus, intuitively the "most precise", has maximal rank. From the representation of the $Q_{r,m,N,s}$ we see much similarity to copy rules, and indeed it is shown in [25] that $Q_{s,m,N,s}$ is nothing but an m^s copy rule obtained from $Q_{0,m,N,s}$. The papers [25, 27], and again the book [67] contain further remarks, results, and numerical experiments related to this subject. Results on component-by-component constructions (see Sect. 4) of embedded lattice rules are outlined in [34]. We also remark that, though sometimes using slightly different techniques, the question of how to extend the number of points of lattice rules has been addressed in numerous papers from theoretical and practical viewpoints, see, for example, [3, 9, 21, 22, 48].

3 Error Bounds for Lattice Rules

Integration rules have to be accompanied by an error analysis in order to be useful in practice. There are, in principle, two ways of establishing an upper bound on the integration error for the lattice rule (3). One method is based on the general Koksma-Hlawka inequality for an integrand f of bounded variation $V(f)$ in the sense of Hardy and Krause (see [31, Chapter 2]). This inequality yields the bound

$$\left| \int_{I^s} f(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} - \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{y}_n) \right| \leq V(f) D_N^*(L),$$

where $D_N^*(L)$ is the *star discrepancy* of the integration nodes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ corresponding to the integration lattice L. Recall that

$$D_N^*(L) = \sup_J \left| \frac{A(J;L)}{N} - \lambda_s(J) \right|,$$

where J runs through all half-open subintervals of I^s with one vertex at the origin, $A(J;L)$ is the number of integers n with $1 \leq n \leq N$ such that $\boldsymbol{y}_n \in J$, and λ_s denotes the s-dimensional Lebesgue measure.

Thus, one arrives at the problem of bounding the star discrepancy $D_N^*(L)$ that was treated by Niederreiter and Sloan [49]. We need some notation in order to state their results. Let $C_s(N)$ be the set of all nonzero $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{Z}^s$ with $-N/2 < h_i \leq N/2$ for $1 \leq i \leq s$. For integers $h \in (-N/2, N/2]$, we put

$$r(h,N) = \begin{cases} N \sin(\pi |h|/N) & \text{if } h \neq 0, \\ 1 & \text{if } h = 0. \end{cases}$$

For $\boldsymbol{h} = (h_1, \ldots, h_s) \in C_s(N)$, we write

$$r(\boldsymbol{h}, N) = \prod_{i=1}^{s} r(h_i, N).$$

Then for an arbitrary s-dimensional integration lattice L with exactly N points in $[0, 1)^s$, we define

$$R(L) = \sum_{\boldsymbol{h} \in C_s(N) \cap L^{\perp}} \frac{1}{r(\boldsymbol{h}, N)}, \tag{5}$$

where $L^{\perp}$ is the dual lattice in Definition 2. According to [49, Proposition 3], the set $C_s(N) \cap L^{\perp}$ is nonempty for $s \geq 2$ and $N \geq 2$. The following discrepancy bound was shown in [49].

Theorem 3 (Niederreiter and Sloan) *Let L be an s-dimensional integration lattice with exactly N points in $[0, 1)^s$, where $s \geq 2$ and $N \geq 2$. Then*

$$D_N^*(L) \leq \frac{s}{N} + R(L).$$

An important quantity related to $R(L)$ is the *figure of merit* $\rho(L)$, sometimes also called the *Zaremba index* of L. For $h \in \mathbb{Z}$ put $r(h) = \max(1, |h|)$, and for $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{Z}^s$ put

$$r(\boldsymbol{h}) = \prod_{i=1}^{s} r(h_i).$$

Then for L as in Theorem 3, $\rho(L)$ is defined by

$$\rho(L) = \min_{\substack{\boldsymbol{h} \in L^{\perp} \\ \boldsymbol{h} \neq \boldsymbol{0}}} r(\boldsymbol{h}).$$

Now we can state the following result from [49].

Theorem 4 (Niederreiter and Sloan) *For L as in Theorem 3, we have*

$$R(L) < \frac{1}{\rho(L)} \left(\frac{2}{\log 2} \right)^{s-1} \left((\log N)^s + \frac{3}{2} (\log N)^{s-1} \right).$$

Lower bounds for $R(L)$ have been proved in [42, 59]. Theorem 4 suggests that the figure of merit $\rho(L)$ should be large for a useful integration lattice L. This is also demonstrated by the following lower bound on the star discrepancy from [49].

Theorem 5 (Niederreiter and Sloan) *For L as in Theorem 3, we have*

$$D_N^*(L) \geq \frac{c_s}{\rho(L)},$$

where c_s is an explicitly known positive constant depending only on s.

An interesting practical problem arises in connection with $R(L)$ in (5), namely how to compute this quantity efficiently. Joe and Sloan [26] use an asymptotic series in order to obtain a very good approximation to $R(L)$ which can be computed in $O(N)$ operations. This should be compared with the definition of $R(L)$ in (5) which requires the summation of $N^{s-1} - 1$ terms.

Now we turn to the second method of bounding the integration error in (3). Here we assume that the integrand f is a sufficiently regular periodic function on $\mathbb{R}^s$ of period 1 in each variable such that f is represented by its absolutely convergent Fourier series

$$f(\boldsymbol{u}) = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \widehat{f}(\boldsymbol{h}) \exp(2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{u}),$$

where $\widehat{f}(\boldsymbol{h}) = \int_{I^s} f(\boldsymbol{u}) \exp(-2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{u}) \mathrm{d}\boldsymbol{u}$ is the $\boldsymbol{h}^{\text{th}}$ Fourier coefficient of f. For a real number $\alpha > 1$, we say that $f \in \mathscr{E}^\alpha$ if there exists a constant $c(f) \geq 0$ such that

$$|\widehat{f}(\boldsymbol{h})| \leq c(f) r(\boldsymbol{h})^{-\alpha} \qquad \text{for all } \boldsymbol{h} \in \mathbb{Z}^s.$$

Let $\boldsymbol{g} \in \mathbb{Z}^s$ be a generating vector of a lattice rule of rank 1. Then for $f \in \mathscr{E}^\alpha$ we have the error bound

$$\Big| \int_{I^s} f(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} - \frac{1}{N} \sum_{n=1}^{N} f\Big(\frac{n-1}{N}\boldsymbol{g}\Big) \Big| \leq c(f) P_\alpha(\boldsymbol{g}, N), \tag{6}$$

where

$$P_\alpha(\boldsymbol{g}, N) = \sum_{\boldsymbol{h}} r(\boldsymbol{h})^{-\alpha}$$

with the summation running over all nonzero $\boldsymbol{h} \in \mathbb{Z}^s$ with $\boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod N$.

The bound (6) raises the question of how small we can make the quantity $P_\alpha(\boldsymbol{g}, N)$. Disney and Sloan [12] have shown that for every $s \geq 3$ and every $\alpha > 1$ we have

$$\min_{\boldsymbol{g}\in\mathbb{Z}^s} P_\alpha(\boldsymbol{g}, N) \leq ((2\mathrm{e}/s)^{\alpha s} + o(1)) \frac{(\log N)^{\alpha s}}{N^\alpha} \quad \text{as } N \to \infty. \tag{7}$$

This implies that with the method of good lattice points we can obtain a convergence rate $O(N^{-\alpha})$ for $f \in \mathscr{E}^\alpha$, up to logarithmic factors.

4 The Search for Good Lattice Rules

So far there exists no general construction principle for lattice rules with excellent properties with respect to a given quality measure such as the star discrepancy, the criteria P_α, R, or the figure of merit ρ. (The only exception is in dimension $s = 2$ where one can use a relation to Diophantine approximation to find explicit constructions, see [2, 47, 67] or [51, Example 4.3.15]. In this context we would like to particularly mention so-called Fibonacci lattice rules.)

For dimensions s larger than two there are mainly averaging arguments which guarantee the existence of lattice rules which behave excellently with respect to a given quality criterion. The simple idea behind this principle is that there must exist a lattice point set for which a given quality criterion is at least as small as the average of the same criterion extended over all possible lattice point sets. However, such results give no information about how to obtain a good lattice point set.

For a long time, good lattice point sets were found by a brute force computer search, see, for example, [18, 19, 45, 46, 61]. But this is infeasible if N and s are not very small. For example for the method of good lattice points for given N one has to search for $\boldsymbol{g}$ within the set $\{0, 1, \ldots, N-1\}^s$ which means that in the worst case one has to check N^s integer vectors to ensure success.

Since the 1980s Ian Sloan has contributed important results to the search for lattice rules. For example, we mention his work together with Linda Walsh [73, 74] (see also [67, Eq. (5.8)]) concerning the effective search for rank-2 lattice rules with small values of the quality criterion P_α. In their work Ian and Walsh restricted their attention to relatively coprime invariants m and n, which allows to write the rank-2 lattice rule in a very symmetric three-sum form. This representation makes it easier to eliminate "geometrically equivalent" rules from the search. Since P_α is invariant for geometrically equivalent rules it suffices to calculate P_α only for one member of a geometrically equivalent family. This is an important observation since there may be a huge number of rules which are distinct but geometrically equivalent.

Back to the nowadays most important case of rank-1 rules: One way to reduce the search space for good lattice points is a method that goes back to Korobov [29]. He suggested considering only lattice points of the form

$$\boldsymbol{g}(g) = (1, g, g^2, \ldots, g^{s-1}) \quad \text{where } g \in \{1, \ldots, N-1\}.$$

This restriction reduces the size of the search space from N^s to $N-1$, since for given N one only has to search for g in the set $\{1, 2, \ldots, N-1\}$. The limitation to good lattice points of so-called "Korobov form" is in many cases justified by averaging results that are basically of the same quality as the averaging results over all lattice points from $\{0, 1, \ldots, N-1\}^s$ (see, for example, [29, 47, 51]). The method is very effective and in fact some of the good lattice points in the tables in [19, 46, 61] are of the specific Korobov form.

In [84] Ian, together with Xiaoqun Wang and Josef Dick, studied Korobov lattice rules in weighted function spaces. As the quality criterion they chose the worst-case integration error in weighted Korobov spaces which is more general than the classical quality measure P_α (in fact, in the unweighted case, P_α of a lattice rule is exactly the squared worst-case error of the same rule). For simplicity we restrict ourselves to the unweighted case and state the following special case of [84, Algorithm 1].

In the following let $Z_N := \{1, 2, \ldots, N-1\}$ and let $\zeta(\beta) = \sum_{j \geq 1} j^{-\beta}$ be the Riemann zeta function.

Algorithm 1 (Korobov Lattice Rule) *Let N be a prime number and let $s \in \mathbb{N}$. The optimal Korobov generator is found by minimizing the measure*

$$P_\alpha(\boldsymbol{g}(g), N) = -1 + \frac{1}{N} \sum_{k=0}^{N-1} \sum_{h_1,\ldots,h_s \in \mathbb{Z}} \prod_{j=1}^{s} \frac{1}{r(h_j)^\alpha} \exp\left(2\pi \mathtt{i} k \frac{h_j g^{j-1}}{N}\right),$$

with respect to $g \in Z_N$.

It is pointed out in [84] that the number of operations needed to find the optimal Korobov lattice rule for even α (in this case one can get rid of the infinite sums in the definition of P_α) and a single dimension s is $O(sN^2)$. This is a remarkable improvement compared to the order $O(N^s)$ for a full search. Now the surprise is that the so found lattice rule is also of good quality. The following result is a specific case of the more general [84, Theorem 4]:

Theorem 6 (Wang, Sloan, and Dick) *Let N be a prime number and assume that g_* was found by Algorithm 1. Then for arbitrary $\tau \in [1, \alpha)$ we have*

$$P_\alpha(\boldsymbol{g}(g_*), N) \leq C_s(\alpha, \tau) \left(\frac{s}{N-1}\right)^\tau,$$

where $C_s(\alpha, \tau) = \exp(2s\tau\zeta(\frac{\alpha}{\tau}))$.

Ian's most important contribution to effective search routines for lattice rules, however, is the so-called component-by-component (CBC) construction of good lattice points. Although already described in the year 1963 by Korobov [30], this method fell into oblivion and it was Ian who re-invented it in a joint paper with Andrew V. Reztsov [72] in the year 2002. With this method good lattice points can be constructed one component at a time. At this time, this was a big surprise. We quote Ian and Reztsov, who wrote: *"The results may be thought surprising, since it is generally accepted that knowledge of a good lattice rule in s dimensions does not help in finding a good rule in $s+1$ dimensions."* (cf. [72, p. 263]).

To be more precise, given N and a quality measure, say, e.g., P_α, one starts with a sufficiently good one-dimensional generator (g_1). To this generator one appends a second component g_2 which is chosen as to minimize the quality criterion, in our case P_α. In a next step, one appends to the now two-dimensional generator (g_1, g_2) a third component g_3 which is again chosen as to minimize P_α. This procedure is repeated until one obtains an s-dimensional generating vector $\boldsymbol{g} = (g_1, g_2, \ldots, g_s)$.

In each of the s steps the search space Z_N has cardinality $N-1$ and hence the overall search space for the CBC method is reduced to a size of order $O(sN)$. Hence this provides a feasible way of finding a generating vector.

The use of the CBC method is justified by the following result which guarantees that the obtained generating vector is of excellent quality. In order to stress the dependence of P_α on the dimension s we will in the following write $P_\alpha^{(s)}$ for the s-dimensional case. We state [72, Theorem 2.1]:

Theorem 7 (Sloan and Reztsov)

i) For arbitrary $\beta > 1$ and prime N, with $N \geq 2\zeta(\beta) + 1$, there exists a sequence $(g_j)_{j=1}^{\infty}$, with $g_j \in Z_N$, such that for all $s \geq 1$ and all $\alpha \geq \beta$

$$P_\alpha^{(s)}((g_1, \ldots, g_s), N) \leq \frac{(1 + 2\zeta(\beta))^{s\alpha/\beta}}{N^{\alpha/\beta}}. \tag{8}$$

ii) The members g_j of a sequence $(g_j)_{j=1}^{\infty}$ satisfying (8) can be determined recursively, by setting $g_1 = 1$ and taking $g_{s+1} \in Z_N$ to be the least value of $g \in Z_N$ that minimizes $P_\beta^{(s+1)}((g_1, \ldots, g_s, g), N)$.

From this result Ian and Reztsov deduced the following:

Theorem 8 (Sloan and Reztsov)

i) Let $\alpha > 1$ and let $s_{\max} \geq 2$ be a fixed positive integer, and let N be a prime number satisfying

$$N > \mathrm{e}^{s_{\max} \frac{\alpha}{\alpha - 1}}.$$

There exists a finite sequence $(g_j)_{j=1}^{s_{\max}}$ such that for any s satisfying $1 \leq s \leq s_{\max}$

$$P_\alpha^{(s)}((g_1, \ldots, g_s), N) \leq D(s, \alpha) \frac{(\log N)^{s\alpha}}{N^\alpha},$$

where

$$D(s, \alpha) := \left(\frac{3}{s_{\max}} \right)^{s\alpha} \mathrm{e}^{s_{\max}\alpha}.$$

ii) The sequence $(g_j)_{j=1}^{s_{\max}}$ can be constructed as in part (ii) of Theorem 7, with β given by

$$\beta := \frac{\log N}{\log N - s_{\max}}.$$

The bound on P_α from Theorem 8 should be compared to that in (7).

The search procedure from Theorem 7 is nowadays called the "CBC algorithm" and can be summarized in concise form as follows:

Algorithm 2 (CBC Algorithm) *Let $s, N \in \mathbb{N}$.*

1. *Choose $g_1 = 1$.*
2. *For $s = 1, 2, \ldots, s_{\max} - 1$, assume that $g_1, \ldots, g_s$ are already found. Choose $g_{s+1} \in Z_N$ to minimize $P_\alpha^{(s+1)}((g_1, \ldots, g_s, g), N)$ as a function of $g \in Z_N$.*

It is pointed out in [72] that the total cost of the CBC construction for P_α is of order of magnitude $O(s^2N^2)$. This construction cost is typical for a straightforward implementation of the CBC algorithm based on P_α and also other quality measures. Sometimes this cost can easily be reduced to $O(sN^2)$ under the assumption of a memory capacity of order $O(N)$. This is comparable to the search for generators of Korobov form and makes the CBC algorithm applicable for moderately large N. However, if one is interested in lattice rules with really large N, one has to further reduce the factor N^2 in the construction cost to get a feasible construction method. A breakthrough with respect to this problem was obtained by Dirk Nuyens and Ronald Cools [57, 58] in 2006 using fast Fourier transform (FFT) methods for the construction of lattice point sets. This way, it is possible to construct, for a given prime number N, an s-dimensional generating vector $\boldsymbol{g}$ in $O(sN \log N)$ operations, compared to $O(sN^2)$ operations for the usual CBC algorithm. The modified CBC construction is commonly known as "fast CBC construction".

The CBC construction is even more suited for the construction of good lattice rules in weighted spaces where the successive coordinates of the integrands have decreasing influence on the integral. This was already indicated in [72] although this paper only deals with P_α, i.e., the unweighted case. We quote from [72, p. 264]:

> Lattice rules obtained by the present component-by-component algorithm may be particularly valuable if the user knows that for the integrand under consideration the first coordinate is more important than the second, the second than the third, and so on. This is because the error bounds in Theorems 2.1 and 4.1 [here Theorems 7 and 8] hold simultaneously for all s up to the dimension of the particular integral. Thus the error bounds hold for all the principal projections obtained by omitting one or more of the later coordinates of the integrand.

The different importance of the coordinate projections is usually modeled by a sequence of weights. For weighted spaces the CBC construction automatically adapts the constructed lattice point to the given sequence of weights. This means that a good lattice point constructed by CBC for a given sequence of weights need not be good for another weight sequence.

Nowadays the fast CBC construction is used with great success also for other quality measures such as the criterion R (see for example [43, Chapter 4]) or the (mean square) worst-case error of (shifted) lattice rules in weighted Sobolev spaces and Korobov spaces. This will be discussed further in Sect. 5.

5 Randomizations and Generalizations of Lattice Rules

Lattice rules are perfectly configured for the numerical integration of smooth and in each variable one-periodic functions. It is well known that they can achieve the optimal convergence rate in the class $\mathscr{E}^{\alpha}$ (see Sect. 3) or, in a more modern setting, in weighted Korobov spaces of smoothness α.

Let $\alpha > 1$ and let $\boldsymbol{\gamma} = (\gamma_j)_{j\in\mathbb{N}}$ be a sequence of positive weights.[1] Then the $\boldsymbol{\gamma}$-weighted Korobov space $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ of smoothness α is a reproducing kernel Hilbert space (see [1] for general information about reproducing kernel Hilbert spaces) of one-periodic, complex valued functions with reproducing kernel

$$K_{s,\alpha,\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \frac{\exp(2\pi\mathtt{i}\boldsymbol{h}\cdot(\boldsymbol{x}-\boldsymbol{y}))}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})},$$

where for $\boldsymbol{h} = (h_1,h_2,\ldots,h_s)$, $r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h}) = \prod_{j=1}^{s} r_{\alpha,\gamma_j}(h_j)$, and for $h \in \mathbb{Z}$ and $\gamma > 0$, $r_{\alpha,\gamma}(h) = \gamma^{-1}|h|^{\alpha}$ if $h \neq 0$ and $r_{\alpha,\gamma}(0) = 1$. The corresponding inner product is

$$\langle f,g\rangle_{s,\alpha,\boldsymbol{\gamma}} = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})\widehat{f}(\boldsymbol{h})\overline{\widehat{g}(\boldsymbol{h})},$$

where $\widehat{f}(\boldsymbol{h}) = \int_{I^s} f(\boldsymbol{x})\exp(-2\pi\mathtt{i}\boldsymbol{h}\cdot\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is the $\boldsymbol{h}^{\text{th}}$ Fourier coefficient of f. The norm in $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ is defined as $\|\cdot\|_{s,\alpha,\boldsymbol{\gamma}} = \langle\cdot,\cdot\rangle_{s,\alpha,\boldsymbol{\gamma}}^{1/2}$.

The worst-case error of a linear integration rule

$$Q_{\mathscr{P},\boldsymbol{a}}(f) = \sum_{j=1}^{N} a_j f(\boldsymbol{x}_j) \quad \text{for } f \in \mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$$

based on a point set $\mathscr{P} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$ in I^s and coefficients $\boldsymbol{a} = (a_1,\ldots,a_N) \in \mathbb{R}^N$ is defined as

$$e(\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}},\mathscr{P},\boldsymbol{a}) = \sup_f \left|\int_{I^s} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - Q_{\mathscr{P},\boldsymbol{a}}(f)\right|,$$

where the supremum is extended over the unit ball of $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$, i.e., over all $f \in \mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ with $\|f\|_{s,\alpha,\boldsymbol{\gamma}} \leq 1$. If $\boldsymbol{a} = (1/N,\ldots,1/N)$, then the integration rule is a quasi-Monte Carlo rule $Q_{\mathscr{P},\boldsymbol{a}}$ and we omit the parameter $\boldsymbol{a}$ in the notation. There is a convenient formula for the worst-case error in $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ of rank-1 lattice rules,

$$e^2(\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}},\mathscr{P}(\boldsymbol{g},N)) = -1 + \sum_{\substack{\boldsymbol{h}\in\mathbb{Z}^s\\ \boldsymbol{g}\cdot\boldsymbol{h}\equiv 0 \pmod N}} \frac{1}{r_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{h})},$$

see [77, Eq. (15)]. This formula should be compared to the definition of P_α in Sect. 3.

[1] For simplicity here we only consider weights of product form, i.e., so-called *product weights*.

Let us first consider the unweighted case, i.e., the weights satisfy $\gamma_j = 1$ for all $j \in \mathbb{N}$ (in this case we also omit $\boldsymbol{\gamma}$ in the notation). Then for prime N one can construct by a CBC algorithm a lattice point $\boldsymbol{g} \in Z_N^s$ such that

$$e^2(\mathscr{H}_{s,\alpha}, \mathscr{P}(\boldsymbol{g}, N)) \leq (1 + 2\zeta(\alpha))^s \frac{1 + 2^{\alpha(s+1)}(1 + \log N)^{\alpha s}}{N^\alpha}.$$

Note that the order of magnitude in N is, up to the $\log N$-factor, best possible since it is known that for arbitrary N-element point sets $\mathscr{P}$ in I^s, $\boldsymbol{a} \in \mathbb{R}^N$, and for every $\alpha > 1$ we have

$$e^2(\mathscr{H}_{s,\alpha}, \mathscr{P}, \boldsymbol{a}) \geq C(s, \alpha) \frac{(\log N)^{s-1}}{N^\alpha},$$

where $C(s, \alpha) > 0$ depends only on α and s, but not on N.

This means that asymptotically, for N tending to infinity, lattice rules can achieve the optimal rate of convergence for the worst-case integration error in (unweighted) Korobov spaces. However, the question is how long one has to wait to see this excellent asymptotic behavior especially when the dimension s is large. This issue is the subject of tractability, a further topic to which Ian made significant contributions during the last two decades. We now switch again to the weighted setting.

The N^{th} *minimal error* in $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ is given by

$$e(N, s) := \inf_{\boldsymbol{a}, \mathscr{P}} e(\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}, \mathscr{P}, \boldsymbol{a}),$$

where the infimum is taken over all N-element point sets $\mathscr{P}$ in I^s and coefficients $\boldsymbol{a}$.

Furthermore, for $\varepsilon \in (0, 1]$ and $s \in \mathbb{N}$ the *information complexity* is the minimal number of information evaluations needed to achieve a minimal error of at most ε, to be more precise

$$N(\varepsilon, s) = \min\{N \in \mathbb{N} : e(N, s) \leq \varepsilon\}.$$

The subject of tractability deals with the question in which way the information complexity depends on ε and s; cf. the three books [52–54].

- The *curse of dimensionality* holds if there exist positive numbers C, τ, and ε_0 such that

$$N(\varepsilon, s) \geq C(1 + \tau)^s \quad \text{for all } \varepsilon \leq \varepsilon_0 \text{ and infinitely many } s.$$

- *Weak tractability* holds if

$$\lim_{\varepsilon^{-1}+s\to\infty} \frac{\log N(\varepsilon, s)}{\varepsilon^{-1} + s} = 0.$$

- *Polynomial tractability* holds if there exist non-negative numbers C, τ_1, τ_2 such that

$$N(\varepsilon, s) \leq C s^{\tau_1} \varepsilon^{-\tau_2} \quad \text{for all } s \in \mathbb{N},\ \varepsilon \in (0, 1).$$

- *Strong polynomial tractability* holds if there exist non-negative numbers C and τ such that

$$N(\varepsilon, s) \leq C \varepsilon^{-\tau} \quad \text{for all } s \in \mathbb{N},\ \varepsilon \in (0, 1).$$

  The *exponent* τ^* *of strong polynomial tractability* is defined as the infimum of τ for which strong polynomial tractability holds.

It follows from a work of Ian together with Henryk Woźniakowski [75] that one has the curse of dimensionality for integration in the unweighted Korobov space. The following result is [75, Theorem 1]:

Theorem 9 (Sloan and Woźniakowski) *Consider integration in the Korobov space* $\mathscr{H}_{s,\alpha}$*. If* $N < 2^s$*, then* $e(N, s) = 1$.

In other words, the number of function evaluations required to achieve a worst case error less than one is exponential in the dimension. To prove this result, the authors constructed for given α and for given integration nodes $\mathscr{P}$ and coefficients $\boldsymbol{a}$ a so-called bump function h which belongs to the unit ball of the Korobov space $\mathscr{H}_{s,\alpha}$ and which satisfies $\int_{I^s} h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$ and $Q_{\mathscr{P},\boldsymbol{a}}(h) = 0$. From this it follows that

$$e(\mathscr{H}_{s,\alpha}, \mathscr{P}, \boldsymbol{a}) \geq \left| \int_{I^s} h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - Q_{\mathscr{P},\boldsymbol{a}}(h) \right| = \left| \int_{I^s} h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| = 1.$$

To obtain positive results for tractability one has to change to the weighted setting. The concept of weighted function spaces was first introduced by Ian and Henryk Woźniakowski in the seminal paper [76] titled “When are quasi Monte Carlo algorithms efficient for high-dimensional problems?” (in fact, this question is programmatic for large parts of Ian’s work during the last 20 years, see also his talk “On the unreasonable effectiveness of QMC” [66] given at the 9th *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* in Warsaw in 2010, or the essay [65]). For the basic idea underlying the weighted setting, which is nowadays standard in this theory, we quote from [77, p. 699]:

> The motivation lies in the idea [...] that it may be useful to order the coordinates $x_1, x_2, \ldots, x_d$ in such a way that x_1 is the most important coordinate, x_2 the next, and so on; and to quantify this by associating non-increasing weights $\gamma_1, \gamma_2, \ldots, \gamma_d$ to the successive coordinate directions.

The problem of tractability in weighted Korobov spaces was tackled by Ian and Henryk Woźniakowski in [77]. The main results of this paper are summarized in the following theorem (we remark that the third item of the following theorem was not explicitly mentioned in [77], but can easily be deduced from the results shown there):

Theorem 10 (Sloan and Woźniakowski) *Numerical integration in $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ is*

1. *strongly polynomially tractable if and only if $\sum_{j=1}^{\infty} \gamma_j < \infty$. If λ_0 is the infimum over all $\lambda \leq 1$ such that $\sum_{j=1}^{\infty} \gamma_j^{\lambda} < \infty$ holds, then the ε-exponent τ^* of strong polynomial tractability lies in the interval $[2\alpha^{-1}, \max(2\alpha^{-1}, 2\lambda_0)]$. In particular, if $\lambda_0 \leq \alpha^{-1}$, then the ε-exponent τ^* of strong polynomial tractability is $2\alpha^{-1}$.*
2. *polynomially tractable if and only if $\limsup_{s\to\infty} \sum_{j=1}^{s} \gamma_j / \log s < \infty$ holds.*
3. *weakly tractable if and only if $\lim_{s\to\infty} \sum_{j=1}^{s} \gamma_j / s = 0$ holds.*

For $s > 1$ and N prime, there exist lattice rules that achieve the corresponding upper bounds on the worst-case error.

The above result guarantees the existence of lattice rules that achieve the upper bounds on the worst-case error. In fact, these lattice rules can be efficiently constructed with the CBC approach (see Sect. 4) as shown by Kuo [32] in 2003 and even by the fast CBC approach according to Nuyens and Cools [57]. Extensions of these results to composite N were given in [4, 33] and results for Korobov spaces with general weights were obtained in [7]. So the problem of numerical integration in weighted Korobov spaces is very well understood. However, these results for Korobov spaces are not often usable in practice because integrands are typically not fully periodic. Since the early 1990s Ian works on the question how lattice rules can also be applied to numerical integration of not necessarily periodic functions. Early works in this direction are [60] and [50]. Nowadays, there are various strategies which can be used in order to apply lattice rules also to not necessarily periodic functions.

One such strategy are methods for periodization, i.e., methods for transforming a sufficiently smooth non-periodic integrand into a periodic integrand without changing its value of the integral. Such periodization techniques can be found, for example, in [10, Section 5.10] or in [47, Section 5.1]. A disadvantage of periodization is that the norm of the transformed integrand can be exponentially large in s. Hence this method is generally only feasible for moderate s.

Another possible strategy are randomly shifted lattice rules. For given quasi-Monte Carlo points $\mathscr{P} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N\}$ in I^s and $\boldsymbol{\Delta} \in I^s$ let

$$\mathscr{P} + \boldsymbol{\Delta} = \{\{\boldsymbol{t}_j + \boldsymbol{\Delta}\} \,:\, j = 1, 2, \ldots, N\}$$

denote the shifted point set. Then the quasi-Monte Carlo rule based on $\mathscr{P} + \boldsymbol{\Delta}$ is called a *shifted quasi-Monte Carlo rule*. In this sense a *shifted lattice rule* is of the form

$$Q_{N,s}(f, \boldsymbol{\Delta}) = \frac{1}{N} \sum_{j=1}^{N} f\left(\left\{\frac{j-1}{N}\boldsymbol{g} + \boldsymbol{\Delta}\right\}\right).$$

In the context of randomly shifted lattice rules one studies the so-called *shift-averaged worst-case error*. Let $(\mathscr{H}, \|\cdot\|_{\mathscr{H}})$ be a normed space of functions defined on the s-dimensional unit cube I^s. Then the worst-case error of a quasi-Monte Carlo rule $Q_{N,s}$ which is based on a point set $\mathscr{P}$ is

$$e(\mathscr{H}, \mathscr{P}) = \sup_f \left| \int_{I^s} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - Q_{N,s}(f) \right|,$$

where the supremum is extended over all $f \in \mathscr{H}$ with $\|f\|_{\mathscr{H}} \leq 1$. The shift-averaged worst-case error of a shifted lattice rule is defined by

$$e^{\mathrm{sh}}(\mathscr{H}, \mathscr{P}(\boldsymbol{g}, N)) = \left(\int_{I^s} e^2(\mathscr{H}, \mathscr{P}(\boldsymbol{g}, N) + \boldsymbol{\Delta}) \mathrm{d}\boldsymbol{\Delta} \right)^{1/2}.$$

This can be used to bound the root-mean-square error over uniformly distributed random shifts $\boldsymbol{\Delta}$ from I^s via the estimate

$$\sqrt{\mathbb{E}\left| \int_{I^s} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - Q_{N,s}(f, \boldsymbol{\Delta}) \right|^2} \leq e^{\mathrm{sh}}(\mathscr{H}, \mathscr{P}(\boldsymbol{g}, N)) \|f\|_{\mathscr{H}}.$$

Randomly shifted lattice rules are applied very successfully in weighted Sobolev spaces which shall be briefly introduced now. We just speak about the Hilbert space case with smoothness one and distinguish between the anchored space $\mathscr{H}_{s,\boldsymbol{\gamma}}^{\pitchfork}$ and the unanchored (or ANOVA) space $\mathscr{H}_{s,\boldsymbol{\gamma}}^{A}$. In both cases the weighted Sobolev spaces are s-fold tensor products of one-dimensional reproducing kernel Hilbert spaces $\mathscr{H}_{s,\boldsymbol{\gamma}}^{\star} = \mathscr{H}_{1,\gamma_1}^{\star} \otimes \ldots \otimes \mathscr{H}_{s,\gamma_s}^{\star}$ where $\star \in \{\pitchfork, A\}$. The one-dimensional building blocks are reproducing kernel Hilbert spaces with reproducing kernel of the form

$$K_{1,\gamma}^{\star}(x, y) = 1 + \gamma \eta_{\star}(x, y)$$

where in the anchored case

$$\eta_{\pitchfork}(x, y) = \begin{cases} \min(x, y) - c & \text{if } x, y > c, \\ c - \max(x, y) & \text{if } x, y < c, \\ 0 & \text{otherwise,} \end{cases}$$

with anchor $c \in [0, 1]$, and in the unanchored case

$$\eta_A(x, y) = \tfrac{1}{2} B_2(|x - y|) + (x - \tfrac{1}{2})(y - \tfrac{1}{2}),$$

where $B_2(x) = x^2 - x + \frac{1}{6}$ is the second Bernoulli polynomial. More detailed information on these spaces can be found, for example, in [10, Sections 4.2 and 4.3].

In [78] Ian, Frances Kuo, and Stephen Joe constructed randomly shifted lattice rules in the weighted anchored Sobolev space with the CBC algorithm. Furthermore, in [79] the three also introduced a CBC algorithm to construct a deterministic shift besides the generating vector. With these constructions one can achieve tractability for the integration problem depending on the decay of the weights. However, in these works only a convergence rate of order $O(N^{-1/2})$ was proved. Later Kuo [32] showed that even the optimal convergence rate of order of magnitude $O(N^{-1})$ can be achieved. Further results in this direction are due to Kuo and Joe [33] and due to Dick [4].

The following result (see [10, Section 5] for a proof) summarizes this development. Let

$$\widetilde{Z}_N := \{g \in \{1, 2, \ldots, N-1\} \ : \ \gcd(g, N) = 1\}.$$

Theorem 11 (Optimal CBC Error Bound) *Let $\star \in \{\pitchfork, A\}$. The generating vector $\boldsymbol{g} \in \widetilde{Z}_N^s$ constructed by the CBC algorithm, minimizing the squared shift-averaged worst-case error $e^{\mathrm{sh}}(\mathscr{H}^{\star}_{s,\boldsymbol{\gamma}}, \mathscr{P}(\boldsymbol{g}, N))$ for the corresponding weighted Sobolev space in each step, satisfies*

$$[e^{\mathrm{sh}}(\mathscr{H}^{\star}_{s,\boldsymbol{\gamma}}, \mathscr{P}(\boldsymbol{g}, N))]^2 \leq \left(\frac{1}{\varphi(N)} \left(-1 + \prod_{j=1}^{s} \left(1 + \gamma_j^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} + \beta_{\star}^{\lambda} \right) \right) \right) \right)^{1/\lambda}$$

for all $\lambda \in (1/2, 1]$, where φ denotes the Euler totient function, $\beta_A = 0$, and $\beta_{\pitchfork} = c^2 - c + 1/3$ in the anchored case with anchor c.

The error estimate in the above theorem guarantees a convergence rate of order of magnitude $O(N^{-1+\varepsilon})$ for $\varepsilon > 0$. Furthermore, under the assumption of sufficiently fast decreasing weights, tractability for the shift-averaged worst-case error is obtained. For example, $\sum_{j=1}^{\infty} \gamma_j < \infty$ implies strong polynomial tractability. If further λ_0 is the infimum over all $\lambda \leq 1$ such that $\sum_{j=1}^{\infty} \gamma_j^{\lambda} < \infty$ holds, then the ε-exponent of strong polynomial tractability is at most $2\lambda_0$.

In practice it may happen that unbounded integrands arise resulting from the use of the cumulative inverse normal transformation in order to map the integral from the unbounded domain $\mathbb{R}^s$ to the unit cube I^s. In [38, 85] Ian and his co-workers studied randomly shifted lattice rules for such types of integrands. Thereby they can achieve a convergence rate close to the optimal order $O(N^{-1})$.

Very briefly we report about a third strategy how to apply lattice rules to non-periodic integrands, namely tent transformed lattice rules. The tent transformation $\phi : [0, 1] \to [0, 1]$ is given by $\phi(x) = 1 - |2x - 1|$. For vectors $\boldsymbol{x}$ the tent transformed point $\phi(\boldsymbol{x})$ is understood componentwise. A tent transformed lattice rule is of the form

$$Q^{\phi}_{N,s}(f) = \frac{1}{N} \sum_{j=1}^{N} f \left(\phi \left(\left\{ \frac{j-1}{N} \boldsymbol{g} \right\} \right) \right).$$

In [11] such rules are used for the integration of cosine series which are not necessarily periodic. The authors obtained convergence rates for the worst-case errors that are optimal in the order of magnitude in N. Furthermore, under certain conditions on the weights one can also achieve tractability. We remark that these results are also valid for the unanchored Sobolev space $\mathscr{H}_{s,\boldsymbol{\gamma}}^{A}$. Via embeddings one can transfer these results also to the anchored Sobolev space $\mathscr{H}_{s,\boldsymbol{\gamma}}^{\pitchfork}$. The first who used the tent transform in the context of lattice rules was Hickernell [20].

Finally, we would like to mention that there is also the concept of polynomial lattice rules which was introduced by Niederreiter, see, for example [47]. Polynomial lattice rules are special instances of digital nets. Their overall structure is very similar to lattice rules. The main difference is that lattice rules are based on number theoretic concepts whereas polynomial lattice point sets are based on algebraic methods (polynomial arithmetic over a finite field). Also polynomial lattice rules work well for the integration of not necessarily periodic functions. A contribution of Ian to this topic is the paper [6]. This paper stood at the beginning of an intensive and fruitful collaboration of the third author with members of Ian's group in Sydney with several visits in both directions. The third author wishes to thank Ian and his colleagues at the University of New South Wales for their warm hospitality during these visits in Sydney.

6 Applications of Lattice Rules

Ian Sloan has always been an influential proponent of making theoretical results applicable to real-world problems, and telling the story of Ian and lattice rules would be incomplete without dealing with his rich work on applications to various topics. We shall try to give an overview of his numerous results on applications of lattice rules, ranging from function approximation to financial mathematics and, recently, partial differential equations.

One of the areas that are most frequently highlighted as fields of application of quasi-Monte Carlo methods is financial mathematics. Ian Sloan has not restricted his research on lattice rules to theoretical results, but has always been very interested in actually applying his methods. One of his co-authors on this subject is Xiaoqun Wang, with whom he published, e.g., the papers [81–83]. The first two of these papers, [81, 82], address an issue that is often not explicitly dealt with in theoretical studies on weighted spaces, namely: how should one, for a given problem from practice, choose the coordinate weights in order to obtain a low error of numerical integration?

The paper [81] first shows that it might be a very bad choice to "blindly" use lattice rules that have been optimized with respect to classical (i.e., non-weighted) error measures. Indeed, applying such rules to problems from financial mathematics can lead to an error behavior that is worse than applying Monte Carlo methods. However, Ian and Wang outline how to choose the weights for a given finance problem, such as pricing arithmetic average Asian options. Numerical evidence is given that, for pricing such derivatives, lattice rules with a suitable choice of weights (which have the form $\gamma_j = a\tau^j$ for some (positive real) parameters a and τ) may outperform crude Monte Carlo and also another quasi-Monte Carlo algorithm based on Sobol' sequences. The weights are chosen according to a matching strategy based on relative sensitivity indices, which are in turn related to the ANOVA decomposition, of the function at hand. Once the weights are fixed, the generating vectors of lattice rules are chosen according to one of the usual algorithms that optimize the integration rules with respect to a certain error criterion (as for example a CBC or acceptance-rejection algorithm). The choice of coordinate weights is also addressed in the paper [82], where not only product weights, but also general weights are allowed. A further focus of [82] is *dimension reduction* methods, such as principal component analysis (PCA) or the Brownian bridge. Indeed, it is shown in [82] for a model problem that dimension reduction techniques combined with a suitable choice of coordinate weights can lead to bounds on the integration error that are independent of the dimension. Ian and Wang refine their dimension reduction strategies even further in [83], where again theoretical and numerical results on option pricing are presented. We also refer to [15] for a survey on quasi-Monte Carlo methods applied to finance problems.

The idea of applying lattice rules to problems such as function approximation can already be found, e.g., in [30] and [24]. In the field of information-based complexity, dealing with high-dimensional problems in general and approximation problems in particular, several authors proposed using similar methods when studying function approximation in certain function spaces. This was for example done in the important paper [55] with Erich Novak and Henryk Woźniakowski, where (among other related questions) the approximation of functions in weighted Korobov spaces by means of lattice rules is studied. The underlying idea of the approximation algorithms analyzed in that paper is to approximate the largest Fourier coefficients of an element of a Korobov space by lattice rules. To be more precise, suppose that a one-periodic s-variate function f which is an element of a (weighted) Korobov space $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ and thus has an absolutely convergent Fourier series expansion,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{h}\in\mathbb{Z}^s} \widehat{f}(\boldsymbol{h}) \exp(2\pi\,\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{x}) \quad \text{for} \quad \boldsymbol{x} \in [0,1)^s,$$

is given. As before, $\widehat{f}(\boldsymbol{h})$ denotes the $\boldsymbol{h}^{\text{th}}$ Fourier coefficient of f. The strategy for function approximation using (rank-1) lattice rules is then to first choose a finite set $\mathscr{A} \subseteq \mathbb{Z}^s$ corresponding to the "largest" Fourier coefficients, and then to approximate the Fourier coefficients $\widehat{f}(\boldsymbol{h})$ of f for which $\boldsymbol{h} \in \mathscr{A}$. Hence, f is approximated by

$$\underbrace{\sum_{\boldsymbol{h}\in\mathscr{A}} \overbrace{\left(\frac{1}{N}\sum_{j=1}^{N} f\left(\left\{\frac{j-1}{N}\boldsymbol{g}\right\}\right) \exp\left(-2\pi\,\mathrm{i}\boldsymbol{h}\cdot\left\{\frac{j-1}{N}\boldsymbol{g}\right\}\right)\right)}^{\approx \widehat{f}(\boldsymbol{h})} \exp(2\pi\,\mathrm{i}\boldsymbol{h}\cdot\boldsymbol{x})}_{\approx f(\boldsymbol{x})}$$

for $\boldsymbol{x} \in I^s$, where $\boldsymbol{g}$ is a suitably chosen generating vector of a lattice rule. By noting that this approximation algorithm is a *linear* algorithm, it is then possible to analyze the approximation problem using the machinery from information-based complexity. Care has to be taken of the difficulty in simultaneously choosing the set $\mathscr{A}$ and its size, as well as the lattice rule and its number of points.

This idea was further pursued in the paper [35] with Kuo and Woźniakowski. In particular, the paper [35] contains a component-by-component construction of generating vectors of lattice rules, by which one can obtain tractability of approximation in the worst-case setting, under conditions on the coordinate weights that are similar to those for high-dimensional integration. To be more precise, one of the main results in [35] is the following.

Theorem 12 (Kuo, Sloan, and Woźniakowski) *Consider function approximation based on function evaluations for weighted Korobov spaces $\mathscr{H}_{s,\alpha,\boldsymbol{\gamma}}$ in the worst-case setting. Then the following statements hold true.*

(a) If the sequence of weights $(\gamma_j)_{j\geq 1}$ in the weighted Korobov space satisfies the condition

$$\sum_{j=1}^{\infty} \gamma_j < \infty,$$

the function approximation problem is strongly polynomially tractable. A generating vector of a lattice rule yielding this result can be found by a component-by-component algorithm.

(b) If the sequence of weights $(\gamma_j)_{j\geq 1}$ in the weighted Korobov space satisfies the condition

$$\limsup_{s\to\infty} \frac{\sum_{j=1}^{s} \gamma_j}{\log(s+1)} < \infty,$$

the function approximation problem is polynomially tractable. A generating vector of a lattice rule yielding this result can be found by a component-by-component algorithm.

While the paper [55] considers the worst-case, randomized, and the quantum settings, [35] focuses on the worst-case setting only. A further paper of Ian with Kuo and Woźniakowski deals with the average-case setting, see [36]. Overall, it can clearly be said that Ian's research in these papers has triggered numerous further results on lattice rules employed for function approximation in modern problem settings.

One of the most recent additions to the fields of application of quasi-Monte Carlo techniques is that of partial differential equations with random coefficients. Ian's paper [16], together with Graham, Kuo, Nuyens, and Scheichl, is one of the first papers addressing quasi-Monte Carlo methods used in combination with finite elements. In this context, randomly shifted lattice rules are (among other quasi-

Monte Carlo methods) used for computing the expectations of nonlinear functionals of solutions of certain elliptic PDEs with random coefficients. The observations in [16] are motivated by an example from physics, namely modeling fluid flow through porous media. After the paper [16], several others on this and related subjects followed, many of them co-authored by Ian (see, e.g., [17, 40, 41]). It is beyond the scope of this article to outline the details of the PDE problem these observations are based on, but we refer to the paper [39] of Ian with Frances Kuo and Christoph Schwab for a summary highlighting some particular issues from the quasi-Monte Carlo point of view. Without going into great detail, it should be mentioned in this context that these papers motivate some of the latest developments in the research on quasi-Monte Carlo methods in general and lattice rules in particular. Among them are increased efforts in deriving results that hold not only for the integration of Hilbert space elements, but more general function classes, integration problems with an unbounded number of variables, and the introduction of new classes of weights. Indeed, the paper [40] introduced new kinds of coordinate weights, so-called *product and order dependent (POD)* weights. These are of the form

$$\gamma_{\mathfrak{u}} := \Gamma_{|\mathfrak{u}|} \prod_{j \in \mathfrak{u}} \gamma_j,$$

for $\mathfrak{u} \subseteq \{1, \ldots, s\}$, for two sequences $(\Gamma_j)_{j \geq 1}$ and $(\gamma_j)_{j \geq 1}$. Note that in the definition of $\gamma_{\mathfrak{u}}$, the factor $\Gamma_{|\mathfrak{u}|}$ only depends on the cardinality of $\mathfrak{u}$, while the other factor has the usual product form. The introduction of this kind of weights in [40] stems from the problem of choosing the weights for a particular integrand in such a way that the overall integration error (or a suitable upper bound on it) is small. The structure of POD weights is such that it allows for a (fast) CBC construction. For further details we refer to the aforementioned papers.

Ian also worked on quasi-Monte Carlo methods applied to integrals appearing in certain problems in statistics. Indeed, the evaluation of the likelihood in generalized response models, as for example occurring in time series regression analysis, leads to multivariate integrals over unbounded regions. This problem in combination with quasi-Monte Carlo methods was first addressed in the paper with Kuo et al. [37], where, due to certain properties of the integrands under consideration, the integration problem over an unbounded region cannot simply be solved by transforming the integral into one over the unit cube. Indeed, the authors of [37] use recentering and rescaling methods before the transformation. The theory in [37] was extended in [62], by using POD weights as they were introduced in [40]. The idea of how to choose POD weights for integration is nicely summarized in [62, p. 633] as follows:

> The integration error is bounded by the product of the worst-case error (depending only on the QMC point set) and the norm of the function (depending only on the integrand). We choose weights that minimise a certain upper bound on this product.

Let us, finally, mention yet another field to which Ian applied the theory of lattice points, namely integral equations. Ian has made many important contributions to the research on integral equations, so for him it seems natural to combine it with the field of quasi-Monte Carlo methods. This idea was already dealt with in, for example, [24, 80], and [86]. The paper [8], however, considers the solution of certain Fredholm integral equations of the second kind in a weighted Korobov setting. Indeed, let us consider the Fredholm integral equation

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + \int_{I^s} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \mathrm{d}\boldsymbol{y},$$

where κ satisfies a convolution assumption, i.e., $\kappa(\boldsymbol{x},\boldsymbol{y}) = k(\boldsymbol{x}-\boldsymbol{y})$. Furthermore k and g are assumed to be elements of a weighted Korobov space, so they are one-periodic and continuous. The solution f is then approximated by a lattice-Nyström method based on lattice points. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N$ be the integration nodes of a lattice rule. Then f is approximated by

$$f_N(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{1}{N}\sum_{n=1}^{N} \kappa(\boldsymbol{x},\boldsymbol{y}_n) f_N(\boldsymbol{y}_n).$$

The function values $f_N(\boldsymbol{y}_1), \dots, f_N(\boldsymbol{y}_N)$ are obtained by solving the linear system $f_N(\boldsymbol{y}_k) = g(\boldsymbol{y}_k) + \frac{1}{N}\sum_{n=1}^{N} \kappa(\boldsymbol{y}_k, \boldsymbol{y}_n) f_N(\boldsymbol{y}_n)$, $1 \le k \le N$. Under suitable assumptions on k and N, existence and stability of a unique solution of the integral equation can be shown. The group structure of lattice points ensures that the linear system from above can be solved using the fast Fourier transform (cf. [86]). The paper [8] then studies the worst-case error of approximating f by f_N, and a CBC construction of lattice rules yielding a small error. Furthermore, conditions for obtaining tractability are discussed. The conditions on the weights in the Korobov space to obtain (strong) polynomial tractability are similar to those for numerical integration.

On a personal note, the paper [8] was partly written when the first author of the current paper visited Australia and Ian Sloan's group at the University of New South Wales for the first time—a trip which resulted in fruitful scientific collaborations and wonderful friendships with Ian and his colleagues.

Acknowledgements The authors would like to thank two anonymous referees for their helpful remarks.

P. Kritzer is supported by the Austrian Science Fund (FWF) Project F5506-N26 and F. Pillichshammer by the Austrian Science Fund (FWF) Project F5509-N26. Both projects are parts of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
2. Bahvalov, N.S.: Approximate computation of multiple integrals. Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. **4**, 3–18 (1959) (Russian)
3. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing embedded lattice rules for multivariate integration. SIAM J. Sci. Comput. **28**, 2162–2188 (2006)

4. Dick, J.: On the convergence rate of the component-by-component construction of good lattice rules. J. Complex. **20**, 493–522 (2004)
5. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
6. Dick, J., Kuo, F.Y., Pillichshammer, F., Sloan, I.H.: Construction algorithms for polynomial lattice rules for multivariate integration. Math. Comput. **74**, 1895–1921 (2005)
7. Dick, J., Sloan, I.H., Wang, X., Woźniakowski, H.: Good lattice rules in weighted Korobov spaces with general weights. Numer. Math. **103**, 63–97 (2006)
8. Dick, J., Kritzer, P., Kuo, F.Y., Sloan, I.H.: Lattice-Nyström method for Fredholm integral equations of the second kind with convolution type kernels. J. Complex. **23**, 752–772 (2007)
9. Dick, J., Pillichshammer, F., Waterhouse, B.J.: The construction of good extensible rank-1 lattices. Math. Comput. **77**, 2345–2373 (2008)
10. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
11. Dick, J., Nuyens, D., Pillichshammer, F.: Lattice rules for nonperiodic smooth integrands. Numer. Math. **126**, 259–291 (2014)
12. Disney, S., Sloan, I.H.: Error bounds for the method of good lattice points. Math. Comput. **56**, 257–266 (1991)
13. Disney, S., Sloan, I.H.: Lattice integration rules of maximal rank formed by copying rank 1 rules. SIAM J. Numer. Anal. **29**, 566–577 (1992)
14. Frolov, K.K.: On the connection between quadrature formulas and sublattices of the lattice of integral vectors. Sov. Math. Dokl. **18**, 37–41 (1977)
15. Giles, M.B., Kuo, F.Y., Sloan, I.H., Waterhouse, B.J.: Quasi-Monte Carlo for finance applications. ANZIAM J. **50**, C308–C323 (2008)
16. Graham, I.G., Kuo, F.Y., Nuyens, D., Scheichl, R., Sloan, I.H.: Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. J. Comput. Phys. **230**, 3668–3694 (2011)
17. Graham, I.G., Kuo, F.Y., Nichols, J.A., Scheichl, R., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. Numer. Math. **131**, 329–368 (2015)
18. Haber, S.: Experiments on optimal coefficients. In: Zaremba, S.K. (ed.) Applications of Number Theory to Numerical Analysis, pp. 11–37. Academic, New York (1972)
19. Haber, S.: Parameters for integrating periodic functions of several variables. Math. Comput. **41**, 115–129 (1983)
20. Hickernell, F.J.: Obtaining $O(n^{-2+\epsilon})$ convergence for lattice quadrature rules. In: Fang, K.T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 274–289. Springer, Berlin (2002)
21. Hickernell, F.J., Niederreiter, H.: The existence of good extensible rank-1 lattices. J. Complex. **19**, 286–300 (2003)
22. Hickernell, F.J., Hong, H.S., L'Ecuyer, P., Lemieux, C.: Extensible lattice sequences for quasi-Monte Carlo quadrature. SIAM J. Sci. Comput. **22**, 1117–1138 (2000)
23. Hlawka, E.: Zur angenäherten Berechnung mehrfacher Integrale. Monatsh. Math. **66**, 140–151 (1962)
24. Hua, L.K. Wang, Y.: Applications of Number Theory to Numerical Analysis. Springer, Berlin (1981)
25. Joe, S., Sloan, I.H.: Imbedded lattice rules for multidimensional integration. SIAM J. Numer. Anal. **29**, 1119–1135 (1992)
26. Joe, S., Sloan, I.H.: On computing the lattice rule criterion *R*. Math. Comput. **59**, 557–568 (1992)
27. Joe, S., Sloan, I.H.: Implementation of a lattice method for numerical multiple integration. ACM Trans. Math. Softw. **19**, 523–545 (1993)
28. Korobov, N.M.: The approximate computation of multiple integrals. Dokl. Akad. Nauk SSSR **124**, 1207–1210 (1959) (Russian)

29. Korobov, N.M.: Properties and calculation of optimal coefficients. Dokl. Akad. Nauk SSSR **132**, 1009–1012 (1960) (Russian)
30. Korobov, N.M.: Number-Theoretic Methods in Approximate Analysis. Fizmatgiz, Moscow (1963) (Russian)
31. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Wiley, New York (1974)
32. Kuo, F.Y.: Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. J. Complex. **19**, 301–320 (2003)
33. Kuo, F.Y., Joe, S.: Component-by-component construction of good lattice rules with a composite number of points. J. Complex. **18**, 943–976 (2002)
34. Kuo, F.Y., Joe, S.: Component-by-component construction of good intermediate-rank lattice rules. SIAM J. Numer. Anal. **41**, 1465–1486 (2003)
35. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Lattice rules for multivariate approximation in the worst case setting. In: Niederreiter, H., Talay, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 289–330. Springer, Berlin (2006)
36. Kuo, F.Y., Sloan, I.H., Woźniakowski, H.: Lattice rule algorithms for multivariate approximation in the average case setting. J. Complex. **24**, 283–323 (2008)
37. Kuo, F.Y., Dunsmuir, W.T.M., Sloan, I.H., Wand, M.P., Womersley, R.S.: Quasi-Monte Carlo for highly structured generalised response models. Methodol. Comput. Appl. Probab. **10**, 239–275 (2008)
38. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Waterhouse, B.J.: Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. J. Complex. **26**, 135–160 (2010)
39. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo methods for high-dimensional integration: the standard (weighted Hilbert space) setting and beyond. ANZIAM J. **53**, 1–37 (2011)
40. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
41. Kuo, F.Y., Schwab, C., Sloan, I.H.: Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. Found. Comput. Math. **15**, 411–449 (2015)
42. Larcher, G.: A best lower bound for good lattice points. Monatsh. Math. **104**, 45–51 (1987)
43. Leobacher, G., Pillichshammer, F.: Introduction to Quasi-Monte Carlo Integration and Applications. Birkhäuser/Springer, Cham (2014)
44. Lyness, J.N., Sloan I.H.: Some properties of rank-2 lattice rules. Math. Comput. **53**, 627–637 (1989)
45. Lyness, J.N., Sørevik, T.: A search program for finding optimal integration lattices. Computing **47**, 103–120 (1991)
46. Maisonneuve, D.: Recherche et utilisation des "bons trellis". Programmation et résultats numériques. In: Zaremba, S.K. (ed.) Applications of Number Theory to Numerical Analysis, pp. 121–201. Academic, New York (1972)
47. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
48. Niederreiter, H., Pillichshammer, F.: Construction algorithms for good extensible lattice rules. Constr. Approx. **30**, 361–393 (2009)
49. Niederreiter, H., Sloan, I.H.: Lattice rules for multiple integration and discrepancy. Math. Comput. **54**, 303–312 (1990)
50. Niederreiter, H., Sloan, I.H.: Integration of nonperiodic functions of two variables by Fibonacci lattice rules. J. Comput. Appl. Math. **51**, 57–70 (1994)
51. Niederreiter, H., Winterhof, A.: Applied Number Theory. Springer, Cham (2015)
52. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume I: Linear Information. European Mathematical Society, Zürich (2008)
53. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. European Mathematical Society, Zürich (2010)

54. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume III: Standard Information for Operators. European Mathematical Society, Zürich (2012)
55. Novak, E., Sloan, I.H., Woźniakowski, H.: Tractability of approximation for weighted Korobov spaces on classical and quantum computers. Found. Comput. Math. **4**, 121–156 (2004)
56. Novak, E., Sloan, I.H., Traub, J.F., Woźniakowski, H.: Essays on the Complexity of Continuous Problems. European Mathematical Society, Zürich (2009)
57. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comput. **75**, 903–920 (2006)
58. Nuyens, D., Cools, R.: Fast component-by-component construction, a reprise for different kernels. In: Niederreiter, H., Talay, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 373–387. Springer, Berlin (2006)
59. Pillichshammer, F.: A lower bound for rank 2 lattice rules. Math. Comput. **73**, 853–860 (2003)
60. Price, J.F., Sloan, I.H.: Pointwise convergence of multiple Fourier series: sufficient conditions and an application to numerical integration. J. Math. Anal. Appl. **169**, 140–156 (1992)
61. Saltykov, A.I.: Tables for computation of multiple integrals using the method of optimal coefficients. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki (1963). English translation: USSR Comput. Math. Math. Phys. **3**, 235–242 (1963)
62. Sinescu, V., Kuo, F.Y., Sloan, I.H.: On the choice of weights in a function space for quasi-Monte Carlo methods for a class of generalised response models in statistics. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, pp. 631–647. Springer, Berlin (2013)
63. Sloan, I.H.: Lattice methods for multiple integration. J. Comput. Appl. Math. **12/13**, 131–143 (1985)
64. Sloan, I.H.: Numerical integration in high dimensions—the lattice rule approach. In: Espelid, T.O., Genz, A.C. (eds.) Genz Numerical Integration: Recent Developments, Software and Applications, pp. 55–69. Kluwer Academic, Dordrecht (1992)
65. Sloan, I.H.: How high is high-dimensional? In: Novak, E., Sloan, I.H., Traub, J.F., Woźniakowski, H. (eds.) Essays on the Complexity of Continuous Problems, pp. 73–87. European Mathematical Society, Zürich (2009)
66. Sloan, I.H.: On the unreasonable effectiveness of QMC. In: Slides of a presentation given at the 9th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing in Warsaw in August 2010. Available at http://mcqmc.mimuw.edu.pl/Presentations/sloan.pdf
67. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Clarendon, Oxford (1994)
68. Sloan, I.H., Kachoyan, P.: Lattices for multiple integration. In: Mathematical Programming and Numerical Analysis Workshop 1983, pp. 147–165. Australian National University Press, Canberra (1984)
69. Sloan, I.H., Kachoyan, P.: Lattice methods for multiple integration: theory, error analysis and examples. SIAM J. Numer. Anal. **24**, 116–128 (1987)
70. Sloan, I.H., Lyness, J.N.: The representation of lattice quadrature rules as multiple sums. Math. Comput. **52**, 81–94 (1989)
71. Sloan, I.H., Lyness, J.N.: Lattice rules: projection regularity and unique representation. Math. Comput. **54**, 649–660 (1990)
72. Sloan, I.H., Reztsov, V.A.: Component-by-component construction of good lattice rules. Math. Comput. **71**, 263–273 (2002)
73. Sloan, I.H., Walsh, L.: Lattice rules—classification and searches. In: Braß, H., Hämmerlin, G. (eds.) Numerical Integration III, pp. 251–260. Birkhäuser, Basel (1988)
74. Sloan, I.H., Walsh, L.: A computer search of rank-2 lattice rules for multidimensional quadrature. Math. Comput. **54**, 281–302 (1990)
75. Sloan, I.H., Woźniakowski, H.: An intractability result for multiple integration. Math. Comput. **66**, 1111–1124 (1997)
76. Sloan, I.H., Woźniakowski, H.: When are quasi Monte Carlo algorithms efficient for high-dimensional problems? J. Complex. **14**, 1–33 (1998)

77. Sloan, I.H., Woźniakowski, H.: Tractability of multivariate integration for weighted Korobov classes. J. Complex. **17**, 697–721 (2001)
78. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. SIAM J. Numer. Anal. **40**, 1650–1665 (2002)
79. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comput. **71**, 1609–1640 (2002)
80. Tichy, R.F.: Über eine zahlentheoretische Methode zur numerischen Integration und zur Behandlung von Integralgleichungen. Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II **196**, 329–358 (1984)
81. Wang, X., Sloan, I.H.: Efficient weighted lattice rules with applications to finance. SIAM J. Sci. Comput. **28**, 728–750 (2006)
82. Wang, X., Sloan, I.H.: Brownian bridge and principal component analysis: towards removing the curse of dimensionality. IMA J. Numer. Anal. **27**, 631–654 (2007)
83. Wang, X., Sloan, I.H.: Quasi-Monte Carlo methods in financial engineering: an equivalence principle and dimension reduction. Oper. Res. **59**, 80–95 (2011)
84. Wang, X., Sloan, I.H., Dick, J.: On Korobov lattice rules in weighted spaces. SIAM J. Numer. Anal. **42**, 1760–1779 (2004)
85. Waterhouse, B.J., Kuo, F.Y., Sloan, I.H.: Randomly shifted lattice rules on the unit cube for unbounded integrands in high dimensions. J. Complex. **22**, 71–101 (2006)
86. Zinterhof, P.: Über die schnelle Lösung von hochdimensionalen Fredholm-Gleichungen vom Faltungstyp mit zahlentheoretischen Methoden. Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II **196**, 159–169 (1987)

Truncation Dimension for Function Approximation

Peter Kritzer, Friedrich Pillichshammer, and Grzegorz W. Wasilkowski

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We consider the approximation of functions of s variables, where s is very large or infinite, that belong to weighted anchored spaces. We study when such functions can be approximated by algorithms designed for functions with only very small number $\dim^{\rm trnc}(\varepsilon, s)$ of variables. Here ε is the error demand and we refer to $\dim^{\rm trnc}(\varepsilon, s)$ as the ε-*truncation dimension*. We show that for sufficiently fast decaying product weights and modest error demand (up to about $\varepsilon \approx 10^{-5}$) the truncation dimension is surprisingly very small.

1 Introduction

In this paper, we consider weighted anchored spaces of s-variate functions with bounded (in L_p norm, $1 \le p \le \infty$) mixed partial derivatives of order one. More precisely, the functions being approximated are from the Banach space $F_{s,p,\boldsymbol{\gamma}}$ whose

P. Kritzer (✉)
Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria
e-mail: peter.kritzer@oeaw.ac.at

F. Pillichshammer
Department of Financial Mathematics and Applied Number Theory, Johannes Kepler University Linz, Linz, Austria
e-mail: friedrich.pillichshammer@jku.at

G. W. Wasilkowski
Computer Science Department, University of Kentucky, Lexington, KY, USA
e-mail: greg@cs.uky.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_34

norm is given by

$$\|f\|_{F_{s,p,\gamma}} = \left(\sum_{\mathfrak{u}} \gamma_{\mathfrak{u}}^{-p} \int_{[0,1]^{|\mathfrak{u}|}} |f^{(\mathfrak{u})}([\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}])|^p \mathrm{d}\boldsymbol{x}_{\mathfrak{u}} \right)^{1/p} .$$

Here, the summation is with respect to the subsets $\mathfrak{u}$ of $[s] = \{1, \dots, s\}$ (when $s = \infty$ the summation is with respect to all finite subsets of $\mathbb{N}$), and $f^{(\mathfrak{u})}([\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}])$ denotes the mixed partial derivatives $\prod_{j \in \mathfrak{u}} \frac{\partial}{\partial x_j}$ of f with values of x_j for $j \notin \mathfrak{u}$ being zero. Note that $f^{(\mathfrak{u})}$ is a mixed partial distributional derivative, i.e., a function in L_p. In [4, Section 2] it is elaborated why the sections $f^{(\mathfrak{u})}([\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}])$ are meaningful (see also [2] for a discussion). A crucial role is played by the weights $\gamma_{\mathfrak{u}}$, which are non-negative real numbers that quantify the importance of sets $\boldsymbol{x}_{\mathfrak{u}} = (x_j)_{j \in \mathfrak{u}}$ of variables.

We continue our considerations from [6], where we dealt with low truncation dimension for numerical integration. We are interested in a very large number s of variables including $s = \infty$. Similar to [6], by *ε-truncation dimension* (or *truncation dimension* for short) we mean (roughly) the smallest number k such that the worst case error (measured in the L_q space) of approximating s-variate functions $f(x_1, \dots, x_s)$ by $f_k = f(x_1, \dots, x_k, 0, \dots, 0)$ is no greater than the error demand ε (see Definition 1 for more). We denote this minimal number k by $\dim^{\mathrm{trnc}}(\varepsilon, s)$.

Note that if the truncation dimension is small, say $\dim^{\mathrm{trnc}}(\varepsilon, s) = 3$, then the s-variate approximation problem can be replaced by the much easier $\dim^{\mathrm{trnc}}(\varepsilon, s)$-variate one, and any efficient algorithm for dealing with functions of only very few variables becomes also efficient for functions of s variables.

The main result of this paper is observing that the ε-truncation dimension is surprisingly small for modest error demand ε and the weights decaying sufficiently fast. For instance, for product weights

$$\gamma_{\mathfrak{u}} = \prod_{j \in \mathfrak{u}} j^{-a}$$

and the parameters $p = q = 2$, we have the following upper bounds $k(\varepsilon)$ for $\dim^{\mathrm{trnc}}(\varepsilon) := \dim^{\mathrm{trnc}}(\varepsilon, \infty)$ and $a = 3, 4, 5$:

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	2	5	12	31	79	198	$a = 3$
$k(\varepsilon)$	2	3	6	11	22	42	$a = 4$
$k(\varepsilon)$	1	2	4	6	11	18	$a = 5$

We stress that our definition of truncation dimension is different from the one proposed in the statistical literature. There the dimension depends on a particular function via its ANOVA decomposition which, in general, cannot be computed. Moreover, for functions from spaces with ANOVA decomposition, small truncation dimension can, according to the state of the art in the literature, only be utilized to some extent, see [1–5, 7]. In particular, there is an equivalence independent of s for product weights that are summable, with the equivalence constant bounded from above by

$$\sum_{\mathfrak{u}} \gamma_{\mathfrak{u}} < \infty.$$

For the product weights mentioned above we have, for $a > 1$,

$$\sum_{\mathfrak{u}} \gamma_{\mathfrak{u}} = \prod_{j=1}^{\infty} (1 + j^{-a}) < \infty$$

and hence the corresponding efficient algorithms for anchored spaces can also be used efficiently for ANOVA spaces.

The paper is structured as follows. In Sect. 2.1, we define the anchored spaces $F_{s,p,\boldsymbol{\gamma}}$, and in Sect. 2.2 we outline the problem setting. We give our results for anchored spaces in Sect. 3, and discuss examples of special kinds of weights in Sects. 3.1 and 3.2.

2 Basic Concepts

2.1 Anchored Spaces

In this section, we briefly recall definitions and basic properties of $\boldsymbol{\gamma}$-weighted anchored Sobolev spaces of s-variate functions. More detailed information can be found in [4, 5, 11].

Here we follow [11, Section 2]: For $p \in [1, \infty]$ let $F = W^1_{p,0}([0, 1])$ be the space of functions defined on $[0, 1]$ that vanish at zero, are absolutely continuous, and have bounded derivative in the L_p norm. We endow F with the norm $\|f\|_F = \|f'\|_{L_p}$ for $f \in F$.

For $s \in \mathbb{N}$ and

$$[s] := \{1, 2, \ldots, s\},$$

we will use $\mathfrak{u}, \mathfrak{v}$ to denote subsets of $[s]$, i.e.,

$$\mathfrak{u}, \mathfrak{v} \subseteq [s].$$

Moreover, for $\boldsymbol{x} = (x_1, x_2, \ldots, x_s) \in [0,1]^s$ and $\mathfrak{u} \subseteq [s]$, $[\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}]$ denotes the s-dimensional vector with all x_j for $j \notin \mathfrak{u}$ replaced by zero, i.e.,

$$[\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}] = (y_1, y_2, \ldots, y_s) \quad \text{with} \quad y_j = \begin{cases} x_j \text{ if } j \in \mathfrak{u}, \\ 0 \ \text{ if } j \notin \mathfrak{u}. \end{cases}$$

We also write $\boldsymbol{x}_{\mathfrak{u}}$ to denote the $|\mathfrak{u}|$-dimensional vector $(x_j)_{j\in\mathfrak{u}}$ and

$$f^{(\mathfrak{u})} = \frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{x}_{\mathfrak{u}}} = \prod_{j\in\mathfrak{u}} \frac{\partial}{\partial x_j} f \quad \text{with} \quad f^{(\emptyset)} = f.$$

For $s \in \mathbb{N}$ and nonempty $\mathfrak{u} \subseteq [s]$, let $F_{\mathfrak{u}}$ be the completion of the space spanned by $f(\boldsymbol{x}) = \prod_{j\in\mathfrak{u}} f_j(x_j)$ for $f_j \in F$ and $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1]^s$, with the norm

$$\|f\|_{F_{\mathfrak{u}}} = \|f^{(\mathfrak{u})}\|_{L_p}.$$

Note that $F_{\mathfrak{u}}$ is a space of functions with domain $[0,1]^s$ that depend only on the variables listed in $\mathfrak{u}$. Moreover, for any $f \in F_{\mathfrak{u}}$ and $\boldsymbol{x} = (x_1, \ldots, x_s)$, $f(\boldsymbol{x}) = 0$ if $x_j = 0$ for some $j \in \mathfrak{u}$. For $\mathfrak{u} = \emptyset$, let $F_{\mathfrak{u}}$ be the space of constant functions with the natural norm.

Consider next a sequence $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq[s]}$ of non-negative real numbers, called *weights*. Since some weights could be zero, we will use

$$\mathfrak{U} = \{\mathfrak{u} \subseteq [s] \ : \ \gamma_{\mathfrak{u}} > 0\}$$

to denote the collection of positive weights. For $p \in [1, \infty]$, we define the corresponding weighted *anchored* space

$$F_{s,p,\boldsymbol{\gamma}} = \operatorname{span}\left(\bigcup_{\mathfrak{u}\in\mathfrak{U}} F_{\mathfrak{u}}\right)$$

with the norm

$$\|f\|_{F_{s,p,\boldsymbol{\gamma}}} = \begin{cases} \left(\sum_{\mathfrak{u}\in\mathfrak{U}} \frac{1}{\gamma_{\mathfrak{u}}^p} \|f^{(\mathfrak{u})}([\cdot_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}])\|_{L_p}^p\right)^{1/p} & \text{if } p < \infty, \\ \max_{\mathfrak{u}\in\mathfrak{U}} \frac{1}{\gamma_{\mathfrak{u}}} \operatorname{ess\,sup}_{\boldsymbol{x}_{\mathfrak{u}}\in[0,1]^{|\mathfrak{u}|}} |f^{(\mathfrak{u})}([\boldsymbol{x}_{\mathfrak{u}}; \boldsymbol{0}_{-\mathfrak{u}}])| & \text{if } p = \infty. \end{cases}$$

Remark 1 Some of the results of this paper are extended to spaces of functions with countably many variables. In such cases, $[s] = \mathbb{N}$, the sets $\mathfrak{u}$ are finite subsets of $\mathbb{N}$, and $\boldsymbol{x} = (x_j)_{j\in\mathbb{N}}$ with $x_j \in [0,1]$. Moreover, the anchored space is the completion of $\operatorname{span}\left(\bigcup_{\mathfrak{u}} F_{\mathfrak{u}}\right)$ with respect to the norm given above.

An important class of weights is provided by product weights

$$\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j$$

for positive reals γ_j. When dealing with them, we will assume without any loss of generality that

$$\gamma_j \ge \gamma_{j+1} > 0 \quad \text{for all } j.$$

Note that for product weights we have $\mathfrak{U} = 2^{[s]} = \{\mathfrak{u} : \mathfrak{u} \subseteq [s]\}$.

For $p = 2$, $F_{s,2,\boldsymbol{\gamma}}$ is a reproducing kernel Hilbert space with kernel

$$K(\boldsymbol{x},\boldsymbol{y}) = \sum_{\mathfrak{u}\in\mathfrak{U}} \gamma_{\mathfrak{u}}^2 \prod_{j\in\mathfrak{u}} \min(x_j, y_j),$$

for $\boldsymbol{x} = (x_1, \ldots, x_s)$ and analogously for $\boldsymbol{y}$, which for product weights reduces to

$$K(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{s} \left(1 + \gamma_j^2 \min(x_j, y_j)\right).$$

2.2 *The Function Approximation Problem*

We follow [11]. Let $q \in [1, \infty]$. For $\mathfrak{u} \in \mathfrak{U}$, let $S_{\mathfrak{u}} : F_{\mathfrak{u}} \to L_q([0,1]^{|\mathfrak{u}|})$ be the embedding operator,

$$S_{\mathfrak{u}}(f_{\mathfrak{u}}) = f_{\mathfrak{u}} \quad \text{for all } f_{\mathfrak{u}} \in F_{\mathfrak{u}}.$$

It is well known that

$$\|S_{\mathfrak{u}}\| = \|S\|^{|\mathfrak{u}|} \tag{1}$$

for the space $F_{\mathfrak{u}}$, where $S : F \to L_q([0,1])$ is the univariate embedding operator.

Let further $\mathscr{L}_q$ be a normed linear space such that $F_{s,p,\boldsymbol{\gamma}}$ is its subspace and the norm $\|\cdot\|_{\mathscr{L}_q}$ is such that

$$\|f_{\mathfrak{u}}\|_{\mathscr{L}_q} = \|f_{\mathfrak{u}}\|_{L_q([0,1]^{|\mathfrak{u}|})} \quad \text{for all } f_{\mathfrak{u}} \in F_{\mathfrak{u}}.$$

Denote by S_s the embedding operator

$$S_s : F_{s,p,\boldsymbol{\gamma}} \to \mathscr{L}_q, \quad S_s(f) = f.$$

In order to make sure that S_s is continuous, we assume from now on that

$$\sum_{\mathfrak{u}\in\mathfrak{U}}\left(\left(\frac{q}{p^*}+1\right)^{-|\mathfrak{u}|/q}\gamma_{\mathfrak{u}}\right)^{p^*}<\infty,$$

where here and throughout this paper p^* denotes the conjugate of p, i.e., $1/p + 1/p^* = 1$. To see that this condition indeed ensures continuity of S_s, we recall the following proposition from [11].

Proposition 1 *We have*

$$\|S\| \le \left(\frac{q}{p^*}+1\right)^{-1/q} \quad \textit{and} \quad \|S_s\| \le \left[\sum_{\mathfrak{u}\in\mathfrak{U}}\left(\left(\frac{q}{p^*}+1\right)^{-|\mathfrak{u}|/q}\gamma_{\mathfrak{u}}\right)^{p^*}\right]^{1/p^*}.$$

Note that for product weights we have

$$\sum_{\mathfrak{u}\in\mathfrak{U}}\left(\left(\frac{q}{p^*}+1\right)^{-|\mathfrak{u}|/q}\gamma_{\mathfrak{u}}\right)^{p^*}=\prod_{j=1}^{s}\left(1+\gamma_j^{p^*}\left(\frac{q}{p^*}+1\right)^{-p^*/q}\right).$$

We are interested in algorithms for approximating$f \in F_{s,p,\boldsymbol{\gamma}}$ of the form

$$A_s(f) = \phi(L_1(f),\dots,L_n(f)),$$

where L_j's are linear or non-linear functionals

$$L_j : F_{s,p,\boldsymbol{\gamma}} \to \mathbb{R}$$

and $\phi : \mathbb{R}^n \to \mathscr{L}_q$. An important class of algorithms is provided by the class of linear algorithms that use *standard information*, which are of the form

$$A_s(f) = \sum_{j=1}^{n} f(\boldsymbol{x}_j)g_j$$

for $g_j \in F_{s,p,\boldsymbol{\gamma}}$. We study the worst case setting, where the error of an algorithm A_s is the operator norm of $S_s - A_s$, i.e.,

$$e(A_s;F_{s,p,\boldsymbol{\gamma}}) := \|S_s - A_s\| = \sup_{\|f\|_{F_{s,p,\boldsymbol{\gamma}}}\le 1}\|f - A_s(f)\|_{\mathscr{L}_q}.$$

Remark 2 The results of this paper can easily be extended to approximation of linear operators that are not necessarily the embeddings from $F_{\mathfrak{u}}$ into $L_q([0,1]^{|\mathfrak{u}|})$.

Indeed, one can consider linear operators $S_{\mathfrak{u}}$ from $F_{\mathfrak{u}}$ into normed spaces $G_{\mathfrak{u}}$ that satisfy the following conditions for every $\mathfrak{u} \in \mathfrak{U}$:

$$\|S_{\mathfrak{u}}\| \;\le\; \|S\|^{|\mathfrak{u}|},$$

$\mathscr{L}_q$ contains the spaces $G_{\mathfrak{u}}$ as subspaces, and

$$\|S_{\mathfrak{u}}(f_{\mathfrak{u}})\|_{\mathscr{L}_q} \;\le\; \|S_{\mathfrak{u}}(f_{\mathfrak{u}})\|_{G_{\mathfrak{u}}}.$$

3 Anchored Decomposition and Truncation Dimension

It is well known, see, e.g., [8], that any $f \in F_{s,p,\boldsymbol{\gamma}}$ has the unique *anchored decomposition*

$$f \;=\; \sum_{\mathfrak{u}\in\mathfrak{U}} f_{\mathfrak{u}}, \tag{2}$$

where $f_{\mathfrak{u}}$ is an element of $F_{\mathfrak{u}}$, depends only on x_j for $j \in \mathfrak{u}$, and

$$f_{\mathfrak{u}}(\boldsymbol{x}) \;=\; 0 \quad \text{if } x_j = 0 \ \text{ for some } \ j \in \mathfrak{u}. \tag{3}$$

For the empty set $\mathfrak{u}$, $f_\emptyset$ is a constant function. We stress that in general we do not know what the elements $f_{\mathfrak{u}}$ are and algorithms are only allowed to evaluate the original function f.

The anchored decomposition has the following important properties, see, e.g., [4]:

$$f^{(\mathfrak{u})}([\cdot_{\mathfrak{u}};\mathbf{0}_{-\mathfrak{u}}]) \;\equiv\; f_{\mathfrak{u}}^{(\mathfrak{u})}, \tag{4}$$

$$f_{\mathfrak{u}} \;\equiv\; 0 \quad \text{iff} \quad f^{(\mathfrak{u})}([\cdot_{\mathfrak{u}};\mathbf{0}_{-\mathfrak{u}}]) \;\equiv\; 0,$$

$$\|f\|_{F_{s,p,\boldsymbol{\gamma}}} \;=\; \left\|\sum_{\mathfrak{u}\in\mathfrak{U}} f_{\mathfrak{u}}\right\|_{F_{s,p,\boldsymbol{\gamma}}} \;=\; \left(\sum_{\mathfrak{u}\in\mathfrak{U}} \gamma_{\mathfrak{u}}^{-p}\, \|f_{\mathfrak{u}}^{(\mathfrak{u})}\|_{L_p}^p\right)^{1/p} \quad \text{for } p < \infty,$$

and

$$\|f\|_{F_{s,\infty,\boldsymbol{\gamma}}} \;=\; \max_{\mathfrak{u}\in\mathfrak{U}} \frac{\|f_{\mathfrak{u}}^{(\mathfrak{u})}\|_{L_\infty}}{\gamma_{\mathfrak{u}}} \quad \text{for } p = \infty.$$

For any $\mathfrak{u} \neq \emptyset$, there exists (unique in the L_p-sense) $g \in L_p([0,1]^{|\mathfrak{u}|})$ such that

$$f_{\mathfrak{u}}(\boldsymbol{x}) = \int_{[0,1]^{|\mathfrak{u}|}} g(\boldsymbol{t}) \prod_{j \in \mathfrak{u}} 1_{[0,x_j)}(t_j) \mathrm{d}\boldsymbol{t} \quad \text{and} \quad f_{\mathfrak{u}}^{(\mathfrak{u})} = g,$$

where $1_J(t)$ is the characteristic function of the set J, i.e., $1_J(t) = 1$ if $t \in J$ and 0 otherwise.

Moreover, for any $\mathfrak{u}$,

$$f([\cdot_{\mathfrak{u}}; \mathbf{0}_{-\mathfrak{u}}]) = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} f_{\mathfrak{v}}.$$

In particular, for $k < s$ we have

$$f([\boldsymbol{x}_{[k]}; \mathbf{0}_{-[k]}]) = f(x_1, \dots, x_k, 0, \dots, 0) = \sum_{\mathfrak{v} \subseteq [k]} f_{\mathfrak{v}}(\boldsymbol{x}) \tag{5}$$

which allows us to compute samples and approximate the *truncated* function

$$f_k(x_1, \dots, x_k) := \sum_{\mathfrak{v} \subseteq [k]} f_{\mathfrak{v}}(\boldsymbol{x}).$$

Moreover, $f_k \in F_{k,p,\boldsymbol{\gamma}} \subseteq F_{s,p,\boldsymbol{\gamma}}$ and

$$\|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} = \|f([\cdot_{[k]}; \mathbf{0}_{-[k]}])\|_{F_{s,p,\boldsymbol{\gamma}}} = \left\| \sum_{\mathfrak{u} \subseteq [k]} f_{\mathfrak{u}} \right\|_{F_{s,p,\boldsymbol{\gamma}}}.$$

This leads to the following concept.

Definition 1 For a given error demand $\varepsilon > 0$, by *ε-truncation dimension for the approximation problem* (or *truncation dimension* for short), denoted by $\dim^{\mathrm{trnc}}(\varepsilon, s)$, we mean the smallest integer k such that

$$\left\| \sum_{\mathfrak{u} \not\subseteq [k]} f_{\mathfrak{u}} \right\|_{\mathscr{L}_q} \leq \varepsilon \left\| \sum_{\mathfrak{u} \not\subseteq [k]} f_{\mathfrak{u}} \right\|_{F_{s,p,\boldsymbol{\gamma}}} \quad \text{for all } f = \sum_{\mathfrak{u} \in \mathfrak{U}} f_{\mathfrak{u}} \in F_{s,p,\boldsymbol{\gamma}}.$$

We also denote $\dim^{\mathrm{trnc}}(\varepsilon) := \dim^{\mathrm{trnc}}(\varepsilon, \infty)$.

A practical estimate for the truncation dimension is the following upper bound.

Theorem 1 *We have*

$$\dim^{\rm trnc}(\varepsilon,s)\ \le\ \min\left\{k\ :\ \left(\sum_{\mathfrak{u}\not\subseteq[k]}\frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*}\ \le\ \varepsilon\right\}\quad for\, p>1,$$

and

$$\dim^{\rm trnc}(\varepsilon,s)\ =\ \min\left\{k\ :\ \max_{\mathfrak{u}\not\subseteq[k]}\gamma_{\mathfrak{u}}\ \le\ \varepsilon\right\}\quad for\, p=1.$$

Here $\sum_{\mathfrak{u}\not\subseteq[k]}$ *means summation over all* $\mathfrak{u}\subseteq[s]$ *with* $\mathfrak{u}\not\subseteq[k]$ *and similarly for* $\max_{\mathfrak{u}\not\subseteq[k]}$.

The proof of the theorem is in the proof of the next, Theorem 2.

For a given $k<s$, let A_k be an algorithm for approximating functions from the space $F_{k,p,\boldsymbol{\gamma}}$. We use it to define the following approximation algorithms for the original space $F_{s,p,\boldsymbol{\gamma}}$,

$$A_{s,k}^{\rm trnc}(f)\ =\ A_k(f([\cdot_{[k]};\mathbf{0}_{-[k]}]))\ =\ A_k(f_k). \tag{6}$$

Since the functions f_k belong to $F_{s,p,\boldsymbol{\gamma}}$, the algorithms $A_{s,k}^{\rm trnc}$ are well defined.

We have the following result.

Theorem 2 *For every* $k<s$, *the worst case error of* $A_{s,k}^{\rm trnc}$ *is bounded by*

$$e(A_{s,k}^{\rm trnc};F_{s,p,\boldsymbol{\gamma}})\ \le\ \left([e(A_k;F_{k,p,\boldsymbol{\gamma}})]^{p^*}+\sum_{\mathfrak{u}\not\subseteq[k]}\frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*}\qquad for\, p>1$$

and by

$$e(A_{s,k}^{\rm trnc};F_{s,1,\boldsymbol{\gamma}})\ \le\ \max\left(e(A_k;F_{k,1,\boldsymbol{\gamma}})\,,\ \max_{\mathfrak{u}\not\subseteq[k]}\gamma_{\mathfrak{u}}\right)\quad for\, p=1.$$

Moreover, if $k\ge\dim^{\rm trnc}(\varepsilon,s)$ *then*

$$e(A_{s,k}^{\rm trnc};F_{s,p,\boldsymbol{\gamma}})\ \le\ \left([e(A_k;F_{k,p,\boldsymbol{\gamma}})]^{p^*}+\varepsilon^{p^*}\right)^{1/p^*}.$$

Proof We prove the theorem for $p>1$ only since the proof for $p=1$ is very similar. Let us first assume that $1<p<\infty$. For any $f\in F_{s,p,\boldsymbol{\gamma}}$ it holds that

$$\begin{aligned}\left\|S_s(f)-A_{s,k}^{\rm trnc}(f)\right\|_{\mathscr{L}_q} &= \|S_s(f)-A_k(f_k)\|_{\mathscr{L}_q}\\ &= \left\|S_k(f_k)-A_k(f_k)+\sum_{\mathfrak{u}\not\subseteq[k]}S_{\mathfrak{u}}(f_{\mathfrak{u}})\right\|_{\mathscr{L}_q}\end{aligned}$$

$$\begin{aligned}
&\leq e(A_k; F_{k,p,\boldsymbol{\gamma}})\, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \left\| \sum_{\mathrm{u} \not\subseteq [k]} S_{\mathrm{u}}(f_{\mathrm{u}}) \right\|_{\mathscr{L}_q} \qquad (7)\\
&\leq e(A_k; F_{k,p,\boldsymbol{\gamma}})\, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \sum_{\mathrm{u} \not\subseteq [k]} \|S_{\mathrm{u}}(f_{\mathrm{u}})\|_{\mathscr{L}_q}\\
&\leq e(A_k; F_{k,p,\boldsymbol{\gamma}})\, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \sum_{\mathrm{u} \not\subseteq [k]} \|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}\, \|S_{\mathrm{u}}\|\,.
\end{aligned}$$

We now have

$$\begin{aligned}
\sum_{\mathrm{u} \not\subseteq [k]} \|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}\, \|S_{\mathrm{u}}\| &= \sum_{\mathrm{u} \not\subseteq [k]} \gamma_{\mathrm{u}}^{-1}\, \|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}\, \gamma_{\mathrm{u}}\, \|S_{\mathrm{u}}\|\\
&\leq \left[\sum_{\mathrm{u} \not\subseteq [k]} \left(\frac{\|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}}{\gamma_{\mathrm{u}}} \right)^p \right]^{1/p} \left[\sum_{\mathrm{u} \not\subseteq [k]} (\gamma_{\mathrm{u}}\, \|S_{\mathrm{u}}\|)^{p^*} \right]^{1/p^*}\\
&\leq \left[\sum_{\mathrm{u} \not\subseteq [k]} \left(\frac{\|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}}{\gamma_{\mathrm{u}}} \right)^p \right]^{1/p} \left[\sum_{\mathrm{u} \not\subseteq [k]} \left(\gamma_{\mathrm{u}}\, \|S\|^{\mathrm{u}}\right)^{p^*} \right]^{1/p^*}\\
&\leq \left[\sum_{\mathrm{u} \not\subseteq [k]} \left(\frac{\|f_{\mathrm{u}}\|_{F_{\mathrm{u}}}}{\gamma_{\mathrm{u}}} \right)^p \right]^{1/p} \left[\sum_{\mathrm{u} \not\subseteq [k]} \left(\gamma_{\mathrm{u}} \left(\frac{q}{p^*} + 1 \right)^{-|\mathrm{u}|/q} \right)^{p^*} \right]^{1/p^*},
\end{aligned}$$

where we used (1) and Proposition 1 in the last two steps, respectively.

Hence, putting together, we get

$$\begin{aligned}
\left\| S_s(f) - A_{s,k}^{\mathrm{trnc}}(f) \right\|_{\mathscr{L}_q} &\leq e(A_k; F_{k,p,\boldsymbol{\gamma}}) \left(\sum_{\mathrm{u} \subseteq [k]} \gamma_{\mathrm{u}}^{-p}\, \|f_{\mathrm{u}}^{(\mathrm{u})}\|_{L_p}^p \right)^{1/p}\\
&\quad + \left(\sum_{\mathrm{u} \not\subseteq [k]} \frac{\gamma_{\mathrm{u}}^{p^*}}{(\frac{q}{p^*} + 1)^{p^*|\mathrm{u}|/q}} \right)^{1/p^*} \left(\sum_{\mathrm{u} \not\subseteq [k]} \gamma_{\mathrm{u}}^{-p}\, \|f_{\mathrm{u}}^{(\mathrm{u})}\|_{L_p}^p \right)^{1/p}.
\end{aligned}$$

Using the Hölder inequality once more, we obtain

$$\begin{aligned}
&\left\| S_s(f) - A_{s,k}^{\mathrm{trnc}}(f) \right\|_{\mathscr{L}_q}\\
&\leq \left(\sum_{\mathrm{u} \subseteq [k]} \gamma_{\mathrm{u}}^{-p}\, \|f_{\mathrm{u}}^{(\mathrm{u})}\|_{L_p}^p + \sum_{\mathrm{u} \not\subseteq [k]} \gamma_{\mathrm{u}}^{-p}\, \|f_{\mathrm{u}}^{(\mathrm{u})}\|_{L_p}^p \right)^{1/p}
\end{aligned}$$

$$\times \left([e(A_k; F_{k,p,\boldsymbol{\gamma}})]^{p^*} + \sum_{\mathfrak{u} \not\subseteq [k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} \right)^{1/p^*}$$

$$= \left(\sum_{\mathfrak{u} \subseteq [s]} \gamma_{\mathfrak{u}}^{-p} \, \|f_{\mathfrak{u}}^{(\mathfrak{u})}\|_{L_p}^p \right)^{1/p} \left([e(A_k; F_{k,p,\boldsymbol{\gamma}})]^{p^*} + \sum_{\mathfrak{u} \not\subseteq [k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} \right)^{1/p^*}$$

$$= \|f\|_{F_{s,p,\boldsymbol{\gamma}}} \left([e(A_k; F_{k,p,\boldsymbol{\gamma}})]^{p^*} + \sum_{\mathfrak{u} \not\subseteq [k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} \right)^{1/p^*} .$$

This shows the result in the first two points of Theorem 2 for $1 < p < \infty$. For $p = \infty$, the result is obtained by letting $p \to \infty$.

Regarding the proof of the last point in Theorem 2, suppose that $k \geq \dim^{\mathrm{trnc}}(\varepsilon, s)$. Then, starting from (7) we can deduce, in a similar way as above,

$$\begin{aligned} \left\| S_s(f) - A_{s,k}^{\mathrm{trnc}}(f) \right\|_{\mathscr{L}_q} &\leq e(A_k; F_{k,p,\boldsymbol{\gamma}}) \, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \left\| \sum_{\mathfrak{u} \not\subseteq [k]} S_{\mathfrak{u}}(f_{\mathfrak{u}}) \right\|_{\mathscr{L}_q} \\ &= e(A_k; F_{k,p,\boldsymbol{\gamma}}) \, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \left\| \sum_{\mathfrak{u} \not\subseteq [k]} f_{\mathfrak{u}} \right\|_{\mathscr{L}_q} \\ &\leq e(A_k; F_{k,p,\boldsymbol{\gamma}}) \, \|f_k\|_{F_{k,p,\boldsymbol{\gamma}}} + \varepsilon \left\| \sum_{\mathfrak{u} \not\subseteq [k]} f_{\mathfrak{u}} \right\|_{F_{s,p,\boldsymbol{\gamma}}} . \end{aligned}$$

Using the latter estimate, we can, in the same way as above deduce that indeed

$$e(A_{s,k}^{\mathrm{trnc}}; F_{s,p,\boldsymbol{\gamma}}) \;\leq\; \left([e(A_k; F_{k,p,\boldsymbol{\gamma}})]^{p^*} + \varepsilon^{p^*} \right)^{1/p^*} ,$$

as claimed.

Regarding the result in Theorem 1, we again show the result for $p > 1$. Suppose that $k \in \mathbb{N}$ is minimal such that

$$\left(\sum_{\mathfrak{u} \not\subseteq [k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*\,|\mathfrak{u}|/q}} \right)^{1/p^*} \;\leq\; \varepsilon.$$

Then we have, again starting from (7), and following the argument above,

$$\begin{aligned}\left\|\sum_{\mathfrak{u}\not\subseteq[k]} S_{\mathfrak{u}}(f_{\mathfrak{u}})\right\|_{\mathscr{L}_q} &\le \left(\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*} \left(\sum_{\mathfrak{u}\not\subseteq[k]} \gamma_{\mathfrak{u}}^{-p}\,\|f_{\mathfrak{u}}^{(\mathfrak{u})}\|_{L_p}^p\right)^{1/p}\\ &= \left(\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*} \left\|\sum_{\mathfrak{u}\not\subseteq[k]} f_{\mathfrak{u}}\right\|_{F_{s,p,\boldsymbol{\gamma}}}\\ &\le \varepsilon \left\|\sum_{\mathfrak{u}\not\subseteq[k]} f_{\mathfrak{u}}\right\|_{F_{s,p,\boldsymbol{\gamma}}}.\end{aligned}$$

Hence we see that $k \ge \dim^{\mathrm{trnc}}(\varepsilon, s)$. This is the result in Theorem 1. □

Remark 3 In view of the last point of Theorem 2, a strategy to obtain a small error of $A_{s,k}^{\mathrm{trnc}}$ in $F_{s,p,\boldsymbol{\gamma}}$ is to choose $k \ge \dim^{\mathrm{trnc}}(\varepsilon, s)$, and n (i.e., the number of functionals used by A_k) large enough to solve the k-variate approximation problem in $F_{k,p,\boldsymbol{\gamma}}$ by A_k within an error threshold of ε, which then yields an overall error of at most $2^{1/p^*}\varepsilon$. Alternatively, we can replace ε by $\varepsilon/(2^{1/p^*})$ to obtain an overall error bound of ε.

Remark 4 In view of the role of the term

$$\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}$$

in the above results, we shall refer to this term as the “truncation error” in the following.

We now apply this theorem to two important classes of weights: product weights and product order-dependent weights.

3.1 Product Weights

We assume in this section that the weights have the following *product* form

$$\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j \quad \text{for} \quad 1 \ge \gamma_j \ge \gamma_{j+1} > 0,$$

introduced in [10]. Here the empty product is considered to be 1, i.e., $\gamma_{\emptyset} = 1$. As already mentioned, for product weights we always have $\mathfrak{U} = 2^{[s]}$.

Proposition 2 *For product weights and $k < s$, the truncation error is bounded by*

$$\left(\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*} \le \prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{1/p^*} \times\left(1-\exp\left(\frac{-1}{(\frac{q}{p^*}+1)^{p^*/q}}\sum_{j=k+1}^{s}\gamma_j^{p^*}\right)\right)^{1/p^*}$$

for $p > 1$, and it is equal to

$$\max_{\mathfrak{u}\not\subseteq[k]} \gamma_{\mathfrak{u}} \quad \textit{for } p = 1.$$

Proof The proof for $p = 1$ is trivial. For $p > 1$, we have

$$\begin{aligned}\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} &= \sum_{\mathfrak{u}\subseteq[s]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} - \sum_{\mathfrak{u}\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\\ &= \prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right) - \prod_{j=1}^{k}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)\\ &= \prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)\left(1-\prod_{j=k+1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{-1}\right).\end{aligned}$$

We have

$$\begin{aligned}1-\prod_{j=k+1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{-1} &= 1-\exp\left(-\sum_{j=k+1}^{s}\log\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)\right)\\ &\le 1-\exp\left(-\sum_{j=k+1}^{s}\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right),\end{aligned}$$

where log denotes the natural logarithm, and the last inequality is due to $\log(1+x) \le x$ for all $x > -1$. This completes the proof. □

We have the following corollaries:

Corollary 1 *Consider product weights. Then* $\dim^{\mathrm{trnc}}(\varepsilon, s)$ *is bounded from above by*

$$\min\left\{k \,:\, 1-\exp\left(\frac{-1}{(\frac{q}{p^*}+1)^{p^*/q}}\sum_{j=k+1}^{s}\gamma_j^{p^*}\right) \le \frac{\varepsilon^{p^*}}{\prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)}\right\}$$

for $p > 1$, *and is equal to*

$$\min\left\{k \,:\, \max_{\mathfrak{u}\not\subseteq[s]}\gamma_{\mathfrak{u}} \le \varepsilon\right\}$$

for $p = 1$.

Corollary 2 *Consider product weights and* $k < s$.

Then the error $e(A_{s,k}^{\mathrm{trnc}}; F_{s,p,\boldsymbol{\gamma}})$ *is bounded from above by*

$$\left([e(A_k; F_{k,p,\boldsymbol{\gamma}})]^{p^*}\right.$$
$$\left.+\prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)\left(1-\exp\left(\frac{-1}{(\frac{q}{p^*}+1)^{p^*/q}}\sum_{j=k+1}^{s}\gamma_j^{p^*}\right)\right)\right)^{1/p^*}$$

for $p > 1$, *and by*

$$\max\left(e(A_k; F_{k,1,\boldsymbol{\gamma}}),\ \max_{\mathfrak{u}\not\subseteq[k]}\gamma_{\mathfrak{u}}\right)$$

for $p = 1$. *Note that since we assumed* $\gamma_j \le 1$ *for all* j *we have that* $\max_{\mathfrak{u}\not\subseteq[k]}\gamma_{\mathfrak{u}} = \gamma_{k+1}$.

Therefore, for the worst case error of $A_{s,k}^{\mathrm{trnc}}$ *not to exceed the error demand* $\varepsilon > 0$, *it is enough to choose* $k = k(\varepsilon)$ *so that*

$$1-\exp\left(\frac{-1}{(\frac{q}{p^*}+1)^{p^*/q}}\sum_{j=k+1}^{s}\gamma_j^{p^*}\right) \le \frac{1}{2}\varepsilon^{p^*}\prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{-1}, \tag{8}$$

(or $\gamma_{k+1} \le \varepsilon$ *for* $p = 1$*), and next to choose the number* $n = n(\varepsilon)$ *of functional evaluations* $L_j(f)$ *used by* A_k *so that*

$$e(A_k; F_{k,p,\boldsymbol{\gamma}}) \le \frac{\varepsilon}{2^{1/p^*}}.$$

Clearly the inequality (8) for $p > 1$ is equivalent to

$$\sum_{j=k+1}^{s} \gamma_j^{p^*} \le -\left(\frac{q}{p^*}+1\right)^{p^*/q} \times \log\left(1-\frac{1}{2}\varepsilon^{p^*}\prod_{j=1}^{s}\left(1+\frac{\gamma_j^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{-1}\right). \tag{9}$$

Example 1 Consider large s including $s=\infty$ and

$$\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} j^{-a} \quad \text{for } a > 1/p^*.$$

For $p=1$, we have

$$\dim^{\text{trnc}}(\varepsilon) = \left\lceil \varepsilon^{-1/a}-1\right\rceil.$$

In particular we have

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$\dim^{\text{trnc}}(\varepsilon)$	3	9	31	99	316	999	$a=2$
$\dim^{\text{trnc}}(\varepsilon)$	2	4	9	21	46	99	$a=3$
$\dim^{\text{trnc}}(\varepsilon)$	1	3	5	9	17	31	$a=4$
$\dim^{\text{trnc}}(\varepsilon)$	1	2	3	6	9	15	$a=5$

Consider next $p > 1$. Unlike in the case $p = 1$, we do not know the exact values of the truncation dimension. However, we have its upper bounds $k(\varepsilon)$ that are relatively small,

$$\dim^{\text{trnc}}(\varepsilon) \le k(\varepsilon).$$

We use the estimate

$$\frac{(k+1)^{-ap^*+1}}{ap^*-1} = \int_{k+1}^{\infty} x^{-ap^*}\,\mathrm{d}x < \sum_{j=k+1}^{\infty} j^{-ap^*} \le \int_{k+1/2}^{\infty} x^{-ap^*}\,\mathrm{d}x$$
$$= \frac{(k+1/2)^{-ap^*+1}}{ap^*-1}.$$

Note that the relative error when using the upper bound to approximate the sum satisfies

$$\frac{\left|\sum_{j=k+1}^{\infty} j^{-ap^*} - \int_{k+1/2}^{\infty} x^{-ap^*}\,\mathrm{d}x\right|}{\int_{k+1/2}^{\infty} x^{-ap^*}\,\mathrm{d}x} = \frac{ap^*-1}{2(k+1)} + O(k^{-2})$$

and is small for large k. To satisfy (9), it is enough to take $k = k(\varepsilon)$ given by

$$k = \left\lceil \left(\frac{-(\frac{q}{p^*} + 1)^{-p^*/q} \, (ap^* - 1)^{-1}}{\log\left(1 - \frac{\varepsilon^{p^*}}{2} \prod_{j=1}^{s} \left(1 + \frac{j^{-ap^*}}{(\frac{q}{p^*}+1)^{p^*/q}}\right)^{-1}\right)} \right)^{1/(ap^*-1)} - \frac{1}{2} \right\rceil .$$

For $p = p^* = 2$, which corresponds to the classical Hilbert space setting, we have

$$k(\varepsilon) = \left\lceil \left(\frac{-(\frac{q}{2} + 1)^{-2/q} \, (2\,a - 1)^{-1}}{\log\left(1 - \frac{\varepsilon^2}{2} \prod_{j=1}^{s} \left(1 + \frac{j^{-2a}}{(\frac{q}{2}+1)^{2/q}}\right)^{-1}\right)} \right)^{1/(2a-1)} - \frac{1}{2} \right\rceil . \qquad (10)$$

If also $q = 2$, then

$$k(\varepsilon) = \left\lceil \left(-2\,(2\,a - 1)\, \log\left(1 - \frac{\varepsilon^2}{2} \prod_{j=1}^{s} \left(1 + \frac{j^{-2a}}{2}\right)^{-1}\right)\right)^{-1/(2a-1)} - \frac{1}{2} \right\rceil .$$

Since s could be huge or $s = \infty$, in calculating the values of $k(\varepsilon)$, we slightly overestimated the product $\prod_{j=1}^{s} \left(1 + \frac{j^{-2a}}{2}\right)$ in the following way:

$$\begin{aligned}
\prod_{j=1}^{s} \left(1 + \frac{j^{-2a}}{2}\right) \le \prod_{j=1}^{\infty} \left(1 + \frac{j^{-2a}}{2}\right) &\le \prod_{j=1}^{1000} \left(1 + \frac{j^{-2a}}{2}\right) \exp\left(\sum_{j=1001}^{\infty} \frac{j^{-2a}}{2} \right) \\
&\le \prod_{j=1}^{1000} \left(1 + \frac{j^{-2a}}{2}\right) \exp\left(\frac{1}{2} \int_{1000.5}^{\infty} x^{-2a} \mathrm{d}x \right) \\
&= \prod_{j=1}^{1000} \left(1 + \frac{j^{-2a}}{2}\right) \exp\left(\frac{1}{2\,(2a-1)}\, 1000.5^{-2a+1} \right).
\end{aligned}$$

This gave us the following estimations for $\prod_{j=1}^{s} \left(1 + \frac{j^{-2a}}{2}\right)$ for $p = 2$:

1.56225 for $a = 2$, 1.51302 for $a = 3$, 1.50306 for $a = 4$, 1.50075 for $a = 5$.

Below we give values of $k(\varepsilon)$ for $a = 2, 3, 4, 5$ and $p = q = 2$ using the estimates above.

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	4	17	80	373	1733	8045	$a = 2$
$k(\varepsilon)$	2	5	12	31	79	198	$a = 3$
$k(\varepsilon)$	2	3	6	11	22	42	$a = 4$
$k(\varepsilon)$	1	2	4	6	11	18	$a = 5$

It is clear that $k(\varepsilon)$ decreases with increasing a. To check whether the estimates above are sharp, we also calculated $k(\varepsilon)$ for $s = 1{,}000{,}000$ directly by computing

$$\prod_{j=1}^{s}\left(1+\frac{j^{-2a}}{2}\right)-\prod_{j=1}^{k}\left(1+\frac{j^{-2a}}{2}\right)$$

and choosing the smallest k for which the difference above is not greater than $\varepsilon^2/2$. The values of $k(\varepsilon)$ obtained this way are exactly the same.

We now consider $p = \infty$ and $q = 2$. By computing

$$\prod_{j=1}^{s}\left(1+\frac{j^{-a}}{\sqrt{3}}\right)-\prod_{j=1}^{k}\left(1+\frac{j^{-a}}{\sqrt{3}}\right)$$

we obtained the following values of $k(\varepsilon)$ for $s = 1{,}000{,}000$, which is the smallest k for which the difference above is not greater than $\varepsilon/2$.

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	26	261	2603	25,433	206,959	722,968	$a = 2$
$k(\varepsilon)$	3	10	32	101	319	1010	$a = 3$
$k(\varepsilon)$	2	4	9	19	40	86	$a = 4$
$k(\varepsilon)$	1	3	5	8	15	26	$a = 5$

Obviously, the values of $k(\varepsilon)$ for $a = 2$ are too large to be of practical interest.

For some particular values of q, the norm of the embedding operator S is known, as for example for $q = 1$, in which case it equals $(1 + p^*)^{-1/p^*}$.

Let us now consider the case $p = p^* = 2$, $s = 1{,}000{,}000$, and $q = 1$. In this case we obtain from (10):

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	4	16	76	354	1643	7628	$a = 2$
$k(\varepsilon)$	2	5	12	30	76	192	$a = 3$
$k(\varepsilon)$	2	3	6	11	21	41	$a = 4$
$k(\varepsilon)$	1	2	4	6	10	17	$a = 5$

For comparison, we also consider the values of $k(\varepsilon)$, by using the precise formula for $\|S_{\mathfrak{u}}\| = \|S\|^{|\mathfrak{u}|} = (1+p^*)^{-|\mathfrak{u}|/p^*}$ in the proof of Theorem 2. This yields, instead of (10):

$$k(\varepsilon) = \left\lceil \left(-3\,(2\,a-1)\,\log\left(1 - \frac{\varepsilon^2}{2}\prod_{j=1}^{s}\left(1+\frac{j^{-2a}}{3}\right)^{-1}\right)\right)^{-1/(2a-1)} - \tfrac{1}{2} \right\rceil. \qquad (11)$$

Then we obtain from (11):

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	3	14	67	312	1449	6727	$a = 2$
$k(\varepsilon)$	2	4	11	28	71	178	$a = 3$
$k(\varepsilon)$	1	3	5	10	20	39	$a = 4$
$k(\varepsilon)$	1	2	4	6	10	17	$a = 5$

We see that the values of $k(\varepsilon)$ computed using the precise value of $\|S\|$ are lower than our general bounds, but not too much.

3.2 Product Order-Dependent Weights

We assume in this section that the weights have the following *product order-dependent (POD)* form,

$$\gamma_{\mathfrak{u}} = c_1\,(|\mathfrak{u}|!)^b \prod_{j\in\mathfrak{u}} \gamma_j,$$

introduced in [9]. Here c_1 is a positive constant. Since the truncation error for $p = 1$ is $\max_{\mathfrak{u}\not\subseteq[k]} \gamma_{\mathfrak{u}}$, we restrict the attention in this section to $p > 1$, i.e., $p^* < \infty$. We will use $[k+1:s]$ to denote

$$[k+1:s] = \{k+1, k+2, \dots, s\} \quad \text{or} \quad \{k+1, k+2, \dots\} \text{ if } s = \infty.$$

Proposition 3 *For POD weights and $k < s$, the truncation error is bounded by*

$$\left(\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}}\right)^{1/p^*} \le \left(\sum_{\mathfrak{v}\subseteq[k]} \frac{\gamma_{\mathfrak{v}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{v}|/q}}\right)^{1/p^*} T(k),$$

where

$$T(k) = \left(\sum_{l=1}^{s-k}\left(\frac{(l+k)!}{k!}\right)^{bp^*} \frac{1}{(\frac{q}{p^*}+1)^{p^*l/q}} \sum_{\substack{\mathfrak{w}\subseteq[k+1:s]\\|\mathfrak{w}|=l}} \prod_{j\in\mathfrak{w}} \gamma_j^{p^*}\right)^{1/p^*}.$$

Proof Of course we have

$$\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} = \sum_{\mathfrak{v}\subseteq[k]} \sum_{\emptyset\neq\mathfrak{w}\subseteq[k+1:s]} \frac{\gamma_{\mathfrak{v}\cup\mathfrak{w}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*(|\mathfrak{v}|+|\mathfrak{w}|)/q}}$$
$$= \sum_{\mathfrak{v}\subseteq[k]} \frac{\gamma_{\mathfrak{v}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{v}|/q}} T(\mathfrak{v},k)^{p^*},$$

where

$$T(\mathfrak{v},k)^{p^*} = \sum_{\emptyset\neq\mathfrak{w}\subseteq[k+1:s]} \left(\frac{(|\mathfrak{v}|+|\mathfrak{w}|)!}{|\mathfrak{v}|!}\right)^{bp^*} \frac{1}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{w}|/q}} \prod_{j\in\mathfrak{w}} \gamma_j^{p^*}$$
$$= \sum_{l=1}^{s-k} \left(\frac{(|\mathfrak{v}|+l)!}{|\mathfrak{v}|!}\right)^{bp^*} \frac{1}{(\frac{q}{p^*}+1)^{p^*l/q}} \sum_{\substack{\mathfrak{w}\subseteq[k+1:s]\\|\mathfrak{w}|=l}} \prod_{j\in\mathfrak{w}} \gamma_j^{p^*}.$$

Since $|\mathfrak{v}| \le k$, we have

$$\frac{(|\mathfrak{v}|+l)!}{|\mathfrak{v}|!} \le \frac{(k+l)!}{k!}.$$

This completes the proof. □

Example 2 Consider large s and

$$\gamma_j = \frac{c_2}{j^a} \quad \text{for } a > \max(1/p^*, b).$$

Clearly

$$\sum_{\substack{\mathfrak{w}\subseteq[k+1:s]\\|\mathfrak{w}|=l}} \prod_{j\in\mathfrak{w}} \gamma_j^{p^*} = c_2^{lp^*} \sum_{j_1=k+1}^{s} j_1^{-ap^*} \sum_{j_2=j_1+1}^{s} j_2^{-ap^*} \dots \sum_{j_l=j_{l-1}+1}^{s} j_l^{-ap^*}$$
$$\le c_2^{lp^*} \int_{k+1/2}^{\infty} x_1^{-ap^*} \int_{x_1}^{\infty} x_2^{-ap^*} \dots \int_{x_{l-1}}^{\infty} x_l^{-ap^*} \mathrm{d}x_l \dots \mathrm{d}x_2 \mathrm{d}x_1$$
$$= \left(\frac{c_2^{p^*}}{(k+1/2)^{ap^*-1}(ap^*-1)}\right)^l \frac{1}{l!}.$$

Therefore

$$\begin{aligned} T(k) &\le \left(\sum_{l=1}^{s-k}\left(\frac{(l+k)!}{k!}\right)^{bp^*}\ \frac{y^l}{l!}\right)^{1/p^*} \\ &= \left(\sum_{l=1}^{s-k}((k+1)\cdots(l+k))^{bp^*}\ \frac{y^l}{l!}\right)^{1/p^*} \end{aligned} \tag{12}$$

with

$$y = \frac{c_2^{p^*}}{(\frac{q}{p^*}+1)^{p^*/q}\,(ap^*-1)\,(k+1/2)^{ap^*-1}}.$$

Hence the upper bound in (12) can be computed efficiently using nested multiplication. We provide now the pseudo-code for doing that:

```
$y := c_2^{p^*}/((q/p^*+1)^{p^*/q}(ap^*-1)(k+1/2)^{ap^*-1})$
$T := y\,s^{bp^*}/(s-k)$
**for** $l = s-k-1$ **to** $1$ **step** $-1$**do**
    $T := (T+1)(l+k)^{bp^*}y/l$
**endfor**
$T := T^{1/p^*}.$
```

Furthermore, for $k \ge 2$,

$$\left(\sum_{\mathfrak{v}\subseteq[k]}\frac{\gamma_{\mathfrak{v}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{v}|/q}}\right)^{1/p^*}$$

$$\le c_1\left(1 + c_2^{p^*}\sum_{j=1}^{k}\frac{j^{-ap^*}}{(q/p^*+1)^{p^*/q}} + \sum_{\substack{\mathfrak{u}\subseteq[k]\\|\mathfrak{u}|\ge 2}}\frac{(|\mathfrak{u}|!)^{bp^*}\,c_2^{|\mathfrak{u}|p^*}}{(q/p^*+1)^{p^*|\mathfrak{u}|/q}}\prod_{j\in\mathfrak{u}}j^{-ap^*}\right)^{1/p^*}.$$

Now we provide an estimate for the last sum in this expression. We have

$$\begin{aligned} &\sum_{\substack{\mathfrak{u}\subseteq[k]\\|\mathfrak{u}|\ge 2}}\frac{(|\mathfrak{u}|!)^{bp^*}\,c_2^{|\mathfrak{u}|p^*}}{(q/p^*+1)^{p^*|\mathfrak{u}|/q}}\prod_{j\in\mathfrak{u}}j^{-ap^*} \\ &= \sum_{\ell=2}^{k}\frac{(\ell!)^{bp^*}\,c_2^{\ell p^*}}{(q/p^*+1)^{p^*\ell/q}}\sum_{\substack{\mathfrak{u}\subseteq[k]\\|\mathfrak{u}|=\ell}}\prod_{j\in\mathfrak{u}}j^{-ap^*} \end{aligned}$$

$$= \sum_{\ell=2}^{k} \frac{(\ell!)^{bp^*}\, c_2^{\ell p^*}}{(q/p^*+1)^{p^*\ell/q}} \sum_{\substack{\mathfrak{u}\subseteq[k]\\ |\mathfrak{u}|=\ell\\ \{1\}\subseteq\mathfrak{u}}} \prod_{j\in\mathfrak{u}} j^{-ap^*} + \sum_{\ell=2}^{k-1} \frac{(\ell!)^{bp^*}\, c_2^{\ell p^*}}{(q/p^*+1)^{p^*\ell/q}} \sum_{\substack{\mathfrak{u}\subseteq[2:k]\\ |\mathfrak{u}|=\ell}} \prod_{j\in\mathfrak{u}} j^{-ap^*}$$

$$= \sum_{\ell=2}^{k} \frac{(\ell!)^{bp^*}\, c_2^{\ell p^*}}{(q/p^*+1)^{p^*\ell/q}} \sum_{\substack{\mathfrak{u}\subseteq[2:k]\\ |\mathfrak{u}|=\ell-1}} \prod_{j\in\mathfrak{u}} j^{-ap^*} + \sum_{\ell=2}^{k-1} \frac{(\ell!)^{bp^*}\, c_2^{\ell p^*}}{(q/p^*+1)^{p^*\ell/q}} \sum_{\substack{\mathfrak{u}\subseteq[2:k]\\ |\mathfrak{u}|=\ell}} \prod_{j\in\mathfrak{u}} j^{-ap^*}.$$

For the two inner sums in the last expression we can now use the same method as in the derivation of (12) for the terms with indices $\ell = 2, \dots, k-1$, and hence obtain

$$\begin{aligned} &\sum_{\substack{\mathfrak{u}\subseteq[k]\\ |\mathfrak{u}|\geq 2}} \frac{(|\mathfrak{u}|!)^{bp^*}\, c_2^{|\mathfrak{u}|p^*}}{(q/p^*+1)^{p^*|\mathfrak{u}|/q}} \prod_{j\in\mathfrak{u}} j^{-ap^*} \\ &\leq \sum_{\ell=2}^{k-1} (\ell!)^{bp^*-1} \left(\frac{c_2^{p^*}}{(q/p^*+1)^{p^*/q}(ap^*-1)1.5^{ap^*-1}} \right)^{\ell} (\ell(ap^*-1)1.5^{ap^*-1}+1) \\ &\quad + \frac{(k!)^{bp^*-ap^*}\, c_2^{kp^*}}{(q/p^*+1)^{p^*k/q}}. \end{aligned} \tag{13}$$

In analogy to product weights and using the upper bounds above, we calculated numbers $k = k(\varepsilon)$ which guarantee that

$$\left(\sum_{\mathfrak{u}\not\subseteq[k]} \frac{\gamma_{\mathfrak{u}}^{p^*}}{(\frac{q}{p^*}+1)^{p^*|\mathfrak{u}|/q}} \right)^{1/p^*} \leq \frac{\varepsilon}{2^{1/p^*}}.$$

Since the upper bound (13) is not sharp for large k, i.e., small ε, we calculated the values of $k(\varepsilon)$ only for $a = 4$. More precisely we did it for $s = 10{,}000$, $b = c_1 = c_2 = 1$, $q = 2$, $p = 2$ and $p = \infty$, and $a = 4$.

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	
$k(\varepsilon)$	3	8	26	81	256	809	$p = \infty$
$k(\varepsilon)$	2	5	12	29	74	185	$p = 2$

Acknowledgements The authors would like to thank two anonymous referees for their remarks that helped improving the presentation of the results in the paper.

P. Kritzer is supported by the Austrian Science Fund (FWF) Project F5506-N26 and F. Pillichshammer by the Austrian Science Fund (FWF) Project F5509-N26. Both projects are parts of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications".

References

1. Gnewuch, M., Hefter, M., Hinrichs, A., Ritter, K.: Embeddings of weighted Hilbert spaces and applications to multivariate and infinite-dimensional integration. J. Approx. Theory **222**, 8–39 (2017)
2. Gnewuch, M., Hefter, M., Hinrichs, A., Ritter, K., Wasilkowski, G.W.: Equivalence of weighted anchored and ANOVA spaces of functions with mixed smoothness of order one in L_p. J. Complex. **40**, 78–99 (2017)
3. Hefter, M., Ritter, K.: On embeddings of weighted tensor product Hilbert spaces. J. Complex. **31**, 405–423 (2015)
4. Hefter, M., Ritter, K., Wasilkowski, G.W.: On equivalence of weighted anchored and ANOVA spaces of functions with mixed smoothness of order one in L_1 and L_∞ norms. J. Complex. **32**, 1–19 (2016)
5. Hinrichs, A., Schneider, J.: Equivalence of anchored and ANOVA spaces via interpolation. J. Complex. **33**, 190–198 (2016)
6. Kritzer, P., Pillichshammer, F., Wasilkowski, G.W.: Very low truncation dimension for high dimensional integration under modest error demand. J. Complex. **35**, 63–85 (2016)
7. Kritzer, P., Pillichshammer, F., Wasilkowski, G.W.: On equivalence of anchored and ANOVA spaces; lower bounds. J. Complex. **38**, 31–38 (2017)
8. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decompositions of multivariate functions. Math. Comput. **79**, 953–966 (2010)
9. Kuo, F.Y., Schwab, C., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **6**, 3351–3374 (2012)
10. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complex. **14**, 1–33 (1998)
11. Wasilkowski, G.W.: Tractability of approximation of ∞-variate functions with bounded mixed partial derivatives. J. Complex. **30**, 325–346 (2014)

On Nonnegativity Preservation in Finite Element Methods for the Heat Equation with Non-Dirichlet Boundary Conditions

Stig Larssonand Vidar Thomée

Dedicated to Ian Sloan on the occasion of his 80th birthday.

Abstract By the maximum principle the solution of the homogeneous heat equation with homogeneous Dirichlet boundary conditions is nonnegative for positive time if the initial values are nonnegative. In recent work it has been shown that this does not hold for the standard spatially discrete and fully discrete piecewise linear finite element methods. However, for the corresponding semidiscrete and Backward Euler Lumped Mass methods, nonnegativity of initial data is preserved, provided the underlying triangulation is of Delaunay type. In this paper, we study the corresponding problems where the homogeneous Dirichlet boundary conditions are replaced by Neumann and Robin boundary conditions, and show similar results, sometimes requiring more refined technical arguments.

1 Introduction

We shall first recall some known results concerning piecewise linear finite element methods for the model problem to find $u = u(x, t)$ for $x \in \Omega,\ t \geq 0$, satisfying the homogeneous heat equation with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} u_t = \Delta u \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \partial\Omega, \qquad \text{for } t \geq 0,\\ u(\cdot, 0) = v \quad \text{in } \Omega, \end{aligned}$$

where Ω is a polygonal domain in $\mathbb{R}^2$, $u_t = \partial u/\partial t$, and Δ the Laplacian. The initial values v are thus the only data of the problem, and its solution may be written

S. Larsson · V. Thomée (✉)
Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
e-mail: stig@chalmers.se; thomee@chalmers.se

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_35

$u(t) = \mathscr{E}(t)v$, for $t \geq 0$, where $\mathscr{E}(t)$ is the solution operator. By the maximum principle, $\mathscr{E}(t)$ is a nonnegative operator, so that

$$v \geq 0 \quad \text{in } \Omega \quad \text{implies} \quad \mathscr{E}(t)v \geq 0 \quad \text{in } \Omega, \quad \text{for } t \geq 0. \tag{1}$$

The basis for the finite element methods studied is the variational formulation of the problem, to find $u = u(\cdot,t) \in H_0^1 = H_0^1(\Omega)$ for $t \geq 0$, such that

$$(u_t, \varphi) + A(u, \varphi) = 0, \quad \forall \varphi \in H_0^1, \quad \text{for } t \geq 0, \quad \text{with } u(0) = v, \tag{2}$$

where $(v, w) = (v, w)_{L_2(\Omega)}$, $A(v, w) = (\nabla v, \nabla w)$. The methods are based on regular triangulations $\mathscr{T}_h = \{\tau\}$ of Ω, with $h = \max_{\tau \in \mathscr{T}_h} \text{diam}(\tau)$, using the finite element spaces

$$S_h = \{\chi \in \mathscr{C}(\bar{\Omega}) : \chi \text{ linear in each } \tau \in \mathscr{T}_h;\ \chi = 0 \text{ on } \partial\Omega\},$$

and the spatially semidiscrete Standard Galerkin method is then to determine $u_h = u_h(\cdot, t) \in S_h$ for $t \geq 0$ such that

$$(u_{h,t}, \chi) + A(u_h, \chi) = 0, \quad \forall \chi \in S_h, \quad \text{for } t \geq 0, \quad \text{with } u_h(0) = v_h, \tag{3}$$

where $v_h \in S_h$ is an approximation to v.

Let now $\{P_j\}_{j=1}^m$ be the interior nodes of $\mathscr{T}_h$, and $\{\Phi_j\}_{j=1}^m \subset S_h$ the corresponding nodal basis, thus with $\Phi_j(P_i) = \delta_{ij}$. We may then write $u_h(t) = \sum_{j=1}^m \alpha_j(t)\Phi_j$ and $v_h = \sum_{j=1}^m \gamma_j \Phi_j$. The problem (3) may then be expressed in matrix form, with $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, as

$$M\alpha' + S\alpha = 0, \quad \text{for } t \geq 0, \quad \text{with } \alpha(0) = \gamma, \tag{4}$$

where the mass matrix $M = (m_{ij})$, $m_{ij} = (\Phi_j, \Phi_i)$, and the stiffness matrix $S = (s_{ij})$, $s_{ij} = A(\Phi_j, \Phi_i)$, are both symmetric positive definite. The solution of (4) is then, with $E(t)$ the solution matrix,

$$\alpha(t) = E(t)\gamma, \quad \text{where } E(t) = \mathrm{e}^{-tH}, \quad H = M^{-1}S, \quad \text{for } t \geq 0. \tag{5}$$

We note that the semidiscrete solution $u_h(t)$ is ≥ 0 (> 0) if and only if $\alpha(t) \geq 0$ (> 0), and that this holds for all $\gamma \geq 0$ (> 0) if and only if $E(t) \geq 0$ (> 0). Here and below, for vectors and matrices, $\geq$ and $>$ denote elementwise inequalities.

It was proved in [10] that, for the semidiscrete Standard Galerkin method, the discrete analogue of (1) is not valid for all $t \geq 0$. However, in the case of the Lumped Mass method, to be discussed below, for which the mass matrix is diagonal, $E(t) \geq 0$ for all $t \geq 0$ if and only if the triangulation is of Delaunay type, so that the sum of the angles opposite any interior edge is $\leq \pi$; for triangulations with all angles $\leq \frac{1}{2}\pi$ this was shown already in [3]. In the case of the Standard Galerkin method, when the solution matrix is not nonnegative for all positive times, the possible nonnegativity of $E(t)$ for larger time was also discussed in [1, 9], with the smallest t_0 such that $E(t) \geq 0$ for $t \geq t_0 > 0$ referred to as the *threshold of nonnegativity*. If $\{\lambda_j\}_{j=1}^m$, with $\lambda_{j+1} \geq \lambda_j$, and $\{\varphi_j\}_{j=1}^m$, $\varphi_j = (\varphi_{j,1}, \dots, \varphi_{j,m})^T$, are the eigenvalues and eigenvectors of H, and if the principal eigenvector $\varphi_1 > 0$, then $E(t) > 0$ if

$$\sum_{j=2}^{m} \mathrm{e}^{-\lambda_j t} \sigma_j^2 < \mathrm{e}^{-\lambda_1 t}, \quad \text{where } \sigma_j = \max_l (|\varphi_{j,l}|/\varphi_{1,l}). \tag{6}$$

As a fully discrete method we consider the Backward Euler method, for which the approximate solution of (5) at $t_n = nk$ is, with k the time step,

$$\alpha^n = E_k^n \gamma, \quad \text{where } E_k = (I + kH)^{-1}. \tag{7}$$

It was shown in [5], under a weak assumption on $\mathscr{T}_h$, that E_k cannot be nonnegative for small $k > 0$. However, if $H^{-1} > 0$, then there exists $k_0 > 0$ such that $E_k > 0$ for $k \geq k_0$, and $H^{-1} \geq 0$ is a necessary condition for this conclusion. If $\mathscr{T}_h$ is strictly Delaunay and quasiuniform, then k_0 may be chosen as $k_0 = \lambda_0 h^2$ for some $\lambda_0 > 0$.

For the Lumped Mass method, the Backward Euler solution matrix E_k is nonnegative for all $k > 0$ if and only if the triangulation is of Delaunay type, and, if this is not the case, it is positive for large k if $S^{-1} > 0$.

Our purpose here is to investigate to what extent these known results for the case of Dirichlet boundary conditions carry over to other boundary conditions. More precisely, in Sects. 2 and 3, respectively, we study Neumann and Robin boundary conditions. In these cases the variational formulation (2) has to be modified which results in different matrix formulations (4). In Sect. 4 we investigate the application to the particular case of a standard uniform triangulation of the unit square, and in Sect. 5 to a uniform partition of a unit one-dimensional interval, in both cases with numerical illustrations.

2 Neumann Boundary Conditions

We now consider the Neumann problem

$$\begin{aligned} u_t &= \Delta u \quad \text{in } \Omega, \quad \text{with } \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega, \qquad \text{for } t \geq 0, \\ u(\cdot, 0) &= v \quad \text{in } \Omega, \end{aligned} \tag{8}$$

where n is the exterior unit normal to $\partial\Omega$. By the maximum principle, the solution operator $\mathscr{E}(t)$, with $\mathscr{E}(t)v = u(\cdot, t)$, is still a nonnegative operator, so that (1) holds. In fact, if u is negative at some point then by the maximum principle it must have a negative minimum on the parabolic boundary, and if $v \geq 0$, then this would be attained at a point (x_0, t_0) with $x_0 \in \partial\Omega$, $t_0 > 0$. But then, by the boundary point lemma, cf. [4, Chapter 3, Theorem 6], we have $\partial u(x_0, t_0)/\partial n < 0$, in conflict with the Neumann boundary condition. Since $\mathscr{E}(t)1 \equiv 1$, we also deduce that

$$\min_{x\in\bar{\Omega}} v(x) \le \mathscr{E}(t)v \le \max_{x\in\bar{\Omega}} v(x), \quad \text{for } t \ge 0.$$

The variational formulation of (8) is to find $u = u(\cdot, t) \in H^1 = H^1(\Omega)$ for $t \ge 0$, such that, again with $A(v, w) = (\nabla v, \nabla w)$,

$$(u_t, \varphi) + A(u, \varphi) = 0, \quad \forall \varphi \in H^1, \quad \text{for } t \ge 0, \quad \text{with } u(0) = v. \tag{9}$$

Note that the boundary conditions on $\partial\Omega$ are now natural boundary conditions, which are not strongly imposed in (9).

2.1 The Standard Galerkin Method

The finite element methods for (8), or (9), now use the piecewise linear finite element spaces $S_h = \{\chi \in \mathscr{C}(\bar{\Omega}) : \chi \text{ linear in each } \tau \in \mathscr{T}_h\}$, where the functions are not required to vanish on $\partial\Omega$. The spatially semidiscrete method uses the analogue of (9) restricted to S_h, i.e., to find $u_h = u_h(t) \in S_h$ for $t \ge 0$ such that

$$(u_{h,t}, \chi) + A(u_h, \chi) = 0, \ \ \forall \chi \in S_h, \ \ t \ge 0, \quad \text{with } u_h(0) = v_h, \tag{10}$$

where $v_h \in S_h$ is an approximation of v.

As in Sect. 1 we may formulate the semidiscrete equation in matrix form (4), with the difference that the $\{P_j\}_{j=1}^m$ are now all the nodes of $\mathscr{T}_h$, including those on the boundary $\partial\Omega$, and correspondingly for the nodal basis functions $\{\Phi_j\}_{j=1}^m \subset S_h$.

Again M is symmetric positive definite, but S is only symmetric positive semidefinite. The solution of (4) may be written, with $E(t)$ the solution matrix,

$$\alpha(t) = E(t)\gamma, \quad \text{where } E(t) = \mathrm{e}^{-tH}, \quad H = M^{-1}S, \quad \text{for } t \ge 0. \tag{11}$$

The values s_{ij} of the elements of S for P_i, P_j neighbors will be of particular interest below, and simple calculations show, see, e.g., [2],

$$s_{ij} = \begin{cases} -\dfrac{\sin(\alpha_{ij} + \beta_{ij})}{2\sin\alpha_{ij}\sin\beta_{ij}}, & \text{if } P_i, P_j \text{ neighbors, not both on } \partial\Omega, \\ -\frac{1}{2}\cot\gamma_{ij}, & \text{if } P_i, P_j \text{ neighbors, both on } \partial\Omega. \end{cases} \tag{12}$$

Here α_{ij} and β_{ij} are the angles opposite the interior edge P_iP_j, and γ_{ij} is the angle opposite the boundary edge P_iP_j. We note that, since $\sum_{j=1}^m \Phi_j = 1$, we have, with $\underline{1} = (1, \dots, 1)^T$,

$$(S\underline{1})_i = \sum_{j=1}^m s_{ij} = \sum_{j=1}^m (\nabla\Phi_i, \nabla\Phi_j) = (\nabla\Phi_i, \nabla 1) = 0, \quad i = 1, \dots, m, \tag{13}$$

and

$$M\underline{1} \cdot \underline{1} = \sum_{i,j=1}^m m_{ij} = \sum_{i,j=1}^m (\Phi_i, \Phi_j) = \Big\| \sum_{i=1}^m \Phi_i \Big\|^2 = \int_\Omega 1 \, \mathrm{d}x = |\Omega|. \tag{14}$$

Again we note that the semidiscrete solution $u_h(t)$ is ≥ 0 if and only if, elementwise, $\alpha(t) \geq 0$, and that this holds for all $\gamma \geq 0$ if and only if $E(t) \geq 0$ elementwise. As for Dirichlet boundary conditions we have the following negative result.

Theorem 1 *$E(t) = \mathrm{e}^{-tH}$, $H = M^{-1}S$, cannot be ≥ 0 for all $t > 0$.*

Proof We show that $E(t) \geq 0$ for all $t > 0$ leads to a contradiction. With $H = (h_{ij})$, we first note that $h_{ij} \leq 0$ for $j \neq i$, as follows from $E(t) = \mathrm{e}^{-tH} = I - tH + O(t^2) \geq 0$, as $t \to 0$. Let P_1 be a node, e.g., a boundary node, such that every neighbor of P_1 has a neighbor which is not a neighbor of P_1. Let P_i be any node $\neq P_1$ which is not a neighbor of P_1. Then since $m_{i1} = s_{i1} = 0$ and $MH = S$, we have $\sum_{j \neq 1} m_{ij} h_{j1} = 0$. But $m_{ij} \geq 0, h_{j1} \leq 0$, so that $m_{ij} h_{j1} \leq 0$, and thus $m_{ij} h_{j1} = 0$ for $j \neq 1$. Hence $h_{j1} = 0$ when $m_{ij} > 0$, i.e., when $j = i$, and also when P_j is a neighbor of P_1, since we can then choose P_i as a neighbor of P_j. Thus $h_{j1} = 0$ for all $j \neq 1$. Thus the first column of H only contains one possible nonzero element, namely h_{11}. But $h_{11} \neq 0$, since otherwise the first column of H and hence that of $S = MH$ would be zero, which it is not. But then the first columns of M and S would be proportional. However, $\sum_{j=1}^m s_{j1} = 0$ and $\sum_{j=1}^m m_{j1} > 0$, which gives a contradiction. □

Note that the proof depends on the fact that $m_{ij} > 0$, for P_i, P_j neighbors, which does not hold for the Lumped Mass method.

We remark that $H = M^{-1}S$ is symmetric positive semidefinite with respect to the inner product $M\alpha \cdot \beta$ and that H therefore has nonnegative eigenvalues $\{\lambda_j\}_{j=1}^m$ with $\lambda_j \leq \lambda_{j+1}$, and orthonormal eigenvectors $\{\varphi_j\}_{j=1}^m$. Here, since $MH\alpha \cdot \alpha = S\alpha \cdot \alpha = \|\nabla(\sum_j \alpha_j \Phi_j)\|^2$, and using also (14), we find that $\lambda_1 = 0$ is a simple eigenvalue, with positive principal eigenvector $\varphi_1 = \omega\underline{1}$, where $\omega = |\Omega|^{-1/2}$.

Any $\gamma \in \mathbb{R}^m$ has the eigenvector expansion

$$\gamma = \sum_{j=1}^m \widetilde{\gamma}_j \varphi_j, \quad \text{where } \widetilde{\gamma}_j = M\gamma \cdot \varphi_j. \tag{15}$$

We note that if $\gamma \geq 0,\ \gamma \neq 0$, we have $\widetilde{\gamma}_1 = M\gamma \cdot \varphi_1 > 0$, and, for $2 \leq j \leq m$,

$$\begin{gathered}|\widetilde{\gamma}_j| = |M\gamma \cdot \varphi_j| \leq (M\gamma \cdot \varphi_1)\sigma_j = \widetilde{\gamma}_1\sigma_j,\\ \text{with } \sigma_j = \max_l(|\varphi_{j,l}|/\varphi_{1,l}) = \omega \max_l |\varphi_{j,l}|.\end{gathered} \tag{16}$$

Hence $\gamma > 0$ if $\sum_{j=2}^m |\widetilde{\gamma}_j|\sigma_j < \widetilde{\gamma}_1$. In fact, we have, for $l = 1, \dots, m$,

$$\gamma_l = \sum_{j=1}^m \widetilde{\gamma}_j\varphi_{j,l} \geq \widetilde{\gamma}_1\varphi_{1,l} - \sum_{j=2}^m |\widetilde{\gamma}_j|\,|\varphi_{j,l}| \geq \varphi_{1,l}\Big(\widetilde{\gamma}_1 - \sum_{j=2}^m |\widetilde{\gamma}_j|\sigma_j\Big) > 0. \tag{17}$$

The following result shows that even though $E(t)$ is not ≥ 0 for all $t \geq 0$, we have $E(t) > 0$ for sufficiently large t.

Theorem 2 $E(t) > 0$ *if* $\sum_{j=2}^m \mathrm{e}^{-\lambda_j t}\sigma_j^2 < 1$.

Proof By (15) we have $E(t)\gamma = \sum_{j=1}^m \mathrm{e}^{-\lambda_j t}\widetilde{\gamma}_j\varphi_j$. We find, if $\gamma \geq 0$, $\gamma \neq 0$, by (16),

$$\sum_{j=2}^m \mathrm{e}^{-\lambda_j t}|\widetilde{\gamma}_j|\sigma_j \leq \widetilde{\gamma}_1 \sum_{j=2}^m e^{-\lambda_j t}\sigma_j^2 < \widetilde{\gamma}_1.$$

Hence, for $l = 1, \dots, m$,

$$(E(t)\gamma)_l = \sum_{j=1}^m \mathrm{e}^{-\lambda_j t}\widetilde{\gamma}_j\varphi_{j,l} \geq \varphi_{1,l}\Big(\widetilde{\gamma}_1 - \sum_{j=2}^m \mathrm{e}^{-\lambda_j t}|\widetilde{\gamma}_j|\sigma_j\Big) > 0.$$

Thus, $E(t)\gamma > 0$ for $\gamma \geq 0,\ \gamma \neq 0$, and hence $E(t) > 0$. □

We now show that, if $E(t) \geq 0$ for some $t \geq 0$, then, as in the continuous case, the maximum and minimum of $E(t)\gamma$ lie between the maximum and minimum at $t = 0$.

Theorem 3 *For* $t \geq 0$, *if* $E(t) \geq 0$, *then* $\min_j \gamma_j \underline{1} \leq E(t)\gamma \leq \max_j \gamma_j \underline{1}$ *for* $\gamma \in \mathbb{R}^m$.

Proof By (13) we have $S\underline{1} = 0$ and hence $H\underline{1} = 0$ and $E(t)\underline{1} = \mathrm{e}^{-tH}\underline{1} = \underline{1}$. With $E(t) = (e_{ij}(t))$, this may be expressed as $\sum_{j=1}^m e_{ij}(t) = 1$ for $i = 1, \dots, m$. Hence, if $E(t) \geq 0$, then $(E(t)\gamma)_i = \sum_{j=1}^m e_{ij}(t)\gamma_j$ is bounded above and below as stated. □

We turn to the Backward Euler method, with solution matrix $E_k = r(kH)$, where $r(\xi) = (1 + \xi)^{-1}$ and $H = M^{-1}S$. As for Dirichlet boundary conditions we have the following negative result.

Theorem 4 E_k *cannot be* ≥ 0 *for all* $k > 0$.

Proof If $E_k \geq 0$ for all k, we conclude that

$$E(t) = \mathrm{e}^{-tH} = \lim_{n\to\infty}\Big(I + \frac{t}{n}H\Big)^{-n} = \lim_{n\to\infty} E^n_{t/n} \geq 0, \quad \text{for any } t > 0,$$

in contradiction to Theorem 1. □

However, E_k is positive for k sufficiently large. In fact, using (15) we have, for $\gamma \geq 0,\ \gamma \neq 0$, since $\widetilde{\gamma}_1\varphi_1 > 0$,

$$E_k\gamma = \sum_{j=1}^{m} r(k\lambda_j)\widetilde{\gamma}_j\varphi_j = \widetilde{\gamma}_1\varphi_1(1 + O(k^{-1})) > 0, \quad \text{for } k \text{ large.}$$

Further, for any k, $E_k^n\gamma$ is positive for large n. More precisely, we have the following result which is shown similarly to Theorem 2.

Theorem 5 $E_k^n > 0$ *if* $\sum_{j=2}^m r(k\lambda_j)^n\sigma_j^2 < 1$.

This shows that $E_k > 0$ for k sufficient large, and also that, for any $k > 0$, $E_k^n > 0$ for n large enough. Recall that in the case of Dirichlet boundary conditions we had to impose the condition $H^{-1} > 0$, whereas for Neumann conditions H is singular.

Under a stronger condition, a more precise result holds:

Theorem 6 *If $s_{ij} < 0$ for all $(i,j) \in \mathscr{N} := \{(i,j) : P_i, P_j$ neighbors$\}$, then $E_k \geq 0$ for $k \geq \max_{\mathscr{N}}(m_{ij}/|s_{ij}|)$.*

Proof For the k stated we have $m_{ij} + ks_{ij} \leq 0$ for $i \neq j$, i.e., the positive definite matrix $M + kS$ is a Stieltjes matrix, and hence $(M + kS)^{-1} \geq 0$, cf. [11], Corollary 3.24. Thus, $E_k = (M + kS)^{-1}M \geq 0$. □

It follows, for example, that if $\{\mathscr{T}_h\}$ is a quasiuniform family, so that $|\tau| \geq ch^2$, with $c > 0$, for all $\tau \in \mathscr{T}_h$, and all angles of the $\mathscr{T}_h$ are uniformly $< \pi/2$, then $m_{ij} \leq Ch^2,\ |s_{ij}| \geq c > 0$, and thus $E_k \geq 0$ for $k \geq \lambda_0 h^2$ with $\lambda_0 = C/c > 0$.

If $E_k \geq 0$, the maximum and minimum of the time discrete solution $E_k^n\gamma$ are attained for $t_n = nk = 0$:

Theorem 7 *For any $k \geq 0$, if $E_k \geq 0$, then* $\min_j \gamma_j \underline{1} \leq E_k\gamma \leq \max_j \gamma_j \underline{1}$ *for* $\gamma \in \mathbb{R}^m$.

Proof Follows as in Theorem 3 from $E_k\underline{1} = \underline{1}$ for $k \geq 0$. □

For the Backward Euler matrix the nonnegativity threshold is the smallest $k_0 \geq 0$ such that $E_{k_0} \geq 0$.

Theorem 8 *If $E_{k_0} \geq 0$, then $E_k \geq 0$ for $k \geq k_0$.*

Proof This is equivalent to saying that if $(\epsilon_0 I + H)^{-1} \geq 0$, then $(\epsilon I + H)^{-1} \geq 0$ for $0 < \epsilon \leq \epsilon_0$. Clearly there is a smallest $\epsilon_1 \geq 0$ such that $(\epsilon I + H)^{-1} \geq 0$ for

$\epsilon_1 \le \epsilon \le \epsilon_0$, and we want to show that $\epsilon_1 = 0$. Assume $\epsilon_1 > 0$ and let $\epsilon = \epsilon_1 - \delta$ with δ small. We may write

$$(\epsilon I + H)^{-1} = (\epsilon_1 I + H - \delta I)^{-1} = (\epsilon_1 I + H)^{-1}(I - K)^{-1}, \quad K = \delta(\epsilon_1 I + H)^{-1}.$$

By assumption $K \ge 0$ and therefore, if δ is so small that, for some matrix norm, $\|K\| < 1$, then we have $(I - K)^{-1} = \sum_{j=0}^{\infty} K^j \ge 0$. Thus $(\epsilon I + H)^{-1} \ge 0$ for δ small, which contradicts the definition of ϵ_1. □

2.2 The Lumped Mass Method

The spatially semidiscrete Lumped Mass method for (8) is to find $u_h(t) \in S_h$, for $t \ge 0$, such that

$$(u_{h,t}, \chi)_h + A(u_h, \chi) = 0, \quad \forall \chi \in S_h, \quad \text{for } t \ge 0, \quad \text{with } u_h(0) = v_h. \tag{18}$$

Here the first term in (9) is evaluated by means of quadrature, using

$$(\psi, \chi)_h = \sum_{\tau \in \mathscr{T}_h} Q_{\tau,h}(\psi\chi), \quad Q_{\tau,h}(f) = \tfrac{1}{3}|\tau| \sum_{j=1}^{3} f(P_{\tau,j}) \approx \int_\tau f \, \mathrm{d}x, \tag{19}$$

where $\{P_{\tau,j}\}_{j=1}^3$ are the vertices of τ, and $|\tau| = \text{area}(\tau)$. In matrix form (18) may be written

$$D\alpha' + S\alpha = 0, \quad \text{for } t \ge 0, \quad \text{with } \alpha(0) = \gamma, \tag{20}$$

where the stiffness matrix is that for the Standard Galerkin method, but the mass matrix is now D whose elements are $d_{ij} = (\Phi_i, \Phi_j)_h$, which vanish for $i \ne j$, so that D is a diagonal matrix. The solution matrix of (20) is $E(t) = \mathrm{e}^{-tH}$, where this time $H = D^{-1}S$. Correspondingly, the solution matrix of the Backward Euler method is $E_k = (I + kH)^{-1}$. In this case we have the following.

Theorem 9 *If $E(t) \ge 0$ for small $t > 0$, or if $E_k \ge 0$ for small $k > 0$, then $s_{ij} \le 0$, $j \ne i$. On the other hand, if $s_{ij} \le 0$ for $j \ne i$, then $E(t) \ge 0$ for all $t \ge 0$, and $E_k \ge 0$ for all $k \ge 0$, and we then have for $\gamma \in \mathbb{R}^m$*

$$\min_j \gamma_j \underline{1} \le E(t)\gamma \le \max_j \gamma_j \underline{1}, \quad \textit{for } t \ge 0; \quad \min_j \gamma_j \underline{1} \le E_k\gamma \le \max_j \gamma_j \underline{1}, \quad \textit{for } k \ge 0.$$

Proof If $E(t) \ge 0$ for small $t > 0$, then $E(t) = I - tH + O(t^2)$ as $t \to 0$, and we then have $h_{ij} \le 0$ for $j \ne i$. The argument for E_k is the same. Hence, $s_{ij} = d_{ii}h_{ij} \le 0$ for $j \ne i$.

Conversely, if $s_{ij} \leq 0$ for $j \neq i$, then, for any $k > 0$, $D + kS$ is a Stieltjes matrix. Hence $(D + kS)^{-1} \geq 0$ and $E_k = (I + kH)^{-1} = (D + kS)^{-1}D \geq 0$. Also,

$$E(t) = \mathrm{e}^{-tH} = \lim_{n\to\infty} (I + \frac{t}{n}H)^{-n} \geq 0, \quad \text{for } t \geq 0.$$

The remaining inequalities are then shown as for the Standard Galerkin method. □

It follows from (12) that $s_{ij} \leq 0$ for $i \neq j$ is equivalent to $\alpha_{ij} + \beta_{ij} \leq \pi$ and $\gamma_{ij} \leq \frac{1}{2}\pi$ for all corresponding neighbors P_i, P_j, and $s_{ij} = 0$ for P_i, P_j not neighbors.

Even if the condition $s_{ij} \leq 0$ for $i \neq j$ does not hold, $E(t)$ is nonnegative for large t, which is shown in the same way as Theorem 2, noting that again $\lambda_1 = 0$ is a simple eigenvalue of H, with positive eigenvector $\varphi_1 = \omega \underline{1}$. Thus, we have:

Theorem 10 $E(t) = \mathrm{e}^{-tH} > 0$ *if* $\sum_{j=2}^{m} \mathrm{e}^{-\lambda_j t}\sigma_j^2 < 1$.

We also have the following analogue of Theorem 5, with the analogous proof.

Theorem 11 $E_k^n = r(kH)^n > 0$ *if* $\sum_{j=2}^{m} r(k\lambda_j)^n \sigma_j^2 < 1$.

Thus also in this case $E_k > 0$ for k sufficiently large, and, for any $k > 0$, $E_k^n > 0$ for n large enough, without special conditions on H.

3 Robin Boundary Conditions

We now consider the model problem with Robin boundary conditions on $\partial\Omega$,

$$\begin{aligned} u_t = \Delta u \quad \text{in } \Omega, \quad \text{with } \frac{\partial u}{\partial n} + \beta u = 0 \quad \text{on } \partial\Omega, \quad \text{for } t \geq 0, \\ u(\cdot, 0) = v \quad \text{in } \Omega, \end{aligned} \tag{21}$$

where the coefficient $\beta(x) > 0$ on $\partial\Omega$. By the maximum principle one shows again that $\mathscr{E}(t)$ is a nonnegative operator, so that (1) holds. In this case, at a negative minimum $u(x_0, t_0)$, with $x_0 \in \partial\Omega$, $t_0 > 0$, we would have $\partial u(x_0, t_0)/\partial n < 0$ in conflict with the Robin boundary condition. We also note for future reference that the principal eigenvalue λ_1 of $-\Delta$ with Robin boundary conditions is simple and positive with a principal eigenfunction $\phi_1 > 0$ in $\bar{\Omega}$, see [6, Theorem 11.10]. This theorem states only that $\phi_1 \geq 0$; the strict positivity then follows from the strong maximum principle and Hopf's boundary lemma [4, Chapter 2, Theorem 7].

The variational formulation of (21) is now to find $u = u(\cdot, t) \in H^1 = H^1(\Omega)$ for $t \geq 0$, such that

$$(u_t, \varphi) + B(u, \varphi) = 0, \quad \forall \varphi \in H^1 \text{ for } t \geq 0, \quad \text{with } u(0) = v, \tag{22}$$

where

$$B(v,w) = (\nabla v, \nabla w) + \langle v, w\rangle_\beta, \quad \text{with } \langle v, w\rangle_\beta = \int_{\partial\Omega} \beta v w \, ds. \tag{23}$$

Note that again the boundary conditions on $\partial\Omega$ are natural boundary conditions which are not strongly imposed.

3.1 The Standard Galerkin Method

The spatially semidiscrete method is now to find $u_h(t) \in S_h$ for $t \geq 0$, such that

$$(u_{h,t}, \chi) + B(u_h, \chi) = 0, \quad \forall \chi \in S_h, \ \text{for } t \geq 0, \quad \text{with } u_h(0) = v_h, \tag{24}$$

where $v_h \in S_h$ is an approximation of v, or, in matrix form (4), where now the mass matrix $M = (m_{ij})$, $m_{ij} = (\Phi_j, \Phi_i)$, and the stiffness matrix $S = (s_{ij})$, $s_{ij} = B(\Phi_j, \Phi_i)$, are both symmetric positive definite. This time we have, for P_i, P_j neighbors,

$$s_{ij} = \begin{cases} -\dfrac{1}{2}\dfrac{\sin(\alpha_{ij}+\beta_{ij})}{\sin\alpha_{ij}\sin\beta_{ij}}, & \text{if } P_i, P_j \text{ not both on } \partial\Omega, \\ -\frac{1}{2}\cot\gamma_{ij} + \Gamma_{ij}, & \text{if } P_i, P_j \ \text{ both on } \partial\Omega, \end{cases} \tag{25}$$

where α_{ij}, β_{ij}, and γ_{ij} are as in (12), and $\Gamma_{ij} = \langle \Phi_i, \Phi_j\rangle_\beta$. The solution of (4) again takes the form (5).

As in the case of Neumann boundary conditions we have the following negative result, which is proved in the same way as Theorem 1.

Theorem 12 *$E(t) = \mathrm{e}^{-tH}$, $H = M^{-1}S$, cannot be ≥ 0 for all $t > 0$.*

Again, the solution matrix could be nonnegative for larger t. The principal eigenvalue λ_1 is now positive, and we shall sometimes need to assume that H has a positive principal eigenvector φ_1. By the Perron–Frobenius theorem, this holds if H^{-1} is eventually positive, i.e., if $H^{-q} > 0$ for some $q > 0$. Another way to justify the assumption that $\varphi_1 > 0$ is to note that it is close to the corresponding continuous positive principal eigenfunction ϕ_1 in the maximum norm for small h. More precisely, the error is $O(h)$ in the energy norm, see [7, Theorem 6.2]. Together with the "almost Sobolev" inequality $\|v_h\|_{L_\infty} \leq C \max(1, \log(1/h))^{1/2}\|v_h\|_{H^1}$ for $v_h \in S_h$, valid under mild assumptions for finite elements in two dimensions, [8, Lemma 6.4], this leads to an $o(1)$ error bound in the maximum norm as $h \to 0$.

The following result, proved as Theorem 2, shows that $E(t) > 0$ for large t. Here and in the following, when $\varphi_1 > 0$, we define $\sigma_j = \max_l(|\varphi_{j,l}|/\varphi_{1,l})$.

Theorem 13 *Assume $\varphi_1 > 0$. Then $E(t) > 0$ if $\sum_{j=2}^m \mathrm{e}^{-\lambda_j t}\sigma_j^2 < \mathrm{e}^{-\lambda_1 t}$.*

We now have the following analogue of Theorem 3.

Theorem 14 *Assume $\varphi_1 > 0$. If $E(t) \geq 0$ for some $t \geq 0$, then*

$$e^{-\lambda_1 t} \min_j(\gamma_j/\varphi_{1,j})\varphi_1 \leq E(t)\gamma \leq e^{-\lambda_1 t} \max_j(\gamma_j/\varphi_{1,j})\varphi_1, \quad \textit{for } \gamma \in \mathbb{R}^m. \tag{26}$$

Proof We enclose the initial vector γ between the largest and smallest possible multiples of φ_1,

$$c_-\varphi_1 = \min_j(\gamma_j/\varphi_{1,j})\varphi_1 \leq \gamma \leq \max_j(\gamma_j/\varphi_{1,j})\varphi_1 = c_+\varphi_1.$$

Since $E(t)(c_+\varphi_1 - \gamma) = e^{-\lambda_1 t}c_+\varphi_1 - E(t)\gamma$, the right hand inequality of (26) follows. The left hand bound is shown analogously. □

As for Dirichlet and Neumann boundary conditions we have the following result for the Backward Euler method, with $E_k = (I + kH)^{-1}$.

Theorem 15 *E_k cannot be ≥ 0 for all $k > 0$.*

In analogy with the Dirichlet case, we have the following result, see Sect. 1 and [5].

Theorem 16 *If $H^{-1} > 0$, then there exists $k_0 > 0$ such that $E_k > 0$ if $k \geq k_0$, and $H^{-1} \geq 0$ is necessary for this to hold.*

Proof Since $E_k = \epsilon(\epsilon I + H)^{-1}$, $\epsilon = k^{-1}$, this follows at once by continuity. □

We also have the following analogue of Theorem 5.

Theorem 17 *Assume $\varphi_1 > 0$. Then $E_k^n > 0$ if $\sum_{j=2}^m (r(k\lambda_j/r(k\lambda_1))^n \sigma_j^2 < 1$.*

For any $k > 0$ this is satisfied for large n. For $n = 1$ a necessary condition for this to hold for large k is $\sum_{j=2}^m (\lambda_1/\lambda_j)\sigma_j^2 \leq 1$.

Finally, we state the following analogues of Theorems 6, 7, and 8; for the proof of Theorem 19 see Theorem 14.

Theorem 18 *If $s_{ij} < 0$ for all $(i,j) \in \mathscr{N} := \{(i,j) : P_i, P_j \text{ neighbors}\}$, then $E_k \geq 0$ for $k \geq \max_{\mathscr{N}}(m_{ij}/|s_{ij}|)$.*

Theorem 19 *Assume $\varphi_1 > 0$. If $E_k \geq 0$ for some $k > 0$, then we have*

$$(1 + k\lambda_1)^{-1} \min_j(\gamma_j/\varphi_{1,j})\varphi_1 \leq E_k\gamma \leq (1 + k\lambda_1)^{-1} \max_j(\gamma_j/\varphi_{1,j})\varphi_1, \quad \textit{for } \gamma \in \mathbb{R}^m.$$

Theorem 20 *If $E_{k_0} \geq 0$, then $E_k \geq 0$ for $k \geq k_0$.*

3.2 The Lumped Mass Method

The spatially semidiscrete Lumped Mass method for (21) is now

$$(u_{h,t}, \chi)_h + B(u_h, \chi) = 0, \quad \forall \chi \in S_h, \quad \text{with } u_h(0) = v_h,$$

where $(\cdot,\cdot)_h$ is defined in (19) and $B(\cdot,\cdot)$ in (23). In matrix form we again have (20) and $H = D^{-1}S$, where the stiffness matrix $S = (s_{ij})$ now has the elements $s_{ij} = B(\Phi_i, \Phi_j)$ in (25). With the same proof as in Theorem 9, we now have:

Theorem 21 *If $E(t) \geq 0$ for small $t > 0$, or if $E_k \geq 0$ for small $k > 0$, then $s_{ij} \leq 0$, $j \neq i$. On the other hand, if $s_{ij} \leq 0$ for $j \neq i$ then $E(t) \geq 0$ for all $t \geq 0$, and $E_k \geq 0$ for all $k \geq 0$. If $\varphi_1 > 0$, we then have for $\gamma \in \mathbb{R}^m$*

$$e^{-\lambda_1 t} \min_j(\gamma_j/\varphi_{1,j})\varphi_1 \leq E(t)\gamma \leq e^{-\lambda_1 t} \max_j(\gamma_j/\varphi_{1,j})\varphi_1, \quad \textit{for } t \geq 0,$$

$$(1 + k\lambda_1)^{-1} \min_j(\gamma_j/\varphi_{1,j})\varphi_1 \leq E_k\gamma \leq (1 + k\lambda_1)^{-1} \max_j(\gamma_j/\varphi_{1,j})\varphi_1, \quad \textit{for } k \geq 0.$$

The condition $s_{ij} \leq 0$ for all neighbors P_i, P_j is now, by (25), equivalent to $\alpha_{ij} + \beta_{ij} \leq \pi$ and $\Gamma_{ij} = \langle \Phi_i, \Phi_j \rangle_\beta \leq \frac{1}{2} \cot \gamma_{ij}$ for the corresponding i, j.

Again, even if the condition $s_{ij} \leq 0$ for $i \neq j$ does not hold, $E(t)$ and E_k may be nonnegative for large t and k, respectively, which is shown as earlier:

Theorem 22 *Assume $\varphi_1 > 0$. Then $E(t) > 0$ if $\sum_{j=2}^m e^{-\lambda_j t}\sigma_j^2 < e^{-\lambda_1 t}$.*

Theorem 23 *If $S^{-1} > 0$, then there exists $k_0 > 0$ such that $E_k > 0$ for $k \geq k_0$, and $S^{-1} \geq 0$ is a necessary condition for this to hold.*

Theorem 24 *Assume $\varphi_1 > 0$. Then $E_k^n > 0$ if $\sum_{j=2}^m (r(k\lambda_j/r(k\lambda_1))^n\sigma_j^2 < 1$.*

4 A Uniform Triangulation of the Unit Square

In this section we consider the above problems in the special case of the unit square $\Omega = (0,1)\times(0,1)$, with a standard uniform triangulation $\mathscr{T}_h$ as follows. Let m_0 be a positive integer, $h_0 = 1/m_0$, and set $x_j = y_j = jh_0$ for $j = 0, \ldots, m_0$. This partitions Ω into squares $[x_j, x_{j+1}]\times[y_l, y_{l+1}]$, and the triangulation is completed by connecting the nodes (x_j, y_{l+1}) and (x_{j+1}, y_l), see Fig. 1. The number of nodes $\{P_i\}_{i=1}^m$ is $m = (m_0+1)^2$, and $h = \sqrt{2}h_0$. We note that this is a Delaunay triangulation, but since the sum of the angles opposite a diagonal edge equals π, the corresponding elements s_{ij} of the stiffness matrix vanish. The case of Dirichlet boundary conditions was studied in [1], where, in particular, it was shown that for the Standard Galerkin method, $E(t) \geq 0$ for t large, and $E_k \geq 0$ for $k \geq 0.46\,h^2$. For the Lumped Mass method, since $\mathscr{T}_h$ is Delaunay, we have $E(t) \geq 0$ for $t \geq 0$ and $E_k \geq 0$ for $k \geq 0$. In this section we shall study the analogous problems with Neumann and Robin boundary conditions.

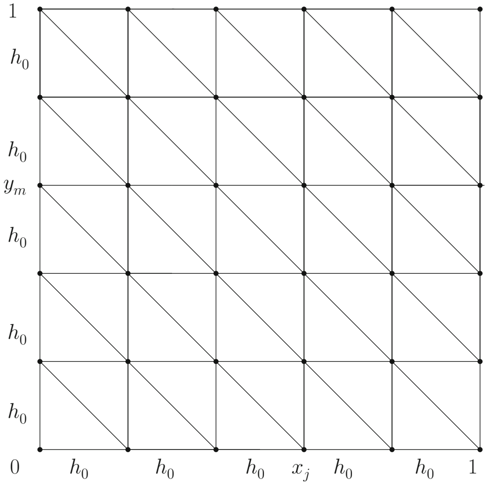


Fig. 1 Uniform triangulation of the unit square

4.1 Neumann Boundary Conditions

We shall say that $P_j \in \Omega$ if P_j is an interior mesh point, that $P_j \in \partial\Omega$ if P_j is a boundary mesh point but not a corner, and we let $Q = Q_1 \cup Q_2$ denote the corners with $Q_1 = \{(1,0),(0,1)\}$ and $Q_2 = \{(0,0),(1,1)\}$. For the stiffness and mass matrices we have for the diagonal elements corresponding to the node P_i:

$$s_{ii} = \|\nabla\Phi_i\|^2 = \begin{cases} 4, \\ 2, \\ 1, \\ 1, \end{cases} \qquad m_{ii} = \|\Phi_i\|^2 = \begin{cases} \frac{1}{4}h^2, & \text{if } P_i \in \Omega, \\ \frac{1}{8}h^2, & \text{if } P_i \in \partial\Omega, \\ \frac{1}{12}h^2, & \text{if } P_i \in Q_1, \\ \frac{1}{24}h^2, & \text{if] } P_i \in Q_2. \end{cases}$$

The off-diagonal elements are zero, except when P_i, P_j are neighbors, in which case P_iP_j is an edge of the triangulation. We distinguish between horizontal and vertical interior edges Z_1, boundary edges Z_2, and diagonal edges Z_3. We then have, cf. (12),

$$s_{ij} = (\nabla\Phi_i, \nabla\Phi_j) = \begin{cases} -1, \\ -\frac{1}{2}, \\ \ \ 0, \end{cases} \qquad m_{ij} = (\Phi_i, \Phi_j) = \begin{cases} \frac{1}{24}h^2, & \text{if } P_iP_j \in Z_1, \\ \frac{1}{48}h^2, & \text{if } P_iP_j \in Z_2, \\ \frac{1}{24}h^2, & \text{if } P_iP_j \in Z_3. \end{cases}$$

In particular, S is an irreducible Stieltjes matrix and thus $S^{-1} > 0$ and $H^{-1} = S^{-1}M > 0$, cf. Corollary 3.24 and Theorem 6.5 in [11]. Hence, the semidiscrete Standard Galerkin solution matrix is nonnegative for large t by Theorem 2 and both the semidiscrete and Backward Euler solution matrices for the Lumped Mass method are nonnegative for $t \geq 0$ and $k \geq 0$, respectively, by Theorem 9. We shall show that, even though the assumptions of Theorem 6 are not satisfied, the Standard Galerkin Backward Euler matrix $E_k = (M + kS)^{-1}M$ is ≥ 0 for $k \geq \lambda_0 h^2$ with $\lambda_0 > 0$ suitably chosen.

Let I_1, I_2, I_3, I_4 be diagonal matrices such that $I_{1,ii} = 1$, if $P_i \in \Omega$, $I_{2,ii} = 1$, if $P_i \in \partial\Omega$, $I_{3,ii} = 1$, if $P_i \in Q_1$, $I_{4,ii} = 1$, if $P_i \in Q_2$, with the remaining elements equal to 0. Let J_1, J_2, J_3 be matrices such that $J_{1,ij} = 1$, if $P_iP_j \in Z_1$, $J_{2,ij} = 1$, if $P_iP_j \in Z_2$, $J_{3,ij} = 1$, if $P_iP_j \in Z_3$, with the remaining elements equal to 0. Then we have

$$\begin{aligned} M &= \tfrac{1}{4}h^2\big(I_1 + \tfrac{1}{2}I_2 + \tfrac{1}{3}I_3 + \tfrac{1}{6}I_4\big) + \tfrac{1}{24}h^2\big(J_1 + \tfrac{1}{2}J_2 + J_3\big),\\ S &= 4I_1 + 2I_2 + I_3 + I_4 - J_1 - \tfrac{1}{2}J_2. \end{aligned}$$

Hence, with $\lambda = k/h^2$, $\Lambda = 4\lambda + \frac{1}{4}$, we find

$$\begin{aligned} M + kS &= (4\lambda + \tfrac{1}{4})h^2I_1 + \tfrac{1}{2}(4\lambda + \tfrac{1}{4})h^2I_2 + (\lambda + \tfrac{1}{12})h^2I_3 + (\lambda + \tfrac{1}{24})h^2I_4\\ &\quad - (\lambda - \tfrac{1}{24})h^2J_1 - \tfrac{1}{2}(\lambda - \tfrac{1}{24})h^2J_2 + \tfrac{1}{24}h^2J_3\\ &= \Lambda h^2\Big\{\Big(I_1 + \tfrac{1}{2}I_2 + (\lambda + \tfrac{1}{12})\Lambda^{-1}I_3 + (\lambda + \tfrac{1}{24})\Lambda^{-1}I_4\\ &\quad - (\lambda - \tfrac{1}{24})\Lambda^{-1}(J_1 + J_2) + \tfrac{1}{24}\Lambda^{-1}J_3\Big\}. \end{aligned}$$

Setting $\mu = (\lambda - \frac{1}{24})\Lambda^{-1}$, $\nu = \frac{1}{24}\Lambda^{-1}$, we thus have

$$M + kS = \Lambda h^2\big(D - \mu J + \nu J_3\big) = \Lambda h^2 D^{1/2}\Big(I - \mu\widetilde{J} + \nu\widetilde{J}_3\Big)D^{1/2},$$

where we have defined

$$\begin{aligned} D &= I_1 + \kappa_2 I_2 + \kappa_3 I_3 + \kappa_4 I_4,\ \text{with}\ \kappa_2 = \tfrac{1}{2},\ \kappa_3 = (\lambda + \tfrac{1}{12})/\Lambda,\ \kappa_4 = (\lambda + \tfrac{1}{24})/\Lambda,\\ J &= J_1 + \tfrac{1}{2}J_2, \quad \widetilde{J} = D^{-1/2}JD^{-1/2}, \quad \widetilde{J}_3 = D^{-1/2}J_3D^{-1/2}. \end{aligned}$$

Our aim is now to show that $\big(I - \mu\widetilde{J} + \nu\widetilde{J}_3\big)^{-1} \geq 0$. We write

$$I - \mu\widetilde{J} + \nu\widetilde{J}_3 = I - \mu\widetilde{J} + \tfrac{1}{2}\nu\widetilde{J}^2 - \tfrac{1}{2}\nu\big(\widetilde{J}^2 - 2\widetilde{J}_3\big) = L - N, \tag{27}$$

where $L = I - \mu\widetilde{J} + \frac{1}{2}\nu\widetilde{J}^2$, $N = \frac{1}{2}\nu\big(\widetilde{J}^2 - 2\widetilde{J}_3\big)$. We will prove that $N \geq 0$ and determine λ so that $L^{-1} \geq 0$ and $\|L^{-1}N\| < 1$. In view of (27) this shows

$$\big(I - \mu\widetilde{J} + \nu\widetilde{J}_3\big)^{-1} = \big(L - N\big)^{-1} = \big(I - L^{-1}N\big)^{-1}L^{-1} = \Big(\sum_{n=0}^{\infty}\big(L^{-1}N\big)^n\Big)L^{-1},$$

and thus that $E_k = (M + kS)^{-1}M \geq 0$ for λ chosen as indicated.

That $L^{-1} \geq 0$ for λ suitably chosen will follow from the following lemma.

Lemma 1 *Let* $\mu > 0$ *and* $0 < 4\omega \leq \mu^2$*. Then the zeros* $x_{1,2}$ *of* $P(x) = 1 - \mu x + \omega x^2$ *satisfy* $0 < x_1 \leq x_2$*, and we have, with* $\gamma_j > 0$*,*

$$\frac{1}{P(x)} = \sum_{j=0}^{\infty} \gamma_j x^j, \quad \textit{for}\; 0 \leq x < x_1 := \frac{\mu}{2\omega}\Big(1 - \sqrt{1 - 4\omega/\mu^2}\Big) > \frac{1}{\mu}.$$

If A *is a nonnegative symmetric matrix with* $\|A\| < x_1$*, for some matrix norm, then* $P(A)^{-1} \geq 0$ *and* $\|\, P(A)^{-1}\| \leq P(\|A\|)^{-1}$*.*

Proof The zeros of $P(x)$ are $x_{1,2} = (\mu/(2\omega))\big(1 \pm \sqrt{1 - 4\omega/\mu^2}\big)$, which shows the first statement. With $\tau = \mu^2/(4\omega)$ we have $x_1 = 2\mu^{-1}\tau\big(1 - \sqrt{1 - 1/\tau}\big) = \mu^{-1}2/\big(1 + \sqrt{1 - 1/\tau}\big) > \mu^{-1}$. Further, if $0 < 4\omega \leq \mu^2$, then by the formula for the product of power series, we have, for $0 \leq x < x_1$,

$$\begin{aligned}
\frac{1}{P(x)} &= \frac{1}{\omega}\frac{1}{(x - x_1)(x - x_2)} = \frac{1}{\omega}\frac{1}{x_1 x_2}\sum_{j=0}^{\infty}\Big(\frac{x}{x_1}\Big)^j \sum_{j=0}^{\infty}\Big(\frac{x}{x_2}\Big)^j \\
&= \frac{1}{\omega}\frac{1}{x_1 x_2}\sum_{j=0}^{\infty}\Big(\sum_{l=0}^{j}\frac{1}{x_1^l x_2^{j-l}}\Big)x^j = \sum_{j=0}^{\infty}\gamma_j x^j.
\end{aligned}$$

The statements about $P(A)$ are now obvious. □

We now note that $L = P(\widetilde{J})$ and apply Lemma 1 with $A = \widetilde{J}$. We set $\omega = \nu/2$ and choose $\lambda > 0$ so that $\mu^2/4\omega = 1$, i.e., the positive root of $(\lambda - \frac{1}{24})^2 - \frac{1}{12}(4\lambda + \frac{1}{24}) = 0$, which is $\lambda = \frac{11}{24}$. With this value of λ we have $\nu = \frac{1}{50}$, $\mu = \frac{1}{5}$, $\Lambda = \frac{25}{12}$, $\kappa_3 = \frac{13}{50}$, $\kappa_4 = \frac{6}{25}$, and $x_1 = \frac{\mu}{\nu} = 10$. Thus, if $\|\widetilde{J}\| := \max_{1 \leq i \leq m}\sum_{j=1}^{m}|\widetilde{J}_{ij}| < 10$, then we may conclude from Lemma 1 that $L^{-1} = P(\widetilde{J})^{-1} \geq 0$.

To bound $\|\widetilde{J}\|$ we note that for non-diagonal neighbors P_i, P_j and with $D = \mathrm{diag}(d_j)$, we have $\widetilde{J}_{ij} = d_i^{-1/2} J_{ij} d_j^{-1/2}$ and considering the various possibilities for P_i, P_j we find for the row sums in $\widetilde{J}$

$$\sum_{j=1}^{m}|\widetilde{J}_{ij}| = \begin{cases}
4\cdot 1\cdot 1\cdot 1 \text{ or } 2(1\cdot 1\cdot 1 + 1\cdot 1\cdot\sqrt{2}) \text{ or } 3\cdot 1\cdot 1\cdot 1 + \sqrt{2}, & \text{if } P_i \in \varOmega,\\
\sqrt{2}\cdot 1\cdot 1 + 2\cdot\sqrt{2}\cdot\frac{1}{2}\cdot\sqrt{2} \text{ or} & \\
\quad \sqrt{2}\cdot 1\cdot 1 + \sqrt{2}\cdot\frac{1}{2}\cdot\sqrt{2} + \sqrt{2}\cdot\frac{1}{2}\cdot\kappa_l^{-\frac{1}{2}},\; l = 3 \text{ or } 4, & \text{if } P_i \in \partial\varOmega,\\
2\cdot\sqrt{2}\cdot\frac{1}{2}\cdot\kappa_l^{-\frac{1}{2}},\; l = 3 \text{ or } 4, & \text{if } P_i \in Q,
\end{cases} \tag{28}$$

and we conclude that $\|\widetilde{J}\| = 2(1 + \sqrt{2}) \approx 4.83 < 10$.

Thus $L^{-1} = P(\widetilde{J})^{-1} \geq 0$. Since it also follows that $\|L^{-1}\| \leq P(\|\widetilde{J}\|)^{-1}$ and that $\|N\| = \omega\|\widetilde{J}^2 - 2\widetilde{J}_3\| \leq \omega\|\widetilde{J}\|^2$, and since $\mu\|\widetilde{J}\| \approx 0.2 \cdot 4.83 < 1$, we find

$$\|L^{-1}N\| \leq \frac{\|N\|}{P(\|\widetilde{J}\|)} \leq \frac{\omega\|\widetilde{J}\|^2}{1 - \mu\|\widetilde{J}\| + \omega\|\widetilde{J}\|^2} < 1. \tag{29}$$

It remains to show that $N = \widetilde{J}^2 - 2\widetilde{J}_3 \geq 0$. This is equivalent to $JD^{-1}J \geq 2J_3$, and to show this we consider the various possibilities for diagonal edges P_iP_j. We note that, with $D = \mathrm{diag}(d_l)$, the non-zero elements of $JD^{-1}J$ are obtained as

$$(JD^{-1}J)_{ij} = \sum_{l=1}^{m} J_{il} d_l^{-1} J_{lj} = J_{il_1} d_{l_1}^{-1} J_{l_1 j} + J_{il_2} d_{l_2}^{-1} J_{l_2 j},$$

where $P_iP_{l_1}P_j$ and $P_iP_{l_2}P_j$ are the two unique paths of non-diagonal edges connecting P_i to P_j. Therefore we find, recalling that $\kappa_4 = \frac{6}{25} \leq \frac{1}{4}$,

$$(JD^{-1}J)_{ij} = \begin{cases} 1 \cdot 1 \cdot 1 + 1 \cdot 1 \cdot 1 = 2, & \text{when } P_i, P_j \in \Omega, \\ \frac{1}{2} \cdot 2 \cdot 1 + 1 \cdot 1 \cdot 1 = 2, & \text{when } P_i \in \partial\Omega, P_j \in \Omega, \\ \frac{1}{2} \cdot 2 \cdot 1 + \frac{1}{2} \cdot 2 \cdot 1 = 2, & \text{when } P_i \in Q_1, P_j \in \Omega, \\ \frac{1}{2} \cdot \kappa_4^{-1} \cdot \frac{1}{2} + 1 \cdot 1 \cdot 1 \geq 2, & \text{when } P_i, P_j \in \partial\Omega, \end{cases}$$

so that $JD^{-1}J \geq 2J_3$. Thus $E_k \geq 0$ for $k \geq \lambda_0 h^2$, with $\lambda_0 = \frac{11}{24} \approx 0.46$.

4.2 Robin Boundary Conditions

We now turn to Robin boundary conditions, with $\beta = 1$, for simplicity. The mass matrix is the same as before. For the stiffness matrix we have

$$s_{ii} = \begin{cases} 4, & \text{if } P_i \in \Omega, \\ 2 + \frac{2}{3}h_0, & \text{if } P_i \in \partial\Omega, \\ 1 + \frac{2}{3}h_0 & \text{if } P_i \in Q_1 \cup Q_2, \end{cases} \quad \text{and} \quad s_{ij} = \begin{cases} -1, & \text{if } P_iP_j \in Z_1, \\ -\frac{1}{2} + \frac{1}{6}h_0, & \text{if } P_iP_j \in Z_2, \\ 0, & \text{if } P_iP_j \in Z_3. \end{cases}$$

As for Neumann boundary conditions S is an irreducible Stieltjes matrix and thus $S^{-1} > 0$, so that $H^{-1} = S^{-1}M > 0$ and the principal eigenvector $\varphi_1 > 0$ by the Perron–Frobenius theorem. Hence the semidiscrete Standard Galerkin solution matrix is nonnegative for large t by Theorem 13 and both the semidiscrete and Backward Euler solution matrices for the Lumped Mass method are nonnegative for $t \geq 0$ and $k \geq 0$, respectively, by Theorem 21. We shall show that, even though the assumptions of Theorem 6 are not satisfied, the Standard Galerkin Backward Euler matrix $E_k = (M + kS)^{-1}M \geq 0$ for $k \geq \lambda_0 h^2$ with $\lambda_0 > 0$ suitably chosen.

Here we find, again with $\lambda = k/h^2$, $\Lambda = 4\lambda + \frac{1}{4}$, $\mu = (\lambda - \frac{1}{24})\Lambda^{-1}$, $\nu = \frac{1}{24}\Lambda^{-1}$,

$$M + kS = \Lambda h^2\Big(\widehat{D} - \mu J_1 - \tfrac{1}{2}(\mu - \tfrac{1}{3}h_0\lambda/\Lambda)J_2 + \nu J_3\Big) = \Lambda h^2\big(\widehat{D} - \mu\widehat{J} + \nu J_3\big),$$

where $\widehat{J} = J_1 + \delta J_2$, with $\delta = \frac{1}{2}(1 - \frac{1}{3}h_0\lambda/(\mu\Lambda))$, and

$$\widehat{D} = I_1 + \widehat{\kappa}_2 I_2 + \widehat{\kappa}_3 I_3 + \widehat{\kappa}_4 I_4, \quad \text{where } \widehat{\kappa}_l = \kappa_l + \tfrac{2}{3}h_0\lambda/\Lambda,\ l = 2, 3, 4.$$

Hence, with $\breve{J} = \widehat{D}^{-1/2}\widehat{J}\widehat{D}^{-1/2}$, $\breve{J}_3 = \widehat{D}^{-1/2}J_3\widehat{D}^{-1/2}$, we have

$$M + kS = \Lambda h^2 \widehat{D}^{1/2}\big(I - \mu\breve{J} + \nu\breve{J}_3\big)\widehat{D}^{1/2}.$$

After some numerical experiments we chose $\lambda = 0.53$ and restrict ourselves to $h_0 \leq \frac{1}{4}$. We then have $\Lambda \approx 2.37$, $\mu \approx 0.21$, $\nu \approx 0.018, \kappa_3 \approx 0.26, \kappa_4 \approx 0.24$, $\widehat{\kappa}_l = \kappa_l + \frac{2}{3}h_0\lambda/\Lambda \approx \kappa_l + 0.15\,h_0$, $l = 2, 3, 4, \delta \approx 0.5 - 0.18\,h_0$. By numerical computation we then find

$$(\widehat{J}\widehat{D}^{-1}\widehat{J})_{ij} = \begin{cases} 1 \cdot 1 \cdot 1 + 1 \cdot 1 \cdot 1 = 2, & \text{when } P_iP_j \in \Omega, \\ \delta \cdot \widehat{\kappa}_2^{-1} \cdot 1 + 1 \cdot 1 \cdot 1 \geq 1.84, & \text{when } P_i \in \partial\Omega, P_j \in \Omega, \\ \delta \cdot \widehat{\kappa}_2^{-1} \cdot 1 + \delta \cdot \widehat{\kappa}_2^{-1} \cdot 1 \geq 1.69, & \text{when } P_i = Q_1, P_j \in \Omega \\ \delta \cdot \widehat{\kappa}_4^{-1} \cdot \delta + 1 \cdot 1 \cdot 1 \geq 1.74, & \text{when } P_i, P_j \in \partial\Omega. \end{cases}$$

We thus have $\widehat{J}\widehat{D}^{-1}\widehat{J} \geq 1.69\,J_3$ and we write, with $\omega = \nu/1.69 \approx 0.010$,

$$I - \mu\breve{J} + \nu\breve{J}_3 = I - \mu\breve{J} + \omega\breve{J}^2 - \omega(\breve{J}^2 - 1.69\,\breve{J}_3).$$

To apply Lemma 1 with this ω and $A = \breve{J}$, we must bound $\|\breve{J}\| = \|\widehat{D}^{-1/2}\widehat{J}\widehat{D}^{-1/2}\|$. It is obvious that all elements of $\breve{J}$ decrease when the parameter $p = h_0\lambda/\Lambda$ increases, so to bound the norm it suffices to consider $p = 0$. We then obtain the same calculations as in (28), with the present $\kappa_l, l = 2, 3, 4$, and, since these increase with λ, it suffices to consider $\lambda = \frac{11}{24}$ as before. We conclude that $\|\breve{J}\| \leq 2(1 + \sqrt{2}) \approx 4.83 < 4.85 \approx 1/\mu < x_1$. In particular, $\mu\|\breve{J}\| < 1$ so that the analogue of (29) holds. Following the proof for Neumann boundary conditions we may thus show that $E_k \geq 0$ for $k \geq 0.53\,h^2$.

4.3 Numerical Illustrations

In order to illustrate our above results in the case of the unit square we have computed nonnegativity thresholds for the different boundary conditions and with various meshsizes, and displayed the results in Table 1. The matrix $E(t) = \mathrm{e}^{-tH}$

Table 1 Nonnegativity thresholds for Dirichlet, Neumann, and Robin boundary conditions

		Dirichlet			Neumann			Robin ($\beta = 1$)		
		$E(t)$	E_k		$E(t)$	E_k		$E(t)$	E_k	
h_0	h	t_0	k_0	k_1	t_0	k_0	k_1	t_0	k_0	k_1
0.100	0.141	0.046	0.0053	0.0092	0.070	0.0053	0.0092	0.070	0.0054	0.0106
0.050	0.070	0.035	0.0013	0.0023	0.042	0.0014	0.0023	0.042	0.0014	0.0027
0.025	0.035	0.021	0.0003	0.0006	0.023	0.0003	0.0006	0.023	0.0003	0.0007

was computed by MATLAB's matrix exponential function and we then determined the visibly unique point t_0 where the minimal element passes from negative to nonnegative values. We also computed upper bounds for the thresholds according to (6) and Theorems 2 and 13. This was done by computing the eigenvalues and eigenvectors of H with MATLAB and finding the zeros of the functions $t \mapsto \sum_{j=2}^{m} \mathrm{e}^{-\lambda_j t}\sigma_j^2 - \mathrm{e}^{-\lambda_1 t}$. These bounds are essentially independent of h and approximately equal to 0.2 for all three boundary conditions. They thus greatly overestimate the thresholds.

Similarly we computed the nonnegativity thresholds k_0 for $E_k = (M+kS)^{-1}M$. In this case we know from the theory that the threshold is bounded above by $k_1 = \lambda_0 h^2$ with $\lambda_0 \approx 0.46$ for Dirichlet and Neumann and $\lambda_0 \approx 0.53$ for Robin boundary conditions ($\beta = 1$), and these bounds are also included in the table.

5 A Special Case in One Space Dimension

In this final section we discuss some analogues in one space dimension of our earlier results. We begin by recalling known results for the case of Dirichlet boundary conditions from [1], and then turn to Neumann and Robin boundary conditions. We shall thus consider the initial-boundary value problem

$$\begin{aligned} u_t = u_{xx}, \quad \text{in } \Omega = (0,1), \quad \text{with } u(0,t) = u(1,t) = 0, \quad \text{for } t > 0,\\ u(x,0) = v(x), \quad \text{in } \Omega. \end{aligned} \tag{30}$$

We partition $\Omega = (0,1)$ uniformly into subintervals $I_j = [x_{j-1}, x_j]$ by $x_j = jh$, $j = 0, \dots, m$, $h = 1/m$, and let S_h be the continuous piecewise linear functions χ on this partition, with $\chi(0) = \chi(1) = 0$. The basis functions $\{\Phi_i\}_{i=1}^{m-1} \subset S_h$, are defined by $\Phi_i(x_j) = \delta_{ij}$.

With $(\cdot,\cdot) = (\cdot,\cdot)_{L_2(\Omega)}$, the spatially semidiscrete analogue of (30) is

$$(u_{h,t}, \chi) + (u_h', \chi') = 0, \quad \forall \chi \in S_h, \ \text{for } t \ge 0, \quad \text{with } u_h(0) = v_h. \tag{31}$$

In matrix form, this may be written as (4), where the $(m-1)\times(m-1)$ mass and stiffness matrices have elements $m_{ij} = (\Phi_i, \Phi_j)$ and $s_{ij} = (\Phi_i', \Phi_j')$, respectively, and we find

$$M = h\begin{pmatrix} 2/3 & 1/6 & 0 & \dots & 0 \\ 1/6 & 2/3 & 1/6 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & 2/3 \end{pmatrix} \quad \text{and} \quad S = h^{-1}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & 2 \end{pmatrix}.$$

The solution matrix is $E(t) = \mathrm{e}^{-tH}$, with $H = M^{-1}S$, and it was shown in [1] that, as in the two-dimensional case, $E(t)$ cannot be nonnegative for all $t > 0$. However, we could have $E(t) \geq 0$ for large t: With $\{\varphi_j\}_{j=1}^{m-1}$ the eigenvectors of H we have a positive principal eigenvector $\varphi_{1,l} = c\sin(\pi x_l)$ with normalizing factor $c > 0$, and it was shown in [1] that $E(t) > 0$ if

$$\sum_{j=2}^{m-1} \mathrm{e}^{-\lambda_j t}\sigma_j^2 < \mathrm{e}^{-\lambda_1 t}, \quad \text{where } \sigma_j = \max_l(|\varphi_{j,l}|/\varphi_{1,l}). \tag{32}$$

For the Backward Euler method, $E_k = (I + kH)^{-1}$, where $H = M^{-1}S$. We quote from [1] that $E_k \geq 0$ if and only if $\lambda = k/h^2 \geq \frac{1}{6}$, i.e., that $k_0 = \frac{1}{6}h^2$ is the nonnegativity threshold.

For the Lumped Mass method we use $(\psi, \chi)_h = h\sum_{j=1}^{m-1}\psi_j\chi_j$ to replace (ψ, χ) in (31). This gives the mass matrix $M = hI$, and thus $H = h^{-1}S$, and we then have that $E(t) \geq 0$ for $t \geq 0$, and $E_k \geq 0$ for $k \geq 0$.

We now turn to the Neumann and Robin boundary conditions and consider the problem

$$u_t = u_{xx} \text{ in } (0,1), \quad u'(0,t) = \beta u(0),\ u'(1,t) = -\beta u(1,t),\ \text{for } t > 0,$$

with given initial values $u(x,0) = v(x)$. Here the constant β is ≥ 0, with $\beta = 0$ for Neumann boundary conditions and $\beta > 0$ for Robin boundary conditions.

This time we use for S_h the $(m+1)$-dimensional space of continuous piecewise linear functions on the I_j, and the basis functions are now $\{\Phi_i\}_{i=0}^m \subset S_h$, defined by $\Phi_i(x_j) = \delta_{ij}$. The semidiscrete problem is then

$$(u_{h,t}, \chi) + (u_h', \chi') + \beta u_h(0)\chi(0) + \beta u_h(1)\chi(1) = 0, \quad \forall \chi \in S_h,\ \text{for } t \geq 0,$$

with $u_h(0) = v_h$. The matrix formulation (4) now uses the $(m+1)\times(m+1)$ mass and stiffness matrices

$$M = h\begin{pmatrix} 1/3 & 1/6 & 0 & \dots & 0 \\ 1/6 & 2/3 & 1/6 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & 1/3 \end{pmatrix} \quad \text{and} \quad S_\beta = h^{-1}\begin{pmatrix} 1+\beta h & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \dots & & 1+\beta h \end{pmatrix},$$

where M and S_β with $\beta > 0$ are positive definite and S_0 is positive semidefinite. Note that S_β is an irreducible Stieltjes matrix for $\beta > 0$ so that $S_\beta^{-1} > 0$, and thus $H^{-1} = S_\beta^{-1}M$ has a positive principal eigenvector. If $\beta = 0$, then $\lambda_1 = 0$ and $\varphi_1 = \underline{1}$ is a positive principal eigenvector. Analogously to Theorem 1 we have the following.

Theorem 25 *$E(t) = \mathrm{e}^{-tH}$, $H = M^{-1}S_\beta$, cannot be ≥ 0 for small $t > 0$.*

Proof If $E(t) \geq 0$ for small $t > 0$, then $h_{ij} \leq 0$ for $j \neq i$, because $E(t) = \mathrm{e}^{-tH} = I - tH + O(t^2)$ as $t \to 0$.

Using the diagonal matrix $D_\beta = \mathrm{diag}(\frac{1}{2}(1+\beta h), 1, \dots, 1, \frac{1}{2}(1+\beta h))$ and the matrix J with elements 1 in the two main bidiagonals and all other elements 0, the mass matrix may be written $M = h(\frac{2}{3}D_0 + \frac{1}{6}J) = \frac{2}{3}hD_0(I + \frac{1}{4}D_0^{-1}J)$, and thus

$$M^{-1} = \tfrac{3}{2}h^{-1}(I + \tfrac{1}{4}D_0^{-1}J)^{-1}D_0^{-1} = \tfrac{3}{2}h^{-1}\sum_{j=0}^{\infty}(-1)^j(\tfrac{1}{4}D_0^{-1}J)^j\,D_0^{-1},$$

where the series converges since, in maximum norm, $\|D_0^{-1}J\| = 2$. Further, $(D_0^{-1}J)^j$ has nonzero elements only in bidiagonals of even order when j is even, and of odd order when j is odd. It follows that the elements of M^{-1} are positive in bidiagonals of even order and negative in those of odd order. Since $S_\beta = h^{-1}(2D_\beta - J)$, the same holds for $H = M^{-1}S_\beta$, in contradiction to $h_{ij} \leq 0$ for $j \neq i$. □

In the same way as in Theorem 2 one shows:

Theorem 26 *$E(t) > 0$ if $\sum_{j=2}^{m+1}\mathrm{e}^{-\lambda_j t}\sigma_j^2 < \mathrm{e}^{-\lambda_1 t}$, where $\sigma_j = \max_l(|\varphi_{j,l}|/\varphi_{1,l})$.*

We now turn to the Backward Euler matrix $E_k = (I + kH)^{-1}$, $H = M^{-1}S_\beta$ and show again that $k_0 = \frac{1}{6}h^2$ is the nonnegativity threshold.

Theorem 27 *$E_k \geq 0$ if and only if $\lambda = k/h^2 \geq 1/6$.*

Proof We have $M = h(\frac{2}{3}D_0 + \frac{1}{6}J)$, $S_\beta = h^{-1}(2D_\beta - J)$, and hence

$$M + kS_\beta = h\big(\tfrac{2}{3}D_0 + 2\lambda D_\beta + (\tfrac{1}{6} - \lambda)J\big).$$

Since this matrix is a Stieltjes matrix for $\lambda \geq 1/6$, it has a nonnegative inverse. Thus $E_k = (M + kS_\beta)^{-1}M \geq 0$.

For the converse, we first note that if $E_{k_0} \geq 0$, then $E_k \geq 0$ for $k \geq k_0$, as follows as in Theorem 8. Let now $\lambda < \frac{1}{6}$ and set $\epsilon = \frac{1}{6} - \lambda$. Then, with $\widetilde{D} = \frac{2}{3}D_0 + 2\lambda D_\beta$,

$$\begin{aligned} E_k &= (M + kS_\beta)^{-1}M = h^{-1}(\widetilde{D} + \epsilon J)^{-1}M = (I + \epsilon\widetilde{D}^{-1}J)^{-1}\widetilde{D}^{-1}(\tfrac{2}{3}D_0 + \tfrac{1}{6}J) \\ &= (I - \epsilon\widetilde{D}^{-1}J)\widetilde{D}^{-1}(\tfrac{2}{3}D_0 + \tfrac{1}{6}J) + O(\epsilon^2) = P - \epsilon Q + O(\epsilon^2), \quad \text{as } \epsilon \to 0, \end{aligned}$$

Table 2 Nonnegativity thresholds for the different boundary conditions

h	Dirichlet	Neumann	Robin
0.020	0.0082	0.0086	0.0086
0.010	0.0044	0.0045	0.0045
0.005	0.0023	0.0023	0.0023

where the elements p_{ij} of P vanish for $|i-j| > 1$ and the elements q_{ij} of Q are positive for $|i-j| = 2$. Hence the elements of E_k in the second bidiagonals are negative for small ϵ, so that $E_k \not\geq 0$ for $\lambda < 1/6$. □

For the Lumped Mass method we replace (ψ, χ) by

$$(\psi, \chi)_h = \tfrac{1}{2}h\psi_0\chi_0 + \tfrac{1}{2}h\psi_m\chi_m + h\sum_{j=1}^{m-1}\psi_j\chi_j,$$

which gives the mass matrix $M = hD_0$ and thus $H = h^{-1}D_0^{-1}S_\beta$. We now have:

Theorem 28 *For the Lumped Mass method we have $E(t) \geq 0$ for all $t \geq 0$ and $E_k \geq 0$ for all $k \geq 0$.*

Proof Since obviously $M + kS_\beta = h(D_0 + 2\lambda D_\beta - \lambda J)$ is a Stieltjes matrix, so that $(M + kS_\beta)^{-1} \geq 0$, we have $E_k = (M + kS_\beta)^{-1}M \geq 0$. Further,

$$E(t) = \mathrm{e}^{-tH} = \lim_{n\to\infty}\Big(I + \frac{t}{n}H\Big)^{-n} = \lim_{n\to\infty} E_{t/n}^n \geq 0, \quad \text{for } t \geq 0.$$

The proof is complete. □

In order to illustrate our results we computed numerically the nonnegativity thresholds for the matrix $E(t) = \mathrm{e}^{-tH}$ for the three boundary conditions, with $\beta = 1$ in the Robin case, and for different h. The computations were carried out in the same way as in Sect. 4.3 and the results are displayed in Table 2. We also computed upper bounds for the thresholds according to (32) and Theorem 26. These bounds are essentially independent of h and approximately 0.052, 0.080, and 0.078, respectively, for the three boundary conditions. We see that they greatly overestimate the thresholds. For E_k we know already from the theory that the threshold is $k_0 = \frac{1}{6}h^2$.

Acknowledgements The authors are grateful to the referees whose comments helped to correct and clarify the exposition.

References

1. Chatzipantelidis, P., Horváth, Z., Thomée, V.: On preservation of positivity in some finite element methods for the heat equation. Comput. Methods Appl. Math. **15**, 417–437 (2015)
2. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comput. **74**, 1–23 (2005) (electronic)

3. Fujii, H.: Some remarks on finite element analysis of time-dependent field problems. In: Theory and Practice in Finite Element Structural Analysis, pp. 91–106. University of Tokyo Press, Tokyo (1973)
4. Protter, M.H., Weinberger, H.F.: Maximum Principles in Differential Equations. Prentice-Hall, Englewood Cliffs (1967)
5. Schatz, A.H., Thomée, V., Wahlbin, L.B.: On positivity and maximum-norm contractivity in time stepping methods for parabolic equations. Comput. Methods Appl. Math. **10**, 421–443 (2010)
6. Smoller, J.: Shock Waves and Reaction-Diffusion Equations, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science], vol. 258. Springer, New York-Berlin (1983)
7. Strang, G., Fix, G.J.: An Analysis of the Finite Element Method. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs (1973).
8. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, Springer Series in Computational Mathematics, vol. 25, 2nd edn. Springer, Berlin (2006)
9. Thomée, V.: On positivity preservation in some finite element methods for the heat equation. In: Numerical Methods and Applications. Lecture Notes in Computer Science, vol. 8962, pp. 13–24. Springer, Cham (2015)
10. Thomée, V., Wahlbin, L.B.: On the existence of maximum principles in parabolic finite element equations. Math. Comput. **77**, 11–19 (2008) (electronic)
11. Varga, R.S.: Matrix Iterative Analysis. Springer Series in Computational Mathematics, vol. 27, expanded edn. Springer, Berlin (2000)

Numerical Solutions of a Boundary Value Problem on the Sphere Using Radial Basis Functions

Quoc T. Le Gia

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Boundary value problems on the unit sphere arise naturally in geophysics and oceanography when scientists model a physical quantity on large scales. Robust numerical methods play an important role in solving these problems. In this article, we construct numerical solutions to a boundary value problem defined on a spherical sub-domain (with a sufficiently smooth boundary) using radial basis functions (RBFs). The error analysis between the exact solution and the approximation is provided. Numerical experiments are presented to confirm theoretical estimates.

1 Introduction

Boundary value problems on the unit sphere arise naturally in geophysics and oceanography when scientists model a physical quantity on large scales. In that situation, the curvature of the Earth cannot be ignored, and a boundary value problem has to be formulated on a subdomain of the unit sphere. For example, the study of planetary-scale oceanographic flows in which oceanic eddies interact with topography such as ridges and land masses or evolve in closed basin lead to the study of point vortices on the surface of the sphere with walls [2, 13]. Such vortex motions can be described as a Dirichlet problem on a subdomain of the sphere for the Laplace-Beltrami operator [4, 21]. Solving the problem exactly via conformal mapping methods onto the complex plane was proposed by Crowdy in [4]. Kidambi and Newton [21] also considered such a problem, assuming the sub-surface of the sphere lent itself to method of images. A boundary integral method for constructing numerical solutions to the problem was discussed in [11]. In this work, we propose

Q. T. Le Gia (✉)
University of New South Wales, Sydney, NSW, Australia
e-mail: qlegia@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_36

a collocation method using spherical radial basis functions. Radial basis functions (RBFs) present a simple and effective way to construct approximate solutions to partial differential equations (PDEs) on spheres, via a collocation method [26] or a Galerkin method [22]. They have been used successfully for solving transport-like equations or semilinear parabolic equations on the sphere [6, 7, 36]. The method does not require a mesh, and is simple to implement.

While meshless methods using RBFs have been employed to derive numerical solutions for PDEs on the sphere only recently, it should be mentioned that approximation methods using RBFs for PDEs on bounded domains have been around for the last two decades. Originally proposed by Kansa [19, 20] for fluid dynamics, approximation methods for many types of PDEs defined on bounded domains in $\mathbb{R}^n$ using RBFs have since been used widely [5, 9, 16, 17].

To the best of our knowledge, approximation methods using RBFs have not been investigated for boundary value problems defined on subdomains of the unit sphere. Given the potential of RBF methods on these problems, the present paper aims to present a collocation method for boundary value problems on the sphere and provide a mathematical foundation for error estimates.

The paper is organized as follows: in Section Preliminaries we review some preliminaries on functions spaces, positive definite kernels, radial basis functions and the generalized interpolation problem on discrete point sets on the unit sphere. In Section Boundary Value Problems on the Sphere we define the boundary value problem on a spherical cap, then present a collocation method using spherical radial basis functions and our main result, Theorem 4. We conclude the paper by giving some numerical experiments in the last section.

Throughout the paper, we denote by $c, c_1, c_2, \ldots$ generic positive constants that may assume different values at different places, even within the same formula.

For two sequences $\{a_\ell\}_{\ell\in\mathbb{N}_0}$ and $\{b_\ell\}_{\ell\in\mathbb{N}_0}$, the notation $a_\ell \sim b_\ell$ means that there exist positive constants c_1 and c_2 such that $c_1 b_\ell \le a_\ell \le c_2 b_\ell$ for all $\ell \in \mathbb{N}_0$.

2 Preliminaries

Let $\mathbb{S}^n$ be the *unit sphere*, i.e. $\mathbb{S}^n := \left\{\boldsymbol{x} \in \mathbb{R}^{n+1} : \|\boldsymbol{x}\| = 1\right\}$ in the Euclidean space $\mathbb{R}^{n+1}$, where $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}\cdot\boldsymbol{x}}$ denotes the Euclidean norm of $\mathbb{R}^{n+1}$, induced by the Euclidean inner product $\boldsymbol{x}\cdot\boldsymbol{y}$ of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^{n+1}$. The surface area of the unit sphere $\mathbb{S}^n$ is denoted by ω_n and is given by

$$\omega_n := |\mathbb{S}^n| = \frac{2\pi^{(n+1)/2}}{\Gamma((n+1)/2)}.$$

The *spherical distance* (or geodesic distance) $\operatorname{dist}_{\mathbb{S}^n}(\boldsymbol{x}, \boldsymbol{y})$ of two points $\boldsymbol{x} \in \mathbb{S}^n$ and $\boldsymbol{y} \in \mathbb{S}^n$ is defined as the length of a shortest geodesic arc connecting the two points. The geodesic distance $\operatorname{dist}_{\mathbb{S}^n}(\boldsymbol{x},\boldsymbol{y})$ is the angle in $[0,\pi]$ between the points $\boldsymbol{x}$ and $\boldsymbol{y}$, thus

$$\operatorname{dist}_{\mathbb{S}^n}(\boldsymbol{x},\boldsymbol{y}) := \arccos(\boldsymbol{x}\cdot\boldsymbol{y}).$$

Let Ω be an open simply connected subdomain of the sphere. For a point set $X := \{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\} \subset \mathbb{S}^n$, the *(global) mesh norm* h_X is given by

$$h_X = h_{X,\mathbb{S}^n} := \sup_{\boldsymbol{x}\in\mathbb{S}^n} \inf_{\boldsymbol{x}_j\in X} \operatorname{dist}_{\mathbb{S}^n}(\boldsymbol{x},\boldsymbol{x}_j),$$

and the *local mesh norm* $h_{X,\Omega}$ with respect to the subdomain Ω is defined by

$$h_{X,\Omega} := \sup_{\boldsymbol{x}\in\Omega} \inf_{\boldsymbol{x}_j\in X\cap\Omega} \operatorname{dist}_{\mathbb{S}^n}(\boldsymbol{x},\boldsymbol{x}_j).$$

The mesh norm $h_{X_2,\partial\Omega}$ of $X_2 \subset \partial\Omega$ along the boundary $\partial\Omega$ is defined by

$$h_{X_2,\partial\Omega} := \sup_{\boldsymbol{x}\in\partial\Omega} \inf_{\boldsymbol{x}_j\in X_2} \operatorname{dist}_{\partial\Omega}(\boldsymbol{x},\boldsymbol{x}_j), \tag{1}$$

where $\operatorname{dist}_{\boldsymbol{x}\in\partial\Omega}$ is here the geodesic distance along the boundary $\partial\Omega$.

2.1 Sobolev Spaces on the Sphere

Let Ω be $\mathbb{S}^n$ or an open measurable subset of $\mathbb{S}^n$. Let $L_2(\Omega)$ denote the Hilbert space of (real-valued) square-integrable functions on Ω with the inner product

$$\langle f,g\rangle_{L_2(\Omega)} := \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\omega_n(\boldsymbol{x})$$

and the induced norm $\|f\|_{L_2(\Omega)} := \langle f,f\rangle^{1/2}_{L_2(\Omega)}$. Here $\mathrm{d}\omega_n$ is the Lebesgue surface area element of the sphere $\mathbb{S}^n$.

The space of continuous functions on the sphere $\mathbb{S}^n$ and on the closed subdomain $\overline{\Omega}$ are denoted by $C(\Omega)$ and $C(\overline{\Omega})$ and are endowed with the supremum norms

$$\|f\|_{C(\mathbb{S}^n)} := \sup_{\boldsymbol{x}\in\mathbb{S}^n} |f(\boldsymbol{x})| \qquad \text{and} \qquad \|f\|_{C(\overline{\Omega})} := \sup_{\boldsymbol{x}\in\overline{\Omega}} |f(\boldsymbol{x})|,$$

respectively.

A *spherical harmonic* of degree $\ell \in \mathbb{N}_0$ (for the sphere $\mathbb{S}^n$) is the restriction of a homogeneous harmonic polynomial on $\mathbb{R}^{n+1}$ of exact degree ℓ to the unit sphere

$\mathbb{S}^n$. The vector space of all spherical harmonics of degree ℓ (and the zero function) is denoted by $\mathbb{H}_\ell(\mathbb{S}^n)$ and has the dimension $Z(n,\ell) := \dim(\mathbb{H}_\ell(\mathbb{S}^n))$ given by

$$Z(n,0) = 1 \qquad \text{and} \qquad Z(n,\ell) = \frac{(2\ell+n-1)\Gamma(\ell+n-1)}{\Gamma(\ell+1)\Gamma(n)} \quad \text{for } \ell \in \mathbb{N}.$$

By $\{\, Y_{\ell,k} \;:\; k = 1,2,\dots,Z(n,\ell)\}$, we will always denote an $L_2(\mathbb{S}^n)$-orthonormal basis of $\mathbb{H}_\ell(\mathbb{S}^n)$ consisting of spherical harmonics of degree ℓ. Any two spherical harmonics of different degree are orthogonal to each other, and the union of all sets $\{\, Y_{\ell,k} \;:\; k = 1,2,\dots,Z(n,\ell)\}$ constitutes a complete orthonormal system for $L_2(\mathbb{S}^n)$. Thus any function $f \in L_2(\mathbb{S}^n)$ can be represented in $L_2(\mathbb{S}^n)$-sense by its *Fourier series* (or Laplace series)

$$f = \sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(n,\ell)} \widehat{f}_{\ell,k} Y_{\ell,k},$$

with the Fourier coefficients $\widehat{f}_{\ell,k}$ defined by

$$\widehat{f}_{\ell,k} := \int_{\mathbb{S}^n} f(\boldsymbol{x}) Y_{\ell,k}(\boldsymbol{x}) \mathrm{d}\omega_n(\boldsymbol{x}).$$

The *space of spherical polynomials of degree* $\leq K$ (that is, the set of the restrictions to $\mathbb{S}^n$ of all polynomials on $\mathbb{R}^{n+1}$ of degree $\leq K$) is denoted by $\mathbb{P}_K(\mathbb{S}^n)$. We have $\mathbb{P}_K(\mathbb{S}^n) = \bigoplus_{\ell=0}^{K} \mathbb{H}_\ell(\mathbb{S}^n)$ and $\dim(\mathbb{P}_K(\mathbb{S}^n)) = Z(n+1,K) \sim (K+1)^n$.

Any orthonormal basis $\{\, Y_{\ell,k} \;:\; k = 1,2,\dots,Z(n,\ell)\}$ of $\mathbb{H}_\ell(\mathbb{S}^n)$ satisfies the *addition theorem* (see [27, p. 10])

$$\sum_{k=0}^{Z(n,\ell)} Y_{\ell,k}(\boldsymbol{x}) Y_{\ell,k}(\boldsymbol{y}) = \frac{Z(n,\ell)}{\omega_n} P_\ell(n+1; \boldsymbol{x}\cdot\boldsymbol{y}), \tag{2}$$

where $P_\ell(n+1;\cdot)$ is the *normalized Legendre polynomial* of degree ℓ in $\mathbb{R}^{n+1}$. The normalized Legendre polynomials $\{\, P_\ell(n+1;\cdot)\}_{\ell\in\mathbb{N}_0}$, form a complete orthogonal system for the space $L_2([-1,1];(1-t^2)^{(n-2)/2})$ of functions on $[-1,1]$ which are square-integrable with respect to the weight function $w(t) := (1-t^2)^{(n-2)/2}$. They satisfy $P_\ell(n+1;1) = 1$ and

$$\int_{-1}^{+1} P_\ell(n+1;t) P_k(n+1;t)(1-t^2)^{(n-2)/2} \mathrm{d}t = \frac{\omega_n}{\omega_{n-1} Z(n,\ell)} \delta_{\ell,k}, \tag{3}$$

where $\delta_{\ell,k}$ is the Kronecker delta (defined to be one if $\ell = k$ and zero otherwise).

The *Laplace-Beltrami operator* Δ^* (for the unit sphere $\mathbb{S}^n$) is the angular part of the Laplace operator $\Delta = \sum_{j=1}^{n+1} \partial^2/\partial x_j^2$ for $\mathbb{R}^{n+1}$. Spherical harmonics of degree ℓ on $\mathbb{S}^n$ are eigenfunctions of $-\Delta^*$, more precisely,

$$-\Delta^* Y_\ell = \lambda_\ell Y_\ell \quad \text{for all } Y_\ell \in \mathbb{H}_\ell(\mathbb{S}^n) \qquad \text{with} \qquad \lambda_\ell := \ell(\ell+n-1).$$

For $s \in \mathbb{R}_0^+$, the *Sobolev space* $H^s(\mathbb{S}^n)$ is defined by (see [24, Chapter 1, Remark 7.6])

$$H^s(\mathbb{S}^n) := \left\{ f \in L_2(\mathbb{S}^n) : \sum_{\ell=0}^{\infty} (1+\lambda_\ell)^s \sum_{k=1}^{Z(n,\ell)} |\widehat{f}_{\ell,k}|^2 < \infty \right\}.$$

The space $H^s(\mathbb{S}^n)$ is a Hilbert space with the inner product

$$\langle f, g \rangle_{H^s(\mathbb{S}^n)} := \sum_{\ell=0}^{\infty} (1+\lambda_\ell)^s \sum_{k=1}^{Z(n,\ell)} \widehat{f}_{\ell,k} \widehat{g}_{\ell,k}$$

and the induced norm

$$\|f\|_{H^s(\mathbb{S}^n)} := \langle f,f \rangle_{H^s(\mathbb{S}^n)}^{1/2} = \sum_{\ell=0}^{\infty} (1+\lambda_\ell)^s \sum_{k=1}^{Z(n,\ell)} |\widehat{f}_{\ell,k}|^2. \qquad (4)$$

If $s > n/2$, then $H^s(\mathbb{S}^n)$ is embedded into $C(\mathbb{S}^n)$, and the Sobolev space $H^s(\mathbb{S}^n)$ is a *reproducing kernel Hilbert space*. This means that the evaluation functional over $H^s(\mathbb{S}^n)$ is a bounded operator. From the Riesz representation theorem, there exists a unique element $K_{\boldsymbol{x}} \in H^s(\mathbb{S}^n)$ with the reproducing property

$$\langle f, K_{\boldsymbol{x}} \rangle_{H^s(\mathbb{S}^n)} = f(\boldsymbol{x}), \quad \forall f \in H^s(\mathbb{S}^n).$$

Since $K_{\boldsymbol{x}}$ is itself a function in $H^s(\mathbb{S}^n)$, for every $\boldsymbol{y} \in \mathbb{S}^n$, there exists a $K_{\boldsymbol{y}} \in H^s(\mathbb{S}^n)$ such that

$$K_{\boldsymbol{x}}(\boldsymbol{y}) = \langle K_{\boldsymbol{x}}, K_{\boldsymbol{y}} \rangle_{H^s(\mathbb{S}^n)}.$$

This allows us to define the reproducing kernel as a function $K_s : \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{R}$ by

$$K_s(\boldsymbol{x}, \boldsymbol{y}) = \langle K_{\boldsymbol{x}}, K_{\boldsymbol{y}} \rangle_{H^s(\mathbb{S}^n)}.$$

From this definition, it is easy to see that K is both symmetric and positive definite.

Sobolev spaces on $\mathbb{S}^n$ can also be defined using local charts (see [24]). Here we use a specific atlas of charts, as in [18].

Let $\boldsymbol{z}$ be a given point on $\mathbb{S}^n$, the spherical cap centered at $\boldsymbol{z}$ of radius θ is defined by

$$G(\boldsymbol{z}, \theta) = \{\boldsymbol{y} \in \mathbb{S}^n : \cos^{-1}(\boldsymbol{z} \cdot \boldsymbol{y}) < \theta\}, \qquad \theta \in (0, \pi),$$

where $\boldsymbol{z} \cdot \boldsymbol{y}$ denotes the Euclidean inner product of $\boldsymbol{z}$ and $\boldsymbol{y}$ in $\mathbb{R}^{n+1}$.

Let $\hat{\boldsymbol{n}}$ and $\hat{\boldsymbol{s}}$ denote the north and south poles of $\mathbb{S}^n$, respectively. Then a simple cover for the sphere is provided by

$$U_1 = G(\hat{\boldsymbol{n}}, \theta_0) \quad \text{and} \quad U_2 = G(\hat{\boldsymbol{s}}, \theta_0), \text{ where } \theta_0 \in (\pi/2, 2\pi/3). \tag{5}$$

The stereographic projection $\sigma_{\hat{\boldsymbol{n}}}$ of the punctured sphere $\mathbb{S}^n \setminus \{\hat{\boldsymbol{n}}\}$ onto $\mathbb{R}^n$ is defined as a mapping that maps $\boldsymbol{x} \in \mathbb{S}^n \setminus \{\hat{\boldsymbol{n}}\}$ to the intersection of the equatorial hyperplane $\{z = 0\}$ and the extended line that passes through $\boldsymbol{x}$ and $\hat{\boldsymbol{n}}$. The stereographic projection $\sigma_{\hat{\boldsymbol{s}}}$ based on $\hat{\boldsymbol{s}}$ can be defined analogously. We set

$$\psi_1 = \frac{1}{\tan(\theta_0/2)} \sigma_{\hat{\boldsymbol{s}}}|_{U_1} \quad \text{and} \quad \psi_2 = \frac{1}{\tan(\theta_0/2)} \sigma_{\hat{\boldsymbol{n}}}|_{U_2}, \tag{6}$$

so that ψ_k, $k = 1, 2$, maps U_k onto $B(0, 1)$, the unit ball in $\mathbb{R}^n$. We conclude that $\mathscr{A} = \{U_k, \psi_k\}_{k=1}^2$ is a C^∞ atlas of covering coordinate charts for the sphere. It is known (see [30]) that the stereographic coordinate charts $\{\psi_k\}_{k=1}^2$ as defined in (6) map spherical caps to Euclidean balls, but in general concentric spherical caps are not mapped to concentric Euclidean balls. The projection ψ_k, for $k = 1, 2$, does not distort too much the geodesic distance between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^n$, as shown in [23].

With the atlas so defined, we define the map π_k which takes a real-valued function g with compact support in U_k into a real-valued function on $\mathbb{R}^n$ by

$$\pi_k(g)(x) = \begin{cases} g \circ \psi_k^{-1}(x), & \text{if } x \in B(0, 1), \\ 0, & \text{otherwise}. \end{cases}$$

Let $\{\chi_k : \mathbb{S}^n \to \mathbb{R}\}_{k=1}^2$ be a partition of unity subordinated to the atlas, i.e., a pair of non-negative infinitely differentiable functions χ_k on $\mathbb{S}^n$ with compact support in U_k, such that $\sum_k \chi_k = 1$. For any function $f : \mathbb{S}^n \to \mathbb{R}$, we can use the partition of unity to write

$$f = \sum_{k=1}^{2} (\chi_k f), \text{ where } (\chi_k f)(\boldsymbol{x}) = \chi_k(\boldsymbol{x}) f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{S}^n.$$

The Sobolev space $H^s(\mathbb{S}^n)$ is defined to be the set

$$\{f \in L_2(\mathbb{S}^n) : \pi_k(\chi_k f) \in H^s(\mathbb{R}^n) \quad \text{for } k = 1, 2\},$$

which is equipped with the norm

$$\|f\|_{H^s(\mathbb{S}^n)} = \left(\sum_{k=1}^{2} \|\pi_k(\chi_k f)\|_{H^s(\mathbb{R}^n)}^2 \right)^{1/2}. \tag{7}$$

This $H^s(\mathbb{S}^n)$ norm is equivalent to the $H^s(\mathbb{S}^n)$ norm given previously in (4) (see [24]).

Let $\Omega \subset \mathbb{S}^n$ be an open connected set with sufficiently smooth boundary. In order to define the Sobolev spaces on Ω, let $D_k = \psi_k(\Omega \cap U_k)$ for $k = 1, 2$. The local Sobolev space $H^s(\Omega)$ is defined to be the set

$$f \in L_2(\Omega) : \pi_k(\chi_k f)|_{D_k} \in H^s(D_k) \text{ for } k = 1, 2,\ D_k \neq \emptyset,$$

which is equipped with the norm

$$\|f\|_{H^s(\Omega)} = \left(\sum_{k=1}^{2} \|\pi_k(\chi_k f)|_{D_k}\|^2_{H^s(D_k)} \right)^{1/2} \tag{8}$$

where, if $D_k = \emptyset$, then we adopt the convention that $\|\cdot\|_{H^s(D_k)} = 0$.

It should be noted that if $s = m$ which is a positive integer, we can define the local Sobolev norm via the following formula

$$\|f\|_{H^m(\Omega)} = \left(\sum_{k=0}^{m} \langle \nabla^k f, \nabla^k f \rangle_{L_2(\Omega)} \right)^{1/2}, \tag{9}$$

where ∇ is the surface gradient on the sphere.

Now we state an extension theorem for a local domain on the sphere. We follow a framework set out in [33, Chapter 4.4]. To this end, let us consider Sobolev spaces $H^s(\mathbb{R}^n_+)$, with $\mathbb{R}^n_+ = \{\boldsymbol{x} \in \mathbb{R}^n : x_1 > 0\}$. For $k \geq 0$ an integer, let

$$H^k(\mathbb{R}^n_+) = \{u \in L^2(\mathbb{R}^n_+) : D^\alpha u \in L^2(\mathbb{R}^n_+) \text{ for } |\alpha| \leq k\}.$$

Here, $D^\alpha u$ is considered as a distribution on the interior $\mathbb{R}^n_+$. We claim that each $u \in H^k(\mathbb{R}^n_+)$ is the restriction to $\mathbb{R}^n_+$ of an element of $H^k(\mathbb{R}^n)$. To see this, fix an integer $N \geq k + 1$, for an u in the Schwartz class $\mathscr{S}(\overline{\mathbb{R}^n_+})$ let

$$Eu(x) = \begin{cases} u(x) & \text{for } x_1 \geq 0, \\ \sum_{j=1}^{N} a_j u(-jx_1, \boldsymbol{x}'), & \text{for } x_1 < 0. \end{cases}$$

Lemma 1 *There exist coefficients $a_1, \ldots, a_N$ such that the map E has a unique continuous extension to*

$$E : H^k(\mathbb{R}^n_+) \to H^k(\mathbb{R}^n), \quad \textit{for } k \leq N - 1.$$

Proof Given $u \in \mathscr{S}(\mathbb{R}^n)$, we get an H^k-estimate on Eu provided all the derivatives of Eu of order $N - 1$ match up at $x_1 = 0$, that is, provided

$$\sum_{j=1}^{N} (-j)^\ell a_j = 1, \text{ for } \ell = 0, 1, \ldots, N - 1. \tag{10}$$

The system (10) is a linear system of N equations for N unknowns a_j; its determinant is a Vandermonde determinant that is non-zero, so a_j can be found. □

Theorem 1 (Extension Theorem) *Let $\Omega \subset \mathbb{S}^n$ be a local region with a sufficiently smooth boundary. For real $s \geq 0$, there exists a bounded extension operator*

$$E : H^s(\Omega) \to H^s(\mathbb{S}^n). \tag{11}$$

Proof For $k \geq 0$ being an integer, let $H^k(\Omega)$ be the space of all $u \in L^2(\Omega)$ such that $Pu \in L^2(\Omega)$ for all differential operators P of order $\leq k$ with coefficients in $C^\infty(\overline{\Omega})$. By covering a neighbourhood of $\partial\Omega \subset \mathbb{S}^n$ with coordinate patches and locally using the extension operator E from above, we get, for each finite N, an extension operator

$$E : H^k(\Omega) \to H^k(\mathbb{S}^n), \quad 0 \leq k \leq N - 1. \tag{12}$$

For real $s \geq 0$, we can use interpolation between Banach spaces (see [33, Chapter 4.2]) to define

$$E : H^s(\Omega) \to H^s(\mathbb{S}^n).$$

□

Theorem 2 (Trace Theorem) *Let $\Omega \subset \mathbb{S}^n$ be a local region with a sufficient smooth boundary. Then, for $s > 1/2$, the restriction of $f \in H^s(\Omega)$ to $\partial\Omega$ is well defined, belongs to $H^{s-1/2}(\partial\Omega)$, and satisfies*

$$\|f\|_{H^{s-1/2}(\partial\Omega)} \leq C\|f\|_{H^s(\Omega)}.$$

Proof The boundary ∂D_k of $D_k = \psi_k(\Omega \cap U_k)$ is given by $\psi_k(\partial\Omega \cap U_k)$ for $k = 1, 2$. Then,

$$\|f\|^2_{H^{s-1/2}(\partial\Omega)} = \sum_{k=1}^{2} \|(\pi_k \chi_k f)|_{\partial D_k}\|^2_{H^{s-1/2}(\partial D_k)}.$$

Using the trace theorem for bounded domains in $\mathbb{R}^n$ [37, Theorem 8.7], there are constants $c_k > 0$ for $k = 1, 2$ so that

$$\|(\pi_k \chi_k f)|_{\partial D_k}\|_{H^{s-1/2}(\partial D_k)} \leq c_k \|(\pi_k \chi_k f)|_{D_k}\|_{H^s(D_k)}.$$

Hence

$$\|f\|^2_{H^{s-1/2}(\partial\Omega)} \leq \max\{c_1^2, c_2^2\} \sum_{k=1}^{2} \|(\pi_k \chi_k f)|_{D_k}\|^2_{H^s(D_k)} = \max\{c_1^2, c_2^2\}\|f\|^2_{H^s(\Omega)}.$$

□

2.2 Positive Definite Kernels on the Sphere and Native Spaces

A continuous real-valued kernel $\phi : \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{R}$ is called *positive definite* on $\mathbb{S}^n$ if (1) $\phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{y},\boldsymbol{x})$ for all $\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^n$ and (2) for every finite set of distinct points $X = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ on $\mathbb{S}^n$, the symmetric matrix $[\phi(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j=1,2,\ldots,N}$ is positive definite.

A kernel $\phi : \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{R}$ defined via $\phi(\boldsymbol{x},\boldsymbol{y}) := \Phi(\boldsymbol{x}\cdot\boldsymbol{y})$, $\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^n$, with a univariate function Φ, is called a *zonal* kernel.

Since the normalized Legendre polynomials $\{P_\ell(n+1;\cdot)\}_{\ell\in\mathbb{N}_0}$, form a complete orthogonal system for $L_2([-1,1];(1-t^2)^{(n-2)/2})$, any function $\Phi \in L_2([-1,1];(1-t^2)^{(n-2)/2})$ can be expanded into a *Legendre series* (see (3) for the normalization)

$$\Phi(t) = \frac{1}{\omega_n}\sum_{\ell=0}^{\infty} a_\ell Z(n,\ell) P_\ell(n+1;t), \tag{13}$$

with the Legendre coefficients

$$a_\ell := \omega_{n-1}\int_{-1}^{+1} \Phi(t) P_\ell(n+1;t)(1-t^2)^{(n-2)/2}\mathrm{d}t.$$

Due to (13) and the addition theorem (2), a zonal kernel $\phi(\boldsymbol{x},\boldsymbol{y}) := \Phi(\boldsymbol{x}\cdot\boldsymbol{y})$, $\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^n$, where $\Phi \in L_2([-1,1];(1-t^2)^{(n-2)/2})$, has the expansion

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\omega_n}\sum_{\ell=0}^{\infty} a_\ell Z(n,\ell) P_\ell(n+1;\boldsymbol{x}\cdot\boldsymbol{y}) = \sum_{\ell=0}^{\infty}\sum_{k=1}^{Z(n,\ell)} a_\ell Y_{\ell,k}(\boldsymbol{x}) Y_{\ell,k}(\boldsymbol{y}). \tag{14}$$

In this paper we will only consider *positive definite zonal continuous kernels* ϕ of the form (14) for which

$$\sum_{\ell=0}^{\infty} |a_\ell| Z(n,\ell) < \infty. \tag{15}$$

This condition implies that the sums in (14) converge uniformly.

In [3], a complete characterization of positive definite kernels is established: a kernel ϕ of the form (14) satisfying the condition (15) is positive definite *if and only if* $a_\ell \geq 0$ for all $\ell \in \mathbb{N}_0$ and $a_\ell > 0$ for infinitely many even values of ℓ and infinitely many odd values of ℓ (see also [32] and [38]).

With each positive definite zonal continuous kernel ϕ of the form (14) and satisfying the condition (15), we associate a *native space*: Consider the linear space

$$F_\phi := \left\{ \sum_{j=1}^{N} \alpha_j \phi(\cdot,\boldsymbol{x}_j) \ : \ \alpha_j \in \mathbb{R},\ \boldsymbol{x}_j \in \mathbb{S}^n,\ j = 1,2,\ldots,N;\ N \in \mathbb{N} \right\},$$

endowed with the inner product

$$\langle \sum_{j=1}^{N} \alpha_j \phi(\cdot, \boldsymbol{x}_j), \sum_{i=1}^{M} \beta_i \phi(\cdot, \boldsymbol{y}_i) \rangle_\phi := \sum_{j=1}^{N} \sum_{i=1}^{M} \alpha_j \beta_i \phi(\boldsymbol{x}_j, \boldsymbol{y}_i)$$

and the associated norm $\|f\|_\phi := \langle f, f \rangle_\phi^{1/2}$. The *native space* $\mathcal{N}_\phi$ associated with ϕ is now defined as the completion of F_ϕ with respect to the norm $\| \cdot \|_\phi$. By construction, the native space $\mathcal{N}_\phi$ is a Hilbert space, and we will denote its inner product and norm also by $\langle \cdot, \cdot \rangle_\phi$ and $\| \cdot \|_\phi$, respectively.

The native space $\mathcal{N}_\phi$ is a (real) *reproducing kernel Hilbert space* with the reproducing kernel ϕ. This means that (1) ϕ is symmetric, (2) $\phi(\cdot, \boldsymbol{y}) \in \mathcal{N}_\phi$ for all (fixed) $\boldsymbol{y} \in \mathbb{S}^n$, and (3) the *reproducing property* holds, that is,

$$\langle f, \phi(\cdot, \boldsymbol{y}) \rangle_\phi = f(\boldsymbol{y}), \qquad \text{for all } f \in \mathcal{N}_\phi \text{ and all } \boldsymbol{y} \in \mathbb{S}^n. \tag{16}$$

It is known that [28, 35] the native space $\mathcal{N}_\phi$ associated with a positive definite continuous zonal kernel ϕ, given by (14) and satisfying the conditions (15) and $a_\ell > 0$ for all $\ell \in \mathbb{N}_0$, can be described by

$$\mathcal{N}_\phi = \left\{ f \in L_2(\mathbb{S}^n) \, : \, \sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(n,\ell)} \frac{|\widehat{f}_{\ell,k}|^2}{a_\ell} < \infty \right\},$$

equipped with the inner product

$$\langle f, g \rangle_\phi = \sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(n,\ell)} \frac{\widehat{f}_{\ell,k} \widehat{g}_{\ell,k}}{a_\ell}$$

and the associated norm

$$\|f\|_\phi = \langle f, f \rangle_\phi^{1/2} = \left(\sum_{\ell=0}^{\infty} \sum_{k=1}^{Z(n,\ell)} \frac{|\widehat{f}_{\ell,k}|^2}{a_\ell} \right)^{1/2}. \tag{17}$$

If $a_\ell > 0$ for all $\ell \in \mathbb{N}_0$, we can conclude, from the assumption (15), that the Fourier series of any $f \in \mathcal{N}_\phi$ converges uniformly and that the native space $\mathcal{N}_\phi$ is embedded into $C(\mathbb{S}^n)$.

Comparing (17) with (4), we see that if $a_\ell \sim (1 + \lambda_\ell)^{-s}$, then $\| \cdot \|_\phi$ and $\| \cdot \|_{H^s(\mathbb{S}^n)}$ are equivalent norms, and hence $\mathcal{N}_\phi$ and $H^s(\mathbb{S}^n)$ are the same space.

2.3 Generalized Interpolation with RBFs

Let $\phi : \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{R}$ be a positive definite zonal continuous kernel given by (14) and satisfying the condition (15). Since the native space $\mathcal{N}_\phi$ is a reproducing kernel Hilbert space with reproducing kernel ϕ, any continuous linear functional $\mathcal{L}$ on $\mathcal{N}_\phi$ has the representer $\mathcal{L}_2\phi(\cdot,\cdot)$. (Here the index 2 in $\mathcal{L}_2\phi(\cdot,\cdot)$ indicates that $\mathcal{L}$ is applied to the kernel ϕ as a function of its second argument. Likewise $\mathcal{L}_1\phi(\cdot,\cdot)$ will indicate that $\mathcal{L}$ is applied to the kernel ϕ as a function of its first argument.)

For a linearly independent set $\Xi = \{\mathcal{L}^1, \mathcal{L}^2, \ldots, \mathcal{L}^N\}$ of continuous linear functionals on $\mathcal{N}_\phi$, the *generalized radial basis function (RBF) interpolation problem* can be formulated as follows: Given the values $\mathcal{L}^1 f, \mathcal{L}^2 f, \ldots, \mathcal{L}^N f$ of a function $f \in \mathcal{N}_\phi$, find the function $\Lambda_\Xi f$ in the N-dimensional approximation space

$$V_\Xi := \operatorname{span}\left\{\mathcal{L}_2^j\phi(\cdot,\cdot) \,:\, j = 1, 2, \ldots, N\right\}$$

such that the conditions

$$\mathcal{L}^i(\Lambda_\Xi f) = \mathcal{L}^i f, \qquad i = 1, 2, \ldots, N, \tag{18}$$

are satisfied. We will call the function $\Lambda_\Xi f \in V_\Xi$ the *radial basis function approximant (RBF approximant)* of f.

Writing the RBF approximant $\Lambda_\Xi f$ as

$$\Lambda_\Xi f(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \mathcal{L}_2^j \phi(\boldsymbol{x}, \cdot), \qquad \boldsymbol{x} \in \mathbb{S}^n,$$

the interpolation conditions (18) can therefore be written as

$$\sum_{j=1}^{N} \alpha_j \langle \mathcal{L}_2^j \phi(\cdot,\cdot), \mathcal{L}_2^i \phi(\cdot,\cdot)\rangle_\phi = \sum_{j=1}^{N} \alpha_j \mathcal{L}_1^i \mathcal{L}_2^j \phi(\cdot,\cdot) = \mathcal{L}^i f,$$
$$i = 1, 2, \ldots, N. \tag{19}$$

Since $f \in \mathcal{N}_\phi$, we have $\mathcal{L}^i f = \langle f, \mathcal{L}_2^i \phi(\cdot,\cdot)\rangle_\phi$, $i = 1, 2, \ldots, N$, and we see that $\Lambda_\Xi f$ is just the *orthogonal projection* of $f \in \mathcal{N}_\phi$ onto the approximation space V_Ξ with respect to $\langle\cdot,\cdot\rangle_\phi$. Therefore,

$$\|f - \Lambda_\Xi f\|_\phi \le \|f\|_\phi. \tag{20}$$

The linear system has always a unique solution, because its matrix

$$[\mathcal{L}_1^i \mathcal{L}_2^j \phi(\cdot,\cdot)]_{i,j=1,2,\ldots,N}$$

is the Gram matrix of the representers of the linearly independent functionals in Ξ.

We observe here that the linear system (19) can be solved for any given data set $\{\mathscr{L}^i f \,:\, i = 1, 2, \ldots, N\}$, where the data does not necessarily has to come from a function in the native space $\mathscr{N}_\phi$, but may come from any function f for which $\mathscr{L}^i f$ is well-defined for all $i = 1, 2, \ldots, N$. Even if f is not in the native space we will use the notation $\Lambda_\Xi f$ for the solution of the generalized RBF interpolation problem (18).

2.4 Sobolev Bounds for Functions with Scattered Zeros

We need the following results from [14] concerning functions with scattered zeros on a subdomain of a Riemannian manifold.

Theorem 3 *Let $\mathbb{M}$ be a Riemannian manifold, $\Omega \subset \mathbb{M}$ be a bounded, Lipschitz domain that satisfies a certain uniform cone condition. Let X be a discrete set with sufficiently small mesh norm h. If $u \in W_p^m(\Omega)$ with $m > d/p$ satisfies $u|_X = 0$, then we have*

$$\|u\|_{W_p^k(\Omega)} \le C_{m,k,p,\mathbb{M}} h^{m-k} \|u\|_{W_p^m(\Omega)}$$

and

$$\|u\|_{L^\infty(\Omega)} \le C_{m,k,p,\mathbb{M}} h^{m-d/p} \|u\|_{W_p^m(\Omega)}.$$

3 Boundary Value Problems on the Sphere

After all these preparations we can formulate a boundary value problem for an elliptic differential operators L. Our standard application (and numerical example in Section Numerical Experiments) will be $L = \kappa^2 I - \Delta^*$, where I is the identity operator and κ is some fixed constant, on simply connected subregion Ω on $\mathbb{S}^n$ with a Lipschitz boundary $\partial\Omega$. This partial differential equation occurs, for example, when solving the heat equation and the wave equation with separation of variables (for $\kappa \neq 0$) or in studying the vortex motion on the sphere (for $\kappa = 0$).

Let $s > 2$, and let Ω be a simply connected subregion with a Lipschitz boundary. Assume that the functions $f \in W_2^{s-2}(\Omega)$ and $g \in C(\partial\Omega)$ are given. We consider the following Dirichlet problem

$$Lu = f \text{ on } \Omega \qquad \text{and } u = g \text{ on } \partial\Omega. \tag{21}$$

The existence and uniqueness of the solution to (21) follows from the general theory of existence and uniqueness of the solution to Dirichlet problems defined on Lipschitz domains in a Riemannian manifold [25].

Lemma 2 *Let $n \geq 2$, and let Ω be a sub-domain on $\mathbb{S}^n$ with a Lipschitz boundary. Let $L = \kappa^2 I - \Delta^*$ for some fixed constant $\kappa \geq 0$ and let $s \geq 2 + n/2$. Then L has the following properties:*

(i) There exists a positive constant c such that

$$\|Lf\|_{H^{s-2}(\Omega)} \leq c\|f\|_{H^s(\Omega)}.$$

(ii) There exists a positive constant c such that

$$\langle Lf, f\rangle_{L_2(\Omega)} \geq c\|f\|^2_{L_2(\Omega)}$$

for all $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ with $f = 0$ on $\partial\Omega$.

(iii) For all $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ which satisfy $Lf = 0$ on Ω, we have

$$\|f\|_{C(\overline{\Omega})} \leq \|f\|_{C(\partial\Omega)}.$$

Proof

(i) Suppose $s = m$, where m is an integer. Using definition (9) and the fact that $\Delta^* = -\nabla^*\nabla$, where ∇^* denote the surface divergent on the sphere, we have

$$\begin{aligned}
\|Lu\|^2_{W_2^{m-2}(\Omega)} &= \sum_{k=0}^{m-2} \langle \nabla^k Lu, \nabla^k Lu\rangle_{L_2(\Omega)} \\
&= \sum_{k=0}^{m-2} \langle \nabla^k(\kappa^2 u - \Delta^* u), \nabla^k(\kappa^2 u - \Delta^* u)\rangle_{L_2(\Omega)} \\
&= \sum_{k=0}^{m-2} \kappa^4 \langle \nabla^k u, \nabla^k u\rangle_{L_2(\Omega)} - 2\kappa^2 \langle \nabla^{k+1} u, \nabla^{k+1} u\rangle_{L_2(\Omega)} \\
&\qquad + \langle \nabla^{k+2} u, \nabla^{k+2} u\rangle_{L_2(\Omega)} \\
&\leq \max\{\kappa^4, 2\kappa^2, 1\} \sum_{k=0}^{m} \langle \nabla^k u, \nabla^k u\rangle_{L_2(\Omega)} \\
&\leq C\|u\|^2_{W_2^s(\Omega)}.
\end{aligned}$$

The case that s is a real number follows from interpolation between bounded operators.

(ii) With the assumption on s, the Sobolev imbedding theorem for functions defined on Riemannian manifolds [15, p.34] implies that $W_2^s(\Omega) \subset C^2(\Omega)$.

From Green's first surface identity [10, (1.2.49)], or more generally, the first Green's formula for compact, connected, and oriented manifolds in $\mathbb{R}^{n+1}$ [1, p.84], we find for any $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ with $f = 0$ on $\partial\Omega$ that

$$\begin{aligned}\langle(\kappa^2 - \Delta^*)f, f\rangle_{L_2(\Omega)} &= \kappa^2\|f\|^2_{L_2(\Omega)} - \langle\Delta^* f, f\rangle_{L_2(\Omega)} \\ &= \kappa^2\|f\|^2_{L_2(\Omega)} + \|\nabla f\|^2_{L_2(\Omega)} - \int_{\partial\Omega} f(\boldsymbol{x})\frac{\partial f(\boldsymbol{x})}{\partial\nu}\mathrm{d}\sigma(\boldsymbol{x}) \\ &= \kappa^2\|f\|^2_{L_2(\Omega)} + \|\nabla f\|^2_{L_2(\Omega)},\end{aligned}$$

where ∇ is the surface gradient, ν the (external) unit normal on the boundary $\partial\Omega$, and $\mathrm{d}\sigma$ the curve element of the boundary (curve) $\partial\Omega$. From the Poincaré inequality for a bounded domain on a Riemannian manifold [31],

$$\|\nabla f\|_{L_2(\Omega)} \geq c\|f\|_{L_2(\Omega)}$$

for all $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ with $f = 0$ on $\partial\Omega$. Thus

$$\langle(\kappa^2 - \Delta^*)f, f\rangle_{L_2(\Omega)} \geq (c + \kappa^2)\|f\|^2_{L_2(\Omega)},$$

from which property (ii) is proved.

(iii) The property (iii) follows from the maximum principle for elliptic PDEs on manifolds. From [29, Theorem 9.3], we know that every $g \in C^1(\Omega)$ which satisfies

$$\Delta^* g - \kappa^2 g \leq 0 \quad \text{on } \Omega \qquad \text{and} \qquad g \geq 0 \quad \text{on } \Omega$$

in a distributional sense satisfies the *strong maximum principle*, that is, if $g(\boldsymbol{y}_0) = 0$ for some $\boldsymbol{y}_0 \in \Omega$ then $g \equiv 0$ in Ω. In particular, this implies if $g \in C^1(\Omega) \cap C(\overline{\Omega})$ that g assumes its zeros on the boundary.

In our case $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$, and since $W_2^s(\Omega) \subset C^2(\Omega)$, we consider (twice differentiable) classical solutions of $\kappa^2 f - \Delta^* f = 0$. From the strong maximum principle we may conclude that every $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ that satisfies $\kappa^2 f - \Delta^* f = 0$ has the property

$$\sup_{\boldsymbol{x}\in\overline{\Omega}} |f(\boldsymbol{x})| = \sup_{\boldsymbol{x}\in\partial\Omega} |f(\boldsymbol{x})|, \tag{22}$$

which establishes property (iii) in the Theorem.

This can be seen as follows: Consider $f \in W_2^s(\Omega) \cap C(\overline{\Omega})$ that satisfies $\kappa^2 f - \Delta^* f = 0$. Let $\boldsymbol{y}_1 \in \overline{\Omega}$ and $\boldsymbol{y}_2 \in \overline{\Omega}$ be such that

$$f(\boldsymbol{y}_1) = \min_{\boldsymbol{y}\in\overline{\Omega}} f(\boldsymbol{y}) \leq f(\boldsymbol{x}) \leq \max_{\boldsymbol{y}\in\overline{\Omega}} f(\boldsymbol{y}) = f(\boldsymbol{y}_2) \qquad \text{for all } \boldsymbol{x} \in \overline{\Omega}.$$

Then

$$\sup_{x\in\overline{\Omega}} |f(\boldsymbol{x})| = \begin{cases} f(\boldsymbol{y}_2) & \text{if } f \geq 0 \text{ on } \overline{\Omega}, \\ -f(\boldsymbol{y}_1) & \text{if } f \leq 0 \text{ on } \overline{\Omega}, \\ \max\{-f(\boldsymbol{y}_1), f(\boldsymbol{y}_2)\} & \text{if } f \text{ assumes negative and positive values} \end{cases} \tag{23}$$

If $f(\boldsymbol{y}_1) \leq 0$, consider $g_1(\boldsymbol{x}) := f(\boldsymbol{x}) - f(\boldsymbol{y}_1)$. Then $g_1(\boldsymbol{y}_1) = 0$ and $g_1(\boldsymbol{x}) \geq 0$ on $\overline{\Omega}$, and we have

$$(\Delta^* - \kappa^2)g_1 = (\Delta^* - \kappa^2)f + \kappa^2 f(\boldsymbol{y}_1) = \kappa^2 f(\boldsymbol{y}_1) \leq 0.$$

Thus the strong maximum principle implies that g_1 assumes its zeros on the boundary and hence $\boldsymbol{y}_1 \in \partial\Omega$. If $f(\boldsymbol{y}_2) \geq 0$, consider $g_2(\boldsymbol{x}) := f(\boldsymbol{y}_2) - f(\boldsymbol{x})$. Then $g_2(\boldsymbol{y}_2) = 0$ and $g_2(\boldsymbol{x}) \geq 0$ on $\overline{\Omega}$, and we find

$$(\Delta^* - \kappa^2)g_2 = -\kappa^2 f(\boldsymbol{y}_2) - (\Delta^* - \kappa^2)f = -\kappa^2 f(\boldsymbol{y}_2) \leq 0.$$

Thus the strong maximum principle implies that g_2 assumes its zeros on the boundary and hence $\boldsymbol{y}_2 \in \partial\Omega$. Thus (23) implies (22). □

We now discuss a method to construct an approximate solution to the Dirichlet problem (21) using radial basis functions. Assume that the values of the functions f and g are given on the discrete sets $X_1 := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\} \subset \Omega$ and $X_2 := \{\boldsymbol{x}_{M+1}, \ldots, \boldsymbol{x}_N\} \subset \partial\Omega$, respectively. Furthermore, assume that the local mesh norm $h_{X_1,\Omega}$ of X_1 and the mesh norm $h_{X_2,\partial\Omega}$ of X_2 along the boundary $\partial\Omega$ (see (1) below) are sufficiently small. We wish to find an approximation of the solution $u \in W_2^s(\Omega) \cap C(\overline{\Omega})$ of the *Dirichlet boundary value problem*

$$Lu = f \quad \text{on } \Omega \qquad \text{and} \qquad u = g \quad \text{on } \partial\Omega.$$

Let $\Xi = \Xi_1 \cup \Xi_2$ with $\Xi_1 := \{\delta_{\boldsymbol{x}_j} \circ L : j = 1, 2, \ldots, M\}$ and $\Xi_2 := \{\delta_{\boldsymbol{x}_j} : j = M+1, \ldots, N\}$.

We choose an RBF ϕ such that $\mathcal{N}_\phi = H^s(\mathbb{S}^n)$ for some $s > 2 + \lfloor n/2 + 1 \rfloor$. Under the assumption that Ξ is a set of linearly independent functionals, we compute the RBF approximant $\Lambda_\Xi u$, defined by

$$\Lambda_\Xi u = \sum_{j=1}^{M} \alpha_j L_2 \phi(\cdot, \boldsymbol{x}_j) + \sum_{j=M+1}^{N} \alpha_j \phi(\cdot, \boldsymbol{x}_j), \tag{24}$$

in which the coefficients α_j, for $j = 1, \ldots, N$, are computed from the *collocation conditions*

$$L(\Lambda_\Xi u)(\boldsymbol{x}_j) = f(\boldsymbol{x}_j), \qquad j = 1, 2, \ldots, M, \tag{25}$$

$$\Lambda_\Xi u(\boldsymbol{x}_j) = g(\boldsymbol{x}_j), \qquad j = M+1, \ldots, N. \tag{26}$$

We want to derive $L_2(\Omega)$-error estimates between the approximation and the exact solution, which is stated in the following theorem.

Theorem 4 *Let $n \geq 2$, and let Ω be a simply connected sub-domain on $\mathbb{S}^n$ with a Lipschitz boundary. Let $L = \kappa^2 I - \Delta^*$ for some fixed constant $\kappa \geq 0$ and let $s \geq 2 + \lfloor n/2 + 1 \rfloor$. Consider the Dirichlet boundary value problem*

$$Lu = f \quad \text{on } \Omega \qquad \text{and} \qquad u = g \quad \text{on } \partial\Omega,$$

where we assume that the unknown solution u is in $W_2^s(\Omega) \cap C(\overline{\Omega})$ and that $f \in W_2^{s-2}(\Omega)$ and $g \in C(\partial\Omega)$. Assume that f is given on the point set $X_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\} \subset \Omega$ with sufficiently small local mesh norm $h_{X_1,\Omega}$, and suppose that g is given on the point set $X_2 = \{\boldsymbol{x}_{M+1}, \ldots, \boldsymbol{x}_N\} \subset \partial\Omega$ with sufficiently small mesh norm $h_{X_2,\partial\Omega}$. Let ϕ be a positive definite zonal continuous kernel of the form (14) for which

$$a_\ell \sim (1 + \lambda_\ell)^{-s}. \tag{27}$$

Let $\Lambda_\Xi u$ denote the RBF approximant (24) which satisfies the collocation conditions (25) and (26). Then

$$\|u - \Lambda_\Xi u\|_{L_2(\Omega)} \leq c \max\{h_{X_1,\Omega}^{s-2}, h_{X_2,\partial\Omega}^{s-n/2}\} \|u\|_{W_2^s(\Omega)}. \tag{28}$$

Our general approach follows the one discussed in [8, 9], and in [35, Chapter 16] for the case of boundary problems on subsets of $\mathbb{R}^n$. In contrast to the approach in [35, Chapter 16], where the error analysis is based on the power function, we also use the results on functions with scattered zeros (see Theorem 3) locally via the charts. Similar problems on bounded flat domains were analyzed in [12] using results on functions with scattered zeros.

Proof

Step 1 First we prove the following inequality using the ideas from [9, Theorem 5.1].

$$\|u - \Lambda_\Xi u\|_{L_2(\Omega)} \leq c \|Lu - L(\Lambda_\Xi u)\|_{L_2(\Omega)} + |\Omega|^{1/2} \|u - \Lambda_\Xi u\|_{C(\partial\Omega)}, \tag{29}$$

where c is a constant from Lemma 2 and $|\Omega|$ denotes the volume of the domain Ω. Since the boundary value problem has a unique solution, there exists a function $w \in W_2^s(\Omega) \cap C(\overline{\Omega})$ such that

$$Lw = Lu \quad \text{on } \Omega \qquad \text{and} \qquad w = \Lambda_\Xi u \quad \text{on } \partial\Omega. \tag{30}$$

From the triangle inequality,

$$\|u - \Lambda_\Xi u\|_{L_2(\Omega)} \leq \|u - w\|_{L_2(\Omega)} + \|w - \Lambda_\Xi u\|_{L_2(\Omega)} \tag{31}$$

Since $L(u-w) = 0$ on Ω (from (30)), the property (iii) from Lemma 2 and (30) imply

$$\begin{aligned} \|u-w\|_{L_2(\Omega)} &\leq |\Omega|^{1/2}\|u-w\|_{C(\overline{\Omega})} \\ &\leq |\Omega|^{1/2}\|u-w\|_{C(\partial\Omega)} = |\Omega|^{1/2}\|u-\Lambda_\Xi u\|_{C(\partial\Omega)}. \end{aligned} \tag{32}$$

Since $w-\Lambda_\Xi u = 0$ on $\partial\Omega$ (from (30)), the property (ii) from Lemma 2 and the Cauchy-Schwarz inequality yield that

$$\begin{aligned} \|w-\Lambda_\Xi u\|^2_{L_2(\Omega)} &\leq c\langle L(w-\Lambda_\Xi u), w-\Lambda_\Xi u\rangle_{L_2(\Omega)} \\ &\leq c\|L(w-\Lambda_\Xi u)\|_{L_2(\Omega)}\|w-\Lambda_\Xi u\|_{L_2(\Omega)}, \end{aligned}$$

thus implying

$$\|w-\Lambda_\Xi u\|_{L_2(\Omega)} \leq c\|Lw - L(\Lambda_\Xi u)\|_{L_2(\Omega)} = c\|Lu - L(\Lambda_\Xi u)\|_{L_2(\Omega)}, \tag{33}$$

where we have used $Lw = Lu$ on Ω in the last step. Applying (32) and (33) in (31) gives (29).

Step 2 In this step, we will estimate the first term in the right hand side of (29). By using Theorem 3, we obtain

$$\begin{aligned} \|Lu - L(\Lambda_\Xi u)\|_{L_2(\Omega)} &\leq ch^{s-2}_{X_1,\Omega}\|Lu - L(\Lambda_\Xi u)\|_{W_2^{s-2}(\Omega)} \\ &\leq ch^{s-2}_{X_1,\Omega}\|u-\Lambda_\Xi u\|_{W_2^s(\Omega)}, \end{aligned} \tag{34}$$

where we have used the fact that $\|Lg\|_{W_2^{s-2}(\Omega)} \leq C\|g\|_{W^s(\Omega)}$, see Lemma 2 part (i).

Next, our assumptions on the region Ω allow us to extend the function $u \in W_2^s(\Omega)$ to a function $Eu \in W_2^s(\mathbb{S}^n)$. Moreover, since $X \subset \Omega$ and $Eu|_\Omega = u|_\Omega$, the generalized interpolant $\Lambda_\Xi u$ coincides with the generalized interpolant $\Lambda_\Xi(Eu)$ on Ω. Finally, the Sobolev space norm on $W_2^s(\mathbb{S}^n)$ is equivalent to the norm induced by the kernel ϕ and the generalized interpolant is norm-minimal. This all gives

$$\begin{aligned} \|u-\Lambda_\Xi u\|_{W_2^s(\Omega)} &= \|Eu - \Lambda_\Xi Eu\|_{W_2^s(\Omega)} \leq \|Eu-\Lambda_\Xi Eu\|_{W_2^s(\mathbb{S}^n)} \\ &\leq \|Eu\|_{W_2^s(\mathbb{S}^n)} \leq C\|u\|_{W_2^s(\mathbb{S}^n)}, \end{aligned} \tag{35}$$

which establishes the stated interior error estimate.

Step 3 In this step, we will estimate the second term in the right hand side of (29). For the boundary estimate, by using Theorem 3 for $\partial\Omega$, which is manifold of dimension $n-1$, we obtain

$$\|u-\Lambda_\Xi u\|_{C(\partial\Omega)} \leq ch^{s-1/2-(n-1)/2}_{X_2,\partial\Omega}\|u-\Lambda_\Xi u\|_{W_2^{s-1/2}(\partial\Omega)}. \tag{36}$$

Using the trace theorem (Theorem 2) and (35), we have

$$\|u - \Lambda_{\Xi} u\|_{W_2^{s-1/2}(\partial\Omega)} \leq C\|u - \Lambda_{\Xi} u\|_{W_2^s(\Omega)} \leq C\|u\|_{W_2^s(\Omega)}. \tag{37}$$

The boundary estimate then follows from (36)–(37).

The $L_2(\Omega)$-error estimate (28) follows from combining (29), (34)–(35) and (36)–(37). □

4 Numerical Experiments

In this section, we consider the following boundary value problem on the spherical rectangular region Ω which is defined in terms of longitudes and latitudes as $[0, 50°N] \times [0, 100°E]$:

$$\begin{cases} Lu(\boldsymbol{x}) & := -\Delta^* u(\boldsymbol{x}) = \delta_{\boldsymbol{p}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) & = 0 \qquad\qquad\qquad\quad \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{38}$$

where

$$\delta_{\boldsymbol{p}}(\boldsymbol{x}) = \frac{1}{\epsilon\sqrt{\pi}} \exp(-(\cos^{-1}(\boldsymbol{p}\cdot\boldsymbol{x})/\epsilon)^2).$$

The example is motivated by the problem of vortex motion on a spherical shell with a solid boundary on its surface [21]. The flow due to a vortex of unit strength located at a point $\boldsymbol{p} \in \Omega$. In the original equation, the right hand side is the Dirac delta function centered at $\boldsymbol{p}$ which we will approximate by $\delta_{\boldsymbol{p}}$ with $\epsilon = 0.1$ and $\boldsymbol{p}$ is a fixed point in Ω with longitude/latitude $(25°N, 26°E)$.

Even though the algorithm allows the collocation points to be scattered freely on the sphere, choosing sets of collocation points distributed roughly uniformly over the whole sphere significantly improves the quality of the approximate solutions and condition numbers. To this end, the sets of points used to construct the approximate solutions are generated using a uniform partition adapted to a spherical rectangle.

The RBF used is Wendland's function [34]

$$\psi(r) = (1-r)_+^8(1 + 8r + 25r^2 + 32r^3)$$

Table 1 Interior errors between the approximation and the reference solution

M	h_{X_1}	$\|e\|$	EOC
264	0.0349	1.0204e−04	
1176	0.0175	3.3285e−06	4.9586
4851	0.0087	1.8440e−07	4.1397

and

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \psi(|\boldsymbol{x}-\boldsymbol{y}|) = \psi(\sqrt{2-2\boldsymbol{x}\cdot\boldsymbol{y}}).$$

It can be shown that ϕ is a kernel which satisfies condition (27) with $s = 9/2$ [28].

The kernel ϕ is a zonal function, i.e. $\phi(\boldsymbol{x},\boldsymbol{y}) = \Phi(\boldsymbol{x}\cdot\boldsymbol{y})$ where $\Phi(t)$ is a univariate function. For zonal functions, the Laplace-Beltrami operator can be computed via

$$\Delta^*\Phi(\boldsymbol{x}\cdot\boldsymbol{y}) = \mathscr{L}\Phi(t), \quad t = \boldsymbol{x}\cdot\boldsymbol{y},$$

where

$$\mathscr{L} = \frac{d}{dt}(1-t^2)\frac{d}{dt}$$

The normalized interior L_2 error $\|e\|$ is approximated by an ℓ_2 error, thus in principle we define

$$\|e\| := \left(\frac{1}{|\Omega|}\int_\Omega |u(\boldsymbol{x}) - \Lambda_\Xi u(\boldsymbol{x})|^2 d\boldsymbol{x}\right)^{1/2}$$

and in practice approximate this by the midpoint rule,

$$\left(\frac{1}{|\mathscr{G}|}\sum_{\boldsymbol{x}(\theta,\phi)\in\mathscr{G}} |u(\theta,\phi) - \Lambda_\Xi u(\theta,\phi)|^2 \sin\theta\right)^{1/2},$$

where $\mathscr{G}$ is a longitude-latitude grid in the interior of Ω containing the centers of rectangles of size 0.9° × 1.8° and $|\mathscr{G}| = 811$.

The errors are computed against a reference solution which are generated by using 30,876 interior points and 750 boundary points.

As can be seen from Table 1 the numerical results show a better convergence rate predicted by Theorem 4 (Figs. 1 and 2).

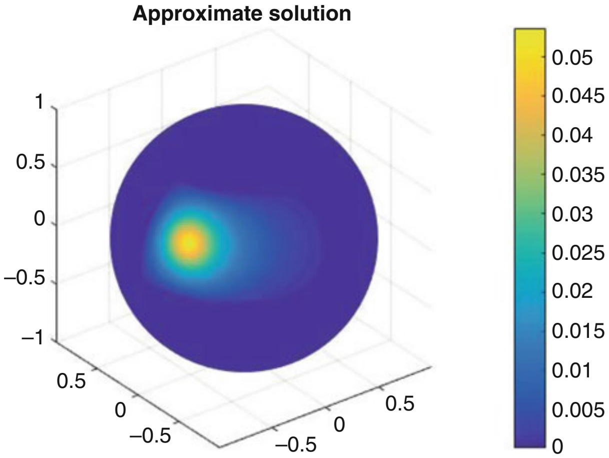


Fig. 1 Approximate solution with $M = 4851$ interior points and $N = 5151$ (total number of points)

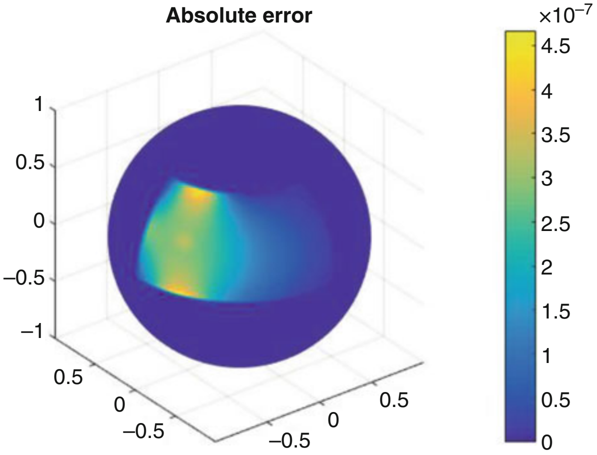


Fig. 2 Absolute errors with $M = 4851$ and $N = 5151$

Acknowledgements The author is grateful to many helpful discussions with Dr. Kerstin Hesse when writing the earlier version of the paper. He would also like to thank Professor Francis Narcowich for pointing out the recent results on Sobolev bounds for functions with scattered zeros on a Riemannian manifold. Many helpful comments from the referees to improve the presentation of the paper are also acknowledged.

References

1. Agricola, I., Friedrich, T.: Global Analysis: Differential forms in Analysis, Geometry and Physics. Graduate Studies in Mathematics, vol. 52. American Mathematical Society, Providence, RI (2002)
2. Chaos, Special Focus Issue: Large long-lived coherent structures out of chaos in planetary atmospheres and oceans. Chaos **4** (1994)
3. Chen, D., Menegatto, V.A., Sun, X.: A necessary and sufficient condition for strictly positive definite functions on spheres. Proc. Am. Math. Soc. **131**, 2733–2740 (2003)
4. Crowdy, D.: Point vortex motion on the surface of a sphere with impenetrable boundaries. Phys. Fluids **18**, 036602 (2006)
5. Fasshauer, G.E.: Solving differential equations with radial basis functions: multilevel methods and smoothing. Adv. Comput. Math. **11**, 139–159 (1999)
6. Flyer, N., Wright, G.: Transport schemes on a sphere using radial basis functions. J. Comput. Phys. **226**, 1059–1084 (2007)
7. Flyer, N., Wright, G.: A radial basis function method for the shallow water equations on a sphere. Proc. R. Soc. A **465**, 1949–1976 (2009)
8. Franke, C., Schaback, R.: Convergence order estimates of meshless collocation methods using radial basis functions. Adv. Comput. Math. **8**, 381–399 (1998)
9. Franke, C., Schaback, R.: Solving partial differential equations by collocation using radial basis functions. Appl. Math. Comput. **93**, 73–82 (1998)
10. Freeden, W., Gervens, T., Schreiner, M.: Constructive Approximation on the Sphere (with Applications to Geomathematics). Clarendon, Oxford (1998)
11. Gemmrich, S., Nigam, N., Steinbach, O.: Boundary integral equations for the Laplace-Beltrami operator. In: Munthe-Kaas, H., Owren, B. (eds.) Mathematics and Computation, a Contemporary View. Proceedings of the Abel Symposium 2006, vol. 3, pp. 21–37. Springer, Heidelberg (2006)
12. Giesl, P., Wendland, H.: Meshless collocation: error estimates with application to dynamical systems. SIAM J. Numer. Anal. **45**, 1723–1741 (2007)
13. Gill, A.E.: Atmosphere-Ocean Dynamics. International Geophysics Series, vol. 30. Academic, New York (1982)
14. Hangelbroek, T., Narcowich, F.J., Ward, J.D.: Polyharmonic and related kernels on manifolds: interpolation and approximation. Found. Comput. Math. **12**, 625–670 (2012)
15. Hebey, E.: Nonlinear Analysis on Manifolds: Sobolev Spaces and Inequalities. Courant Lecture Notes in Mathematics. American Mathematical Society, Providence, RI (2000)
16. Hon, Y.C., Mao, X.Z.: An efficient numerical scheme for Burgers' equation. Appl. Math. Comput. **95**, 37–50 (1998)
17. Hon, Y.C., Schaback, R.: On unsymmetric collocation by radial basis functions. Appl. Math. Comput. **119**, 177–186 (2001)
18. Hubbert, S., Morton, T.M.: A Duchon framework for the sphere. J. Approx. Theory **129**, 28–57 (2004)
19. Kansa, E.J.: Multiquadrics - a scattered data approximation scheme with applications to computational fluid-dynamics I. Comput. Math. **19**, 127–145 (1990)
20. Kansa, E.J.: Multiquadrics - a scattered data approximation scheme with applications to computational fluid-dynamics II: solutions to parabolic, hyperbolic and elliptic partial differential equations. Comput. Math. **19**, 147–161 (1990)
21. Kidambi, R., Newton, P.K.: Point vortex motion on a sphere with solid boundaries. Phys. Fluids **12**(3), 581–588 (2000)
22. Le Gia, Q.T.: Galerkin approximation for elliptic PDEs on spheres. J. Approx. Theory **130**, 123–147 (2004)
23. Le Gia, Q.T., Narcowich, F.J., Ward, J.D., Wendland, H.: Continuous and discrete least-squares approximation by radial basis functions on spheres. J. Approx. Theory **143**, 124–133 (2006)

24. Lions, J.L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vol. I. Springer, New York (1972)
25. Mitrea, M., Taylor, M.: Boundary layer methods for Lipschitz domains in Riemannian manifolds. J. Funct. Anal. **163**, 181–251 (1999)
26. Morton, T.M., Neamtu, M.: Error bounds for solving pseudodifferential equations on spheres by collocation with zonal kernels. J. Approx. Theory **114**, 242–268 (2002)
27. Müller, C.: Spherical Harmonics. Lecture Notes in Mathematics, vol. 17. Springer, New York (1966)
28. Narcowich, F.J., Ward, J.D.: Scattered data interpolation on spheres: error estimates and locally supported basis functions. SIAM J. Math. Anal. **33**(6), 1393–1410 (2002)
29. Pucci, P., Serrin, J.: Review: the strong maximum principle revisited, J. Differ. Equ. **196**, 1–66 (2004)
30. Ratcliffe, J.G.: Foundations of Hyperbolic Manifolds. Springer, New York (1994)
31. Saloff-Coste, L: Pseudo-Poincaré inequalities and applications to Sobolev inequalities. In: Laptev, A. (ed.) Around the Research of Vladimir Maz'ya I, pp. 349–372. Springer, New York (2010)
32. Schoenberg, I.J.: Positive definite function on spheres, Duke Math. J. **9**, 96–108 (1942)
33. Taylor, M.E.: Partial Differential Equations I, 2nd edn. Springer, New York (2011)
34. Wendland, H: Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. **4**, 389–396 (1995)
35. Wendland, H.: Scattered Data Approximation. Cambridge University Press, Cambridge (2005)
36. Wendland, H: A high-order approximation method for semilinear parabolic equations on spheres. Math. Comput. **82**, 227–245 (2013)
37. Wloka, J.: Partial Differential Equations, Cambridge University Press, Cambridge (2005)
38. Xu, Y., Cheney, E.W.: Strictly positive definite functions on spheres. Proc. Am. Math. Soc. **116**, 977–981 (1992)

Approximate Boundary Null Controllability and Approximate Boundary Synchronization for a Coupled System of Wave Equations with Neumann Boundary Controls

Tatsien Li, Xing Lu, and Bopeng Rao

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In this paper, for a coupled system of wave equations with Neumann boundary controls, the approximate boundary null controllability, the approximate boundary synchronization and the approximate boundary synchronization by groups are taken into account, respectively. Like in the case with Dirichlet boundary controls, the corresponding conditions of compatibility, and the criteria of Kalman's type as necessary conditions are obtained. The sufficiency of Kalman's criteria is further discussed in one dimensional space.

T. Li (✉)
School of Mathematical Sciences, Fudan University, Shanghai, China

Shanghai Key Laboratory for Contemporary Applied Mathematics, Shanghai, China

Nonlinear Mathematical Modeling and Methods Laboratory, Shanghai, China
e-mail: dqli@fudan.edu.cn

X. Lu
School of Mathematical Sciences, Fudan University, Shanghai, China
e-mail: xinglu12@fudan.edu.cn

B. Rao
Institut de Recherche Mathématique Avancée, Université de Strasbourg, Strasbourg, France

School of Mathematical Sciences, Fudan University, Shanghai, China
e-mail: bopeng.rao@math.unistra.fr

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_37

1 Introduction and Preliminaries

Consider the following coupled system of wave equations with Neumann boundary controls:

$$\begin{cases} U'' - \Delta U + AU = 0 & \text{in } (0, +\infty) \times \Omega, \\ U = 0 & \text{on } (0, +\infty) \times \Gamma_0, \\ \partial_\nu U = DH & \text{on } (0, +\infty) \times \Gamma_1 \end{cases} \tag{1}$$

and the corresponding initial data

$$t = 0: \quad U = U_0, \quad U' = U_1 \quad \text{in} \quad \Omega, \tag{2}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded domain with smooth boundary $\Gamma = \Gamma_1 \cup \Gamma_0$, ∂_ν denotes the outward normal derivative on the boundary, $A = (a_{ij})$ is a matrix of order N, D is a full column-rank $N \times M (M \le N)$ matrix, both A and D having constant elements, $U = (u^{(1)}, \cdots, u^{(N)})^T$ and $H = (h^{(1)}, \cdots, h^{(M)})^T$ denote the state variables and the boundary controls, respectively.

Denote

$$\mathscr{H}_0 = L^2(\Omega), \quad \mathscr{H}_1 = H^1_{\Gamma_0}(\Omega), \quad \mathscr{L} = L^2(0, T; L^2(\Gamma_1)), \tag{3}$$

where $H^1_{\Gamma_0}(\Omega)$ is the subspace of $H^1(\Omega)$, composed of all the functions with null trace on Γ_0, while $T > 0$ is a given constant.

If $\Gamma_0 = \emptyset$, we modify the space $\mathscr{H}_0$ by $\mathscr{H}_0 = \{f \in L^2(\Omega) : \quad \int_\Omega f(x)dx = 0\}$. However, in order to simplify the presentation, we only consider the case that $\text{mes}(\Gamma_0) > 0$ in this paper. The hypothesis $\overline{\Gamma}_1 \cap \overline{\Gamma}_0 = \emptyset$ is necessary to guarantee the smoothness of solutions to differential equations with mixed boundary conditions. Moreover, we assume that Ω satisfies the usual geometric control condition [1], namely, there exists an $x_0 \in \mathbb{R}^n$, such that for $m = x - x_0$ we have

$$(m, \nu) \le 0, \quad \forall x \in \Gamma_0, \quad (m, \nu) > 0, \quad \forall x \in \Gamma_1, \tag{4}$$

where $(\cdot, \cdot)$ denotes the inner product in $\mathbb{R}^n$.

By Li and Rao [7], if the number of boundary controls is adequate, namely, if $M = N$, then for any given initial data $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, there exists a boundary control H in $\mathscr{L}^N$, such that system (1) is exactly boundary null controllable. However, if there is a lack of boundary controls: $M < N$, then system (1) does not possess the exact boundary null controllability. Thus, it is natural to ask whether there are some kinds of controllabilities in a weak sense when there is a lack of boundary controls under the geometric control condition. In [5] and [6], Li and Rao have proposed the concepts of the approximate boundary null controllability and the approximate boundary synchronization for a coupled system of wave equations with Dirichlet boundary controls and discussed them thoroughly. In this paper, we will take the consideration on the approximate boundary null controllability and the approximate boundary synchronization of the coupled system (1) with Neumann boundary controls, and will get similar results to that of the system with Dirichlet boundary controls.

It is well-known that the solution to the wave equation with Neumann boundary condition does not possess the hidden regularity on the boundary, which is different from that with Dirichlet boudary condition, it requires us to consider the inhomogeneous problem in a smoother function space of initial data. To be specific, the approximate controllability and the approximate synchronization no longer stay in function space of $L^2(\Omega) \times H^{-1}(\Omega)$ as in the case of Dirichlet boundary controls. Moreover, under Neumann boundary controls, up to now, the Kalman's criterion is known to be sufficient only for diagonalizable systems in the one-space-dimensional case (see Sect. 3). However, under Dirichlet boundary controls, this criterion is sufficient not only for diagonalizable systems in the one-space-dimensional case, but also for $N \times N$ cascade systems, and some specific 2×2 systems.

Let

$$\Phi = (\phi^{(1)}, \cdots, \phi^{(N)})^T.$$

Consider the following adjoint problem:

$$\begin{cases} \Phi'' - \Delta\Phi + A^T\Phi = 0 & \text{in } (0, +\infty) \times \Omega, \\ \Phi = 0 & \text{on } (0, +\infty) \times \Gamma_0, \\ \partial_\nu \Phi = 0 & \text{on } (0, +\infty) \times \Gamma_1, \\ t = 0: \quad \Phi = \Phi_0, \quad \Phi' = \Phi_1 & \text{in } \Omega, \end{cases} \tag{5}$$

where A^T is the transpose of A. Define the linear unbounded operator $-\Delta$ in $\mathscr{H}_0$ by

$$D(-\Delta) = \{\Phi \in H^2(\Omega) : \quad \Phi|_{\Gamma_0} = 0, \;\; \partial_\nu \Phi|_{\Gamma_1} = 0\}.$$

Clearly, $-\Delta$ is a positively definite self-adjoint operator, then, for any given $s \in \mathbb{R}$, we can define the operator $(-\Delta)^{\frac{s}{2}}$ with the domain

$$\mathscr{H}_s = D((-\Delta)^{\frac{s}{2}})$$

which, endowed with the norm $\|\Phi\|_s = \|(-\Delta)^{\frac{s}{2}}\Phi\|$ (where $\|\cdot\|$ is the norm in $\mathscr{L}^2(\Omega)$), constitutes a Hilbert space, and its dual space is $\mathscr{H}_s' = \mathscr{H}_{-s}$. In particular, we have

$$\mathscr{H}_1 = D(\sqrt{-\Delta}) = \{\Phi \in H^1(\Omega) : \quad \Phi|_{\Gamma_0} = 0\}.$$

Let $C^0_{loc}([0, +\infty); \mathscr{H}_s)$ stand for the space of continuous functions of t, defined on $[0, +\infty)$ with the values in $\mathscr{H}_s$, and equipped with the uniform norm for any finite $T > 0$:

$$\|f\|_{C^0([0,T];\mathscr{H}_s)} = \sup_{0\le t\le T} \|f(t)\|_{\mathscr{H}_s}. \tag{6}$$

Similarly, let $C^1_{loc}([0,+\infty);\mathscr{H}_{s-1})$ stand for the space of continuously differentiable functions of t, defined on $[0,+\infty)$ with the values and the derivatives in $\mathscr{H}_{s-1}$, and equipped with the uniform norm for any finite $T > 0$:

$$\|f\|_{C^1([0,T];\mathscr{H}_{s-1})} = \sup_{0\le t\le T} \big(\|f(t)\|_{\mathscr{H}_{s-1}} + \|f'(t)\|_{\mathscr{H}_{s-1}}\big). \tag{7}$$

By Li and Rao [7], we have

Lemma 1 *For any given initial data $(\Phi_0,\Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$ with $s \in \mathbb{R}$, the adjoint problem (5) admits a unique solution*

$$\Phi \in \big(C^0_{loc}([0,+\infty);\mathscr{H}_s)\big)^N \cap \big(C^1_{loc}([0,+\infty);\mathscr{H}_{s-1})\big)^N,$$

Lemma 2 *For any given initial data $(U_0,U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ with $s > \frac{1}{2}$ and for any given boundary function $H \in \mathscr{L}^M$, the mixed initial-boundary value problem (1)–(2) admits a unique weak solution in the sense of duality, such that*

$$U \in \big(C^0_{loc}([0,+\infty);\mathscr{H}_{1-s})\big)^N \cap \big(C^1_{loc}([0,+\infty);\mathscr{H}_{-s})\big)^N,$$

and the linear mapping

$$\mathscr{R}:\ (U_0,U_1,H) \to (U,U') \tag{8}$$

is continuous with respect to the corresponding topologies.

2 Approximate Boundary Null Controllability

Definition 1 Let $s > \frac{1}{2}$. System (1) is approximately null controllable at the time $T > 0$, if for any given initial data $(U_0,U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0,T]$, such that the sequence $\{U_n\}$ of the solutions to the corresponding mixed initial-boundary value problem (1)–(2) satisfies

$$\big(U_n(T),U_n'(T)\big) \longrightarrow 0 \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad n \to +\infty \tag{9}$$

at $t = T$, or, equivalently,

$$\big(U_n,U_n'\big) \longrightarrow 0 \quad \text{in } \big(C^0_{loc}([T,+\infty);\mathscr{H}_{1-s}\times\mathscr{H}_{-s})\big)^N \quad \text{as} \quad n \to +\infty. \tag{10}$$

Obviously, the exact boundary null controllability implies the approximate boundary null controllability. However, since we can not get the convergence of the sequence $\{H_n\}$ of boundary controls from Definition 1, generally speaking, the approximate boundary null controllability does not lead to the exact boundary null controllability.

Similarly to the coupled system of wave equations with Dirichlet boundary controls [5, 6], we give the following

Definition 2 For $(\Phi_0, \Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$ $(s > \frac{1}{2})$, the adjoint problem (5) is D-observable on $[0, T]$, if

$$D^T\Phi \equiv 0 \quad \text{on} \quad [0,T] \times \Gamma_1 \Rightarrow (\Phi_0, \Phi_1) \equiv 0, \quad \text{then} \quad \Phi \equiv 0. \tag{11}$$

In order to find the equivalence between the approximate boundary null controllability of the original system (1) and the D-observability of the adjoint problem (5), let $\mathscr{C}$ be the set of all the initial states $(V(0), V'(0))$ of the following backward problem:

$$\begin{cases} V'' - \Delta V + AV = 0 & \text{in } (0,T) \times \Omega, \\ V = 0 & \text{on } (0,T) \times \Gamma_0, \\ \partial_\nu V = DH & \text{on } (0,T) \times \Gamma_1, \\ V(T) = 0, \quad V'(T) = 0 & \text{in } \Omega \end{cases} \tag{12}$$

with all admissible boundary controls $H \in \mathscr{L}^M$.

By Lemma 2, we have

Lemma 3 *For any given $T > 0$, for any given final data $(V(T), V'(T)) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, where $s > \frac{1}{2}$, and any given boundary function H in $\mathscr{L}^M$, the backward problem* (12) *(in which the null final condition is replaced by the given final data) admits a unique weak solution in the sense of duality, such that*

$$V \in \left(C^0([0,T]; \mathscr{H}_{1-s})\right)^N \cap \left(C^1([0,T]; \mathscr{H}_{-s})\right)^N,$$

and a result similar to (8) *holds, too.*

Lemma 4 *System* (1) *possesses the approximate boundary null controllability in the sense of Definition 1, if and only if*

$$\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N. \tag{13}$$

Proof Assume that $\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$. By the definition of $\mathscr{C}$, for any given $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, such that the sequence $\{V_n\}$ of the solutions to the corresponding backward problem (12) satisfies

$$\big(V_n(0), V_n'(0)\big) \to (U_0, U_1) \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad n \to +\infty. \tag{14}$$

Let $\mathscr{R}$ be the continuous linear mapping defined by (8). We have

$$\mathscr{R}(U_0, U_1, H_n) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0) + \mathscr{R}(V_n(0), V_n'(0), H_n). \tag{15}$$

On the other hand, by the definition of V_n, we have

$$\mathscr{R}(V_n(0), V_n'(0), H_n)(T) = 0. \tag{16}$$

Therefore,

$$\mathscr{R}(U_0, U_1, H_n)(T) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(T). \tag{17}$$

By Lemma 2, and noting (14), we then get

$$\begin{aligned} &\|\mathscr{R}(U_0, U_1, H_n)(T)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\ \leq &c\|(U_0 - V_n(0), U_1 - V_n'(0))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty. \end{aligned} \tag{18}$$

Here and hereafter, c always denotes a positive constant. Thus, system (1) is approximately null controllable.

Inversely, assume that system (1) is approximately null controllable. For any given $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, such that the sequence $\{U_n\}$ of the solutions to the corresponding mixed problem (1)–(2) satisfies

$$\big(U_n(T), U_n'(T)\big) = \mathscr{R}(U_0, U_1, H_n)(T) \to (0, 0) \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \tag{19}$$

as $n \to +\infty$. Taking such H_n as the boundary control, we solve the backward problem (12) and get the corresponding solution V_n. By the linearity of the mapping $\mathscr{R}$, we have

$$\mathscr{R}(U_0, U_1, H_n) - \mathscr{R}(V_n(0), V_n'(0), H_n) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0). \tag{20}$$

By Lemma 3 and noting (19), we have

$$\begin{aligned} &\|\mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(0)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\ \leq &c\|(U_n(T) - V_n(T), U_n'(T) - V_n'(T)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\ = &c\|(U_n(T), U_n'(T))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty. \end{aligned} \tag{21}$$

Combining (20), we then get

$$\begin{aligned}
&\|(U_0, U_1) - (V_n(0), V_n'(0))\|_{(\mathscr{H}_{1-s})^N\times(\mathscr{H}_{-s})^N} \\
=&\|\mathscr{R}(U_0, U_1, H_n)(0) - \mathscr{R}(V_n(0), V_n'(0), H_n)(0)\|_{(\mathscr{H}_{1-s})^N\times(\mathscr{H}_{-s})^N} \\
\leq& c\|\mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(0)\|_{(\mathscr{H}_{1-s})^N\times(\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty,
\end{aligned} \tag{22}$$

which shows that $\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$. □

Theorem 1 *System* (1) *is approximately null controllable at the time* $T > 0$ *in the sense of Definition 1, if and only if the adjoint problem* (5) *is D-observable on* $[0, T]$ *in the sense of Definition 2.*

Proof Assume that system (1) is not approximately null controllable at the time $T > 0$. By Lemma 4, there is a nontrivial vector $(-\Phi_1, \Phi_0) \in \mathscr{C}^\perp$. Here and hereafter, the orthogonal complement space is always defined in the sense of duality. Thus, $(-\Phi_1, \Phi_0) \in (\mathscr{H}_{s-1})^N \times (\mathscr{H}_s)^N$. Taking (Φ_0, Φ_1) as the initial data, we solve the adjoint problem (5) and get the solution $\Phi \not\equiv 0$. Multiplying Φ on the both sides of the backward problem (12) and integrating by parts, we get

$$\begin{aligned}
&\langle V(0), \Phi_1\rangle_{(\mathscr{H}_{1-s})^N;(\mathscr{H}_{s-1})^N} - \langle V'(0), \Phi_0\rangle_{(\mathscr{H}_{-s})^N;(\mathscr{H}_s)^N} \\
&= \int_0^T \int_{\Gamma_1} (DH, \Phi) d\Gamma dt.
\end{aligned} \tag{23}$$

The righthand side of (23) is meaningful due to

$$\Phi \in \big(C^0([0, T]; \mathscr{H}_s)\big)^N \hookrightarrow \big(L^2(0, T; L^2(\Gamma_1))\big)^N, \quad s > \frac{1}{2}. \tag{24}$$

Noticing $(V(0), V'(0)) \in \mathscr{C}$ and $(-\Phi_1, \Phi_0) \in \mathscr{C}^\perp$, it is easy to see from (23) that for any given H in $\mathscr{L}^M$, we have

$$\int_0^T \int_{\Gamma_1} (DH, \Phi) d\Gamma dt = 0.$$

Then, it follows that

$$D^T\Phi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1. \tag{25}$$

But $\Phi \not\equiv 0$, which implies that the adjoint problem (5) is not D-observable on $[0, T]$.

Inversely, assume that the adjoint problem (5) is not D-observable in $[0, T]$, then there exists a nontrivial initial data $(\Phi_0, \Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$, such that the solution Φ to the corresponding adjoint problem (5) satisfies (25). For any given $(U_0, U_1) \in \bar{\mathscr{C}}$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$, such that the solution V_n to the corresponding backward problem (12) satisfies

$$V_n(0) \to U_0, \quad V_n'(0) \to U_1 \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad n \to +\infty. \tag{26}$$

Similar to (23), multiplying Φ on the both sides of the backward problem (12) and noting (25), we get

$$\langle V_n(0), \Phi_1 \rangle_{(\mathscr{H}_{1-s})^N;(\mathscr{H}_{s-1})^N} - \langle V_n'(0), \Phi_0 \rangle_{(\mathscr{H}_{-s})^N;(\mathscr{H}_s)^N} = 0. \tag{27}$$

Let $n \to +\infty$, noting (26) we get

$$\langle (U_0, U_1), (-\Phi_1, \Phi_0) \rangle_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N;(\mathscr{H}_{s-1})^N \times (\mathscr{H}_s)^N} = 0, \quad \forall (U_0, U_1) \in \bar{\mathscr{C}},$$

which indicates that $(-\Phi_1, \Phi_0) \in \bar{\mathscr{C}}^\perp$, thus $\bar{\mathscr{C}} \neq (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$. □

Theorem 2 *If for any given initial data $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, system (1) is approximately null controllable in the sense of Definition 1 for some s ($> \frac{1}{2}$), then for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, system (1) possesses the same approximate boundary null controllability, too.*

Proof For any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ ($s > \frac{1}{2}$), by the density of $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$ in $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, we can find a sequence $\{(U_0^n, U_1^n)\}_{n \in \mathbb{N}}$ in $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, satisfying

$$(U_0^n, U_1^n) \to (U_0, U_1) \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad n \to +\infty. \tag{28}$$

By the assumption, for any fixed $n \geq 1$, there exists a sequence $\{H_k^n\}_{k \in \mathbb{N}}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, such that the solution $\{U_k^n\}$ to the corresponding mixed problem (1)–(2) satisfies

$$(U_k^n(T), (U_k^n)'(T)) \to (0, 0) \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad k \to +\infty. \tag{29}$$

For any given $n \geq 1$, let k_n be an integer such that

$$\begin{aligned} &\|\mathscr{R}(U_0^n, U_1^n, H_{k_n}^n)(T)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\ =&\|(U_{k_n}^n(T), (U_{k_n}^n)'(T))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \leq \frac{1}{2^n}. \end{aligned} \tag{30}$$

Thus, we get a sequence $\{k_n\}$ with $k_n \to +\infty$ as $n \to +\infty$. For the sequence $\{H_{k_n}\}$ of the boundary controls in $\mathscr{L}^M$, we have

$$\mathscr{R}(U_0^n, U_1^n, H_{k_n}^n)(T) \to 0 \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \ n \to +\infty. \tag{31}$$

Therefore, by the linearity of $\mathscr{R}$, the combination of (28) and (31) gives

$$\mathscr{R}(U_0, U_1, H^n_{k_n}) = \mathscr{R}(U_0 - U^n_0, U_1 - U^n_1, 0) + \mathscr{R}(U^n_0, U^n_1, H^n_{k_n}) \to 0 \tag{32}$$

in $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ as $n \to +\infty$, which indicates that the sequence of boundary controls $\{H^n_{k_n}\}$ realize the approximate boundary null controllability for (U_0, U_1), then for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, system (1) is approximately null controllable, too. □

Remark 1 Since $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N \subset (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, if system (1) is approximately null controllable in the sense of Definition 1 for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, then it possesses the same approximate boundary null controllability for any given initial data $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, too.

Remark 2 Theorem 2 and Remark 1 indicate that for system (1), the approximate boundary null controllability for the initial data in $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$ is equivalent to the approximate boundary null controllability for the initial data in $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$ with the same convergence space $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$.

Remark 3 In a similar way, we can prove that for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, system (1) is approximately null controllable in the sense of Definition 1, then for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s'})^N \times (\mathscr{H}_{-s'})^N$ $(s' > s)$, system (1) is still approximately null controllable.

Corollary 1 *If $M = N$, then system* (1) *is approximately null controllable.*

Proof Since $M = N$, the observations given by (11) become

$$\Phi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1.$$

By Holmgren's uniqueness theorem [9], we then get the D-observability of the adjoint problem (5). Hence, by Theorem 1 we get the approximate boundary null controllability of system (1). □

By means of Theorem 1, we can give a necessary condition for the approximate boundary null controllability.

Theorem 3 *If system* (1) *is approximately null controllable at the time $T > 0$, then we have the following Kalman's criterion:*

$$rank(D, AD, \cdots, A^{N-1}D) = N. \tag{33}$$

In order to prove Theorem 3, we first give the following

Lemma 5 (See [6]) *Let A be a matrix of order N, D an $N \times M$ matrix. Let $d \geq 0$ be an integer. Then the rank condition*

$$rank(D, AD, \cdots, A^{N-1}D) \geq N - d$$

holds, if and only if Ker(D^T) *does not contain any invariant subspace* V *of* A^T*, such that dim*$(V) > d$.

Proof (Proof of Theorem 3) By Lemma 5 (in which we take $d = 0$), we only need to prove that $\text{Ker}(D^T)$ does not contain any nontrivial invariant subspace of A^T.

If not, let e_n be the solution to the following eigenvalue problem:

$$\begin{cases} -\Delta e_n = \mu_n^2 e_n & \text{in } \Omega, \\ e_n = 0 & \text{on } \Gamma_0, \\ \partial_\nu e_n = 0 & \text{on } \Gamma_1. \end{cases} \tag{34}$$

Since $\text{mes}(\Gamma_0) > 0$, we have $\mu_n > 0$. Assume that A^T possesses a nontrivial invariant subspace V such that $V \subseteq \text{Ker}(D^T)$. For any fixed integer $n > 0$, define

$$W = \{e_n w : \quad w \in V\}.$$

Clearly, W is a finitely dimensional invariant subspace of $-\Delta + A^T$, then, we can solve the adjoint problem (5) in W and express the solutions as $\Phi = e_n w(t)$, where $w(t) \in V$ satisfies

$$w'' + (\mu_n^2 I + A^T)w = 0, \quad w(0) = w_0 \in V, \quad w'(0) = w_1 \in V.$$

Since $w(t) \in V$ for any given $t \geq 0$, Φ satisfies the condition of D-observation:

$$D^T\Phi = e_n D^T w(t) \equiv 0 \quad \text{on} \quad [0,T] \times \Gamma_1.$$

However, $\Phi \not\equiv 0$, then the adjoint problem (5) is not D-observable on $[0,T]$, which is a contradiction. The proof is complete. □

3 Sufficiency of Kalman's Criterion in One Dimensional Space

In this section, we discuss the sufficiency of Kalman's criterion (33). Generally speaking, similar to the approximate boundary null controllability for a coupled system of wave equations with Dirichlet boundary controls, Kalman's criterion is not sufficient. The reason is that Kalman's criterion does not depend on T, then if it is sufficient, the approximate boundary null controllability of the original system (1), or the D-observability of the adjoint problem (5) could be immediately realized, however, this is impossible since the wave propagates with a finite speed.

Theorem 4 *Let* μ_n^2 *and* e_n *be defined by* (34). *Assume that the set*

$$\Lambda = \{(m,n) : \quad \mu_n \neq \mu_m, \quad e_m = e_n \ \text{on} \ \Gamma_1\} \tag{35}$$

is not empty. For any given $(m,n) \in \Lambda$*, setting*

$$\varepsilon = \frac{\mu_m^2 - \mu_n^2}{2}, \tag{36}$$

the adjoint problem

$$\begin{cases} \phi'' - \Delta\phi + \varepsilon\psi = 0 & \textit{in } (0,+\infty) \times \Omega, \\ \psi'' - \Delta\psi + \varepsilon\phi = 0 & \textit{in } (0,+\infty) \times \Omega, \\ \phi = \psi = 0 & \textit{on } (0,+\infty) \times \Gamma_0, \\ \partial_\nu\phi = \partial_\nu\psi = 0 & \textit{on } (0,+\infty) \times \Gamma_1 \end{cases} \tag{37}$$

admits a nontrivial solution $(\phi,\psi) \not\equiv (0,0)$*, such that*

$$\phi \equiv 0 \quad \textit{on} \quad [0 + \infty) \times \Gamma_1. \tag{38}$$

Proof Let

$$\phi = (e_n - e_m), \quad \psi = (e_n + e_m), \quad \lambda^2 = \frac{\mu_m^2 + \mu_n^2}{2}. \tag{39}$$

It is easy to check that (ϕ,ψ) satisfies the following system:

$$\begin{cases} \lambda^2\phi + \Delta\phi - \varepsilon\psi = 0 & \text{in } (0,+\infty) \times \Omega, \\ \lambda^2\psi + \Delta\psi - \varepsilon\phi = 0 & \text{in } (0,+\infty) \times \Omega, \\ \phi = \psi = 0 & \text{on } (0,+\infty) \times \Gamma_0, \\ \partial_\nu\phi = \partial_\nu\psi = 0 & \text{on } (0,+\infty) \times \Gamma_1. \end{cases} \tag{40}$$

Moreover, noting the definition (35) of Λ, we have

$$\phi = 0 \quad \text{on} \quad \Gamma_1. \tag{41}$$

Then, let

$$\phi_\lambda = e^{i\lambda t}\phi, \quad \psi_\lambda = e^{i\lambda t}\psi. \tag{42}$$

It is easy to see that $(\phi_\lambda,\psi_\lambda)$ is a nontrivial solution to the adjoint problem (37), which satisfies condition (38). □

In order to illustrate the validity of the assumptions given in Theorem 4, we may examine the following situations, in which the set Λ is indeed not empty.

1. $\Omega = (0, \pi)$, $\Gamma_1 = \{\pi\}$. In this case, we have

$$\mu_n = n + \frac{1}{2}, \quad e_n = (-1)^n \sin(n + \frac{1}{2})x, \quad e_n(\pi) = e_m(\pi) = 1. \tag{43}$$

Thus, $(m, n) \in \Lambda$ for all $m \neq n$.

2. $\Omega = (0, \pi) \times (0, \pi)$, $\Gamma_1 = \{\pi\} \times [0, \pi]$. Let

$$\mu_{m,n} = \sqrt{(m + \frac{1}{2})^2 + n^2}, \quad e_{m,n} = (-1)^m \sin(m + \frac{1}{2})x \sin ny. \tag{44}$$

We have

$$e_{m,n}(\pi, y) = e_{m',n}(\pi, y) = \sin ny, \quad 0 \leq y \leq \pi. \tag{45}$$

Thus, $(\{m, n\}, \{m', n\}) \in \Lambda$ for all $m \neq m'$ and $n \geq 1$.

Remark 4 Theorem 4 implies that Kalman's criterion (33) is not sufficient in general. As a matter of fact, for the adjoint problem (37) which satisfies the condition of observation (38), we have $N = 2$,

$$A = \begin{pmatrix} 0 & \varepsilon \\ \varepsilon & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad (D, AD) = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \tag{46}$$

and the corresponding Kalman's criterion is satisfied. Theorem 4 shows that Kalman's criterion can not guarantee the D-observability of the adjoint problem (37). Nevertheless, for some special system (1), Kalman's criterion is still sufficient for the approximate boundary null controllability.

Similar to a coupled system of wave equations with Dirichlet boundary controls, in the case of one dimensional space, for coupling matrix A under some assumptions, Kalman's criterion is also sufficient for the approximate boundary null controllability of the original system.

By an Ingham's inequality given in [3], we first give some lemmas.

Let $\mathbb{Z}$ denote the set of all the integers, $\{\beta_n^{(l)}\}_{1 \leq l \leq m, n \in \mathbb{Z}}$ be a strictly increasing real sequence:

$$\cdots \beta_{-1}^{(1)} < \cdots < \beta_{-1}^{(m)} < \beta_0^{(1)} < \cdots < \beta_0^{(m)} < \beta_1^{(1)} < \cdots < \beta_1^{(m)} < \cdots \tag{47}$$

Definition 3 The sequence $\{e^{i\beta_n^{(l)} t}\}_{1 \leq l \leq m; n \in \mathbb{Z}}$ is ω-linearly independent in $\mathscr{L}^2(0, T)$, if for $T > 0$ large enough, the following conditions

$$\sum_{n \in \mathbb{Z}} \sum_{l=1}^{m} a_n^{(l)} e^{i\beta_n^{(l)} t} = 0 \quad \text{on} \quad [0, T], \quad \text{with} \quad \sum_{n \in \mathbb{Z}} \sum_{l=1}^{m} |a_n^{(l)}|^2 < +\infty$$

imply that

$$a_n^{(l)} = 0, \quad n \in \mathbb{Z}, \quad 1 \le l \le m.$$

By Li and Rao [6], we have

Lemma 6 *Assume that* (47) *holds, and there exist positive constants* c, s *and* γ, *such that*

$$\beta_{n+1}^{(l)} - \beta_n^{(l)} \ge m\gamma, \tag{48}$$

$$\frac{c}{|n|^s} \le \beta_n^{(l+1)} - \beta_n^{(l)} \le \gamma \tag{49}$$

for all l *with* $1 \le l \le m$ *and all* $n \in \mathbb{Z}$ *with* $|n|$ *large enough. Then the sequence* $\{e^{i\beta_n^{(l)}t}\}_{1\le l\le m; n\in\mathbb{Z}}$ *is* ω*-linearly independent in* $\mathscr{L}^2(0,T)$*, provided that* $T > 2\pi D^+$*, where* D^+ *is the upper density of the sequence* $\{\beta_n^{(l)}\}_{1\le l\le m; n\in\mathbb{Z}}$*, defined by*

$$D^+ = \limsup_{R\to+\infty} \frac{N(R)}{2R}, \tag{50}$$

where $N(R)$ *denotes the number of* $\{\beta_n^{(l)}\}$ *contained in the interval* $[-R, R]$.

Corollary 2 *For*

$$\delta_1 < \delta_2 < \cdots < \delta_m, \tag{51}$$

we define

$$\begin{cases} \beta_n^{(l)} = \sqrt{(n+\frac{1}{2})^2 + \delta_l \varepsilon}, & l = 1, 2, \cdots, m, \quad n \ge 0, \\ \beta_{-n}^{(l)} = -\beta_n^{(l)}, & l = 1, 2, \cdots, m, \quad n \ge 1, \end{cases} \tag{52}$$

where $|\varepsilon| > 0$ *is small enough. Then the sequence* $\{e^{i\beta_n^{(l)}t}\}_{1\le l\le m; n\in\mathbb{Z}}$ *is* ω*-linearly independent in* $\mathscr{L}^2(0,T)$*, provided that* $T > 2m\pi$.

Proof For $|\varepsilon| > 0$ small enough, it is easy to see that the sequence $\{\beta_n^{(l)}\}_{1\le l\le m; n\in\mathbb{Z}}$ satisfies (47). On the other hand, for $|\varepsilon| > 0$ small enough and for $|n| > 0$ large enough, we have

$$\beta_{n+1}^{(l)} - \beta_n^{(l)} = O(1). \tag{53}$$

In fact, for $n \geq 0$, by the first equation of (52), we have

$$\begin{aligned}\beta_{n+1}^{(l)} - \beta_n^{(l)} =& \sqrt{(n+\frac{3}{2})^2 + \delta_l \varepsilon} - \sqrt{(n+\frac{1}{2})^2 + \delta_l \varepsilon} \\ =& \frac{(n+\frac{3}{2})^2 - (n+\frac{1}{2})^2}{\sqrt{(n+\frac{3}{2})^2 + \delta_l \varepsilon} + \sqrt{(n+\frac{1}{2})^2 + \delta_l \varepsilon}} \\ =& \frac{2n+2}{\sqrt{(n+\frac{3}{2})^2 + \delta_l \varepsilon} + \sqrt{(n+\frac{1}{2})^2 + \delta_l \varepsilon}}.\end{aligned} \tag{54}$$

Therefore, (53) holds as $n > 0$ is large enough. Then, by the second equation of (52), we can finally prove (53).

Moreover, for $|\varepsilon| > 0$ small enough and for $|n| > 0$ large enough, we have

$$\beta_n^{(l+1)} - \beta_n^{(l)} = O\Big(\Big|\frac{\varepsilon}{n}\Big|\Big). \tag{55}$$

In fact, for $n \geq 0$, by the first equation of (52), we have

$$\beta_n^{(l+1)} - \beta_n^{(l)} = \frac{(\delta_{l+1} - \delta_l)\varepsilon}{\sqrt{(n+\frac{1}{2})^2 + \delta_{l+1}\varepsilon} + \sqrt{(n+\frac{1}{2})^2 + \delta_l \varepsilon}}.$$

Then, (55) holds for $n > 0$ large enough. Again, using the second equation of (52), we can finally prove (55).

Thus, the sequence $\{\beta_n^{(l)}\}_{1\leq l\leq m; n\in\mathbb{Z}}$ satisfies all the assumptions given in Lemma 6, in which $s = 1$. Moreover, by definition (50), a computation shows $D^+ = m$. This complete the proof. □

Now consider the following one dimensional adjoint problem:

$$\begin{cases} \Phi'' - \Delta\Phi + \varepsilon A^T \Phi = 0, & t > 0, \quad 0 < x < \pi, \\ \Phi(t, 0) = 0, & t > 0, \\ \partial_\nu \Phi(t, \pi) = 0, & t > 0, \\ t = 0: \quad \Phi = \Phi_0, \quad \Phi' = \Phi_1, & 0 < x < \pi \end{cases} \tag{56}$$

with the observation on $x = \pi$:

$$D^T \Phi(t, \pi) = 0 \quad \text{on} \quad [0, T], \tag{57}$$

where $|\varepsilon| > 0$ is small enough.

Assume that the $N \times N$ matrix A^T is diagonalizable with m ($\leq N$) distinct real eigenvalues:

$$\delta_1 < \delta_2 < \cdots < \delta_m, \tag{58}$$

and the corresponding eigenvectors $w^{(l,\mu)}$:

$$A^T w^{(l,\mu)} = \delta_l w^{(l,\mu)}, \quad 1 \leq l \leq m, \quad 1 \leq \mu \leq \mu_l, \tag{59}$$

we have

$$\sum_{l=1}^{m} \mu_l = N. \tag{60}$$

Let

$$e_n = (-1)^n \sin(n + \frac{1}{2})x, \quad n \geq 1. \tag{61}$$

e_n is an eigenfunction of $-\Delta$ in $\mathscr{H}_1$, satisfying

$$\begin{cases} -\Delta e_n = \mu_n^2 e_n, & \text{in} \quad 0 < x < \pi, \\ e_n = 0 & \text{on} \quad x = 0, \\ \partial_\nu e_n = 0 & \text{on } x = \pi, \end{cases} \tag{62}$$

in which $\mu_n = (n + \frac{1}{2})$. Thus, $e_n w^{(l,\mu)}$ is an eigenvector of $-\Delta + \varepsilon A^T$, corresponding to the eigenvalue $(n + \frac{1}{2})^2 + \delta_l \varepsilon$. Still defining $\{\beta_n^{(l)}\}_{1 \leq l \leq m; n \in \mathbb{Z}}$ by (52), it is clear that in $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$ the eigenvector of the corresponding system (56) is

$$E_n^{(l,\mu)} = \begin{pmatrix} \frac{e_n w^{(l,\mu)}}{i\beta_n^{(l)}} \\ e_n w^{(l,\mu)} \end{pmatrix}, \quad 1 \leq l \leq m, \quad 1 \leq \mu \leq \mu_l, \quad n \in \mathbb{Z}, \tag{63}$$

where we define $e_{-n} = e_n$ for all $n \geq 1$. Thus, $\{E_n^{(l,\mu)}\}_{1 \leq l \leq m, 1 \leq \mu \leq \mu_l, n \in \mathbb{Z}}$ forms a Riesz basis in $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$ (see [2]).

By Remark 1, we only need to consider any given initial data in $(\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$:

$$\begin{pmatrix} \Phi_0 \\ \Phi_1 \end{pmatrix} = \sum_{n \in \mathbb{Z}} \sum_{l=1}^{m} \sum_{\mu=1}^{\mu_l} \alpha_n^{(l,\mu)} E_n^{(l,\mu)}, \tag{64}$$

the solution to the corresponding adjoint problem (56) is given by

$$\begin{pmatrix} \Phi \\ \Phi' \end{pmatrix} = \sum_{n\in\mathbb{Z}} \sum_{l=1}^{m} \sum_{\mu=1}^{\mu_l} \alpha_n^{(l,\mu)} e^{i\beta_n^{(l)} t} E_n^{(l,\mu)}. \tag{65}$$

In particular, we have

$$\Phi = \sum_{n\in\mathbb{Z}} \sum_{l=1}^{m} \sum_{\mu=1}^{\mu_l} \frac{\alpha_n^{(l,\mu)}}{i\beta_n^{(l)}} e^{i\beta_n^{(l)} t} e_n w^{(l,\mu)}, \tag{66}$$

and the corresponding condition of observation (57) leads to

$$\sum_{n\in\mathbb{Z}} \sum_{l=1}^{m} D^T \Big(\sum_{\mu=1}^{\mu_l} \frac{\alpha_n^{(l,\mu)}}{i\beta_n^{(l)}} w^{(l,\mu)} \Big) e^{i\beta_n^{(l)} t} = 0 \quad \text{on} \quad [0, T]. \tag{67}$$

Theorem 5 *Assume that A and D satisfy Kalman's criterion* (33)*. Assume furthermore that A^T is diagonalizable with* (58)–(59)*. Then the adjoint problem* (56) *is D-observable for $|\varepsilon| > 0$ small enough, provided that $T > 2m\pi$.*

Proof Since A^T satisfies (58)–(59), as $T > 2m\pi$, applying Corollary 2 to each component of the relation (67), we get

$$D^T \Big(\sum_{\mu=1}^{\mu_l} \frac{\alpha_n^{(l,\mu)}}{i\beta_n^{(l)}} w^{(l,\mu)} \Big) = 0, \quad 1 \le l \le m, \quad n \in \mathbb{Z}. \tag{68}$$

Since Kalman's criterion (33) holds, by the case $d = 0$ in Lemma 5, Ker(D^T) does not contain any nontrivial invariant subspace of A^T, then it follows that

$$\sum_{\mu=1}^{\mu_l} \frac{\alpha_n^{(l,\mu)}}{i\beta_n^{(l)}} w^{(l,\mu)} = 0, \quad 1 \le l \le m, \quad n \in \mathbb{Z}. \tag{69}$$

Then

$$\alpha_n^{(l,\mu)} = 0, \quad 1 \le \mu \le \mu_l, \quad 1 \le l \le m, \quad n \in \mathbb{Z}, \tag{70}$$

hence, we get $\Phi \equiv 0$, namely, the adjoint problem (56) is D-observable. □

Similar to a coupled system of wave equations with Dirichlet boundary controls, the control time for the approximate boundary null controllability could be further reduced in the case that A^T possesses N distinct real eigenvalues.

Theorem 6 *Under the assumptions of Theorem 5, assume furthermore that A^T possesses N distinct real eigenvalues:*

$$\delta_1 < \delta_2 < \cdots < \delta_N. \tag{71}$$

Then the adjoint problem (56) is D-observable for $|\varepsilon| > 0$ small enough, provided that $T > 2\pi(N - M + 1)$, where $M = rank(D)$.

Proof Let $w^{(1)}, w^{(1)}, \cdots, w^{(N)}$ be the eigenvectors corresponding to the eigenvalues $\delta_1, \delta_2, \cdots, \delta_N$, respectively. In the present case, (67) can be written as

$$\sum_{n\in\mathbb{Z}} \sum_{l=1}^{N} D^T \frac{\alpha_n^{(l)}}{i\beta_n^{(l)}} w^{(l)} e^{i\beta_n^{(l)} t} = 0 \quad \text{on} \quad [0, T]. \tag{72}$$

Since rank$(D) = M$, without loss of generality, we may assume that $D^T w^{(1)}, \cdots$, $D^T w^{(M)}$ are linearly independent, then, there exists an invertible matrix S of order M, such that

$$SD^T(w^{(1)}, \cdots, w^{(M)}) = (e_1, \cdots e_M), \tag{73}$$

where $e_1, \cdots e_M$ are the canonical basis vectors in $\mathbb{R}^N$. Multiplying S to (72) from the left, we get

$$\sum_{n\in\mathbb{Z}} \Big\{ \sum_{l=1}^{M} \frac{\alpha_n^{(l)}}{i\beta_n^{(l)}} e_l e^{i\beta_n^{(l)} t} + \sum_{k=M+1}^{N} \frac{\alpha_n^{(k)}}{i\beta_n^{(k)}} SD^T w^{(k)} e^{i\beta_n^{(k)} t} \Big\} = 0 \quad \text{on} \quad [0, T]. \tag{74}$$

Noting the specific form of the canonical basis $(e_1, \cdots e_M)$, it is easy to check that for each $1 \leq l \leq M$, the upper density D_l^+ of the corresponding sequence $\{\beta_n^{(l)}, \beta_n^{(k)}\}_{M+1\leq k\leq N; n\in\mathbb{Z}}$ is equal to $(N - M + 1)$. Then, applying Corollary 2 to each equality of (74), we get (70), provided that $T > 2\pi(N - M + 1)$. The proof is complete. □

4 Approximate Boundary Synchronization

Definition 4 Let $s > \frac{1}{2}$. System (1) is approximately synchronizable at the time $T > 0$, if for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, and a function $u \in C^0_{loc}([0, \infty); \mathscr{H}_{1-s}) \cap C^1_{loc}([0, \infty); \mathscr{H}_{-s})$, such that the sequence $\{U_n\}$ of the solutions to the corresponding mixed initial-boundary value problem (1)–(2) satisfies

$$\big(u_n^{(k)}(T), (u_n^{(k)})'(T)\big) \to (u(T), u'(T)) \quad \text{in} \quad \mathscr{H}_{1-s} \times \mathscr{H}_{-s} \quad \text{as } n \to +\infty \tag{75}$$

for $1 \leq k \leq N$ at the time $t = T$, or, equivalently,

$$\big(u_n^{(k)}, (u_n^{(k)})'\big) \to (u, u') \quad \text{in } C^0_{loc}([T, +\infty); \mathscr{H}_{1-s} \times \mathscr{H}_{-s}) \text{ as } n \to +\infty \tag{76}$$

for $1 \leq k \leq N$. Here, u, being unknown a priori, is called the corresponding approximately synchronizable state.

Similarly to [5], we can easily get the following

Theorem 7 *Assume that system* (1) *is approximately synchronizable at the time* $T > 0$ *in the sense of Definition* 4. *Assume furthermore that at least for one initial data* $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, *the corresponding approximately synchronizable state* $u \not\equiv 0$. *Then the coupling matrix* $A = (a_{ij})$ *should satisfy the following condition of compatibility:*

$$\sum_{j=1}^{N} a_{ij} \overset{\text{def.}}{=} a \quad (i = 1, \cdots, N), \tag{77}$$

where a *is a constant independent of* $i = 1, \cdots, N$. *This condition of compatibility* (77) *is called the row sum condition.*

Proof If system (1) is approximately synchronizable, by the definition, we have

$$u'' - \Delta u + \Big(\sum_{j=1}^{N} a_{ij}\Big) u = 0, \quad i = 1, \cdots, N \quad \text{in} \quad \mathfrak{D}'((T, +\infty) \times \Omega) \tag{78}$$

as $t \geq T$. Hence, for $i, k = 1, \cdots, N$, we have

$$\Big(\sum_{j=1}^{N} a_{kj} - \sum_{j=1}^{N} a_{ij}\Big) u = 0 \quad \text{in} \quad \mathfrak{D}'((T, +\infty) \times \Omega). \tag{79}$$

Noticing that $u \not\equiv 0$ for at least one initial data (U_0, U_1), we have

$$\sum_{j=1}^{N} a_{kj} = \sum_{j=1}^{N} a_{ij}, \quad i, k = 1, \cdots, n, \tag{80}$$

which is the desired condition of compatibility (77). □

Remark 5 If for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, the corresponding approximately synchronizable state $u \equiv 0$, we get the trivial case of approximate boundary null controllability. Therefore, we need to exclude this situation beforehand.

Now let $\mathscr{C}$ be the set of all the initial states $(V(0), V'(0))$ of the solutions to the backward problem

$$\begin{cases} V'' - \Delta V + AV = 0 & \text{in } (0,T)\times\Omega, \\ V = 0 & \text{on } (0,T)\times\Gamma_0, \\ \partial_\nu V = DH & \text{on } (0,T)\times\Gamma_1, \\ V(T) = v_0 e, \quad V'(T) = v_1 e & \text{in } \Omega \end{cases} \tag{81}$$

as (v_0, v_1) varies in $\mathscr{H}_{1-s} \times \mathscr{H}_{-s}$ $(s > \frac{1}{2})$ and H varies in $\mathscr{L}^M$, where

$$e = (1, 1, \cdots, 1)^T. \tag{82}$$

Lemma 7 *Under the condition of compatibility* (77)*, system* (1) *possesses the approximate boundary synchronization in the sense of Definition 4, if and only if*

$$\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N. \tag{83}$$

Proof Assume that $\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$. Then for any given $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, by the definition of $\mathscr{C}$, there exist the final state $(v_{0n}e, v_{1n}e) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ at $t = T$ and a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, such that the solution V_n to the corresponding backward problem (81) satisfies

$$(V_n(0), V_n'(0)) \to (U_0, U_1) \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as } n \to +\infty. \tag{84}$$

Let $\mathscr{R}$ be the continuous linear mapping defined by (8):

$$\mathscr{R}: \quad (U_0, U_1, H) \to (U, U').$$

Thus, we have

$$\mathscr{R}(U_0, U_1, H_n) - \mathscr{R}(V_n(0), V_n'(0), H_n) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0). \tag{85}$$

By the definition of V_n, we have

$$\mathscr{R}(V_n(0), V_n'(0), H_n)(T) = (v_{0n}e, v_{1n}e),$$

then

$$\mathscr{R}(U_0, U_1, H_n)(T) - (v_{0n}e, v_{1n}e) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(T). \tag{86}$$

By Lemma 2, we get

$$
\begin{aligned}
&\|\mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(T)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\
\leq &c\|(U_0 - V_n(0), U_1 - V_n'(0))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N}.
\end{aligned} \tag{87}
$$

Noting (84), it follows from (86)–(87) that

$$
\|\mathscr{R}(U_0, U_1, H_n)(T) - (v_{0n}e, v_{1n}e)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty,
$$

which indicates that for any given $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$, such that system (1) is approximately synchronizable.

Inversely, assume that system (1) is approximately synchronizable. Then for any given $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ and $(v_0e, v_1e) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, such that the solution to the mixed problem (1)–(2) satisfies

$$
\mathscr{R}(U_0, U_1, H_n)(T) \to (v_0e, v_1e) \text{ in } (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \text{ as } n \to +\infty. \tag{88}
$$

Taking the boundary condition corresponding to such boundary control H_n and (v_0e, v_1e) as the final state at $t = T$, we solve the backward problem (81) and get its solution V_n. By the linearity of $\mathscr{R}$, we have

$$
\mathscr{R}(U_0, U_1, H_n) - \mathscr{R}(V_n(0), V_n'(0), H_n) = \mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0). \tag{89}
$$

Noting (88), by the wellposedness of the corresponding backward problem of system (1), we have

$$
\begin{aligned}
&\|\mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(0)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\
\leq &c\|(U_n(T) - V_n(T), U_n'(T) - V_n'(T))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\
= &c\|(U_n(T) - v_0e, U_n'(T) - v_1e\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty.
\end{aligned} \tag{90}
$$

Then, by (89) we get

$$
\begin{aligned}
&\|(U_0, U_1) - (V_n(0), V_n'(0))\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\
= &\|\mathscr{R}(U_0, U_1, H_n)(0) - \mathscr{R}(V_n(0), V_n'(0), H_n)(0)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \\
= &\|\mathscr{R}(U_0 - V_n(0), U_1 - V_n'(0), 0)(0)\|_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N} \to 0 \quad \text{as} \quad n \to +\infty,
\end{aligned} \tag{91}
$$

which indicates that $\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$. □

Lemma 8 *Under the condition of compatibility* (77)*, system* (1) *is approximately synchronizable at the time* $T > 0$ *in the sense of Definition* 4*, if and only if* *the solution Φ to the adjoint problem* (5) *possesses the following synchronizable observability: for* $(\Phi_0, \Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$ $(s > \frac{1}{2})$, *if*

$$D^T \Phi \equiv 0 \quad on \quad [0, T] \times \Gamma_1 \tag{92}$$

and

$$(\Phi(T), e) \equiv 0, \quad (\Phi'(T), e) \equiv 0 \quad in \quad \Omega, \tag{93}$$

then $\Phi \equiv 0$, *where* e *is given by* (82).

Proof First, let Φ be the solution to the adjoint problem (5). Multiplying Φ on the both sides of the backward problem (81) and integrating by parts, we get

$$\begin{aligned}&\langle \Phi(T), v_1 e\rangle_{(\mathscr{H}_s)^N \times (\mathscr{H}_{-s})^N} - \langle \Phi'(T), v_0 e\rangle_{(\mathscr{H}_{s-1})^N \times (\mathscr{H}_{1-s})^N} \\ =&\langle V'(0), \Phi_0\rangle_{(\mathscr{H}_{-s})^N \times (\mathscr{H}_s)^N} - \langle V(0), \Phi_1\rangle_{(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{s-1})^N} + \int_0^T \int_{\Gamma_1} (DH, \Phi) d\Gamma dt,\end{aligned} \tag{94}$$

where $s > \frac{1}{2}$.

Assume that system (1) possesses the synchronizable observability, but is not approximately synchronizable. By Lemma 7, there exists a nontrivial initial data $(-\Phi_1, \ \Phi_0) \in \mathscr{C}^{\perp}$ with $(-\Phi_1, \Phi_0) \in (\mathscr{H}_{s-1})^N \times (\mathscr{H}_s)^N$. Taking such (Φ_0, Φ_1) as the initial data, we solve the corresponding adjoint problem (5), and its solution $\Phi \not\equiv 0$. Thus, noting (94) and the definition of $\mathscr{C}$, for any given $(v_0, v_1) \in \mathscr{H}_{1-s} \times \mathscr{H}_{-s}$ and $H \in \mathscr{L}^M$, we have

$$\begin{aligned}&\langle \Phi(T), v_1 e\rangle_{(\mathscr{H}_s)^N \times (\mathscr{H}_{-s})^N} - \langle \Phi'(T), v_0 e\rangle_{(\mathscr{H}_{s-1})^N \times (\mathscr{H}_{1-s})^N} \\ &= \int_0^T \int_{\Gamma_1} (DH, \Phi) d\Gamma dt.\end{aligned} \tag{95}$$

The righthand side of the above formula is meaningful because of (24). Thus the conditions of observation (92)–(93) hold, but $\Phi \not\equiv 0$, hence, the synchronizable observability fails, which is a contradiction.

Inversely, assume that system (1) is approximately synchronizable. Assume that the solution Φ to the adjoint problem (5) with the initial data $(\Phi_0, \Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$ $(s > \frac{1}{2})$ satisfies the conditions of observation (92)–(93). For any given $(U_0, U_1) \in \overline{\mathscr{C}}$, by the definition of $\overline{\mathscr{C}}$, there exist $(v_0 e, v_1 e) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ and a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$, such that the solution V_n to the corresponding backward problem (81) satisfies

$$V_n(0) \to U_0, \quad V_n'(0) \to U_1 \quad \text{in} \quad (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N \quad \text{as} \quad n \to +\infty. \tag{96}$$

Noting the conditions of observation (92)–(93), by (94) we have

$$\langle V_n'(0), \Phi_0\rangle_{(\mathscr{H}_{-s})^N\times(\mathscr{H}_s)^N} - \langle V_n(0), \Phi_1\rangle_{(\mathscr{H}_{1-s})^N\times(\mathscr{H}_{s-1})^N} = 0. \tag{97}$$

Taking $n \to +\infty$, we get

$$\langle (U_0, U_1), (-\Phi_1, \Phi_0)\rangle_{(\mathscr{H}_{1-s})^N\times(\mathscr{H}_{-s})^N;(\mathscr{H}_{s-1})^N\times(\mathscr{H}_s)^N} = 0,\ \ \forall (U_0, U_1) \in \overline{\mathscr{C}}, \tag{98}$$

then $(-\Phi_1, \Phi_0) \in \overline{\mathscr{C}}^{\perp}$. Thus, $\overline{\mathscr{C}} \neq (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$. By Lemma 7, system (1) is not approximately synchronizable, which is a contradiction. The proof is complete. □

We know that the exact boundary synchronization of system (1) is equivalent to the exact boundary null controllability of its reduced system (see [4]). For the approximate boundary synchronization, similar result holds, too.

Lemma 9 (See [8]) *Let*

$$C = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}_{(N-1)\times N} \tag{99}$$

be the corresponding synchronization matrix. C is a full row-rank matrix, and $Ker(C) = Span\{e\}$, where e is given by (82). Then the following properties are equivalent:

(1) *The condition of compatibility (77) holds;*
(2) *$e = (1, 1, \cdots, 1)^T$ is a right eigenvector of the matrix A, corresponding to the eigenvalue a;*
(3) *$Ker(C)$ is an one dimensional invariant subspace of A:*

$$AKer(C) \subseteq Ker(C); \tag{100}$$

(4) *There exists a unique matrix $\overline{A}$ of order $(N-1)$, such that*

$$CA = \overline{A}C. \tag{101}$$

By Lemma 9, under the condition of compatibility (77), setting $W = CU$, we get the following reduced problem of the original problem (1)–(2):

$$\begin{cases} W'' - \Delta W + \overline{A}W = 0 & \text{in } (0, +\infty) \times \Omega, \\ W = 0 & \text{on } (0, +\infty) \times \Gamma_0, \\ \partial_\nu W = \overline{D}H & \text{on } (0, +\infty) \times \Gamma_1, \\ t = 0:\ W = CU_0 \overset{\text{def.}}{=} W_0, \quad W' = CU_1 \overset{\text{def.}}{=} W_1 & \text{in } \Omega, \end{cases} \tag{102}$$

where $\overline{D} = CD$.

Let

$$\Psi = (\psi^{(1)}, \cdots, \psi^{(N-1)})^T.$$

Consider the adjoint problem of the reduced problem (102):

$$\begin{cases} \Psi'' - \Delta\Psi + \overline{A}^T\Psi = 0 & \text{in } (0, +\infty) \times \Omega, \\ \Psi = 0 & \text{on } (0, +\infty) \times \Gamma_0, \\ \partial_\nu \Psi = 0 & \text{on } (0, +\infty) \times \Gamma_1, \\ t = 0: \quad \Psi = \Psi_0, \quad \Psi' = \Psi_1 & \text{in } \Omega. \end{cases} \tag{103}$$

Definition 5 For $(\Psi_0, \Psi_1) \in (\mathscr{H}_s)^{N-1} \times (\mathscr{H}_{s-1})^{N-1}$ $(s > \frac{1}{2})$, the reduced adjoint problem (103) is CD-observable on $[0, T]$, if

$$(CD)^T\Psi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1 \quad \Longrightarrow \quad (\Psi_0, \Psi_1) \equiv 0, \quad \text{then} \quad \Psi \equiv 0. \tag{104}$$

Theorem 8 *Under the condition of compatibility* (77)*, system* (1) *is approximately synchronizable, if and only if the reduced adjoint problem* (103) *is CD-observable on* $[0, T]$.

Proof By Lemma 8, we only need to prove that the CD-observability of the reduced adjoint problem (103) is equivalent to the synchronizable observability of the adjoint problem (5). By Lemma 9, $e = (1, 1, \cdots, 1)^T$ is a right eigenvector of the matrix A, corresponding to the eigenvalue a, where a is given by (77). Let $E \in \mathbb{R}^N$ be the corresponding left eigenvector, such that

$$Ae = ae, \quad E^TA = aE^T.$$

Assume that the reduced adjoint problem (103) is CD-observable. Let Φ be the solution to the adjoint problem (5), which satisfies the conditions of observation (92)–(93). Let

$$\widetilde{\Phi} = (\Phi, e)E.$$

Using the coupled system of wave equations in (5), we have

$$\widetilde{\Phi}'' - \Delta\widetilde{\Phi} + A^T\widetilde{\Phi} = (\Phi'' - \Delta\Phi + A^T\Phi, e)E = 0. \tag{105}$$

Moreover, by the condition of observation (93), we have

$$\widetilde{\Phi}(T) = \widetilde{\Phi}'(T) = 0. \tag{106}$$

By (105)–(106), noting that $\widetilde{\Phi}$ is subjected to the homogeneous boundary condition, we get

$$\widetilde{\Phi} \equiv 0,$$

then, noting $\mathrm{Ker}(C) = \mathrm{Span}\{e\}$, we have

$$\Phi \in \{\mathrm{Span}\{e\}\}^{\perp} = \mathrm{Im}(C^T).$$

Thus, there exists a Ψ such that

$$\Phi = C^T \Psi. \tag{107}$$

Substituting it into the coupled system of wave equations in (5), and noting (101), we get

$$C^T \Psi'' - C^T \Delta\Psi + A^T C^T \Psi = C^T(\Psi'' - \Delta\Psi + \overline{A}^T \Psi) = 0.$$

Since C^T is an injection, we have

$$\Psi'' - \Delta\Psi + \overline{A}^T \Psi = 0 \quad \text{in} \quad \Omega. \tag{108}$$

Furthermore, by the condition of observation (92), we have

$$(CD)^T \Psi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1. \tag{109}$$

Therefore, it follows from the CD-observability of the reduced adjoint problem (103) that

$$\Psi \equiv 0, \quad \text{then} \quad \Phi \equiv 0, \tag{110}$$

thus, system (1) is approximately synchronizable.

Inversely, assume that system (1) is approximately synchronizable. Let Ψ be the solution to the reduced adjoint problem (103), and Φ be defined by (107). Noting (101), it is easy to get that

$$\Phi'' - \Delta\Phi + A^T \Phi = 0. \tag{111}$$

On the other hand, by Definition 5, the boundary observation gives

$$D^T \Phi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1. \tag{112}$$

Noticing $Ce = 0$, it is easily seen that as $t \geq 0$ we have

$$(\Phi, e) \equiv 0, \quad (\Phi', e) \equiv 0. \tag{113}$$

Hence, Φ is the solution to the adjoint problem (5) and satisfies the conditions of observation (92)–(93). By Lemma 8, we have

$$\Phi \equiv 0. \tag{114}$$

Thus, since C^T is an injection, we have

$$\Psi \equiv 0, \tag{115}$$

that is to say, the reduced adjoint problem (103) is CD-observable. □

Remark 6 Theorem 8 indicates that under the condition of compatibility (77), the approximate boundary synchronization of system (1) is equivalent to the approximate boundary null controllability of the reduced system (102).

Theorem 9 *Under the condition of compatibility* (77)*, if for any given initial data* $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$*, system* (1) *is approximately synchronizable for some* $s\ (> \frac{1}{2})$*, then system* (1) *possesses the same approximate boundary synchronization for any given initial data* $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$*, too.*

Proof Suppose that for any given initial data $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, system (1) is approximately synchronizable. Then, under the condition of compatibility (77), the reduced system (102) is approximately null controllable for any given initial data in $(\mathscr{H}_1)^{N-1} \times (\mathscr{H}_0)^{N-1}$, thus by Theorem 2, the reduced system (102) is approximately null controllable for any given initial data in $(\mathscr{H}_{1-s})^{N-1} \times (\mathscr{H}_{-s})^{N-1}$, too. Then, by Theorem 1, the reduced adjoint problem (103) is CD-observable on $[0, T]$ in $(\mathscr{H}_{1-s})^{N-1} \times (\mathscr{H}_{-s})^{N-1}$. Finally, by Theorem 8, the original system (1) is approximately synchronizable in $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, too. □

Remark 7 Obviously, if for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, system (1) is approximately synchronizable, then for any given initial data $(U_0, U_1) \in (\mathscr{H}_1)^N \times (\mathscr{H}_0)^N$, system (1) is also approximately synchronizable for this $s\ (> \frac{1}{2})$.

Corollary 3 *Under the condition of compatibility* (77)*, if rank*$(CD) = N - 1$*, then system* (1) *is approximately synchronizable.*

Proof Since $(CD)^T$ is an injection, the condition of observation in Definition 5 gives

$$\Psi \equiv 0 \quad \text{on} \quad [0, T] \times \Gamma_1.$$

By means of Holmgren's uniqueness theorem [9], we get the CD-observability of the reduced adjoint problem (103) on $[0, T]$, then by Theorem 8, we get the approximate boundary synchronization of system (1). □

Noting (101), by Theorem 3, we get the following criterion of Kalman's type for the approximate boundary synchronization of system (1).

Theorem 10 *Under the condition of compatibility* (77)*, if system* (1) *is approximately synchronizable at* $T > 0$*, then we have the following criterion of Kalman's type:*

$$rank(CD, CAD, \cdots, CA^{N-1}D) = N - 1. \tag{116}$$

Proof Since system (1) is approximately synchronizable, its reduced problem (102) is approximately null controllable, thus, applying Theorem 3, we have

$$\text{rank}(\bar{D}, \bar{A}\bar{D}, \cdots, \bar{A}^{N-2}\bar{D}) = N - 1. \tag{117}$$

By Cayley-Hamilton's Theorem, we get

$$\text{rank}(\bar{D}, \bar{A}\bar{D}, \cdots, \bar{A}^{N-2}\bar{D}) = \text{rank}(\bar{D}, \bar{A}\bar{D}, \cdots, \bar{A}^{N-1}\bar{D}). \tag{118}$$

Noting (101), we have

$$(CD, CAD, \cdots, CA^{N-1}D) = (\bar{D}, \bar{A}\bar{D}, \cdots, \bar{A}^{N-1}\bar{D}). \tag{119}$$

Then, combining (117)–(119), we get (116). □

5 Approximate Boundary Synchronization by Groups

When $\text{rank}(D, AD, \cdots, A^{N-1}D)$ is further reduced, we can consider the approximate boundary synchronization by p (≥ 1)-groups (when $p = 1$, it is just the approximate boundary synchronization).

Let p (≥ 1) be an integer, and

$$0 = m_0 < m_1 < m_2 < \cdots < m_p = N.$$

The approximate boundary synchronization by p-groups means that the components of U are divided into p groups:

$$(u^{(1)}, \cdots, u^{(m_1)}), \quad (u^{(m_1+1)}, \cdots, u^{(m_2)}), \cdots, (u^{(m_{p-1}+1)}, \cdots, u^{(m_p)}), \tag{120}$$

and each group possesses the corresponding approximate boundary synchronization, respectively.

Definition 6 Let $s > \frac{1}{2}$. System (1) is approximately synchronizable by p-groups at $T > 0$, if for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, there exists a sequence $\{H_n\}$ of boundary controls in $\mathscr{L}^M$ with compact support in $[0, T]$ and functions $u_r \in C^0_{loc}([T,+\infty);\mathscr{H}_{1-s}) \cap C^1_{loc}([T,+\infty);\mathscr{H}_{-s})$ $(1 \le r \le p)$, such that the corresponding sequence $\{U_n\}$ of the solutions to the mixed initial-boundary value problem (1)–(2) satisfies

$$\big(u_n^{(k)}(T), (u_n^{(k)})'(T)\big) \to (u_r(T), u_r'(T)) \text{ in } \mathscr{H}_{1-s} \times \mathscr{H}_{-s} \text{ as } n \to +\infty \tag{121}$$

for $m_{r-1}+1 \le k \le m_r$ ($1 \le r \le p$), or, equivalently,

$$u_n^{(k)} \to u_r \text{ in } C^0_{loc}([T,+\infty);\mathscr{H}_{1-s}) \cap C^1_{loc}([T,+\infty);\mathscr{H}_{-s}) \text{ as } n \to +\infty \tag{122}$$

for $m_{r-1}+1 \le k \le m_r$ ($1 \le r \le p$). Here, $(u_1,\cdots,u_p)$, being unknown a prior, is called the approximately synchronizable state by p-groups.

Let

$$C_r = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}, \quad 1 \le r \le p \tag{123}$$

be an $(m_r - m_{r-1} - 1) \times (m_r - m_{r-1})$ matrix with full row-rank, and

$$C = \begin{pmatrix} C_1 & & & \\ & C_2 & & \\ & & \ddots & \\ & & & C_p \end{pmatrix} \tag{124}$$

be the $(N-p) \times N$ matrix of synchronization by p-groups. Clearly,

$$\text{Ker}(C) = \text{Span}\{e_1,\cdots,e_p\}, \tag{125}$$

where

$$(e_r)_j = \begin{cases} 1, & m_{r-1}+1 \le j \le m_r, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, (122) can be written as

$$(U_n, U_n') \to \Big(\sum_{r=1}^p u_r e_r, \ \sum_{r=1}^p u_r' e_r\Big) \quad \text{in} \quad (C^0_{loc}([T,+\infty);\mathscr{H}_{1-s} \times \mathscr{H}_{-s}))^N \tag{126}$$

as $n \to +\infty$.

Theorem 11 *Assume that system (1) is approximately synchronizable by p-groups in the sense of Definition 6. Assume furthermore that at least for one initial data* $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$*, the components of* $(u_1, \cdots, u_p)$ *are mutually linear independent. Then the coupling matrix* $A = (a_{ij})$ *should satisfy the following condition of compatibility:*

$$AKer(C) \subseteq Ker(C), \tag{127}$$

or, equivalently, there exists a unique matrix $\overline{A}$ *of order* $(N-p)$*, such that*

$$CA = \overline{A}C. \tag{128}$$

Proof Let $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ $(s > \frac{1}{2})$, $\{H_n\}$ be the sequence of boundary controls which realizes the approximate boundary synchronization by p-groups for system (1), and $\{U_n\}$ be the sequence of solutions to the corresponding problem (1)–(2). Multiplying C from the left to (1), and taking $n \to +\infty$, it follows from (122) and (125) that

$$\sum_{r=1}^{p} u_r CAe_r = 0 \quad \text{in} \quad \mathfrak{D}'((T, +\infty) \times \Omega). \tag{129}$$

By the second hypothesis of Theorem 11, at least for one initial data (U_0, U_1), the components $u_1, u_2, \cdots, u_p$ of the corresponding approximately synchronizable state by p-groups are linearly independent, then we have

$$CAe_r = 0, \quad 1 \le r \le p, \tag{130}$$

namely, (127) holds. □

Remark 8 If for any given initial data $(U_0, U_1) \in (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$, the components of $(u_1, \cdots, u_p)$ are linearly dependent, then, through a suitable reversible linear transformation of the state variables U, the corresponding approximately synchronizable state $(u_1, \cdots, u_p)$ by p-groups contains at least one null component. We exclude this special situation beforehand.

Now, let $\mathscr{C}$ be the set of all the initial states $(V(0), V'(0))$ of the backward problem

$$\begin{cases} V'' - \Delta V + AV = 0 & \text{in } (0, T) \times \Omega, \\ V = 0 & \text{on } (0, T) \times \Gamma_0, \\ \partial_\nu V = DH & \text{on } (0, T) \times \Gamma_1, \\ V(T) = V_0, \quad V'(T) = V_1 & \text{in } \Omega \end{cases} \tag{131}$$

as (V_0, V_1) varies in $(\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N$ ($s > \frac{1}{2}$) with $CV_0 = CV_1 = 0$, and H varies in $\mathscr{L}^M$.

Lemma 10 *Under the condition of compatibility* (127)*, system* (1) *is approximately synchronizable by p-groups in the sense of Definition 6, if and only if*

$$\bar{\mathscr{C}} = (\mathscr{H}_{1-s})^N \times (\mathscr{H}_{-s})^N. \tag{132}$$

Proof Similar to the proof of Lemma 7. □

Lemma 11 *Under the condition of compatibility* (127)*, system* (1) *is approximately synchronizable by p-groups in the sense of Definition 6, if and only if the solution Φ to the adjoint problem* (5) *possesses the following synchronizable observability by p-groups: for any given initial data $(\Phi_0, \Phi_1) \in (\mathscr{H}_s)^N \times (\mathscr{H}_{s-1})^N$ $(s > \frac{1}{2})$, if*

$$D^T \Phi = 0 \quad on \quad [0, T] \times \Gamma_1 \tag{133}$$

and

$$\Phi(T),\ \Phi'(T) \in \{Ker(C)\}^{\perp}, \tag{134}$$

then $\Phi \equiv 0$, where C is the matrix of synchronization by p-groups defined by (124)*.*

Proof Similar to the proof of Lemma 8. □

Similarly, using the matrix C of synchronization by p-groups, defined by (124), we can get the corresponding reduced adjoint problem (103) and its CD-observability (see Definition 5).

Theorem 12 *Under the condition of compatibility* (127)*, system* (1) *is approximately synchronizable by p-groups in the sense of Definition 6, if and only if the reduced adjoint problem* (103) *is CD-observable on* $[0, T]$.

Proof By Lemma 11, we only need to prove that the CD-observability of the reduced adjoint problem (103) is equivalent to the observability of synchronization by p-groups of the adjoint problem (5).

Assume that the reduced adjoint problem (103) is CD-observable on $[0, T]$. Let Φ be the solution to the adjoint problem (5) with (133)–(134). By the condition of compatibility (127), there exist real coefficients α_{rk} such that

$$Ae_k = \sum_{r=1}^{p} \alpha_{rk} e_r, \quad 1 \le k \le p. \tag{135}$$

Let $\phi_k = (e_k, \Phi)$. Taking the inner product of (5) with e_k, and noting (134) and (125), for $1 \le k \le p$ we get

$$\begin{cases} \phi_k'' - \Delta\phi_k + \sum_{r=1}^{p} \alpha_{rk}\phi_r = 0 & \text{in } (0, +\infty) \times \Omega, \\ \phi_k = 0 & \text{on } (0, +\infty) \times \Gamma_0, \\ \partial_\nu \phi_k = 0 & \text{on } (0, +\infty) \times \Gamma_1, \\ t = T: \quad \phi_k = \phi_k' = 0 & \text{in } \Omega. \end{cases} \tag{136}$$

Then we get

$$t \ge 0: \quad \phi_1 = \phi_2 = \cdots = \phi_p \equiv 0, \tag{137}$$

hence

$$\Phi \in \{\mathrm{Ker}(C)\}^\perp = \mathrm{Im}(C^T). \tag{138}$$

Therefore, there exists $\Psi = (\psi^{(1)}, \cdots, \psi^{(N-p)})^T$ such that

$$\Phi = C^T \Psi. \tag{139}$$

Substituting (139) into (5) and (133), and noting (128) and the fact that C^T is an injection, we have

$$\Psi'' - \Delta\Psi + \overline{A}^T \Psi = 0 \quad \text{in} \quad \Omega \tag{140}$$

and

$$(CD)^T \Psi = 0 \quad \text{on} \quad [0, T] \times \Gamma_1. \tag{141}$$

Thus, by the CD-observability of problem (103) on $[0, T]$, we get

$$\Psi \equiv 0, \quad \text{then} \quad \Phi \equiv 0. \tag{142}$$

This implies that the adjoint problem (5) possesses the observability of synchronization by p-groups on $[0, T]$.

Inversely, assume that the adjoint problem (5) possesses the observability of synchronization by p-groups. Let Ψ be the solution to the reduced adjoint problem (103), and Φ be defined by (139). Noting (125) and (128), we have

$$\Phi'' - \Delta\Phi + A^T \Phi = 0, \tag{143}$$

$$D^T \Phi = 0 \quad \text{on} \quad [0, T] \times \Gamma_1 \tag{144}$$

and

$$t \geq 0 : \quad \Phi \in \{\mathrm{Ker}(C)\}^{\perp}, \tag{145}$$

which shows that Φ is the solution to the adjoint problem (5), and satisfies the conditions of observation by p-groups (133)–(134). By the observability of synchronization by p-groups of the adjoint problem (5), we have

$$\Phi = C^T \Psi \equiv 0, \quad \text{then} \quad \Psi \equiv 0, \tag{146}$$

namely, the reduced adjoint system (103) is CD-observable on $[0, T]$. □

Noting (128), by Theorem 3, we get the following criterion of Kalman's type for the approximate boundary synchronization by p-groups for system (1).

Theorem 13 *Under the condition of compatibility* (127)*, if system* (1) *is approximately synchronizable by p-groups at* $T > 0$*, then we have the following criterion of Kalman's type:*

$$rank(CD, CAD, \cdots, CA^{N-1}D) = N - p, \tag{147}$$

where the matrix C of synchronization by p-groups is defined by (124)*.*

Proof Applying Theorem 3 to the corresponding reduced adjoint problem (103), which is CD-observable by Theorem 12, we get

$$\mathrm{rank}(CD, \overline{A}CD, \cdots, \overline{A}^{N-p-1}CD) = N - p. \tag{148}$$

The remaining part of the proof is similar to that of Theorem 10. □

Acknowledgements Projet supported by the National Basic Research Program of China (No 2013CB834100), and the National Natural Science Foundation of China (No 11121101).

References

1. Bardos, C., Lebeau, G., Rauch, J.: Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J. Control Optim. **30**, 1024–1065 (1992)
2. Gohberg, I.C., Krein, M.G.: Introduction to the Theory of Linear Nonselfadjoint Operators. American Mathematical Society, Providence, RI (1969)
3. Komornik, V., Loreti, P.: Fourier Series in Control Theory. Springer Monographs in Mathematics. Springer, Berlin (2005)
4. Li, T-T., Rao, B.: Exact synchronization for a coupled system of wave equation with Dirichlet boundary controls. Chin. Ann. Math. Ser. B **34**(1), 139–160 (2013)
5. Li, T-T., Rao, B.: Asymptotic controllability and asymptotic synchronization for a coupled system of wave equations with Dirichlet boundary controls. Asymptot. Anal. **86**, 199–226 (2014)

6. Li, T-T., Rao, B.: Criteria of Kalman's type to the approximate controllability and the approximate synchronization for a coupled system of wave equations with Dirichlet boundary controls. SIAM J. Control Optim. **54**, 49–72 (2016)
7. Li, T-T., Rao, B.: Exact boundary controllability for a coupled system of wave equations with Neumann controls. Chin. Ann. Math. Ser. B **38**(2), 473–488 (2017)
8. Li, T-T., Rao, B., Wei, Y.: Generalized exact boundary synchronization for a coupled system of wave equations. Discrete Contin. Dyn. Syst. **34**, 2893–2905 (2014)
9. Lions, J.-L.: Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, vol. 1. Masson, Paris (1988)

Sparse Support Vector Machines in Reproducing Kernel Banach Spaces

Zheng Li, Yuesheng Xu, and Qi Ye

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We present a novel approach for support vector machines in reproducing kernel Banach spaces induced by a finite basis. In particular, we show that the support vector classification in the 1-norm reproducing kernel Banach space is mathematically equivalent to the *sparse* support vector machine. Finally, we develop fixed-point proximity algorithms for finding the solution of the non-smooth minimization problem that describes the sparse support vector machine. Numerical results are presented to demonstrate that the sparse support vector machine outperforms the classical support vector machine for the binary classification of simulation data.

1 Introduction

Although reproducing kernel Hilbert spaces (RKHS) are a classical subject [2], reproducing kernels became popular only recently in the field of machine learning [1, 5, 14, 17]. Since Banach spaces have a richer geometric structure compared

Z. Li
Guangdong Province Key Lab of Computational Science, School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong, People's Republic of China

Y. Xu (✉)
School of Data and Computer Science, Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University, Guangzhou, Guangdong, People's Republic of China

Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA
e-mail: xuyuesh@mail.sysu.edu.cn; yxu06@syr.edu

Q. Ye
School of Mathematical Sciences, South China Normal University, Guangzhou, Guangdong, People's Republic of China
e-mail: yeqi@m.scnu.edu.cn

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_38

to Hilbert spaces, there has been much interest in generalizing results for Hilbert spaces to Banach spaces in machine learning. The paper [20] proposed the new concept of reproducing kernel Banach spaces (RKBSs) in the context of machine learning. Based on the new theory of RKBSs, sampling, embedding probability measures, and vector quantization were generalized from RKHSs to RKBSs respectively in [7, 15] and [18]. Moreover, the recent papers [6, 19] showed how to construct explicit representations of support vector machines in RKBSs, which can be easily computed and coded in a similar way as for RKHSs.

RKBSs (see Definition 1) are a natural generalization of RKHSs. We present the underlying idea of RKBSs using the simple framework provided by a finite basis. The goal of this article is to extend the classical support vector machines in RKHSs to RKBSs using finite bases. Specifically, we construct the p-norm RKBSs for $1 \leq p \leq \infty$ using a finite basis. Because a finite-dimensional normed space is always reflexive and complete, we can obtain the same theoretical results as [19] in this simple case without complex assumptions. Moreover, the support vector machine in the 1-norm RKBSs is mathematically equivalent to the sparse support vector machine developed in [16, 21]. By Theorem 3, the special support vector machines in the p_m-norm RKBSs can be solved explicitly when $p_m = \frac{2m}{2m-1}$ (see Eqs. (13) and (14) below). Theorem 4 assures that the p_m-norm support vector machine solution is convergent to the 1-norm support vector machine solution when p_m tends to 1. This shows that the sparse support vector machines are constructed well in the 1-norm RKBSs.

We organize this paper in six sections. In Sect. 2 we present a sparse support vector machine. Section 3 is devoted to a description of RKBSs. We introduce, in Sect. 4, support vector machines in p-norm RKBSs. In Sect. 5 we use the fixed-point proximity algorithm to solve both the classical support vector machines and sparse support vector machines. Numerical experiments are presented in Sect. 5 to demonstrate that the sparse support vector machines outperform the classical support vector machines for simulated data.

2 Sparse Support Vector Machines

The goal of this section is to motivate the idea of sparse support vector machines. We begin with the standard binary classification problem: We choose a hyperplane

$$H := \left\{ \boldsymbol{x} \in \mathbb{R}^d : f_D(\boldsymbol{x}) := \boldsymbol{a}_D^T \boldsymbol{x} + b_D = 0 \right\},$$

to separate a training data site $D := \{(\boldsymbol{x}_k, y_k) : k \in \mathbb{N}_N\}$ composed of input data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \Omega \subseteq \mathbb{R}^d$ and output data values $y_1, \ldots, y_N \in \{\pm 1\}$, where $\mathbb{N}_N := \{1, 2, \ldots, N\}$. Let $X := \{\boldsymbol{x}_k : k \in \mathbb{N}_N\}$. Using the training data D, we can obtain a classification rule induced by the hyperplane H or the decision function f_D to predict labels at unknown locations, that is,

$$r(\boldsymbol{x}) := \operatorname{sign}\left(f_D(\boldsymbol{x})\right), \quad \text{for } \boldsymbol{x} \in \Omega.$$

The coefficients $\boldsymbol{a}_D, b_D$ of the optimal hyperplane are obtained by the hard margin support vector machine

$$\begin{aligned} &\min_{(\boldsymbol{a},b)\in\mathbb{R}^d\times\mathbb{R}} \|\boldsymbol{a}\|_2^2 \\ &\text{s.t. } y_k\left(\boldsymbol{a}^T\boldsymbol{x}_k + b\right) \geq 1, \quad \text{for all } k \in \mathbb{N}_N. \end{aligned} \tag{1}$$

It is also called the maximal margin classifier. To relax the constraints of the optimization problem (1), we introduce some slack variables $\xi_k \geq 0$ and obtain the coefficients $\boldsymbol{a}_D, b_D$ of the optimal hyperplane by the soft margin support vector machine

$$\begin{aligned} &\min_{(\boldsymbol{a},b)\in\mathbb{R}^d\times\mathbb{R}} C\sum_{k\in\mathbb{N}_N} \xi_k + \frac{1}{2}\|\boldsymbol{a}\|_2^2 \\ &\text{s.t. } y_k\left(\boldsymbol{a}^T\boldsymbol{x}_k + b\right) \geq 1-\xi_k, \quad \xi_k \geq 0, \quad \text{for all } k \in \mathbb{N}_N, \end{aligned} \tag{2}$$

where the constant C is a free positive parameter for balancing the margins and the errors. Introducing the positive parameter $\sigma := (2NC)^{-1}$, the optimization problem (2) is then equivalent to

$$\min_{(\boldsymbol{a},b)\in\mathbb{R}^d\times\mathbb{R}} \frac{1}{N}\sum_{k\in\mathbb{N}_N}\left(1-y_k\left(\boldsymbol{a}^T\boldsymbol{x}_k+b\right)\right)_+ + \sigma\|\boldsymbol{a}\|_2^2, \tag{3}$$

where $(\cdot)_+$ is the cutoff function, that is, $(z)_+ := z$ when $z \geq 0$ otherwise $(z)_+ := 0$.

The linear classification may be extended to the nonlinear classification in the sense that the hyperplane is replaced by a manifold. To be more precise, the manifold is given by a decision function f_D composed of a basis $\{\phi_k : \Omega \to \mathbb{R} : k \in \mathbb{N}_n\}$, that is,

$$M := \left\{\boldsymbol{x} \in \Omega : f_D(\boldsymbol{x}) := \boldsymbol{a}_D^T\boldsymbol{\phi}(\boldsymbol{x}) = 0\right\},$$

where $\boldsymbol{\phi} := (\phi_k : k \in \mathbb{N}_n)$. To simplify the discussion and the notation in this article, we focus on finite bases only. However, we do not exclude large basis sizes in the sense $n \gg N$. Analogously to the optimization problem (3), the coefficients $\boldsymbol{a}_D$ of the optimal manifold are obtained by solving the optimization problem

$$\min_{\boldsymbol{a}\in\mathbb{R}^n} \frac{1}{N}\sum_{k\in\mathbb{N}_N}\left(1-y_k\boldsymbol{a}^T\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right)_+ + \sigma\|\boldsymbol{a}\|_2^2. \tag{4}$$

Next, we transfer the optimization problem (4) to another low-dimensional optimization problem. Let the linear space

$$\mathscr{H} := \left\{f := \boldsymbol{a}^T\boldsymbol{\phi} : \boldsymbol{a} \in \mathbb{R}^n\right\},$$

be equipped with the norm

$$\|f\|_{\mathscr{H}} := \|\boldsymbol{a}\|_2 .$$

Clearly, the space $\mathscr{H}$ is a Hilbert space. We also find that $\mathscr{H}$ is a RKHS with the reproducing kernel $K(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{y})$ such that

$$\text{(i) } K(\boldsymbol{x},\cdot) \in \mathscr{H} \text{ and (ii)} f(\boldsymbol{y}) = (f, K(\boldsymbol{x},\cdot))_{\mathscr{H}} ,$$

for all $\boldsymbol{x} \in \Omega$ and all $f \in \mathscr{H}$, where $(\cdot,\cdot)_{\mathscr{H}}$ is an inner product of $\mathscr{H}$. More precisely,

$$(f,g)_{\mathscr{H}} := \boldsymbol{a}^T\boldsymbol{b}, \quad \text{for} f = \boldsymbol{a}^T\boldsymbol{\phi}, g = \boldsymbol{b}^T\boldsymbol{\phi} \in \mathscr{H}.$$

With respect to support vector machines, we can view the RKHS $\mathscr{H}$ as a feature space induced from a feature map $\Phi : \Omega \to \mathscr{H}$ such that $\Phi(\boldsymbol{x}) = K(\boldsymbol{x},\cdot)$ and

$$K(\boldsymbol{x},\boldsymbol{y}) = (K(\cdot,\boldsymbol{y}), K(\boldsymbol{x},\cdot))_{\mathscr{H}} = (\Phi(\boldsymbol{y}), \Phi(\boldsymbol{x}))_{\mathscr{H}} .$$

Here, we can roughly see $\boldsymbol{\phi}$ as an equivalent representer of the feature map Φ in the sense $\Phi(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}$.

Remark 1 Usually, the basis $\{\phi_k : k \in \mathbb{N}_n\}$ is constructed by the positive *eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_n$ and orthonormal *eigenfunctions* $\{e_n : k \in \mathbb{N}_n\} \subseteq \mathrm{L}_2(\Omega)$ of a positive definite kernel K, that is,

$$\phi_k := \sqrt{\lambda_k} e_k, \quad \text{for } k \in \mathbb{N}_n,$$

and

$$\lambda_k e_k(\boldsymbol{x}) = \int_\Omega K(\boldsymbol{x},\boldsymbol{y}) e_k(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}, \quad \text{for all } k \in \mathbb{N}_n.$$

According to the Mercer theorem [17, Theorem 4.49], the positive definite kernel K can possess countable infinite eigenvalues and eigenfunctions such that we can obtain a countable infinite basis.

The construction of the reproducing kernel Hilbert space $\mathscr{H}$ assures that the optimization problem (4) is equivalent to

$$\min_{f\in\mathscr{H}} \frac{1}{N} \sum_{k\in\mathbb{N}_N} (1 - y_k f(\boldsymbol{x}_k))_+ + \sigma \|f\|^2_{\mathscr{H}} . \tag{5}$$

This shows that the support vector machines can be formulated in the feature spaces. By the representer theorems of RKHSs [17, Theorem 5.5], the solution of the optimization problem (5) can be represented by a linear combination of the kernel basis, that is,

$$f_D(\boldsymbol{x}) := \boldsymbol{c}_D^T \boldsymbol{k}_X(\boldsymbol{x}),$$

where $\boldsymbol{c}_D \in \mathbb{R}^N$ is a suitable parameter vector and $\boldsymbol{k}_X(\boldsymbol{x}) := (K(\boldsymbol{x}, \boldsymbol{x}_k) : k \in \mathbb{N}_N)$. Moreover, since

$$\left\| \boldsymbol{c}^T \boldsymbol{k}_X \right\|_{\mathscr{H}} = \|\boldsymbol{c}\|_{\mathsf{K}_X} := \sqrt{\boldsymbol{c}^T \mathsf{K}_X \boldsymbol{c}},$$

where the matrix $\mathsf{K}_X := \left(K(\boldsymbol{x}_j, \boldsymbol{x}_k) : j, k \in \mathbb{N}_N \right) \in \mathbb{R}^{N \times N}$ is positive definite, the optimization problem (5) can be equivalently formulated as

$$\min_{\boldsymbol{c} \in \mathbb{R}^N} \frac{1}{N} \sum_{k \in \mathbb{N}_N} \left(1 - y_k \boldsymbol{c}^T \boldsymbol{k}_X \left(\boldsymbol{x}_k \right) \right)_+ + \sigma \|\boldsymbol{c}\|_{\mathsf{K}_X}^2 . \tag{6}$$

Moreover, the optimal coefficients $\boldsymbol{a}_D$ of f_D based on $\boldsymbol{\phi}$ can be represented by the optimal coefficients $\boldsymbol{c}_D$ of f_D based on $\boldsymbol{k}_X$ in the sense

$$\boldsymbol{a}_D = \Phi_X \boldsymbol{c}_D$$

with the matrix $\Phi_X := (\boldsymbol{\phi}(\boldsymbol{x}_k) : k \in \mathbb{N}_N) \in \mathbb{R}^{n \times N}$. Here, both $\boldsymbol{a}_D$ and $\boldsymbol{c}_D$ are the optimal coefficients of the decision function while $\boldsymbol{a}_D$ and $\boldsymbol{c}_D$ depend on different bases of the reproducing kernel Hilbert spaces. Comparing the equivalent optimization problems (4) and (6), we find that the optimization problem (6) reduces the dimension of the original optimization problem (4).

Fast numerical algorithms for sparse sampling play a central role in the area of signal and image processing. Thus, people are interested in investigating whether support vector machines possess sparsity such as discussed in [16]. A simple idea is to generalize the optimization problem (4) to the 1-norm support vector machines in [21] such as

$$\min_{\boldsymbol{a} \in \mathbb{R}^n} \frac{1}{N} \sum_{k \in \mathbb{N}_N} \left(1 - y_k \boldsymbol{a}^T \boldsymbol{\phi} \left(\boldsymbol{x}_k \right) \right)_+ + \sigma \|\boldsymbol{a}\|_1 . \tag{7}$$

Remark 2 The 1-norm regularization is a rigorous tool to obtain sparse solutions. Generally speaking, sparse solutions of the 0-norm regularization are affected by the sparsity of the matrix Φ_X, that is, the smallest number of the linearly dependent columns from Φ_X. By relaxing the 0-norm to the 1-norm, we obtain sparse solutions by convex relaxation techniques. In many practical applications, optimally sparse learning solutions may not be necessary and we only need certain sparsity in the learning solutions. Many theoretical results of sparse modeling can be found in the well-known review papers [3, 4].

One may expect that the optimization problem (7) could be mathematically transferred into

$$\min_{\boldsymbol{c}\in\mathbb{R}^N} \frac{1}{N} \sum_{k\in\mathbb{N}_N} \left(1 - y_k \boldsymbol{c}^T \boldsymbol{k}_X\left(\boldsymbol{x}_k\right)\right)_+ + \sigma \left\|\boldsymbol{c}\right\|_1. \tag{8}$$

However, the optimization problem (8) then loses the connection to the feature maps. In this article, we shall show that sparse support vector machines such as the optimization problem (7) can also be formulated in feature spaces. To be more precise, we will define the sparse support vector machine in the 1-norm RKBSs. However, solutions of the sparse support vector machines will be different from those of the optimization problem (8).

3 Reproducing Kernel Banach Spaces

In this section, we describe theory of RKBSs defined by a finite basis.

Let $\mathscr{B}$ be a Banach space with the dual space $\mathscr{B}'$. The dual bilinear product $\langle\cdot,\cdot\rangle_{\mathscr{B}}$ is defined on $\mathscr{B}$ and $\mathscr{B}'$, that is,

$$\langle f, G\rangle_{\mathscr{B}} := G(f), \quad \text{for all } f \in \mathscr{B} \text{ and all } G \in \mathscr{B}'.$$

This shows that the norm of $\mathscr{B}'$ can be written as

$$\|G\|_{\mathscr{B}'} = \sup_{\|f\|_{\mathscr{B}}=1} \langle f, G\rangle_{\mathscr{B}}.$$

We say that a normed space $\mathscr{F}$ is isometrically equivalent to the dual space $\mathscr{B}'$ if there is a bijective continuous map T from $\mathscr{F}$ onto $\mathscr{B}'$ such that $\|g\|_{\mathscr{F}} = \|T(g)\|_{\mathscr{B}'}$. Hence, $\mathscr{F}$ can be viewed as a duplicate space of $\mathscr{B}'$ and $g \in \mathscr{F}$ can be seen as a duplicate element of $G = T(g) \in \mathscr{B}'$. We denote that $\mathscr{F} \cong \mathscr{B}'$. Thus, the dual bilinear product $\langle\cdot,\cdot\rangle_{\mathscr{B}}$ can be defined on $\mathscr{B}$ and $\mathscr{F}$ as

$$\langle f, g\rangle_{\mathscr{B}} := \langle f, T(g)\rangle_{\mathscr{B}}, \quad \text{for all } f \in \mathscr{B} \text{ and all } g \in \mathscr{F}.$$

Now we recall the definition of the RKBSs introduced in [19].

Definition 1 Let $\Omega \subset \mathbb{R}^d$ be a domain. Let $\mathscr{B}$ be a Banach space composed of functions $f : \Omega \to \mathbb{R}$. Let $\mathscr{F}$ be a normed space composed of functions $g : \Omega \to \mathscr{R}$ such that the dual space $\mathscr{B}'$ of $\mathscr{B}$ is isometrically equivalent to $\mathscr{F}$. Let $K : \Omega \times \Omega \to \mathbb{R}$ be a kernel function. We call $\mathscr{B}$ a *reproducing kernel Banach space* and K its *reproducing kernel* if

$$\text{(i) } K(\boldsymbol{x},\cdot) \in \mathscr{F}, \ \text{(ii) } \langle f, K(\boldsymbol{x},\cdot)\rangle_{\mathscr{B}} = f(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \Omega \text{ and all } f \in \mathscr{B},$$

$$\text{(iii) } K(\cdot,\boldsymbol{y}) \in \mathscr{B}, \ \text{(iv) } \langle K(\cdot,\boldsymbol{y}), g\rangle_{\mathscr{B}} = g(\boldsymbol{y}), \quad \text{for all } \boldsymbol{y} \in \Omega \text{ and all } g \in \mathscr{F}.$$

Remark 3 The RKBS given in Definition 1 is called a two-sided RKBS in [19, Definition 2.1]. In [6, 19], the RKBSs can be divided into right-sided and left-sided RKBSs and the right-sided and left-sided domains of the reproducing kernels can be different. For convenience, we only discuss the two-sided RKBSs in this article and let the right-sided and left-sided domains of the reproducing kernels be the same. Thus, we shorten the terms two-sided RKBSs and two-sided reproducing kernels by using RKBSs and the reproducing kernels instead. Moreover, the RKBSs given in [20] can still be viewed as the typical case of the RKBSs discussed here. Clearly, the reflexivity and smoothness are not the necessary conditions of the RKBSs.

Now we construct the p-norm RKBS for $1 \leq p \leq \infty$ and the reproducing kernel K by a sequence of functions $\phi_k : \Omega \to \mathbb{R}$ for $k \in \mathbb{N}_n$. Consider the vector $\boldsymbol{\phi} := (\phi_k : k \in \mathbb{N}_n)$ and the kernel

$$K(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{y}), \quad \text{for } \boldsymbol{x},\boldsymbol{y} \in \Omega.$$

In this article, these functions $\{\phi_k : k \in \mathbb{N}_n\}$ will be called a basis of the RKBSs. Thus, we suppose that $\{\phi_k : k \in \mathbb{N}_n\}$ are *linearly independent*, that is, $\boldsymbol{a}^T\boldsymbol{\phi} = 0$ implies that $\boldsymbol{a} = \boldsymbol{0}$. Then we can use the basis vector $\boldsymbol{\phi}$ to construct the linear space

$$\mathscr{B}^p := \left\{ f := \boldsymbol{a}^T\boldsymbol{\phi} : \boldsymbol{a} \in \mathbb{R}^n \right\},$$

equipped with the norm

$$\|f\|_{\mathscr{B}^p} := \|\boldsymbol{a}\|_p .$$

Remark 4 In this article, we only look at RKBSs set up by a finite basis. It is clear that any finite-dimensional normed space is always reflexive and complete. As a result, for this case the complex assumptions and the technical restrictions as stated in [19] are not necessary to construct the p-norm RKBSs. Being different from [19], all theorems for this case can be proved easily by the construction of finite bases.

Theorem 1 *The space $\mathscr{B}^p$ for $1 \leq p \leq \infty$ is a reproducing kernel Banach space with the reproducing kernel K.*

Proof Clearly, $\mathscr{B}^p$ is a Banach space because it is a finite-dimensional normed space. Moreover, the linear independence of $\{\phi_k : k \in \mathbb{N}_n\}$ assures that $\mathscr{B}^p$ and l_p are isometrically isomorphic, where l_p is the collection of all sequences of n scalars with the standard norm $\|\cdot\|_p$. Since the dual space of l_p is isometrically equivalent to l_q, the dual space of $\mathscr{B}^p$ is isometrically equivalent to $\mathscr{B}^q$ where q is the conjugate of p. This assures that the dual bilinear product $\langle \cdot, \cdot \rangle_{\mathscr{B}^p}$ can be represented as

$$\langle f, g \rangle_{\mathscr{B}^p} := \boldsymbol{a}^T\boldsymbol{b}, \quad \text{for } f = \boldsymbol{a}^T\boldsymbol{\phi} \in \mathscr{B}^p \text{ and } g = \boldsymbol{b}^T\boldsymbol{\phi} \in \mathscr{B}^q \cong (\mathscr{B}^p)' .$$

Hence, we can verify for all $\boldsymbol{x},\boldsymbol{y} \in \Omega$ that

$$\langle f, K(\boldsymbol{x},\cdot)\rangle_{\mathscr{B}^p} = \boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \langle K(\cdot,\boldsymbol{y}), g\rangle_{\mathscr{B}^p} = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{y}) = g(\boldsymbol{y}).$$

This completes the proof. □

Remark 5 In this article, we only focus on the finite basis to reduce the mathematical complexity. But, in [19], the RKBSs are introduced by the infinite dimensional basis and the infinite-dimensional basis requires additional conditions for the construction of the p-norm RKBSs in [19, Chapter 3].

Next, we study the map $\varPhi : \varOmega \to \mathscr{B}^p$ defined by $\varPhi(\boldsymbol{x}) := \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}$. Clearly, $\varPhi$ is also a map from $\varOmega$ to $\mathscr{B}^q$. The reproducing properties of $\mathscr{B}^p$ assure that

$$K(\boldsymbol{x},\boldsymbol{y}) = \langle K(\cdot,\boldsymbol{y}), K(\boldsymbol{x},\cdot)\rangle_{\mathscr{B}^p} = \langle \varPhi(\boldsymbol{y}), \varPhi(\boldsymbol{x})\rangle_{\mathscr{B}^p}\,.$$

This shows that the feature map $\varPhi$ is also well-posed for the p-norm RKBSs and we can call the RKBS $\mathscr{B}^p$ a feature space induced by the feature map $\varPhi$.

In the following, we prove advanced properties of the p-norm RKBSs.

Theorem 2

(i) The basis $\{\phi_k : k \in \mathbb{N}_n\}$ *is orthonormal in* $\mathscr{B}^p$ *such that* $\langle \phi_k, \phi_l\rangle_{\mathscr{B}^p} = \delta_{kl}$ *for all* $k,l \in \mathbb{N}_n$*, where* δ_{kl} *is the Kronecker delta function.*

(ii) The linear span $\{\varPhi(\boldsymbol{x}) : \boldsymbol{x} \in \varOmega\}$ *is equal to* $\mathscr{B}^p$*.*

(iii) The point evaluation functional $\delta_{\boldsymbol{x}}$ *is continuous on* $\mathscr{B}^p$ *for any* $\boldsymbol{x} \in \varOmega$*.*

(iv) The RKBS $\mathscr{B}^2$ *is also a RKHS with the reproducing kernel* K*.*

(v) If $\{\phi_k : k \in \mathbb{N}_n\} \subseteq \mathrm{C}(\varOmega)$*, then* $\mathscr{B}^p \subseteq \mathrm{C}(\varOmega)$ *and* $K \in \mathrm{C}(\varOmega \times \varOmega)$*.*

Proof By the construction of $\mathscr{B}^p$, we can verify the properties (i)–(viii) by the same methods of [19, Chapter 2]. □

To close this section, we present several examples of the basis $\left\{\phi_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{N}_m^d\right\}$ induced by the eigenvalues and eigenfunctions of the classical positive definite kernels. Here $\boldsymbol{k} := (k_1,\ldots,k_d)^T$ is an integer vector in the integer set $\mathbb{N}_m^d := \otimes_{j=1}^d \mathbb{N}_m$ and the total number of the basis is $n := m^d$.

(I) *Min Kernels.*

$$\boldsymbol{\phi}_{\boldsymbol{k}}\,(\boldsymbol{x}) := \prod_{j=1}^d \frac{1}{k_j\pi}\sin(k_j\pi x_j),$$

for $\boldsymbol{x} := (x_1,\ldots,x_d)^T \in [0,1]^d$ and $\boldsymbol{k} \in \mathbb{N}_m^d$.

(II) *Gaussian Kernels with the shape parameters* $\theta > 0$.

$$\boldsymbol{\phi}_{\boldsymbol{k}}\,(\boldsymbol{x}) := \prod_{j=1}^d \sqrt{\rho_{\theta,k_j}}e_{\theta,k_j}(x_j), \tag{9}$$

for $\boldsymbol{x} := (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ and $\boldsymbol{k} \in \mathbb{N}_m^d$, where

$$\rho_{\theta,k} := (1 - w_\theta)w_\theta^{k-1}, e_{\theta,k}(x) := \left(\frac{\left(1+4\theta^2\right)^{1/4}}{2^{k-1}(k-1)!}\right)^{1/2} e^{-u_\theta x^2} H_{k-1}\left(\left(1+4\theta^2\right)^{1/4} x\right),$$

$$w_\theta := \frac{2\theta^2}{1 + (1 + 4\theta^2)^{1/2} + 2\theta^2}, \quad u_\theta := \frac{2\theta^2}{1 + (1 + 4\theta^2)^{1/2}}.$$

Here H_k is the Hermite polynomial of degree k, that is,

$$H_k(x) := (-1)^k e^{x^2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(e^{-x^2}\right).$$

(III) *Power Series Kernels.*

$$\boldsymbol{\phi}_{\boldsymbol{k}}(\boldsymbol{x}) := \prod_{j=1}^{d} \sqrt{c_{k_j}} x_j^{k_j},$$

for $\boldsymbol{x} := (x_1, \ldots, x_d)^T \in (-1, 1)^d$ and $\boldsymbol{k} \in \mathbb{N}_m^d$, where the coefficients c_k are the positive coefficients of an analytic function η, that is, $\eta(z) = \sum_{k\in\mathbb{N}_0} c_k z^k$. For example,

$$\eta(z) := e^z, \quad \eta(z) := \frac{1}{1-\theta z} \text{ with } 0 < \theta < 1, \quad \eta(z) := I_0(2z^{1/2}),$$

where I_0 is the modified Bessel function of the first kind of order 0. For this typical basis, the analytic coefficients and the power functions are not the eigenvalues and eigenfunctions of the power series kernels. Moreover, the elements of $\boldsymbol{k}$ can be endowed with 0.

4 Support Vector Machines in *p*-Norm Reproducing Kernel Banach Spaces

In this section, we study support vector machines induced by loss functions.

According to [17], we suppose that loss functions $L(y, \cdot)$ are convex for any $y \in \{\pm 1\}$. Interesting examples of loss functions include

$$\begin{aligned}&\text{the hinge loss: } L(y,t) := (1-yt)_+\,,\\ &\text{the squared hinge loss: } L(y,t) := (1-yt)_+^2\,,\\ &\text{the least square loss: } L(y,t) := (1-yt)^2\,.\end{aligned} \tag{10}$$

For given training data $D \subseteq \Omega \times \{\pm 1\}$, the decision function will be obtained by the support vector machine defined in the p-norm RKBSs $\mathscr{B}^p$, for $1 \leq p \leq \infty$. That is,

$$\min_{f \in \mathscr{B}^p} \frac{1}{N} \sum_{k \in \mathbb{N}_N} L\left(y_k, f(\boldsymbol{x}_k)\right) + \sigma \|f\|_{\mathscr{B}^p}, \tag{11}$$

where the regularization parameter σ is positive. We call the optimization problem (11) the *p-norm support vector machine*. By the construction of the RKBS $\mathscr{B}^p$ (see Theorem 1), the optimization problem (11) can be equivalently stated as

$$\min_{\boldsymbol{a} \in \mathbb{R}^n} \frac{1}{N} \sum_{k \in \mathbb{N}_N} L\left(y_k, \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}_k)\right) + \sigma \|\boldsymbol{a}\|_p, \tag{12}$$

where $\boldsymbol{\phi}$ is composed of the orthonormal basis of $\mathscr{B}^p$. Obviously, when L is the hinge loss and $p = 1$, the optimization problem (12) is the same as the 1-norm support vector machine described in Eq. (7). For convenience, we further suppose that $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_N)$ are linearly independent. This ensures that the feature elements $\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_N)$ are linearly independent.

We shall mainly focus on the sparse support vector machines such as the 1-norm support vector machines. Before the discussion of the sparse support vector machines, we study the typical support vector machines defined in the strictly convex and smooth RKBSs first. Let

$$p_m := \frac{2m}{2m-1} \text{ and } q_m := 2m, \quad \text{for } m \in \mathbb{N}.$$

Then, $p_m \to 1$ when $m \to \infty$, and the p_m-norm RKBS $\mathscr{B}^{p_m}$ is strictly convex and smooth because l_{p_m} is strictly convex and smooth when $p_m > 1$. By the strict convexity and smoothness of the p_m-norm RKBS $\mathscr{B}^{p_m}$, [19, Theorem 5.10] guarantees that the optimization problem

$$\min_{f \in \mathscr{B}^{p_m}} \frac{1}{N} \sum_{k \in \mathbb{N}_N} L\left(y_k, f(\boldsymbol{x}_k)\right) + \sigma \|f\|_{\mathscr{B}^{p_m}}, \tag{13}$$

has the unique solution

$$f_D^{*m}(\boldsymbol{x}) := \boldsymbol{\beta}_N^{*m} (\boldsymbol{c}_D^{*m})^T \boldsymbol{k}_X^{*m}(\boldsymbol{x}), \quad \text{for } \boldsymbol{x} \in \Omega, \tag{14}$$

where $\boldsymbol{c}_D^{*m} \in \mathbb{R}^N$ is the suitable parameter vector and the multivariate kernel vector

$$\boldsymbol{k}_X^{*m}(\boldsymbol{x}) := \left(K^{*m}\left(\boldsymbol{x}, \boldsymbol{x}_{k_1}, \ldots, \boldsymbol{x}_{k_{2m-1}}\right) : k_1, \ldots, k_{2m-1} \in \mathbb{N}_N\right).$$

We define the vector-valued function

$$\boldsymbol{\beta}_N^{*m}(\boldsymbol{c}) := \left(\prod_{j=1}^{2m-1} c_{k_j} : k_1, \ldots, k_{2m-1} \in \mathbb{N}_N\right), \quad \text{for } \boldsymbol{c} := (c_k : k \in \mathbb{N}_N) \in \mathbb{R}^N,$$

and the multivariate kernel $K^{*m} : \otimes_{j=1}^{2m-1} \Omega \to \mathbb{R}$

$$K^{*m}(\boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2m-1}) := \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}^{*m}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2m-1}), \quad \text{for } \boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2m-1} \in \Omega,$$

where

$$\boldsymbol{\phi}^{*m}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2m-1}) := \left(\prod_{j=1}^{2m-1} \phi_k(\boldsymbol{y}_j) : k \in \mathbb{N}_n\right).$$

Clearly, we have that $\boldsymbol{\phi}^{*1} = \boldsymbol{\phi}$, $K^{*1} = K$, $\boldsymbol{k}_X^{*1} = \boldsymbol{k}_X$, and $\boldsymbol{\beta}_N^{*1}(\boldsymbol{c}_D^{*1}) = \boldsymbol{c}_D^{*1} = \boldsymbol{c}_D$. This indicates that f_D^{*1} is consistent with the classical decision function f_D obtained in the reproducing kernel Hilbert space $\mathscr{H} = \mathscr{B}^2$.

Remark 6 For convenience of presentation, we change the notation of the paper [19] slightly. The main technique of the representation of the decision function in the p_m-norm RKBS $\mathscr{B}^{p_m}$ is based on the representer theorem in the RKBSs which verifies that the Gâteaux derivatives of the p_m-norm at the decision functions are the linear combinations of the feature elements $\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_N)$. More details of the proof can be found in [19, Section 2.6].

Next, we discuss how a suitable parameter vector $\boldsymbol{c}_D^{*m}$ is obtained. Note that [19, Theorem 5.10] also provides a representation of the p_m-norm of the decision function, that is,

$$\left\|\boldsymbol{\beta}_N^{*m}(\boldsymbol{c})^T \boldsymbol{k}_X^{*m}\right\|_{\mathscr{B}^{p_m}} = \|\boldsymbol{c}\|_{\mathsf{K}_X^{*m}} := \left(\boldsymbol{c}^T \mathsf{K}_X^{*m} \boldsymbol{\beta}_N^{*m}(\boldsymbol{c})\right)^{1-1/2m}, \tag{15}$$

where the matrix $\mathsf{K}_X^{*m} := \left(\boldsymbol{k}_X^{*m}(\boldsymbol{x}_k) : k \in \mathbb{N}\right)^T \in \mathbb{R}^{N \times N^{2m-1}}$.

Remark 7 It is clear that $\mathsf{K}_X^{*1} = \mathsf{K}_X$. Actually, K_X^{*m} can be viewed as a $2m$-dimensional positive definite matrix or a tensor. More precisely, the matrix K_X^{*m} can be rewritten as

$$\left(K^{*m}(\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}, \ldots, \boldsymbol{x}_{k_{2m}}) : k_1, k_2, \ldots, k_{2m} \in \mathbb{N}_N\right) \in \mathbb{R}^{N \times \ldots \times N}.$$

Since the $2m$-dimensional positive definite matrix is a new concept, we do not consider the high-dimensional format of K_X^{*m} here and just view K_X^{*m} as a regular matrix.

Theorem 3 *The p_m-norm support vector machine* (13) *has a unique solution f_D^{*m} given in Eq.* (14)*, where the parameter vector $\boldsymbol{c}_D^{*m}$ uniquely solves the optimization problem*

$$\min_{\boldsymbol{c}\in\mathbb{R}^N} \frac{1}{N} \sum_{k\in\mathbb{N}_N} L\left(y_k, \boldsymbol{\beta}_N^{*m}(\boldsymbol{c})^T \boldsymbol{k}_X^{*m}(\boldsymbol{x}_k)\right) + \sigma \left\|\boldsymbol{c}\right\|_{\mathsf{K}_X^{*m}}. \tag{16}$$

Proof Combining Eqs. (14) and (15), the optimization problem (13) can be equivalently stated as the optimization problem (16). □

When $m = 1$, $\boldsymbol{c}_D^{*1} = \boldsymbol{c}_D$ are the coefficients of the decision function $f_D^{*1} = f_D$ based on the kernel basis $\boldsymbol{k}_X^{*1} = \boldsymbol{k}_X$. But $\boldsymbol{c}_D^{*m}$ are not the coefficients of another decision function f_D^{*m} based on the multivariate kernel basis $\boldsymbol{k}_X^{*m}$ when $m > 1$. Generally, the coefficients $\boldsymbol{\beta}_N^{*m}(\boldsymbol{c}_D^{*m})$ of f_D^{*m} based on $\boldsymbol{k}_X^{*m}$ are obtained by the vector-valued function $\boldsymbol{\beta}_N^{*m}$ at $\boldsymbol{c}_D^{*m}$. Moreover, the parameter vector $\boldsymbol{c}_D^{*m}$ will be different for various m and the decision function f_D^{*m} needs to be solved case-by-case for m.

We next look at the coefficients of the decision function f_D^{*m} based on $\boldsymbol{\phi}$. By expansion of Eq. (14), the decision function f_D^{*m} can be rewritten as

$$f_D^{*m} = \boldsymbol{a}_X^{*m}(\boldsymbol{c}_D^{*m})^T \boldsymbol{\phi},$$

where

$$\boldsymbol{a}_X^{*m}(\boldsymbol{c}_D^{*m}) := \left(\Phi_X \boldsymbol{c}_D^{*m}\right)^{2m-1} = \operatorname{diag}\left(\Phi_X \boldsymbol{c}_D^{*m}\right)^{2m-2} \left(\Phi_X \boldsymbol{c}_D^{*m}\right).$$

Here, the power m of a vector $\boldsymbol{a} := (a_k : k \in \mathbb{N}_n) \in \mathbb{R}^n$ is defined by

$$\boldsymbol{a}^m := \left(a_k^m : k \in \mathbb{N}_n\right),$$

and the diag is defined by

$$\operatorname{diag}(\boldsymbol{a}) := \begin{pmatrix} a_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & a_n \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Then, $\boldsymbol{a}_X^{*1}(\boldsymbol{c}_D^{*m}) = \Phi_X \boldsymbol{c}_D$ is the same as the coefficients of the classical decision function f_D based on $\boldsymbol{\phi}$. This shows that the support vector machines defined in Banach spaces can also be reduced to the low-dimensional optimization problems.

Now we investigate the 1-norm support vector machines

$$\min_{\boldsymbol{a}\in\mathbb{R}^n} \frac{1}{N} \sum_{k\in\mathbb{N}_N} L\left(y_k, \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}_k)\right) + \sigma \left\|\boldsymbol{a}\right\|_1. \tag{17}$$

The geometrical structure of the 1-norm assures that the optimization problem (17) possesses sparse solutions for $n \gg N$. By the construction of the 1-norm RKBS $\mathscr{B}^1$, optimization problem (17) is equivalent to

$$\min_{f\in\mathscr{B}^1} \frac{1}{N} \sum_{k\in\mathbb{N}_N} L\left(y_k, f(\boldsymbol{x}_k)\right) + \sigma \left\|f\right\|_{\mathscr{B}^1}. \tag{18}$$

We have the following approximation theorem.

Theorem 4 *The* 1*-norm support vector machine* (18) *has a solution* f_D^* *which is a weak* accumulation point of the sequence* $\left\{f_D^{*m} : m \in \mathbb{N}\right\}$ *given in Theorem 3.*

Proof According to [19, Theorem 5.9], the decision function f_D^* given by the optimization problem (18) can be approximated by the decision function f_D^{*m} given by the optimization problems (13) such that f_D^{*m} is weakly* convergent to f_D^*. □

Here, the weak* accumulation point in Theorem 4 means that there exists a subsequence $\left\{f_D^{*m_k} : k \in \mathbb{N}\right\}$ of $\left\{f_D^{*m} : m \in \mathbb{N}\right\}$ such that $f_D^{*m_k}$ converges to the decision function f_D^* in the weak* topology, that is,

$$\langle g, f_D^{*m_k}\rangle_{\mathscr{B}^\infty} \to \langle g, f_D^*\rangle_{\mathscr{B}^\infty}, \quad \text{when } m_k \to \infty,$$

for all $g \in \mathscr{B}^\infty$. This indicates that

$$f_D^{*m_k}(\boldsymbol{x}) = \langle \Phi(\boldsymbol{x}), f_D^{*m_k}\rangle_{\mathscr{B}^\infty} \to \langle \Phi(\boldsymbol{x}), f_D^{*\infty}\rangle_{\mathscr{B}^\infty} = f_D^*(\boldsymbol{x}), \quad \text{when } m_k \to \infty,$$

for all $\boldsymbol{x} \in \Omega$. Let $\boldsymbol{a}^*$ be the coefficients of f_D^* based on $\boldsymbol{\phi}$, that is, $f_D^* = \boldsymbol{a}^{*T}\boldsymbol{\phi}$. If we take $g := \phi_j$ for all $j \in \mathbb{N}_n$, then

$$\boldsymbol{a}_X^{*m_k}(\boldsymbol{c}_D^{*m_k}) \to \boldsymbol{a}^*, \quad \text{when } m_k \to \infty;$$

hence

$$\left\|\boldsymbol{a}_X^{*m_k}(\boldsymbol{c}_D^{*m_k}) - \boldsymbol{a}^*\right\|_1 \to 0, \quad \text{when } m_k \to \infty,$$

and

$$\left\|f_D^{*m_k} - f_D^*\right\|_{\mathscr{B}^1} \to 0, \quad \text{when } m_k \to \infty.$$

Remark 8 In this article, we only focus on the finite-dimensional RKBSs. As a result, the basis vector $\boldsymbol{\phi}$ is finite. If $\boldsymbol{\phi}$ is infinite, the weak* convergence may not imply the 1-norm convergence because the feature elements $\{\Phi(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}$ are not the Schauder bases of the infinite-dimensional RKBSs and we can only determine that span $\{\Phi(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}$ is dense in the infinite-dimensional RKBSs. However, the multivariate kernel bases of the decision functions in the finite-dimensional RKBSs are still finite. More details of the learning problems in the infinite-dimensional RKBSs may be found in [19].

Since the 1-norm geometry is not strictly convex, solutions of the 1-norm optimization problems may not be unique. Thus, there will be different subsequences of $\left\{f_D^{*m} : m \in \mathbb{N}\right\}$ which converge to various optimal solutions. If the optimization problem (18) has the unique solution f_D^* (e.g., $L(y, \cdot)$ is strictly convex for any $y \in \{\pm 1\}$), then the sequence $\left\{f_D^{*m} : m \in \mathbb{N}\right\}$ converges to f_D^*, that is, $f_D^{*m} \to f_D^*$ when $m \to \infty$. In fact, we only need one sparse solution to construct the support vector classifiers. In the following section, we shall investigate numerical algorithms of the sparse support vector machines.

5 Numerical Experiments

In this section, we develop efficient algorithms to solve the optimization problems (4) and (7). Because the optimization problems (4) and (7) are equivalent to the support vector machines in the RKHS and the 1-norm RKBS, respectively, we call the optimization problem (4) the 2-norm SVM and (7) the 1-norm SVM.

As we see, the hinge loss function and the ℓ^1-norm are non-differentiable, which poses a big challenge for solving the problems (4) and (7). However, we shall use the recently developed fixed-point proximity algorithm [8–13] to directly solve these problems. We first follow the same idea in [10] to rewrite the hinge loss function as an equivalent compact form for the convenience of developing proximity algorithms. We define a $N \times n$ matrix

$$\mathsf{B} := \operatorname{diag}(y_i : i \in \mathbb{N}_N)\Phi_X.$$

For $z \in \mathbb{R}^N$, let

$$\mathscr{L}(z) := \frac{1}{N} \sum_{i \in \mathbb{N}_N} (1 - z_i)_+. \tag{19}$$

Then the hinge loss function can be rewritten as $\mathscr{L}(\mathsf{B}\boldsymbol{a})$. We remark that the squared hinge loss function can thus be rewritten as $\mathscr{L}_s(\mathsf{B}\boldsymbol{a})$ with

$$\mathscr{L}_s(z) := \frac{1}{N} \sum_{i \in \mathbb{N}_N} (1 - z_i)_+^2. \tag{20}$$

Since the proximity operator is essential in developing fixed-point proximity algorithms, we first recall it. The proximity operator of a convex function g is defined for $z \in \mathbb{R}^n$ by

$$\operatorname{prox}_g(z) := \operatorname{argmin}\left\{\frac{1}{2}\|s - z\|_2^2 + g(s) : s \in \mathbb{R}^n\right\}.$$

Now we present a fixed-point proximity algorithm to solve the 2-norm SVM. Note that by the above reformulation, the objective function of the 2-norm SVM consists of two terms including the differentiable 2-norm regularization term and the non-differentiable term. A class of proximity algorithms were developed in [12] to handle optimization problems of this kind. We follow the same idea to derive an algorithm to solve the problem (4).

Algorithm 1 Fixed-point proximity algorithm to solve the 2-norm SVM

```
Given: the identity matrix I, the reformulation matrix B, the positive parameters $\sigma$, $\mu$, and $\kappa \in (0, 1)$;
Initialization: $\boldsymbol{v}^0$;
for $k = 1, 2, \ldots,$ do
    $\boldsymbol{v}^{k+1} = (1-\kappa)\boldsymbol{v}^k + \kappa(\mathsf{I} - \operatorname{prox}_{\frac{1}{\mu}\mathscr{L}})(\boldsymbol{v}^k - \frac{\mu}{2\sigma}\mathsf{B}\mathsf{B}^\top \boldsymbol{v}^k);$
end for
Denote by $\boldsymbol{v}$ the convergent point, compute $\boldsymbol{a} = -\frac{\mu}{2\sigma}\mathsf{B}^\top \boldsymbol{v}$.
```

We next consider the 1-norm SVM (7). Let

$$\mathscr{R}(\boldsymbol{z}) := \sigma\|\boldsymbol{z}\|_1 \text{ for any } \boldsymbol{z} \in \mathbb{R}^n. \tag{21}$$

We observe from the above reformulation that the objective function of the 1-norm SVM consists of two non-differentiable terms. The fixed-point proximity algorithms [8–11, 13] have shown their nice performance for solving non-differentiable optimization problems. In particular, a class of two-step fixed-point proximity algorithms for specifically solving the 1-norm SVM problems were developed in [10]. Therefore, we apply one of the efficient algorithms in [10] to solve problem (7).

Algorithm 2 Two-step fixed-point proximity algorithm for solving the 1-norm SVM

```
Given: the identity matrix I, the reformulation matrix B, the positive parameters $\sigma$, $\beta$, $\epsilon$, and $\omega$;
Initialization: $\boldsymbol{a}^0$, $\boldsymbol{v}^0$;
for $k = 1, 2, \ldots,$ do
    $\boldsymbol{v}^{k+1} = (\mathsf{I} - \operatorname{prox}_{\frac{1}{\epsilon}\mathscr{L}})(\boldsymbol{v}^k + \mathsf{B}(\boldsymbol{a}^k + \omega(\boldsymbol{a}^k - \boldsymbol{a}^{k-1})))$
    $\boldsymbol{a}^{k+1} = \operatorname{prox}_{\frac{1}{\sigma\beta}\mathscr{R}}(\boldsymbol{a}^k - \frac{\epsilon}{\sigma\beta}\mathsf{B}^\top(\boldsymbol{v}^{k+1} + (1-\omega)(\boldsymbol{v}^{k+1} - \boldsymbol{v}^k))).$    (22)
end for
```

Remark 9 The convergence of Algorithms 1 and 2 can be found in [10].

We summarize the relationships of the above support vector machines in the following table:

Support vector machines	Equivalent optimization problems
General p-norm, general loss	(11), (12)
Typical p_m-norm, general loss	(13), (16)
1-norm, general loss	(17), (18)
1-norm, hinge loss	(7)
2-norm, hinge loss	(4), (5)
Hyperplane, hinge loss	(2), (3) $\approx$ (1)

The 2-norm, p_m-norm, and 1-norm RKBSs have the same reproducing kernels K, feature maps Φ, and basis $\boldsymbol{\phi}$. But, their support vector machines induced by the same loss functions are still different such as

Support vector machines	2-norm RKBSs	p_m-norm RKBSs	1-norm RKBSs
Decision functions	$f_D = f_D^{*1}$	f_D^{*m}	f^*
Kernel basis	$\boldsymbol{k}_X = \boldsymbol{k}_X^{*1}$, $K = K^{*1}$	$\boldsymbol{k}_X^{*m}$, K^{*m}	Non
Coefficients of kernel basis	$\boldsymbol{c}_D = \boldsymbol{c}_D^{*1}$	$\boldsymbol{\beta}_N^{*m}\left(\boldsymbol{c}_D^{*m}\right)$	Non
Coefficients of $\boldsymbol{\phi}$	$\Phi_X \boldsymbol{c}_D = \boldsymbol{a}_X^{*1}\left(\boldsymbol{c}_D^{*1}\right)$	$\boldsymbol{a}_X^{*m}\left(\boldsymbol{c}_D^{*m}\right)$	$\boldsymbol{a}^*$

Below, we present numerical results to demonstrate advantages of learning in RKBS over that in RKHS. Specifically, we implement Algorithms 1 and 2 for problem (4) and (7), respectively, with both the hinge loss function and the squared hinge loss function. All the numerical experiments are implemented on a personal computer with a 2.6 GHz Intel Core i5 CPU and an 8G RAM memory.

We compare these methods on two artificial data sets. The first data set contains 500 instances. They are randomly generated on the domain $[0, 1] \times [0, 1]$. The instance lying inside the circle $x_1^2 + x_2^2 = 0.25^2$ or $(x_1 - 0.75)^2 + (x_2 - 0.5)^2 = 0.25^2$ has the positive label $+1$, others have the negative label -1. We randomly choose 200 instances as training data, and 300 instances as testing data. We use the Gaussian kernel bases introduced in (9) in Sect. 3 with $d = 2$ and $m = 10$ for both of the algorithms. Thus, the total number of the basis functions is $m^2 = 100$, see (9) in Sect. 3 for more details. We apply Algorithm 1 to train the 2-norm SVM and Algorithm 2 to train the 1-norm SVM on this data set. The hinge loss function and the squared hinge loss function are used as the fidelity term for each of the models. In all cases, the parameters are tuned to approximately achieve the best test accuracy performance. We compare the numbers of support vectors, the training accuracy and the test accuracy of training by the 1-norm SVM and the 2-norm SVM. We remark that the support vector is valid if the value of its coefficient is greater than 10^{-8}. The stopping criterion of each algorithm is set to be the relative error between the successive iterations less than a given tolerance, which we set as 10^{-8} in these experiments. The numerical results are presented in Figs. 1, 2, and Table 1.

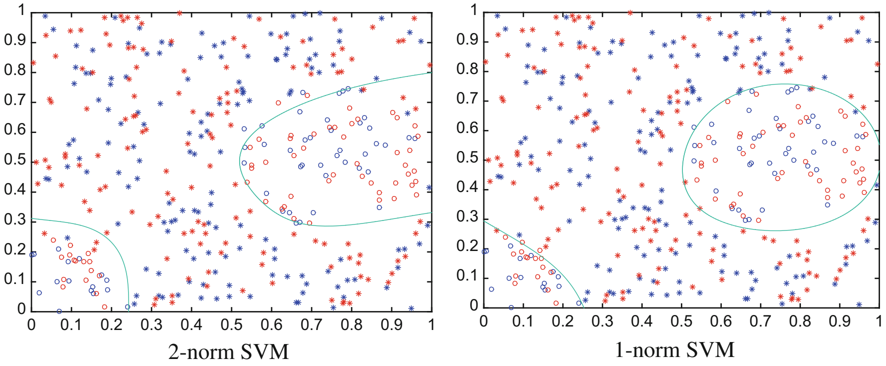


Fig. 1 The result of training the 2-norm SVM and the 1-norm SVM with hinge loss function for binary classification in the domain $[0, 1]^2$. An instance with label $+1$ is denoted by a circle, and an instance with label -1 is denoted by a star. The blue symbols are training data and the red symbols are testing data. The left figure is the result of training the 2-norm SVM with the model parameter $\sigma = 1$ and the Gaussian bases parameters $\theta = 10, m = 10$. The right figure shows that result of training the 1-norm SVM with the model parameter $\sigma = 0.0001$ and the Gaussian bases parameters $\theta = 2, m = 10$

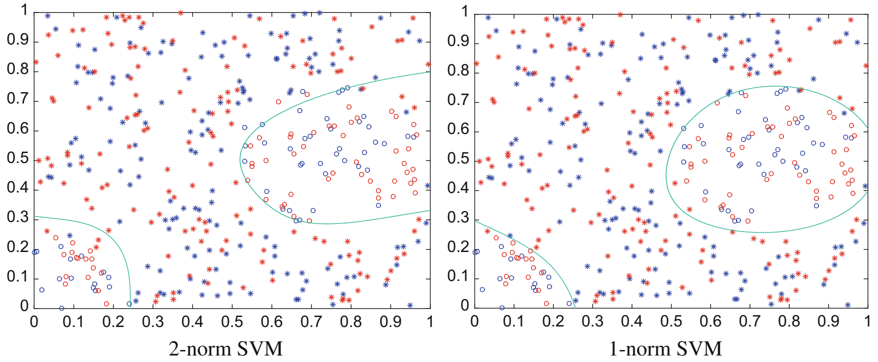


Fig. 2 The result of training the 2-norm SVM and the 1-norm SVM with squared hinge loss function for binary classification in the domain $[0, 1]^2$. An instance with label $+1$ is denoted by a circle, and an instance with label -1 is denoted by a star. The blue symbols are training data and the red symbols are testing data. The left figure is the result of training the 2-norm SVM with the model parameter $\sigma = 1$ and the Gaussian bases parameters $\theta = 10, m = 10$. The right figure shows that result of training the 1-norm SVM with the model parameter $\sigma = 0.0001$ and the Gaussian bases parameters $\theta = 2, m = 10$

The second data set is a little more complex than the first one. It contains 400 instances, and is also randomly generated on the domain $[0, 1] \times [0, 1]$. The instance lying inside the circle $x_1^2 + x_2^2 = 0.25^2$, $(x_1 - 0.75)^2 + (x_2 - 0.5)^2 = 0.25^2$, or $x_1^2 + (x_2 - 1)^2 = 0.25^2$ has the positive label $+1$, others have the negative label -1. Therefore, this data set has three separate regions for the positive data. We set 200 instances as training data, and 200 instances as testing data. We also apply

Table 1 Comparison of 2-norm SVM and 1-norm SVM in the numbers of support vectors, training accuracy and testing accuracy

	Data set 1 HL			Data set 1 SHL			Data set 2 HL		
	SVs	Train (%)	Test (%)	SVs	Train (%)	Test (%)	SVs	Train (%)	Test (%)
ℓ^2-SVM	100	96.00	96.00	100	96.00	96.00	100	92.00	92.50
ℓ^1-SVM	35	98.50	97.00	42	98.00	98.00	68	100.00	95.00

Data Set 1 HL is the first data set we mentioned above with hinge loss function used as the fidelity term, Data Set 1 SHL is also the first data set and squared hinge loss function is used as the fidelity term, and Data Set 2 HL is the second data set and the hinge loss function is used as the fidelity term

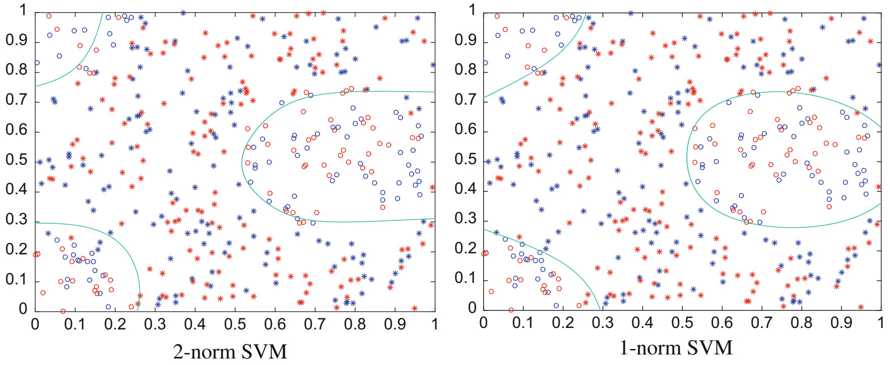


Fig. 3 The result of training 2-norm SVM and 1-norm SVM with hinge loss function for binary classification in the domain $[0, 1]^2$. This figure is different from Fig. 1; it has three separate region for label $+1$. An instance with label $+1$ is denoted by a circle, and an instance with label -1 is denoted by a star. The blue symbols are training data and the red symbols are testing data. The left figure is the result of training 2-norm SVM with the model parameter $\sigma = 0.1$ and the Gaussian bases parameters $\theta = 10, m = 10$. The right figure shows that result of training 1-norm SVM with the model parameter $\sigma = 0.00001$ and the Gaussian bases parameters $\theta = 2, m = 10$

Algorithm 1 to train the 2-norm SVM and Algorithm 2 to train the 1-norm SVM on this data set. The hinge loss function is used as the fidelity term. We also compare the numbers of support vectors, the training accuracy and the testing accuracy of training by the 1-norm SVM and the 2-norm SVM. We present the numerical result in Fig. 3 and Table 1.

We observe from Figs. 1, 2, 3 and Table 1 that in the two datasets, 1-norm SVM achieves higher training and testing accuracy, while using much less support vectors. This strongly shows the advantage of learning in RKBSs over that in RKHSs.

Acknowledgements The first author is supported in part by the Special Project on High-performance Computing under the National Key R&D Program (No. 2016YFB0200602), and by the Natural Science Foundation of China under grants 11471013 and 91530117. The third author would like to express his gratitude to the grant of the "Thousand Talents Program" for junior scholars of China, the grant of the Natural Science Foundation of China (11601162), and the grant of South China Normal University (671082, S80835, and S81031).

References

1. Alpaydin, E.: Introduction to Machine Learning. MIT, Cambridge, MA (2010)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
3. Bruckstein, A., Donoho, D., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
4. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
5. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. (N.S.) **39**(1), 1–49 (2002) (electronic)
6. Fasshauer, G., Hickernell, F., Ye, Q.: Solving support vector machines in reproducing kernel Banach spaces with positive definite functions. Appl. Comput. Harmon. Anal. **38**, 115–139 (2015)
7. García, A., Portal, A.: Sampling in reproducing kernel Banach spaces. Mediterr. J. Math. **10**, 1401–1417 (2013)
8. Li, Q., Micchelli, C.A., Shen, L., Xu, Y.: A proximity algorithm accelerated by Gauss-Seidel iterations for L1/TV denoising models. Inverse Prob. **28**(9), 095003 (2012)
9. Li, Q., Shen, L., Xu, Y., Zhang, N.: Multi-step proximity algorithms for solving a class of convex optimization problems. Adv. Comput. Math. **41**(2), 387–422 (2014)
10. Li, Z., Song, G., Xu, Y.: Fixed-point proximity algorithms for solving sparse machine learning models. Int. J. Numer. Anal. Model. **15**(1–2), 154–169 (2018)
11. Li, Q., Xu, Y., Zhang, N.: Two-step fixed-point proximity algorithms for multi-block separable convex problems. J. Sci. Comput. **70**, 1204–1228 (2017)
12. Micchelli, C.A., Shen, L., Xu, Y.: Proximity algorithms for image models: denoising. Inverse Prob. **27**(4), 045009, 30 (2011)
13. Micchelli, C.A., Shen, L., Xu, Y., Zeng, X.: Proximity algorithms for the L1/TV image denoising model. Adv. Comput. Math. **38**(2), 401–426 (2013)
14. Schaback, R., Wendland, H.: Kernel techniques: from machine learning to meshless methods. Acta Numer. **15**, 543–639 (2006)
15. Sriperumbudur, B., Fukumizu, K., Lanckriet, G.: Learning in Hilbert vs.Banach spaces: a measure embedding viewpoint. In: Advances in Neural Information Processing Systems, pp. 1773–1781. MIT, Cambridge (2011)
16. Steinwart, I.: Sparseness of support vector machines. J. Mach. Learn. Res. **4**, 1071–1105 (2003)
17. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
18. Villmann, T., Haase, S., Kästner, M.: Gradient based learning in vector quantization using differentiable kernels. In: Advances in Self-Organizing Maps, pp. 193–204. Springer, Santiago (2013)
19. Xu, Y., Ye, Q.: Generalized Mercer kernels and reproducing kernel Banach spaces. Mem. AMS (accepted). arXiv:1412.8663
20. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. J. Mach. Learn. Res. **10**, 2741–2775 (2009)
21. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-norm support vector machines. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) The Annual Conference on Neural Information Processing Systems 16, pp. 1–8 (2004)

Mean Convergence of Interpolation at Zeros of Airy Functions

Doron S. Lubinsky

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The classical Erdős-Turán theorem established mean convergence of Lagrange interpolants at zeros of orthogonal polynomials. A non-polynomial extension of this was established by Ian Sloan in 1983. Mean convergence of interpolation by entire functions has been investigated by Grozev, Rahman, and Vértesi. In this spirit, we establish an Erdős-Turán theorem for interpolation by entire functions at zeros of the Airy function.

1 Introduction

The classical Erdős-Turán theorem involves a weight w on an compact interval, which we take as $[-1, 1]$. We assume that $w \geq 0$ and is positive on a set of positive measure. Let p_n denote the corresponding orthonormal polynomial of degree $n \geq 0$, so that for $m, n \geq 0$,

$$\int_{-1}^{1} p_n p_m w = \delta_{mn}.$$

Let us denote the zeros of p_n in $[-1, 1]$ by

$$-1 < x_{nn} < x_{n-1,n} < \cdots < x_{2n} < x_{1n} < 1.$$

D. S. Lubinsky (✉)
Georgia Institute of Technology, School of Mathematics, Atlanta, GA, USA
e-mail: lubinsky@math.gatech.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_39

Given $f : [-1, 1] \to \mathbb{R}$, let $L_n [f]$ denote the Lagrange interpolation polynomial to f at $\{x_{jn}\}_{j=1}^{n}$, so that $L_n [f]$ has degree at most $n-1$ and

$$L_n [f] \left(x_{jn}\right) = f \left(x_{jn}\right), \quad 1 \leq j \leq n.$$

Theorem 1 (Erdős-Turán Theorem) *Let $f : [-1, 1] \to \mathbb{R}$ be continuous. For $n \geq 1$, let $L_n [f]$ denote the Lagrange interpolation polynomial to f at the zeros of p_n. Then*

$$\lim_{n\to\infty} \int_{-1}^{1} (f - L_n [f])^2 \, w = 0.$$

The ramifications of this result continue to be explored to this day. It has been extended in numerous directions: for example, rather than requiring f to be continuous, we can allow it to be Riemann integrable. We may replace w by a positive measure μ, which may have non-compact support. In addition, convergence in L_2 may be replaced, under additional conditions on w, by convergence in L_p. There is a very large literature on all of this. See [4, 10–13, 22, 23] for references and results.

Ian Sloan and his coauthor Will Smith ingeniously used results on mean convergence of Lagrange interpolation in various L_p norms, to establish the definitive results on convergence of product integration rules [14, 16, 18–21]. This is a subject of substantial practical importance, for example in numerical solution of integral equations.

One can speculate that it was this interest in product integration that led to Ian Sloan extending the Erdős-Turán theorem to non-polynomial interpolation. Here is an important special case of his general result [17, p. 99]:

Theorem 2 (Sloan's Erdös-Turán Theorem on Sturm-Liouville Systems) *Consider the eigenvalue problem*

$$p(x) \, u''(x) + q(x) \, u'(x) + [r(x) + \lambda] \, u(x) = 0$$

with boundary conditions

$$(\cos\alpha) \, u(a) + (\sin\alpha) \, u'(a) = 0;$$
$$(\cos\beta) \, u(b) + (\sin\beta) \, u'(b) = 0.$$

Assume that p'', q', r are continuous and real valued on $[a, b]$, that $p > 0$ there, while α, β are real. Let $\{u_n\}_{n\geq 0}$ be the eigenfunctions, ordered so that the corresponding eigenvalues $\{\lambda_n\}$ are increasing. Given continuous $f : [a, b] \to \mathbb{R}$, let $\mathscr{L}_n [f]$ denote the linear combination of $\{u_j\}_{j=0}^{n}$ that coincides with f at the $n + 1$ zeros of u_{n+1} in the open interval (a, b). Let

$$w(x) = \frac{1}{p(x)} \exp\left(\int_a^x \frac{q(t)}{p(t)} dt\right), \quad x \in [a, b].$$

Then

$$\lim_{n\to\infty}\int_a^b \left(f(x)-\mathscr{L}_n[f](x)\right)^2 w(x)\,dx = 0,$$

provided $f(a) = 0$ *if* $\sin\alpha = 0$ *and* $f(b) = 0$ *if* $\sin\beta = 0$. *Moreover, there is a constant c independent of n and f such that for all such f,*

$$\int_a^b \left(f(x)-\mathscr{L}_n[f](x)\right)^2 w(x)\,dx$$
$$\le C \inf_{c_0,c_1,\dots,c_n}\int_a^b \left(f(x)-\sum_{j=0}^n c_j u_j(x)\right)^2 w(x)\,dx.$$

As a specific example, Sloan considers the Bessel equation. His general theorem [17, p. 102], from which the above result is deduced, involves orthonormal functions, associated reproducing kernels, and interpolation points satisfying two boundedness conditions. In 1988, M. R. Akhlaghi [2] extended Sloan's result to convergence in L_p for $p \ge 1$.

Interpolation by trigonometric polynomials is closely related to that by algebraic polynomials, in as much as every even trigonometric polynomial has the form $P(\cos\theta)$ where P is an algebraic polynomial. From trigonometric polynomials, one can pass via scaling limits to entire functions of exponential type, and the latter have a long and gloried history associated with sampling theory. However, to this author's knowledge, the first general result on mean convergence of entire interpolants at equispaced points is due to Rahman and Vértesi [15, Theorem 1, p. 304]. Define the classic sinc kernel

$$\mathbb{S}(t) = \begin{cases} \frac{\sin\pi t}{\pi t}, & \mathrm{t} \ne 0, \\ 1, & \mathrm{t} = 0. \end{cases}$$

Given a function $f:\mathbb{R}\to\mathbb{R}$, and $\tau > 0$, define the (formal) Lagrange interpolation series

$$L_\tau[f;x] = \sum_{k=-\infty}^{\infty} f\left(\frac{k\pi}{\tau}\right)\mathbb{S}\left(\tau\left(x-\frac{k\pi}{\tau}\right)\right).$$

It is easily seen that this series converges uniformly in compact sets if for some $p > 1$, we have

$$\sum_{k=-\infty}^{\infty}\left|f\left(\frac{k\pi}{\tau}\right)\right|^p < \infty.$$

Theorem 3 (Theorem of Rahman and Vértesi) *Let $f : \mathbb{R} \to \mathbb{R}$ be Riemann integrable over every finite interval and satisfy for some $\beta > \frac{1}{p}$,*

$$|f(x)| \leq C(1+|x|)^{-\beta}, \quad x \in \mathbb{R}.$$

Then

$$\lim_{\tau\to\infty} \int_{-\infty}^{\infty} |f(x) - L_\tau[f;x]|^p \, dx = 0.$$

Butzer, Higgins and Stens later showed that this result is equivalent to the classical sampling theorem, and as such is an example of an approximate sampling theorem [3]. Of course there are sampling theorems at nonequally spaced points (see for example [7, 25]), and in the setting of de Branges spaces, there are more general expansions involving interpolation series. However, as far as this author is aware, there are no analogues of the Rahman-Vértesi theorem in that more general setting. Ganzburg [5] and Littman [9] have explored other aspects of convergence of Lagrange interpolation by entire functions.

One setting where mean convergence has been explored, is interpolation at zeros of Bessel functions, notably by Grozev and Rahman [6, Theorem 1, p. 48]. Let $\alpha > -1$ and

$$J_\alpha(z) = \left(\frac{z}{2}\right)^\alpha \sum_{k=0}^{\infty} (-1)^k \frac{\left(\frac{z}{2}\right)^{2k}}{k!\Gamma(\alpha+k+1)}$$

denote the Bessel function of order α. It is often convenient to instead use its entire cousin,

$$G_\alpha(z) = z^{-\alpha} J_\alpha(z).$$

J_α has positive zeros

$$j_{\alpha,1} < j_{\alpha,2} < j_{\alpha,3} < \cdots,$$

and matching negative zeros

$$j_{\alpha,-k} = -j_{\alpha,k}, \quad k \geq 1,$$

so for $f : \mathbb{R} \to \mathbb{C}$, and $\tau > 0$, one can define the formal interpolation series

$$L_{\alpha,\tau}[f;x] = \sum_{k=-\infty, k\neq 0}^{\infty} f\left(\frac{j_{\alpha,k}}{\tau}\right) \ell_{\alpha,k}(\tau z),$$

where for $k \neq 0$,

$$\ell_{\alpha,k}(z) = \frac{G_\alpha(z)}{G_\alpha'(j_{\alpha,k})(z - j_{\alpha,k})}.$$

Theorem 4 (Theorem of Rahman and Grozev) *Let* $\alpha \geq -\frac{1}{2}$ *and* $p > 1$*, or let* $-1 < \alpha < -\frac{1}{2}$ *and* $1 < p < \frac{2}{|2\alpha+1|}$. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be Riemann integrable over every finite interval and satisfy for some* $\delta > 0$*,*

$$|f(x)| \leq C(1 + |x|)^{-\alpha-\frac{1}{2}-\frac{1}{p}-\delta}, \quad x \in \mathbb{R}.$$

Then

$$\lim_{\tau\to\infty} \int_{-\infty}^{\infty} \left| |x|^{\alpha+\frac{1}{2}} (f(x) - L_\tau[f;x]) \right|^p dx = 0.$$

Note that $p = 2$ is always included. The proof of this theorem involves a lot of tools: detailed properties of entire functions of exponential type and of Bessel functions, and a converse Marcinkiewicz-Zygmund inequality that is itself of great interest.

In this paper, we explore convergence of interpolation at scaled zeros of Airy functions. Recall that the Airy function Ai is given on the real line by [1, 10.4.32, p. 447]

$$Ai(x) = \frac{1}{\pi} \int_0^\infty \cos\left(\frac{1}{3}t^3 + xt\right) dt.$$

The Airy function Ai is an entire function of order $\frac{3}{2}$, with only real negative zeros $\{a_j\}$, where

$$0 > a_1 > a_2 > a_3 > \cdots.$$

Ai satisfies the differential equation

$$Ai''(z) - zAi(z) = 0.$$

The Airy kernel $\mathbb{A}i(\cdot,\cdot)$, much used in random matrix theory, is defined [8] by

$$\mathbb{A}i(a,b) = \begin{cases} \frac{Ai(a)Ai'(b)-Ai'(a)Ai(b)}{a-b}, & \text{a} \neq \text{b}, \\ Ai'(a)^2 - aAi(a)^2, & \text{a=b}. \end{cases}$$

Observe that

$$\ell_j(z) = \frac{\mathbb{A}i(z, a_j)}{\mathbb{A}i(a_j, a_j)} = \frac{Ai(z)}{Ai'(a_j)(z - a_j)},$$

is the Airy analogue of a fundamental of Lagrange interpolation, satisfying

$$\ell_j(a_k) = \delta_{jk}.$$

There is an analogue of sampling series and Lagrange interpolation series involving $\{\ell_j\}$:

Definition 1 Let $\mathcal{G}$ be the class of all functions $g : \mathbb{C} \to \mathbb{C}$ with the following properties:

(a) g is an entire function of order at most $\frac{3}{2}$;

(b) There exists $L > 0$ such that for $\delta \in (0, \pi)$, some $C_\delta > 0$, and all $z \in \mathbb{C}$ with $|\arg z| \leq \pi - \delta$,

$$|g(z)| \leq C_\delta (1 + |z|)^L \left| \exp\left(-\frac{2}{3} z^{\frac{3}{2}}\right) \right|;$$

(c)

$$\sum_{j=1}^{\infty} \frac{|g(a_j)|^2}{|a_j|^{1/2}} < \infty.$$

In [8, Corollary 1.3, p. 429], it was shown that each $g \in \mathcal{G}$ admits the expansion

$$g(z) = \sum_{j=1}^{\infty} g(a_j) \frac{\mathbb{A}i(z, a_j)}{\mathbb{A}i(a_j, a_j)}.$$

Moreover, for $f, g \in \mathcal{G}$, there is the quadrature formula [8, Corollary 1.4, p. 429]

$$\int_{-\infty}^{\infty} f(x) g(x)\, dx = \sum_{j=1}^{\infty} \frac{(fg)(a_j)}{\mathbb{A}i(a_j, a_j)}.$$

In analogy with the entire interpolants of Grozev-Rahman, we define for $f : \mathbb{R} \to \mathbb{R}$, the formal series

$$\mathbb{L}_\tau[f; z] = \sum_{j=1}^{\infty} f\left(\frac{a_j}{\tau}\right) \ell_j(\tau z) = \sum_{j=1}^{\infty} f\left(\frac{a_j}{\tau}\right) \frac{\mathbb{A}i(\tau z, a_j)}{\mathbb{A}i(a_j, a_j)}. \tag{1}$$

Note that it samples f only in $(-\infty, 0)$.

We prove:

Theorem 5 *Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and Riemann integrable in each finite interval, with $f(x) = 0$ in $[0, \infty)$. Assume in addition that for some $\beta > \frac{1}{2}$, and $x \in \mathbb{R}$,*

$$|f(x)| \leq C(1 + |x|)^{-\beta}. \tag{2}$$

Then

$$\lim_{\tau\to\infty}\int_{-\infty}^{\infty}\left(f(x)-\mathbb{L}_\tau[f;x]\right)^2dx=0. \tag{3}$$

Observe that the integration is over the whole real line. We expect that there is an analogue of this theorem at least for all $p > 1$. However, this seems to require a converse Marcinkiewicz-Zygmund inequality estimating L_p norms of appropriate classes of entire functions in terms of their values at Airy zeros. This is not available, so we content ourselves with a weaker result for the related operator

$$\mathbb{L}_\tau^*[f;z]=\sum_{j=1}^{\infty}f\left(\frac{a_{2j-1}}{\tau}\right)\left[\ell_{2j-1}(\tau z)+\ell_{2j}(\tau z)\right].$$

This interpolates f at each $\frac{a_{2j-1}}{\tau}$, but not at $\frac{a_{2j}}{\tau}$.

Theorem 6

(a) For bounded functions $f:\mathbb{R}\to\mathbb{R}$, and $\tau\geq|a_1|$,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{L}_\tau^*[f;x]\right|\leq C\sup_j\left|f\left(\frac{a_{2j-1}}{\tau}\right)\right|, \tag{4}$$

where C is independent of $\tau\geq1$ and f.

(b) Let $\frac{4}{5}<p<\infty$. Let $f:\mathbb{R}\to\mathbb{R}$ be bounded and Riemann integrable in each finite interval, with $f(x)=0$ in $[0,\infty)$. Assume in addition that for some $\beta>\frac{1}{p}$, and $x\in\mathbb{R}$, we have (2). Then

$$\lim_{\tau\to\infty}\int_{-\infty}^{\infty}\left|f(x)-\mathbb{L}_\tau^*[f;x]\right|^p dx=0. \tag{5}$$

Note that $\mathbb{L}_\tau^*\left[f;\frac{z}{\tau}\right]\in\mathcal{G}$, so this also establishes density of that class of functions in a suitable space of functions containing those in Theorem 6. The usual approach to Erdős-Turan theorems is via quadrature formulae and density of polynomials, or entire functions of exponential type, in appropriate spaces. The latter density is not available for $\mathcal{G}$. So in Sect. 2, we establish convergence for characteristic functions of intervals. We prove Theorems 5 and 6 in Sect. 3. Throughout $C, C_1, C_2, \ldots$ denote positive constants independent of n, x, z, t, τ, and possibly other specified quantities. The same symbol does not necessarily denote the same constant in different occurrences, even when used in the same line.

2 Interpolation of Step Functions

We prove:

Theorem 7 *Let $r > 0$, and f denote the characteristic function $\chi_{[-r,0]}$ of the interval $[-r, 0]$. Then for $p > \frac{4}{5}$,*

$$\lim_{\tau\to\infty} \int_{-\infty}^{\infty} |\mathbb{L}_\tau [f; x] - f(x)|^p \, dx = 0. \tag{6}$$

and

$$\lim_{\tau\to\infty} \int_{-\infty}^{\infty} |\mathbb{L}_\tau^* [f; x] - f(x)|^p \, dx = 0. \tag{7}$$

This section is organized as follows: we first recall some asymptotics associated with Airy functions. Then we prove some estimates on integrals involving the fundamental polynomials ℓ_j. Next we prove the case $p = 2$ of Theorem 7. Then we estimate a certain sum and finally prove the general case of Theorem 7.

Firstly, the following asymptotics and estimates are listed on pp. 448–449 of [1]: see (10.4.59–61) there.

$$Ai(x) = \frac{1}{2\pi^{1/2}} x^{-1/4} \exp\left(-\frac{2}{3}x^{\frac{3}{2}}\right)(1 + o(1)), \quad x \to \infty; \tag{8}$$

$$Ai(-x) = \pi^{-1/2} x^{-1/4} \left[\sin\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right) + O\left(x^{-\frac{3}{2}}\right)\right], \quad x \to \infty. \tag{9}$$

Then as Ai is entire, for $x \in [0, \infty)$,

$$|Ai(x)| \leq C(1+x)^{-1/4} \exp\left(-\frac{2}{3}x^{\frac{3}{2}}\right) \text{ and} \tag{10}$$
$$|Ai(-x)| \leq C(1+x)^{-1/4};$$

$$Ai'(-x) = -\pi^{-1/2} x^{1/4} \cos\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right)\left(1 + O\left(x^{-\frac{4}{3}}\right)\right) + O\left(x^{-\frac{2}{3}}\right), \quad x \to \infty. \tag{11}$$

Next, the zeros $\{a_j\}$ of Ai satisfy [1, p. 450, (10.4.94,96)]

$$a_j = -[3\pi(4j-1)/8]^{2/3}\left(1 + O\left(\frac{1}{j^2}\right)\right) \tag{12}$$
$$= -\left(\frac{3\pi j}{2}\right)^{2/3}(1 + o(1)).$$

Consequently,

$$\left|a_{j+1}\right| - \left|a_j\right| = \frac{\pi}{\left|a_j\right|^{1/2}}\left(1 + o\left(1\right)\right). \tag{13}$$

In addition,

$$\begin{aligned} Ai'\left(a_j\right) &= \left(-1\right)^{j-1}\pi^{-1/2}\left(\frac{3\pi}{8}\left(4j-1\right)\right)^{1/6}\left(1 + O\left(j^{-2}\right)\right) \\ &= \left(-1\right)^{j-1}\pi^{-1/2}\left|a_j\right|^{1/4}\left(1 + o\left(1\right)\right). \end{aligned} \tag{14}$$

A calculation shows that

$$\left|Ai'\left(a_j\right)\right| - \left|Ai'\left(a_{j-1}\right)\right| = C_0 j^{-5/6}(1 + O\left(j^{-1}\right)), \quad C_0 = \frac{1}{6}\left(\frac{3}{2\pi^2}\right)^{1/6}. \tag{15}$$

Define the Scorer function [1, p. 448, (10.4.42)]

$$Gi\left(x\right) = \frac{1}{\pi}\int_0^\infty \sin\left(\frac{t^3}{3} + xt\right)dt. \tag{16}$$

We shall use an identity for the Hilbert transform of the Airy function [24, p. 71, eqn. (4.4)]:

$$\frac{1}{\pi}PV\int_{-\infty}^{\infty}\frac{Ai\left(t\right)}{t-x}dt = -Gi\left(x\right). \tag{17}$$

Here PV denotes Cauchy principal value integral. We also use [1, p. 450, eqn. (10.4.87)]

$$Gi\left(-x\right) = \pi^{-1/2}x^{-1/4}\left[\cos\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right) + o\left(1\right)\right], \quad x \to \infty. \tag{18}$$

Finally the Airy kernel $\mathbb{A}i\left(a,b\right)$ satisfies [8, p. 432]

$$\int_{-\infty}^{\infty}\mathbb{A}i\left(a_j,s\right)\mathbb{A}i\left(s,a_k\right)ds = \delta_{jk}\mathbb{A}i\left(a_j,a_j\right) = \delta_{jk}Ai'\left(a_j\right)^2.$$

Thus

$$\int_{-\infty}^{\infty}\ell_j\left(s\right)\ell_k\left(s\right)ds = \delta_{jk}\frac{1}{\mathbb{A}i\left(a_j,a_j\right)} = \delta_{jk}\frac{1}{Ai'\left(a_j\right)^2}. \tag{19}$$

Lemma 1

(a) As $j \to \infty$,

$$\int_{-\infty}^{\infty} \ell_j(t)\,dt = \frac{\pi}{|a_j|^{1/2}}(1 + o(1)). \tag{20}$$

(b) Uniformly in j, r with $r > |a_j|$,

$$\int_{-r}^{0} \ell_j(t)\,dt = \frac{\pi}{|a_j|^{1/2}}\left(1 + o(1) + O\left(\frac{1}{|a_j|^{1/4}\, r^{3/4}\,(r - |a_j|)}\right)\right). \tag{21}$$

Proof

(a) Now (17) yields

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \ell_j(t)\,dt = -\frac{Gi(a_j)}{Ai'(a_j)} \tag{22}$$

Here using (12),

$$\cos\left(\frac{2}{3}|a_j|^{\frac{3}{2}} + \frac{\pi}{4}\right) = (-1)^j + O\left(\frac{1}{j}\right),$$

so from (18),

$$Gi(a_j) = \pi^{-1/2}|a_j|^{-1/4}(-1)^j(1 + o(1)).$$

Substituting this and (14) into (22) gives

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \ell_j(t)\,dt = \frac{1}{|a_j|^{1/2}}(1 + o(1)).$$

(b) Using the bound (10),

$$\begin{aligned}\int_0^{\infty} |\ell_j(t)|\,dt &\leq \frac{C}{|Ai'(a_j)|}\int_0^{\infty} \frac{t^{-1/4}\exp\left(-\frac{2}{3}t^{\frac{3}{2}}\right)}{t - a_j}dt \\ &\leq \frac{C}{|a_j Ai'(a_j)|} \leq C|a_j|^{-5/4},\end{aligned} \tag{23}$$

recall (14). Next,

$$\left|\int_{-\infty}^{-r} \ell_j(t)\, dt\right| = \frac{1}{\left|Ai'\left(a_j\right)\right|} \left|\int_r^{\infty} \frac{Ai(-x)}{x+a_j} dx\right| \tag{24}$$

$$= \frac{1}{\left|Ai'\left(a_j\right)\right|} \left| \pi^{-1/2} \int_r^{\infty} \frac{x^{-1/4} \sin\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right)}{x+a_j} dx \right.$$

$$\left. + O\left(\int_r^{\infty} \frac{x^{-7/4}}{\left|x+a_j\right|} dx\right) \right|,$$

by (9). Here,

$$I = \int_r^{\infty} \frac{x^{-1/4} \sin\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right)}{x+a_j} dx$$

$$= \int_r^{\infty} \frac{-x^{-\frac{3}{4}} \frac{d}{dx}\left[\cos\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right)\right]}{x+a_j} dx$$

$$= \frac{\cos\left(\frac{2}{3}r^{3/2} + \frac{\pi}{4}\right)}{r^{3/4}\left(r - \left|a_j\right|\right)} + \int_r^{\infty} \cos\left(\frac{2}{3}x^{\frac{3}{2}} + \frac{\pi}{4}\right) \frac{d}{dx}\left[\frac{1}{x^{3/4}(x-\left|a_j\right|)}\right] dx$$

$$= O\left(\frac{1}{r^{3/4}\left(r-\left|a_j\right|\right)}\right) + O\left(\int_r^{\infty} \left|\frac{d}{dx}\left[\frac{1}{x^{3/4}(x-\left|a_j\right|)}\right]\right| dx\right)$$

$$= O\left(\frac{1}{r^{3/4}\left(r-\left|a_j\right|\right)}\right), \tag{25}$$

as $\frac{1}{x^{3/4}(x-\left|a_j\right|)}$ is decreasing in $[r, \infty)$. Next,

$$\int_r^{\infty} \frac{x^{-7/4}}{\left|x+a_j\right|} dx \le \frac{1}{r-\left|a_j\right|} \int_r^{\infty} x^{-7/4} dx \le \frac{C}{r^{3/4}\left(r-\left|a_j\right|\right)}.$$

Thus, using also (25) in (24),

$$\left|\int_{-\infty}^{-r} \ell_j(t)\, dt\right| \le \frac{C}{\left|a_j\right|^{1/4} r^{3/4}\left(r-\left|a_j\right|\right)}.$$

Together with (20) and (23), this gives the result (21).

□

Lemma 2 *Let* $L \geq 1$, *and*

$$S_L(x) = \sum_{j=1}^{L} \ell_j(x) = \sum_{j=1}^{L} \frac{\mathbb{A}i\left(x, a_j\right)}{\mathbb{A}i\left(a_j, a_j\right)}. \tag{26}$$

Then

$$\lim_{L\to\infty} \frac{1}{|a_{L+1}|} \int_{-\infty}^{\infty} \left(S_L(x) - \chi_{[a_{L+1},0]}(x)\right)^2 dx = 0. \tag{27}$$

Proof Using (19), and then (14),

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left(S_L(x) - \chi_{[a_{L+1},0]}(x)\right)^2 dx \\
&= \sum_{j=1}^{L} \frac{1}{Ai'\left(a_j\right)^2} - 2\sum_{j=1}^{L} \int_{a_{L+1}}^{0} \ell_j(x)\, dx + |a_{L+1}| \\
&= \sum_{j=1}^{L} \frac{\pi}{\left|a_j\right|^{1/2}} (1 + o(1)) \\
&\quad - 2\sum_{j=1}^{L} \left\{ \frac{\pi}{\left|a_j\right|^{1/2}} \left(1 + o(1) + O\left(\frac{1}{|a_{L+1}|^{3/4} \left|a_j\right|^{1/4} \left(|a_{L+1}| - \left|a_j\right|\right)}\right)\right)\right\} + |a_{L+1}| \\
&= |a_{L+1}| - \sum_{j=1}^{L} \frac{\pi}{\left|a_j\right|^{1/2}} (1 + o(1)) \\
&\quad + O\left(|a_{L+1}|^{-3/4} \sum_{j=1}^{L} \frac{1}{\left|a_j\right|^{1/4} \left(|a_{L+1}| - \left|a_j\right|\right)}\right),
\end{aligned} \tag{28}$$

by (14) and (21). Here using (13),

$$\sum_{j=1}^{L} \frac{\pi}{\left|a_j\right|^{1/2}} = \sum_{j=1}^{L} \left(\left|a_{j+1}\right| - \left|a_j\right|\right)(1 + o(1)) = |a_{L+1}|(1 + o(1)). \tag{29}$$

Also, from (12),

$$\begin{aligned}
&|a_{L+1}|^{-3/4} \sum_{j=1}^{L} \frac{1}{\left|a_j\right|^{1/4} \left(|a_{L+1}| - \left|a_j\right|\right)} \\
&\qquad \leq C\, |a_{L+1}|^{-7/4} \sum_{j=1}^{L} \frac{1}{j^{1/6} \left(1 - \frac{j}{L+1}\right)}
\end{aligned}$$

$$\leq C\left|a_{L+1}\right|^{-7/4}\int_0^L \frac{1}{x^{1/6}\left(1-\frac{x}{L+1}\right)}dx$$
$$\leq C\left|a_{L+1}\right|^{-7/4}L^{5/6}\log L$$
$$\leq C\left|a_{L+1}\right|^{-1/2}\log L. \tag{30}$$

Substituting this and (29) into (28), gives

$$\int_{-\infty}^{\infty}\left(S_L\left(x\right)-\chi_{\left[a_{L+1},0\right]}\left(x\right)\right)^2 dx$$
$$= o\left(\left|a_{L+1}\right|\right)+O\left(C\left|a_{L+1}\right|^{-1/2}\log L\right) = o\left(\left|a_{L+1}\right|\right).$$

□

Proof of Theorem 7 for $p = 2$ Given $\tau \geq \left|a_1\right|/r$, choose $L = L\left(\tau\right)$ by the inequality

$$\left|a_L\right| \leq \tau r < \left|a_{L+1}\right|. \tag{31}$$

Then

$$\mathbb{L}_\tau\left[f;x\right] = \sum_{a_j/\tau\in[-r,0]}\ell_j\left(\tau x\right) = \sum_{j=1}^{L}\ell_j\left(\tau x\right). \tag{32}$$

By Lemma 2,

$$\int_{-\infty}^{\infty}\left(\mathbb{L}_\tau\left[f;x\right]-\chi_{\left[a_{L+1},0\right]}\left(\tau x\right)\right)^2 dx$$
$$= \frac{1}{\tau}\int_{-\infty}^{\infty}\left(\sum_{j=1}^{L}\ell_j\left(t\right)-\chi_{\left[a_{L+1},0\right]}\left(t\right)\right)^2 dt$$
$$= \frac{1}{\tau}o\left(\left|a_{L+1}\right|\right) = o\left(1\right),$$

as $\tau \to \infty$. Also, as $\tau \to \infty$,

$$\int_{-\infty}^{\infty}\left(\chi_{[r,0]}\left(x\right)-\chi_{\left[a_{L+1},0\right]}\left(\tau x\right)\right)^2 dx$$
$$= \int_{-\infty}^{\infty}\chi_{\left[\frac{a_{L+1}}{\tau},r\right]}\left(x\right)^2 dx$$
$$= \frac{\left|a_{L+1}\right|}{\tau}-r \leq \frac{\left|a_{L+1}\right|-\left|a_L\right|}{\tau} = O\left(L^{-1/3}\tau^{-1}\right) = o\left(1\right).$$

Then (6) follows. Since $\mathbb{L}_\tau [f; x] = \mathbb{L}_\tau^* [f; x]$ if L above is even, (7) also follows. The case of odd L is easily handled by estimating separately the single extra term. □

Next we bound a generalization of $S_L(x)$:

Lemma 3 *Let $A > 0$, $0 \leq \beta < \frac{5}{4}$ and $\left\{c_j\right\}_{j=1}^\infty$ be real numbers such that for $j \geq 1$,*

$$c_{2j} = c_{2j-1}, \tag{33}$$

and

$$\left|c_{2j-1}\right| \leq A\left(1 + \left|a_{2j}\right|\right)^{-\beta}. \tag{34}$$

Let

$$\hat{S}(x) = \sum_{j=1}^\infty c_j \ell_j(x).$$

Then the series converges and for all real x,

$$\left|\hat{S}(x)\right| \leq CA(1 + |x|)^{-\beta}. \tag{35}$$

Here C is independent of x, A, and $\left\{c_j\right\}_{j=1}^\infty$.

Proof We may assume that $A = 1$. We assume first that $x \in (-\infty, 0)$, as this is the most difficult case. Set $a_0 = 0$. Choose an even integer $j_0 \geq 2$ such that $x \in [a_{j_0}, a_{j_0-2})$. Let us first deal with central terms: assume that $j \geq 1$ and $|j - j_0| \leq 3$. Then

$$\left|\ell_j(x)\right| = \frac{1}{\left|Ai'\left(a_j\right)\right|}\left|\frac{Ai(x) - A_i\left(a_j\right)}{x - a_j}\right| = \left|\frac{Ai'(t)}{Ai'\left(a_j\right)}\right|,$$

for some t between x and a_j, so $(|t| + 1) / \left|a_j\right| \sim 1$. Using (11), (14), and the continuity of Ai', we see that

$$\left|\ell_j(x)\right| \leq C\frac{(1 + |t|)^{1/4}}{\left|a_j\right|^{1/4}} \leq C.$$

Thus as $|x| + 1 \sim \left|a_{j_0}\right|$, and (34) holds,

$$\sum_{j:|j-j_0|\leq 3} \left|c_j \ell_j(x)\right| \leq C(1 + |x|)^{-\beta}. \tag{36}$$

Again, we emphasize that C is independent of L and x and $\{c_j\}$. We turn to the estimation of

$$S^*(x) = \left(\sum_{j=1}^{j_0-4} + \sum_{j=j_0+3}^{\infty}\right) c_j \ell_j(x).$$

Recall from (14) that $Ai'(a_j)$ has sign $(-1)^{j-1}$. Then

$$\begin{aligned}
&S^*(x)\\
&= Ai(x)\left(\sum_{k=1}^{(j_0-4)/2} + \sum_{k=j_0/2+2}^{\infty}\right) c_{2k-1}\left[\frac{1}{|Ai'(a_{2k-1})|(x-a_{2k-1})} - \frac{1}{|Ai'(a_{2k})|(x-a_{2k})}\right]\\
&= Ai(x)(\Sigma_1 + \Sigma_2),
\end{aligned} \tag{37}$$

where

$$\begin{aligned}
\Sigma_1 &= \left(\sum_{k=1}^{(j_0-4)/2} + \sum_{k=j_0/2+2}^{\infty}\right) \frac{c_{2k-1}}{|Ai'(a_{2k-1})|}\left(\frac{1}{x-a_{2k-1}} - \frac{1}{x-a_{2k}}\right)\\
&= \left(\sum_{k=1}^{(j_0-4)/2} + \sum_{k=j_0/2+2}^{\infty}\right) \frac{c_{2k-1}}{|Ai'(a_{2k-1})|}\frac{a_{2k-1}-a_{2k}}{(x-a_{2k-1})(x-a_{2k})}
\end{aligned} \tag{38}$$

and

$$\Sigma_2 = \left(\sum_{k=1}^{(j_0-4)/2} + \sum_{k=j_0/2+2}^{\infty}\right) c_{2k-1}\left(\frac{1}{|Ai'(a_{2k-1})|} - \frac{1}{|Ai'(a_{2k})|}\right)\frac{1}{x-a_{2k}}.$$

Then if $\mathscr{I}$ denotes the set of integers k with either $1 \leq k \leq (j_0-4)/2$ or $k \geq j_0/2+2$, we see that $\frac{|x-a_{2k}|}{|x-a_{2k-1}|}$, and $\frac{|a_{2k-1}-a_{2k}|}{|a_{2k-2}-a_{2k}|}$ are bounded above and below by positive constants independent of k, x, so using (14) and (34), so

$$\begin{aligned}
\Sigma_1 &\leq C\sum_{k\in\mathscr{I}} \frac{1}{|a_{2k}|^{1/4+\beta}}\frac{|a_{2k-2}-a_{2k}|}{(x-a_{2k})^2}\\
&\leq C\int_{[|a_1|,\infty)\setminus[|a_{j_0-3}|,|a_{j_0+1}|]} \frac{1}{t^{1/4+\beta}}\frac{1}{(|x|-t)^2}dt\\
&= \frac{C}{|a_{j_0}|^{5/4+\beta}}\int_{[\frac{|a_1|}{|a_{j_0}|},\infty)\setminus\left[\frac{|a_{j_0-3}|}{|a_{j_0}|},\frac{|a_{j_0+1}|}{|a_{j_0}|}\right]} \frac{1}{s^{1/4+\beta}}\frac{1}{\left(\frac{|x|}{|a_{j_0}|}-s\right)^2}ds
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{C}{\left|a_{j_0}\right|^{5/4+\beta}} \left(\left\{ \begin{array}{cc} \left|a_{j_0}\right|^{\beta-3/4}, & \beta > 3/4 \\ \log\left|a_{j_0},\right| & \beta = 3/4 \\ 1, & \beta < 3/4 \end{array} \right\} + \frac{\left|a_{j_0}\right|}{\left||x| - \left|a_{j_0-3}\right|\right|} + \frac{\left|a_{j_0}\right|}{\left||x| - \left|a_{j_0+1}\right|\right|} \right) \\
&\leq C \left(\frac{1}{\left|a_{j_0}\right|^{-1/4+\beta}} + \frac{\left|a_{j_0}\right|^{1/2}}{\left|a_{j_0}\right|^{1/4+\beta}} \right) \\
&\leq C \left(1 + |x|\right)^{1/4-\beta}, \qquad (39)
\end{aligned}$$

recall that $\beta < \frac{5}{4}$, and that $\left||x| - \left|a_{j_0-3}\right|\right| \geq \left|\left|a_{j_0}\right| - \left|a_{j_0-2}\right|\right| \geq C \left|a_{j_0}\right|^{-1/2}$, by (13). Next, using (13)–(15) and (34),

$$\begin{aligned}
\left|\Sigma_2\right| &\leq C \sum_{k \in \mathscr{I}} \frac{k^{-5/6}}{\left|a_{2k}\right|^{\beta} \left|Ai'\left(a_{2k-1}\right)\right| \left|Ai'\left(a_{2k}\right)\right|} \frac{1}{\left|x - a_{2k}\right|} \\
&\leq C \sum_{k \in \mathscr{I}} \frac{\left(\left|a_{2k}\right| - \left|a_{2k-2}\right|\right)}{\left|a_{2k}\right|^{5/4+\beta}} \frac{1}{\left|x - a_{2k}\right|} \\
&\leq C \int_{\left[\left|a_1\right|, \infty\right) \backslash \left[\left|a_{j_0-3}\right|, \left|a_{j_0+1}\right|\right]} \frac{1}{t^{5/4+\beta} \left||x| - t\right|} dt \\
&\leq \frac{C}{\left|a_{j_0}\right|^{5/4+\beta}} \int_{\left[\frac{\left|a_1\right|}{\left|a_{j_0}\right|}, \infty\right) \backslash \left[\frac{\left|a_{j_0-3}\right|}{\left|a_{j_0}\right|}, \frac{\left|a_{j_0+1}\right|}{\left|a_{j_0}\right|}\right]} \frac{1}{s^{5/4+\beta}} \frac{1}{\left|\frac{|x|}{\left|a_{j_0}\right|} - s\right|} ds \\
&\leq \frac{C}{\left|a_{j_0}\right|^{5/4+\beta}} \left(\left|a_{j_0}\right|^{1/4+\beta} + \left|\log\left|\frac{|x|}{\left|a_{j_0}\right|} - \frac{\left|a_{j_0-3}\right|}{\left|a_{j_0}\right|}\right|\right| + \left|\log\left|\frac{|x|}{\left|a_{j_0}\right|} - \frac{\left|a_{j_0+1}\right|}{\left|a_{j_0}\right|}\right|\right| \right) \\
&\leq C \left(\left|a_{j_0}\right|^{-1} + \left|a_{j_0}\right|^{-5/4-\beta} \log j_0 \right) \leq C \left(1 + |x|\right)^{1/4-\beta},
\end{aligned}$$

recall $\beta < \frac{5}{4}$. Substituting this and (39) into (37) gives

$$\left|S^*(x)\right| \leq C \left|Ai(x)\right| \left(1 + |x|\right)^{1/4-\beta} \leq C \left(1 + |x|\right)^{-\beta},$$

in view of (10). This and (36) gives (35). Finally, the case where $x \geq 0$ is easier. □

We deduce:

Proof of Theorem 7 for the general case Recall that $f = \chi_{[-r,0]}$. Assume first $p > 2$. The previous lemma (with all $c_j = 1$ and $\beta = 0$) shows that

$$\left|\mathbb{L}_\tau\left[f; x\right] - f(x)\right| \leq \sup_{x \in \mathbb{R}} \left|\mathbb{L}_\tau\left[f; x\right]\right| + 1 \leq C,$$

where C is independent of τ. We can then apply the case $p = 2$ of Theorem 7:

$$\limsup_{\tau\to\infty} \int_{-\infty}^{\infty} |\mathbb{L}_\tau[f;x] - f(x)|^p \, dx$$
$$\leq C^{p-2} \limsup_{\tau\to\infty} \int_{-\infty}^{\infty} |\mathbb{L}_\tau[f;x] - f(x)|^2 \, dx = 0.$$

Next, if $\frac{4}{5} < p < 2$, and $s \geq 2r$, Hölder's inequality gives

$$\limsup_{\tau\to\infty} \int_{-s}^{s} |\mathbb{L}_\tau[f;x] - f(x)|^p \, dx$$
$$\leq \limsup_{\tau\to\infty} \left(\int_{-s}^{s} |\mathbb{L}_\tau[f;x] - f(x)|^2 \, dx \right)^{\frac{p}{2}} (2s)^{1-\frac{p}{2}} = 0. \tag{40}$$

Next, for $|x| \geq s > 2r$, and all τ, we have

$$|\mathbb{L}_\tau[f;x] - f(x)| \leq \sum_{a_j \in [-\tau r, 0]} \frac{|Ai(\tau x)|}{|Ai'(a_j)| \, |\tau x - a_j|}$$
$$\leq C \frac{|Ai(\tau x)|}{\tau |x|} \sum_{a_j \in [-\tau r, 0]} \frac{1}{|a_j|^{1/4}}$$
$$\leq C(\tau |x|)^{-5/4} \sum_{j \leq C(\tau r)^{3/2}} \frac{1}{j^{1/6}}$$
$$\leq C(\tau |x|)^{-5/4} (\tau r)^{5/4} = C r^{5/4} |x|^{-5/4}.$$

Thus

$$\limsup_{\tau\to\infty} \int_{|x|\geq s} |\mathbb{L}_\tau[f;x] - f(x)|^p \, dx \leq C \int_{|x|\geq s} |x|^{-5p/4} \, dx \leq C s^{1-5p/4} \to 0$$

as $s \to \infty$, recall $p > 4/5$. Together with (40), this gives (6). Of course (7) also follows as $\mathbb{L}_\tau^*[f;x]$ differs from $\mathbb{L}_\tau[f;x]$ in at most one term, which can easily be estimated. □

3 Proof of Theorems 5 and 6

Proof of Theorem 5 Suppose first that f is bounded and Riemann integrable, and supported in $(-r, 0]$, some $r > 0$. Let $\varepsilon > 0$. Then we can find a (piecewise constant) step function g also compactly supported in $(-r, 0]$ such that both

$$g \geq f \text{ in } \mathbb{R} \text{ and } \int_{-\infty}^{\infty} (f - g)^2 < \varepsilon^2.$$

This follows directly from the theory of Riemann sums and the boundedness of f. Theorem 7 implies that for any such step function g,

$$\lim_{\tau\to\infty}\left(\int_{-\infty}^{\infty}\left(g\left(x\right)-\mathbb{L}_{\tau}\left[g;x\right]\right)^{2}dx\right)^{1/2}=0.$$

Then using also the orthonormality relation (19),

$$\begin{aligned}
&\limsup_{\tau\to\infty}\left(\int_{-\infty}^{\infty}\left(f\left(x\right)-\mathbb{L}_{\tau}\left[f;x\right]\right)^{2}dx\right)^{1/2}\\
&\quad\leq\left(\int_{-\infty}^{\infty}\left(f\left(x\right)-g\left(x\right)\right)^{2}dx\right)^{1/2}+\limsup_{\tau\to\infty}\left(\int_{-\infty}^{\infty}\left(g\left(x\right)-\mathbb{L}_{\tau}\left[g;x\right]\right)^{2}dx\right)^{1/2}\\
&\qquad+\limsup_{\tau\to\infty}\left(\int_{-\infty}^{\infty}\mathbb{L}_{\tau}\left[g-f;x\right]^{2}dx\right)^{1/2}\\
&\quad\leq\varepsilon+0+\limsup_{\tau\to\infty}\left(\frac{1}{\tau}\int_{-\infty}^{\infty}\left(\sum_{j=1}^{\infty}\left(f-g\right)\left(\frac{a_j}{\tau}\right)\ell_j\left(x\right)\right)^{2}dx\right)^{1/2}\\
&\quad=\varepsilon+\limsup_{\tau\to\infty}\left(\frac{1}{\tau}\sum_{a_j\in(-\tau r,0)}\frac{\left(f-g\right)^{2}\left(\frac{a_j}{\tau}\right)}{Ai'\left(a_j\right)^{2}}\right)^{1/2}\\
&\quad=\varepsilon+C\limsup_{\tau\to\infty}\left(\sum_{\frac{a_j}{\tau}\in(-r,0)}\left(\frac{\left|a_j\right|}{\tau}-\frac{\left|a_{j-1}\right|}{\tau}\right)\left(f-g\right)^{2}\left(\frac{a_j}{\tau}\right)\right)^{1/2}\\
&\quad=\varepsilon+C\left(\int_{-r}^{0}\left|f-g\right|^{2}\left(x\right)dx\right)^{1/2}\leq C\varepsilon.
\end{aligned}$$

Here C is independent of ε, g and f, and we have used (13), (14), that $\max_j\left(\frac{|a_j|}{\tau}-\frac{|a_{j-1}|}{\tau}\right)\to 0$ as $\tau\to\infty$, and the theory of Riemann sums. So we have the result for such compactly supported f. Now assume that f is supported in $(-\infty,-r)$ and for some $\beta>\frac{1}{2}$, (2) holds. Then using (19) again,

$$\begin{aligned}
&\left(\int_{-\infty}^{\infty}\left(f\left(x\right)-\mathbb{L}_{\tau}\left[f;x\right]\right)^{2}dx\right)^{1/2}\\
&\quad\leq\left(\int_{r}^{\infty}f^{2}\left(x\right)dx\right)^{1/2}+\left(\frac{1}{\tau}\sum_{a_j\in(-\infty,-\tau r)}\frac{f^{2}\left(\frac{a_j}{\tau}\right)}{Ai'\left(a_j\right)^{2}}\right)^{1/2}
\end{aligned}$$

$$\leq C\left(\int_r^\infty (1+|x|)^{-2\beta}\,dx\right)^{1/2} + C\left(\frac{1}{\tau}\sum_{a_j\in(-\infty,-\tau r)} \frac{\left(\frac{|a_j|}{\tau}\right)^{-2\beta}}{|a_j|^{1/2}}\right)$$

$$\leq Cr^{\frac{1}{2}-\beta} + C\left(\frac{1}{\tau}\sum_{a_j\in(-\infty,-\tau r)} \left(\frac{|a_j|}{\tau}\right)^{-2\beta} \left(|a_j| - |a_{j-1}|\right)\right)^{1/2}$$

$$\leq Cr^{\frac{1}{2}-\beta} + C\left(\tau^{-1+2\beta}\int_{\tau r}^\infty t^{-2\beta}dt\right)^{1/2} \leq Cr^{1/2-\beta},$$

where C is independent of r and τ. So

$$\limsup_{\tau\to\infty}\left(\int_{-\infty}^\infty (f(x) - \mathbb{L}_\tau[f;x])^2\,dx\right)^{1/2} \leq Cr^{1/2-\beta}.$$

This can be made arbitrarily small for large enough r. Together with the case above, this easily implies the result. □

Proof of Theorem 6

(a) From Lemma 3 with $\beta = 0$,

$$\sup_{x\in\mathbb{R}}\left|\sum_{j=1}^{\infty} f\left(\frac{a_{2j-1}}{\tau}\right)\left(\ell_{2j-1}(\tau x)) + \ell_{2j}(\tau x)\right)\right| \leq C\sup_j\left|f\left(\frac{a_{2j-1}}{\tau}\right)\right|.$$

Here C is independent of f, τ.

(b) For $p = 2$, the exact same proof as of Theorem 5 gives

$$\lim_{\tau\to\infty}\int_{-\infty}^\infty \left(\mathbb{L}_\tau^*[f;x] - f(x)\right)^2 dx = 0.$$

If $p > 2$, we can use the boundedness of the operators, to obtain

$$\limsup_{\tau\to\infty}\int_{-\infty}^\infty |\mathbb{L}_\tau^*[f;x] - f(x)|^p\,dx$$

$$\leq \limsup_{\tau\to\infty}\left(\sup_{x\in\mathbb{R}}|\mathbb{L}_\tau^*[f;x] - f(x)|\right)^{p-2}\int_{-\infty}^\infty |\mathbb{L}_\tau^*[f;x] - f(x)|^2\,dx$$

$$\leq C\|f\|_{L_\infty(\mathbb{R})}^{p-2}\limsup_{\tau\to\infty}\int_{-\infty}^\infty |\mathbb{L}_\tau^*[f;x] - f(x)|^2\,dx = 0.$$

Now let $\frac{4}{5} < p < 2$. Let $s > 0$. Hölder's inequality gives

$$\begin{aligned}&\limsup_{\tau\to\infty}\int_{-s}^{s}\left|\mathbb{L}_\tau^*[f;x]-f(x)\right|^p dx\\&\quad\le\limsup_{\tau\to\infty}\left(\int_{-s}^{s}\left|\mathbb{L}_\tau^*[f;x]-f(x)\right|^2 dx\right)^{\frac{p}{2}}(2s)^{1-\frac{p}{2}}=0,\end{aligned}\tag{41}$$

by the case $p = 2$. Next, our bound (2) on f and Lemma 3 show that for all x,

$$\left|\mathbb{L}_\tau^*[f;x]-f(x)\right|\le C(1+|x|)^{-\beta}.$$

Then

$$\begin{aligned}&\int_{|x|\ge s}\left|\mathbb{L}_\tau^*[f;x]-f(x)\right|^p dx\\&\quad\le C\int_{s}^{\infty}(1+|x|)^{-p\beta}dx\le Cs^{1-p\beta}\to 0\text{ as }s\to\infty,\end{aligned}$$

as $\beta > \frac{1}{p}$.

□

Acknowledgements Research supported by NSF grant DMS1362208.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1965)
2. Akhlaghi, M.R.: On L_p-convergence of nonpolynomial interpolation. J. Approx. Theory **55**, 194–204 (1988)
3. Butzer, P.L., Higgins, J.R., Stens, R.L.: Classical and approximate sampling theorems: studies in the $L^p(\mathbb{R})$ and the uniform norm. J. Approx. Theory **137**, 250–263 (2005)
4. Erdös, P., Turán, P.: On interpolation, I. Quadrature and mean convergence in the Lagrange interpolation. Ann. Math. **38**, 142–155 (1937)
5. Ganzburg, M.: Polynomial interpolation and asymptotic representations for zeta functions. Diss. Math. (Rozprawy Mat.) **496**, 117 pp. (2013)
6. Grozev, G.R., Rahman, Q.I.: Lagrange interpolation in the zeros of Bessel functions by entire functions of exponential type and mean convergence. Methods Appl. Anal. **3**, 46–79 (1996)
7. Higgins, J.R.: Sampling Theory in Fourier and Signal Analysis: Foundations. Oxford University Press, Oxford (1996)
8. Levin, E., Lubinsky, D.S.: On the Airy reproducing kernel, sampling series, and quadrature formula. Integr. Equ. Oper. Theory **63**, 427–438 (2009)
9. Littman, F.: Interpolation and approximation by entire functions. In: Approximation Theory XI, pp. 243–255. Nashboro Press, Brentwood (2008)
10. Lubinsky, D.S.: A Taste of Erdős on Interpolation. In: Paul Erdős and His Mathematics, L. Bolyai Society Mathematical Studies II, pp. 423–454. Bolyai Society, Budapest (2000)

11. Mastroianni, G., Milovanovic, G.V.: Interpolation Processes Basic Theory and Applications. Springer Monographs in Mathematics, vol. XIV. Springer, Berlin (2009)
12. Nevai, P.: Lagrange interpolation at zeros of orthogonal polynomials. In: Lorentz, G.G., et al. (eds.) Approximation Theory II, pp. 163–201. Academic, New York (1976)
13. Nevai, P.: Geza Freud, orthogonal polynomials and Christoffel functions: a case study. J. Approx. Theory **48**, 3–167 (1986)
14. Rabinowitz, P., Sloan, I.H.: Product integration in the presence of a singularity. SIAM J. Numer. Anal. **21**, 149–166 (1984)
15. Rahman, Q.I., Vértesi, P.: On the L^p convergence of Lagrange interpolating entire functions of exponential type. J. Approx. Theory **69**, 302–317 (1992)
16. Sloan, I.H.: On choosing the points in product integration. J. Math. Phys. **21**, 1032–1039 (1979)
17. Sloan, I.H.: Nonpolynomial interpolation. J. Approx. Theory **39**, 97–117 (1983)
18. Sloan, I.H., Smith, W.E.: Product integration with the Clenshaw-Curtis points: implementation and error estimates, Numer. Math. **34**, 387–401 (1980)
19. Sloan, I.H., Smith, W.E.: Properties of interpolatory product integration rules. SIAM J. Numer. Anal. **19**, 427–442 (1982)
20. Smith, W.E., Sloan, I.H.: Product-integration rules based on the zeros of Jacobi polynomials. SIAM J. Numer. Anal. **17**, 1–13 (1980)
21. Smith, W.E., Sloan, I.H., Opie, A.: Product integration over infinite intervals I. rules based on the zeros of Hermite polynomials. Math. Comput. **40**, 519–535 (1983)
22. Szabados, J., Vértesi, P.: Interpolation of Functions. World Scientific, Singapore (1990)
23. Szabados, J., Vértesi, P.: A survey on mean convergence of interpolatory processes. J. Comput. Appl. Math. **43**, 3–18 (1992)
24. Vallée, O., Soares, M.: Airy Functions and Applications to Physics. World Scientific, Singapore (2004)
25. Zayed, A.I., El-Sayed, M.A., Annaby, M.H.: On Lagrange interpolations and Kramer's sampling theorem associated with self-adjoint boundary value problems. J. Math. Anal. Appl. **158**, 269–284 (1991)

Exponential Sum Approximations for $t^{-\beta}$

William McLean

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Given $\beta > 0$ and $\delta > 0$, the function $t^{-\beta}$ may be approximated for t in a compact interval $[\delta, T]$ by a sum of terms of the form we^{-at}, with parameters $w > 0$ and $a > 0$. One such an approximation, studied by Beylkin and Monzón (Appl. Comput. Harmon. Anal. 28:131–149, 2010), is obtained by applying the trapezoidal rule to an integral representation of $t^{-\beta}$, after which Prony's method is applied to reduce the number of terms in the sum with essentially no loss of accuracy. We review this method, and then describe a similar approach based on an alternative integral representation. The main difference is that the new approach achieves much better results *before* the application of Prony's method; after applying Prony's method the performance of both is much the same.

1 Introduction

Consider a Volterra operator with a convolution kernel,

$$\mathscr{K}u(t) = (k * u)(t) = \int_0^t k(t-s)u(s)\,ds \quad \text{for } t > 0, \tag{1}$$

and suppose that we seek a numerical approximation to $\mathscr{K}u$ at the points of a grid $0 = t_0 < t_1 < t_2 < \cdots < t_{N_t} = T$. For example, if we know $U^n \approx u(t_n)$

W. McLean (✉)
School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW, Australia
e-mail: w.mclean@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_40

and define (for simplicity) a piecewise-constant interpolant $\tilde{U}(t) = U^n$ for $t \in I_n = (t_{n-1}, t_n)$, then

$$\mathscr{K}u(t_n) \approx \mathscr{K}\tilde{U}(t_n) = \sum_{j=1}^{n} \omega_{nj} U^j \quad \text{where} \quad \omega_{nj} = \int_{I_j} k(t_n - s)\, ds.$$

The number of operations required to compute this sum in the obvious way for $1 \le n \le N_t$ is proportional to $\sum_{n=1}^{N_t} n \approx N_t^2/2$, and this quadratic growth can be prohibitive in applications where each U^j is a large vector and not just a scalar. Moreover, it might not be possible to store U^j in active memory for all time levels j.

These problems can be avoided using a simple, fast algorithm if the kernel k admits an exponential sum approximation

$$k(t) \approx \sum_{l=1}^{L} w_l e^{b_l t} \quad \text{for } \delta \le t \le T, \tag{2}$$

provided sufficient accuracy is achieved using only a moderate number of terms L, for a choice of $\delta > 0$ that is smaller than the time step $\Delta t_n = t_n - t_{n-1}$ for all n. Indeed, if $\Delta t_n \ge \delta$ then $\delta \le t_n - s \le T$ for $0 \le s \le t_{n-1}$ so

$$\sum_{j=1}^{n-1} \omega_{nj} U^j = \int_0^{t_{n-1}} k(t_n - s)\tilde{U}(s)\, ds \approx \int_0^{t_{n-1}} \sum_{l=1}^{L} w_l e^{b_l(t_n - s)} \tilde{U}(s)\, ds = \sum_{l=1}^{L} \Theta_l^n,$$

where

$$\Theta_l^n = w_l \int_0^{t_{n-1}} e^{b_l(t_n-s)} \tilde{U}(s)\, ds = \sum_{j=1}^{n-1} \kappa_{lnj} U^j \quad \text{and} \quad \kappa_{lnj} = w_l \int_{I_j} e^{b_l(t_n-s)}\, ds.$$

Thus,

$$\mathscr{K}\tilde{U}(t_n) \approx \omega_{nn} U^n + \sum_{l=1}^{L} \Theta_l^n, \tag{3}$$

and by using the recursive formula

$$\Theta_l^n = \kappa_{ln,n-1} U^{n-1} + e^{b_l \Delta t_n} \Theta_l^{n-1} \quad \text{for } n \ge 2, \quad \text{with } \Theta_l^1 = 0,$$

we can evaluate $\mathscr{K}\tilde{U}(t_n)$ for $1 \le n \le N$ to an acceptable accuracy with a number of operations proportional to LN_t—a substantial saving if $L \ll N_t$. In addition, we may overwrite Θ_l^{n-1} with Θ_l^n, and overwrite U^{n-1} with U^n, so that the active storage requirement is proportional to L instead of N_t.

In the present work, we study two exponential sum approximations to the kernel $k(t) = t^{-\beta}$ with $\beta > 0$. Our starting point is the integral representation

$$\frac{1}{t^\beta} = \frac{1}{\Gamma(\beta)} \int_0^\infty e^{-pt} p^\beta \, \frac{dp}{p} \quad \text{for } t > 0 \text{ and } \beta > 0, \tag{4}$$

which follows easily from the integral definition of the Gamma function via the substitution $p = y/t$ (if y is the original integration variable). Section 2 discusses the results of Beylkin and Monzón [3], who used the substitution $p = e^x$ in (4) to obtain

$$\frac{1}{t^\beta} = \frac{1}{\Gamma(\beta)} \int_{-\infty}^\infty \exp(-te^x + \beta x)\, dx. \tag{5}$$

Applying the infinite trapezoidal rule with step size $h > 0$ leads to the approximation

$$\frac{1}{t^\beta} \approx \frac{1}{\Gamma(\beta)} \sum_{n=-\infty}^{\infty} w_n e^{-a_n t} \tag{6}$$

where

$$a_n = e^{hn} \quad \text{and} \quad w_n = he^{\beta nh}. \tag{7}$$

We will see that the relative error,

$$\rho(t) = 1 - \frac{t^\beta}{\Gamma(\beta)} \sum_{n=-\infty}^{\infty} w_n e^{-a_n t}, \tag{8}$$

satisfies a uniform bound for $0 < t < \infty$. If t is restricted to a compact interval $[\delta, T]$ with $0 < \delta < T < \infty$, then we can similarly bound the relative error in the *finite* exponential sum approximation

$$\frac{1}{t^\beta} \approx \frac{1}{\Gamma(\beta)} \sum_{n=-M}^{N} w_n e^{-a_n t} \quad \text{for } \delta \le t \le T, \tag{9}$$

for suitable choices of $M > 0$ and $N > 0$.

The exponents $a_n = e^{nh}$ in the sum (9) tend to zero as $n \to -\infty$. In Sect. 3 we see how, for a suitable threshold exponent size a^*, Prony's method may be used to replace $\sum_{a_n \le a^*} w_n e^{-a_n t}$ with an exponential sum having fewer terms. This idea again follows Beylkin and Monzón [3], who discussed it in the context of approximation by Gaussian sums.

Section 4 introduces an alternative approach based on the substitution $p = \exp(x - e^{-x})$, which transforms (4) into the integral representation

$$\frac{1}{t^\beta} = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{\infty} \exp\bigl(-\varphi(x,t)\bigr)(1 + e^{-x})\,dx, \tag{10}$$

where

$$\varphi(x,t) = tp - \beta \log p = t\exp(x - e^{-x}) - \beta(x - e^{-x}). \tag{11}$$

Applying the infinite trapezoidal rule again leads to an approximation of the form (6), this time with

$$a_n = \exp\bigl(nh - e^{-nh}\bigr) \quad \text{and} \quad w_n = h(1 + e^{-nh})\exp\bigl(\beta(nh - e^{-nh})\bigr). \tag{12}$$

As $x \to \infty$, the integrands in both (5) and (10) decay like $\exp(-te^x)$. However, they exhibit different behaviours as $x \to -\infty$, with the former decaying like $e^{\beta x} = e^{-\beta|x|}$ whereas the latter decays much faster, like $\exp(-\beta e^{-x}) = \exp(-\beta e^{|x|})$, as seen in Fig. 1 (note the differing scales on the vertical axis).

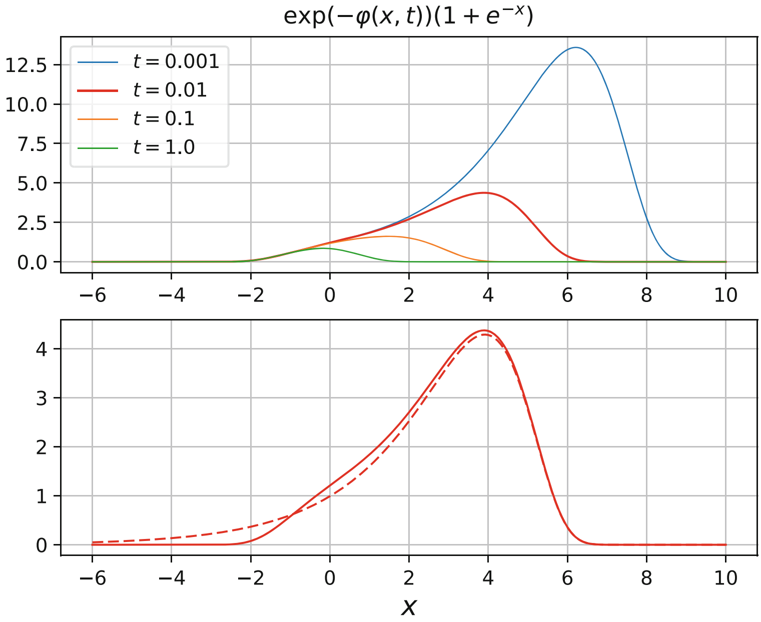


Fig. 1 Top: the integrand from (10) when $\beta = 1/2$ for different t. Bottom: comparison between the integrands from (5) and (10) when $t = 0.01$; the dashed line is the former and the solid line the latter

Li [5] summarised several alternative approaches for fast evaluation of a fractional integral of order α, that is, for an integral operator of the form (1) with kernel

$$k(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)} = \frac{\sin \pi\alpha}{\pi} \int_0^\infty e^{-pt} p^{-\alpha}\, dp \quad \text{for } 0 < \alpha < 1 \text{ and } t > 0, \tag{13}$$

where the integral representation follows from (4), with $\beta = 1-\alpha$, and the reflection formula for the Gamma function, $\Gamma(\alpha)\Gamma(1-\alpha) = \pi/\sin \pi\alpha$. She developed a quadrature approximation,

$$\int_0^\infty e^{-pt} p^{-\alpha}\, dp \approx \sum_{j=1}^{Q} w_j e^{-p_j t} p_j^{-\alpha} \quad \text{for } \delta \le t < \infty, \tag{14}$$

which again provides an exponential sum approximation, and showed that the error can be made smaller than ϵ for all $t \in [\delta, \infty)$ with Q of order $(\log \epsilon^{-1} + \log \delta^{-1})^2$.

More recently, Jiang et al. [4] developed an exponential sum approximation for $t \in [\delta, T]$ using composite Gauss quadrature on dyadic intervals, applied to (5), with Q of order

$$(\log \epsilon^{-1}) \log\bigl(T\delta^{-1} \log \epsilon^{-1}\bigr) + (\log \delta^{-1}) \log\bigl(\delta^{-1} \log \epsilon^{-1}\bigr).$$

In other applications, the kernel $k(t)$ is known via its Laplace transform,

$$\hat{k}(z) = \int_0^\infty e^{-zt} k(t)\, dt,$$

so that instead of the exponential sum (2) it is natural to seek a sum-of-poles approximation,

$$\hat{k}(z) \approx \sum_{l=1}^{L} \frac{w_l}{z - b_l}$$

for z in a suitable region of the complex plane; see, for instance, Alpert et al. [2] and Xu and Jian [7].

2 Approximation Based on the Substitution $p = e^x$

The nature of the approximation (6) is revealed by a remarkable formula for the relative error [3, Section 2]. For completeness, we outline the proof.

Theorem 1 *If the exponents and weights are given by* (7)*, then the relative error* (8) *has the representation*

$$\rho(t) = -2\sum_{n=1}^{\infty} R(n/h)\cos\bigl(2\pi(n/h)\log t - \Phi(n/h)\bigr) \tag{15}$$

where $R(\xi)$ and $\Phi(\xi)$ are the real-valued functions defined by

$$\frac{\Gamma(\beta+i2\pi\xi)}{\Gamma(\beta)} = R(\xi)e^{i\Phi(\xi)} \quad \textit{with } R(\xi) > 0 \textit{ and } \Phi(0) = 0.$$

Moreover, $R(\xi) \le e^{-2\pi\theta|\xi|}/(\cos\theta)^\beta$ for $0 \le \theta < \pi/2$ and $-\infty < \xi < \infty$.

Proof For each $t > 0$, the integrand $f(x) = \exp(-te^x + \beta x)$ from (5) belongs to the Schwarz class of rapidly decreasing C^∞ functions, and we may therefore apply the Poisson summation formula to conclude that

$$h\sum_{n=-\infty}^{\infty} f(nh) = \sum_{n=-\infty}^{\infty} \tilde{f}(n/h) = \int_{-\infty}^{\infty} f(x)\,dx + \sum_{n\neq 0}\tilde{f}(n/h),$$

where the Fourier transform of f is

$$\tilde{f}(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi\xi x} f(x)\,dx = \int_{-\infty}^{\infty} \exp\bigl(-te^x + (\beta - i2\pi\xi)x\bigr)\,dx.$$

The substitution $p = te^x$ gives

$$\tilde{f}(\xi) = \frac{1}{t^{\beta-i2\pi\xi}}\int_0^\infty e^{-p}p^{\beta-i2\pi\xi}\,\frac{dp}{p} = \frac{\Gamma(\beta-i2\pi\xi)}{t^{\beta-i2\pi\xi}},$$

so, with a_n and w_n defined by (7),

$$\frac{1}{\Gamma(\beta)}\sum_{n=-\infty}^{\infty} w_n e^{-a_n t} = \frac{1}{t^\beta} + \frac{1}{t^\beta}\sum_{n\neq 0}\frac{\Gamma(\beta - i2\pi n/h)}{\Gamma(\beta)}\,t^{i2\pi n/h}.$$

The formula for $\rho(t)$ follows after noting that $\overline{\Gamma(\beta+i2\pi\xi)} = \Gamma(\beta - i2\pi\xi)$ for all real ξ; hence, $R(-\xi) = R(\xi)$ and $\Phi(-\xi) = -\Phi(\xi)$.

To estimate $R(\xi)$, let $y > 0$ and define the ray $\mathscr{C}_\theta = \{se^{i\theta} : 0 < s < \infty\}$. By Cauchy's theorem,

$$\Gamma(\beta+iy) = \int_{\mathscr{C}_\theta} e^{-p}p^{\beta+iy}\,\frac{dp}{p} = \int_0^\infty e^{-se^{i\theta}}(se^{i\theta})^{\beta+iy}\,\frac{ds}{s}$$

and thus

$$|\Gamma(\beta+iy)| \le \int_0^\infty e^{-s\cos\theta}e^{-\theta y}s^\beta\,\frac{ds}{s} = \frac{e^{-\theta y}}{(\cos\theta)^\beta}\int_0^\infty e^{-s}s^\beta\,\frac{ds}{s} = \frac{e^{-\theta y}}{(\cos\theta)^\beta}\,\Gamma(\beta),$$

implying the desired bound for $R(\xi)$. □

In practice, the amplitudes $R(n/h)$ decay so rapidly with n that only the first term in the expansion (15) is significant. For instance, since [1, 6.1.30]

$$\left|\Gamma(\tfrac{1}{2}+iy)\right|^2 = \frac{\pi}{\cosh(\pi y)},$$

if $\beta = 1/2$ then $R(\xi) = (\cosh 2\pi^2\xi)^{-1/2} \le \sqrt{2}e^{-\pi^2\xi}$ so, choosing $h = 1/3$, we have $R(1/h) = 1.95692 \times 10^{-13}$ and $R(2/h) = 2.70786 \times 10^{-26}$. In general, the bound $R(n/h) \le e^{-2\pi\theta n/h}/(\cos\theta)^\beta$ from the theorem is minimized by choosing $\tan\theta = 2\pi n/(\beta h)$, implying that

$$R(n/h) \le \bigl(1 + (r_n/\beta)^2\bigr)^{\beta/2}\exp\bigl(-r_n\arctan(r_n/\beta)\bigr) \quad \text{where } r_n = 2\pi n/h.$$

Since we can evaluate only a *finite* exponential sum, we now estimate the two tails of the infinite sum in terms of the upper incomplete Gamma function,

$$\Gamma(\beta,q) = \int_q^\infty e^{-p}p^\beta\,\frac{dp}{p} \quad \text{for } \beta > 0 \text{ and } q > 0.$$

Theorem 2 *If the exponents and weights are given by* (7)*, then*

$$t^\beta \sum_{n=N+1}^{\infty} w_n e^{-a_n t} \le \Gamma(\beta, te^{Nh}) \quad \textit{provided } te^{Nh} \ge \beta,$$

and

$$t^\beta \sum_{n=-\infty}^{-M-1} w_n e^{-a_n t} \le \Gamma(\beta) - \Gamma(\beta, te^{-Mh}) \quad \textit{provided } te^{-Mh} \le \beta.$$

Proof For each $t > 0$, the integrand $f(x) = \exp(-te^x + \beta x)$ from (5) decreases for $x > \log(\beta/t)$. Therefore, if $Nh \ge \log(\beta/t)$, that is, if $te^{Nh} \ge \beta$, then

$$t^\beta h \sum_{n=N+1}^{\infty} f(nh) \le t^\beta \int_{Nh}^\infty f(x)\,dx = \int_{te^{Nh}}^\infty e^{-p}p^\beta\,\frac{dp}{p} = \Gamma(\beta, te^{Nh}),$$

where, in the final step, we used the substitution $p = te^x$. Similarly, the function $f(-x) = \exp(-te^{-x} - \beta x)$ decreases for $x > \log(t/\beta)$ so if $Mh \ge \log(t/\beta)$, that is, if $te^{-Mh} \le \beta$, then

$$t^\beta h \sum_{n=-\infty}^{-M-1} f(nh) = t^\beta h \sum_{n=M+1}^{\infty} f(-nh) \le t^\beta \int_{Mh}^{\infty} f(-x)\,dx = \int_0^{te^{-Mh}} e^{-p} p^\beta \frac{dp}{p},$$

where, in the final step, we used the substitution $p = te^{-x}$. □

Given $\epsilon_{\mathrm{RD}} > 0$ there exists $h > 0$ such that

$$2\sum_{n=1}^{\infty} |\Gamma(\beta + i2\pi n/h)| = \epsilon_{\mathrm{RD}}\Gamma(\beta), \tag{16}$$

and by Theorem 1,

$$|\rho(t)| \le \epsilon_{\mathrm{RD}} \quad \text{for } 0 < t < \infty,$$

so ϵ_{RD} is an upper bound for the *relative discretization* error. Similarly, given a sufficiently small $\epsilon_{\mathrm{RT}} > 0$, there exist $x_\delta > 0$ and $X_T > 0$ such that $\delta e^{x_\delta} \ge \beta$ and $Te^{-X_T} \le \beta$ with

$$\Gamma(\beta, \delta e^{x_\delta}) = \epsilon_{\mathrm{RT}}\Gamma(\beta) \quad \text{and} \quad \Gamma(\beta) - \Gamma(\beta, Te^{-X_T}) = \epsilon_{\mathrm{RT}}\Gamma(\beta). \tag{17}$$

Thus, by Theorem 2,

$$\frac{t^\beta}{\Gamma(\beta)} \sum_{n=N+1}^{\infty} w_n e^{-a_n t} \le \epsilon_{\mathrm{RT}} \quad \text{for } t \ge \delta \text{ and } Nh \ge x_\delta,$$

and

$$\frac{t^\beta}{\Gamma(\beta)} \sum_{n=-\infty}^{-M-1} w_n e^{-a_n t} \le \epsilon_{\mathrm{RT}} \quad \text{for } t \le T \text{ and } Mh \ge X_T,$$

showing that $2\epsilon_{\mathrm{RT}}$ is an upper bound for the *relative truncation* error. Denoting the overall relative error for the finite sum (9) by

$$\rho_M^N(t) = 1 - \frac{t^\beta}{\Gamma(\beta)} \sum_{n=-M}^{N} w_n e^{-a_n t}, \tag{18}$$

we therefore have

$$|\rho_M^N(t)| \le \epsilon_{\mathrm{RD}} + 2\epsilon_{\mathrm{RT}} \quad \text{for } \delta \le t \le T, Nh \ge x_\delta \text{ and } Mh \ge X_T. \tag{19}$$

The estimate for $R(\xi)$ in Theorem 1, together with the asymptotic behaviours

$$\Gamma(\beta, q) \sim q^{\beta-1}e^{-q} \quad \text{as } q \to \infty,$$

and

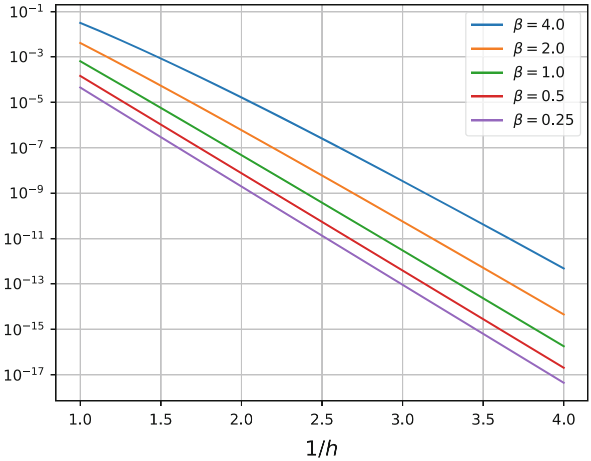
$$\Gamma(\beta) - \Gamma(\beta, q) \sim \frac{q^\beta}{\beta} \quad \text{as } q \to 0,$$

imply that (19) can be satisfied with

$$h^{-1} \ge C \log \epsilon_{\mathrm{RD}}^{-1}, \qquad N \ge Ch^{-1} \log(\delta^{-1} \log \epsilon_{\mathrm{RT}}^{-1}), \qquad M \ge Ch^{-1} \log(T\epsilon_{\mathrm{RT}}^{-1}).$$

Figure 2 shows the relation between ϵ_{RD} and $1/h$ given by (16), and confirms that $1/h$ is approximately proportional to $\log \epsilon_{\mathrm{RT}}^{-1}$. In Fig. 3, for each value of ϵ we computed h by solving (16) with $\epsilon_{\mathrm{RD}} = \epsilon/3$, then computed x_δ and X_T by solving (17) with $\epsilon_{\mathrm{RT}} = \epsilon/3$, and finally put $M = \lceil X_T/h \rceil$ and $N = \lceil x_\delta/h \rceil$.

Fig. 2 The bound ϵ_{RD} for the relative discretization error, defined by (16), as a function of $1/h$ for various choices of β



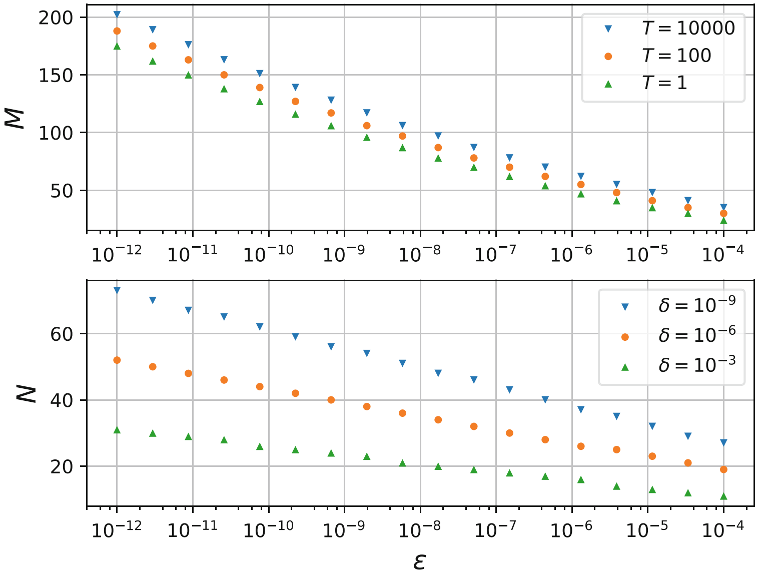


Fig. 3 The growth in M and N as the upper bound for the overall relative error (18) decreases, for different choices of T and δ

3 Prony's Method

The construction of Sect. 2 leads to an exponential sum approximation (9) with many small exponents a_n. We will now explain how the corresponding terms can be aggregated to yield a more efficient approximation.

Consider more generally an exponential sum

$$g(t) = \sum_{l=1}^{L} w_l e^{-a_l t},$$

in which the weights and exponents are all strictly positive. Our aim is to approximate this function by an exponential sum with fewer terms,

$$g(t) \approx \sum_{k=1}^{K} \tilde{w}_k e^{-\tilde{a}_k t}, \quad 2K - 1 < L,$$

whose weights $\tilde{w}_k$ and exponents $\tilde{a}_k$ are again all strictly positive. To this end, let

$$g_j = (-1)^j g^{(j)}(0) = \sum_{l=1}^{L} w_l a_l^j.$$

We can hope to find $2K$ parameters $\tilde{w}_k$ and $\tilde{a}_k$ satisfying the $2K$ conditions

$$g_j = \sum_{k=1}^{K} \tilde{w}_k \tilde{a}_k^j \quad \text{for } 0 \le j \le 2K-1, \tag{20}$$

so that, by Taylor expansion,

$$g(t) \approx \sum_{j=0}^{2K-1} g_j \frac{(-t)^j}{j!} = \sum_{k=1}^{K} \tilde{w}_k \sum_{j=0}^{2K-1} \frac{(-\tilde{a}_k t)^j}{j!} \approx \sum_{k=1}^{K} \tilde{w}_k e^{-\tilde{a}_k t}.$$

The approximations here require that the g_j and the $\tilde{a}_k t$ are nicely bounded, and preferably small.

In Prony's method, we seek to satisfy (20) by introducing the monic polynomial

$$Q(z) = \prod_{k=1}^{K} (z - \tilde{a}_k) = \sum_{k=0}^{K} q_k z^k,$$

and observing that the unknown coefficients q_k must satisfy

$$\sum_{m=0}^{K} g_{j+m} q_m = \sum_{m=0}^{K} \sum_{k=1}^{K} \tilde{w}_k \tilde{a}_k^{j+m} q_m = \sum_{k=1}^{K} \tilde{w}_k \tilde{a}_k^j \sum_{m=0}^{K} q_m \tilde{a}_k^m = \sum_{k=1}^{K} \tilde{w}_k \tilde{a}_k^j Q(\tilde{a}_k) = 0,$$

for $0 \le j \le K-1$ (so that $j + m \le 2K-1$ for $0 \le m \le K$), with $q_K = 1$. Thus,

$$\sum_{m=0}^{K-1} g_{j+m} q_m = b_j, \quad \text{where } b_j = -g_{j+K}, \quad \text{for } 0 \le j \le K-1,$$

which suggests the procedure *Prony* defined in Algorithm 1. We must, however, beware of several potential pitfalls:

1. the best choice for K is not clear;
2. the $K \times K$ matrix $[g_{j+k}]$ might be badly conditioned;
3. the roots of the polynomial $Q(z)$ might not all be real and positive;

Algorithm 1 *Prony*$(a_1, \ldots, a_L, w_1, \ldots w_L, K)$

```
Require: $2K - 1 \le L$
  Compute $g_j = \sum_{l=1}^{L} w_l a_l^j$ for $0 \le j \le 2K-1$
  Find $q_0, \ldots, q_{K-1}$ satisfying $\sum_{m=0}^{K-1} g_{j+m} q_m = -g_{j+K}$ for $0 \le j \le K-1$, and put $q_K = 1$
  Find the roots $\tilde{a}_1, \ldots, \tilde{a}_K$ of the polynomial $Q(z) = \sum_{k=0}^{K} q_k z^k$
  Find $\tilde{w}_1, \ldots, \tilde{w}_K$ satisfying $\sum_{k=1}^{K} \tilde{a}_k^j \tilde{w}_k \approx g_j$ for $0 \le j \le 2K-1$
  return $\tilde{a}_1, \ldots, \tilde{a}_K, \tilde{w}_1, \ldots, \tilde{w}_K$
```

4. the linear system for the $\tilde{w}_k$ is overdetermined, and the least-squares solution might have large residuals;
5. the $\tilde{w}_k$ might not all be positive.

We will see that nevertheless the algorithm can be quite effective, even when $K = 1$, in which case we simply compute

$$g_0 = \sum_{l=1}^{L} w_l, \quad g_1 = \sum_{l=1}^{L} w_l a_l, \quad \tilde{a}_1 = g_1/g_0, \quad \tilde{w}_1 = g_0.$$

Example 1 We took $\beta = 3/4$, $\delta = 10^{-6}$, $T = 10$, $\epsilon = 10^{-8}$, $\epsilon_{\mathrm{RD}} = 0.9 \times 10^{-8}$ and $\epsilon_{RT} = 0.05 \times 10^{-8}$. The methodology of Sect. 2 led to the choices $h = 0.47962$, $M = 65$ and $N = 36$, and we confirmed via direct evaluation of the relative error that $|\rho_M^N(t)| \le 0.92 \times 10^{-8}$ for $\delta \le t \le T$. We applied Prony's method to the first L terms of the sum in (9), that is, those with $-M \le n \le L - M$, thereby reducing the total number of terms by $L - K$. Table 1 lists, for different choices of L and K, the additional contribution to the relative error, that is, $\max_{1 \le p \le P} |\eta(t_p)|$ where

$$\eta(t) = \frac{t^\beta}{\Gamma(\beta)} \left(\sum_{k=1}^{K} \tilde{w}_k e^{-\tilde{a}_k t} - \sum_{l=1}^{L} w_{l'} e^{-a_{l'} t} \right), \qquad l' = l - M + 1, \tag{21}$$

Table 1 Performance of Prony's method for different L and K using the parameters of Example 1

L	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
66	9.64e−01	4.30e−01	6.15e−02	3.02e−03	4.77e−05	2.29e−07
65	8.11e−01	1.69e−01	9.89e−03	1.80e−04	9.98e−07	**1.66e−09**
64	5.35e−01	4.59e−02	1.03e−03	6.85e−06	1.35e−08	7.96e−12
63	2.72e−01	9.17e−03	7.76e−05	1.89e−07	**1.36e−10**	2.74e−14
62	1.12e−01	1.46e−03	4.64e−06	**4.19e−09**	1.11e−12	3.58e−16
61	3.99e−02	1.98e−04	2.38e−07	8.05e−11	8.28e−15	3.52e−16
60	1.28e−02	2.43e−05	1.10e−08	1.41e−12	4.63e−16	2.24e−16
59	3.82e−03	2.78e−06	**4.81e−10**	2.36e−14	4.63e−16	1.25e−16
58	1.10e−03	3.05e−07	2.02e−11	4.46e−16	1.23e−16	6.27e−17
57	3.07e−04	3.27e−08	8.25e−13	5.60e−17	8.40e−17	
56	8.43e−05	**3.44e−09**	3.32e−14	8.96e−17	5.60e−17	
55	2.29e−05	3.59e−10	1.32e−15	4.48e−17	4.48e−17	
48	**2.30e−09**	3.98e−17	2.58e−18			
47	6.16e−10	3.92e−18	1.54e−18			

For each K, we seek the largest L for which the maximum relative error (shown in bold) is less than $\epsilon = 10^{-8}$

and we use a geometric grid in $[\delta, 1]$ given by $t_p = T^{(p-1)/(P-1)}\delta^{(P-p)/(P-1)}$ for $1 \le p \le P$ with $P = 751$. The largest reduction consistent with maintaining overall accuracy was when $L = 65$ and $K = 6$, and Fig. 4 (Top) plots $|\eta(t)|$ in this case, as well as the overall relative error (Bottom) for the resulting approximation,

$$\frac{1}{t^\beta} \approx \frac{1}{\Gamma(\beta)}\left(\sum_{k=1}^{K} \tilde{w}_k e^{-\tilde{a}_k t} + \sum_{n=L-M}^{N} w_n e^{-a_n t}\right) \quad \text{for } 10^{-6} \le t \le 10. \tag{22}$$

In this way, the number of terms in the exponential sum approximation was reduced from $M+1+N = 102$ to $(M+K-L)+1+N = 43$, with the maximum absolute value of the relative error growing only slightly to 1.07×10^{-8}. Figure 4 (Bottom) shows that the relative error is closely approximated by the first term in (15), that is, $\rho_N^M(t) \approx -2R(h^{-1})\cos\bigl(2\pi h^{-1}\log t - \Phi(h^{-1})\bigr)$ for $\delta \le t \le T$.

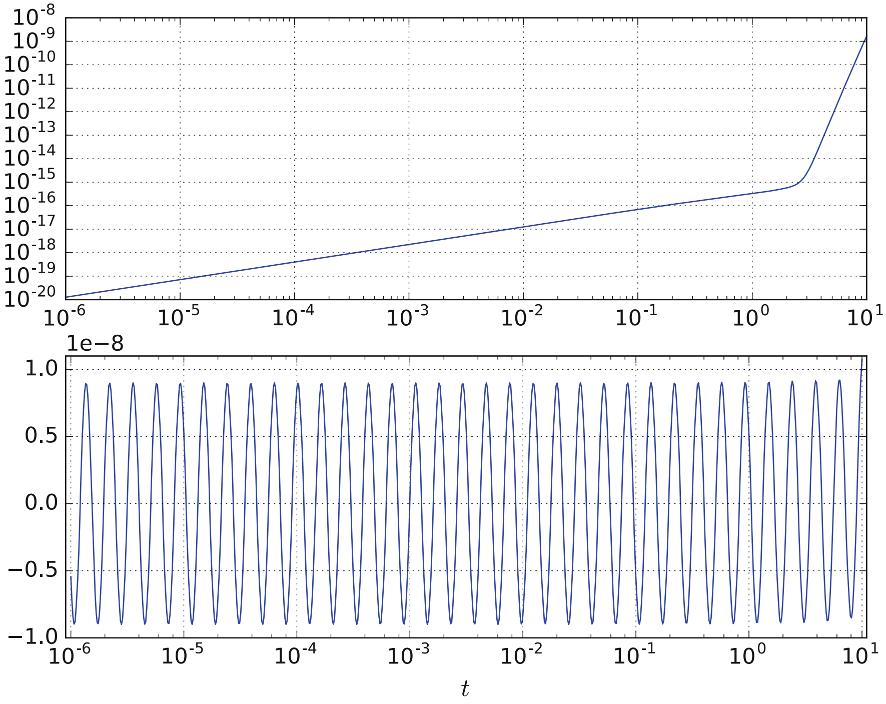


Fig. 4 Top: the additional contribution $|\eta(t)|$ to the relative error from applying Prony's method in Example 1 with $L = 65$ and $K = 6$. Bottom: the overall relative error for the resulting approximation (22) of $t^{-\beta}$ requiring $L - K = 59$ fewer terms

4 Approximation Based on the Substitution $p = \exp(x - e^{-x})$

We now consider the alternative exponents and weights given by (12). A different approach is needed for the error analysis, and we define

$$\mathscr{I}(f) = \int_{-\infty}^{\infty} f(x)\,dx \quad \text{and} \quad \mathscr{Q}(f,h) = h\sum_{n=-\infty}^{\infty} f(nh) \quad \text{for } h > 0,$$

so that $\mathscr{Q}(f,h)$ is an infinite trapezoidal rule approximation to $\mathscr{I}(f)$. Recall the following well-known error bound.

Theorem 3 *Let $r > 0$. Suppose that $f(z)$ is continuous on the closed strip $|\Im z| \le r$, analytic on the open strip $|\Im z| < r$, and satisfies*

$$\int_{-\infty}^{\infty} \bigl(|f(x+ir)| + |f(x-ir)|\bigr)\,dx \le A_r$$

with

$$\int_{-r}^{r} |f(x \pm iy)|\,dy \to 0 \quad \text{as } |x| \to \infty.$$

Then, for all $h > 0$,

$$|\mathscr{Q}(f,h) - \mathscr{I}(f)| \le \frac{A_r e^{-2\pi r/h}}{1 - e^{-2\pi r/h}}.$$

Proof See McNamee et al. [6, Theorem 5.2]. □

For $t > 0$, we define the entire analytic function of z,

$$f(z) = \exp\bigl(-\varphi(z,t)\bigr)(1 + e^{-z}), \tag{23}$$

where $\varphi(z,t)$ is the analytic continuation of the function defined in (11). In this way, $t^{-\beta} = \mathscr{I}(f)/\Gamma(\beta)$ by (10).

Lemma 1 *If $0 < r < \pi/2$, then the function f defined in (23) satisfies the hypotheses of Theorem 3 with $A_r \le Ct^{-\beta}$ for $0 < t \le 1$, where the constant $C > 0$ depends only on β and r.*

Proof A short calculation shows that

$$\Re\varphi(x \pm iy, t) = t\exp(x - e^{-x}\cos y)\cos(y + e^{-x}\sin y) - \beta(x - e^{-x}\cos y),$$

and that if $0 < \epsilon < \pi/2 - r$, then

$$0 \le y + e^{-x} \sin y \le \frac{\pi}{2} - \epsilon \quad \text{for } x \ge x^* = \log \frac{\sin r}{\pi/2 - r - \epsilon} \text{ and } 0 \le y \le r. \tag{24}$$

Thus, if $x \ge x^*$ then $\cos(r + e^{-x} \sin r) \ge \cos(\pi/2 - \epsilon) = \sin\epsilon$ so

$$\Re\varphi(x \pm ir, t) \ge t \exp(x - e^{-x^*} \cos r) \sin\epsilon - \beta x + \beta e^{-x} \cos r \ge cte^x - \beta x,$$

where $c = \exp(-e^{-x^*} \cos r) \sin\epsilon > 0$. If necessary, we increase x^* so that $x^* > 0$. Since $|1 + e^{-(x \pm ir)}| \le 1 + e^{-x}$,

$$\begin{aligned}\int_{x^*}^{\infty} |f(x \pm ir)|\, dx &= \int_{x^*}^{\infty} \exp\bigl(-\Re\varphi(x \pm ir, t)\bigr) \bigl|1 + e^{-(x \pm ir)}\bigr|\, dx \\ &\le \int_{x^*}^{\infty} \exp(-cte^x + \beta x)(1 + e^{-x})\, dx,\end{aligned}$$

and the substitution $p = e^x$ then yields, with $p^* = e^{x^*}$,

$$\begin{aligned}\int_{x^*}^{\infty} |f(x \pm ir)|\, dx &\le \int_{p^*}^{\infty} e^{-ctp} p^{\beta} (1 + p^{-1}) \frac{dp}{p} \le \bigl(1 + (p^*)^{-1}\bigr) \int_{p^*}^{\infty} e^{-ctp} p^{\beta} \frac{dp}{p} \\ &= \frac{1 + (p^*)^{-1}}{(ct)^{\beta}} \int_{ctp^*}^{\infty} e^{-p} p^{\beta} \frac{dp}{p} \le \frac{1 + (p^*)^{-1}}{(ct)^{\beta}} \int_{0}^{\infty} e^{-p} p^{\beta} \frac{dp}{p} \equiv Ct^{-\beta}.\end{aligned}$$

Also, if $x \ge 0$ then

$$\Re\varphi(x \pm ir, t) \ge -t \exp(x - e^{-x} \cos r) - \beta(x - e^{-x} \cos r) \ge -te^x - \beta x$$

so

$$\int_{0}^{x^*} |f(x \pm ir)|\, dx \le \int_{0}^{x^*} \exp(te^x + \beta x)(1 + e^{-x})\, dx \le 2x^* \exp\bigl(te^{x^*} + \beta x^*\bigr),$$

which is bounded for $0 < t \le 1$. Similarly, if $x \le 0$ then $\exp(x - e^{-x} \cos r) \le 1$ so $\Re\varphi(x \pm ir, t) \ge -t + \beta e^{-x} \cos r$ and therefore, using again the substitution $p = e^x$,

$$\begin{aligned}\int_{-\infty}^{0} |f(x \pm ir)|\, dx &\le \int_{-\infty}^{0} \exp(t - \beta e^{-x} \cos r)(1 + e^{-x})\, dx \\ &= \int_{0}^{\infty} \exp(t - \beta e^{x} \cos r)(1 + e^{x})\, dx = e^t \int_{1}^{\infty} e^{-\beta p \cos r} (1 + p) \frac{dp}{p},\end{aligned}$$

which is also bounded for $0 < t \le 1$. The required estimate for A_r follows.

If $x \ge x^*$, then the preceding inequalities based on (24) show that

$$\int_{-r}^{r} |f(x+iy)|\,dy \le 2r \max_{|y|\le r} |f(x+iy)| \le 2r\exp(-cte^{x} + \beta x)(1+e^{-x}),$$

which tends to zero as $x \to \infty$ for any $t > 0$. Similarly, if $x \le 0$, then $\Re\varphi(x \pm iy) \ge -t + \beta e^{-x}\cos r$ for $|y| \le r$, so

$$\int_{-r}^{r} |f(x+iy)|\,dy \le 2r\exp(t - \beta e^{-x}\cos r)(1+e^{-x}),$$

which again tends to zero as $x \to -\infty$. □

Together, Theorem 3 and Lemma 1 imply the following bound for the relative error (8) in the infinite exponential sum approximation (6).

Theorem 4 *Let $h > 0$ and define a_n and w_n by (12). If $0 < r < \pi/2$, then there exists a constant C_1 (depending on β and r) such that*

$$|\rho(t)| \le C_1 e^{-2\pi r/h} \quad \textit{for}\ 0 < t \le 1.$$

Proof The definitions above mean that $hf(nh) = w_n e^{-a_n t}$. □

Thus, a relative accuracy ϵ is achieved by choosing h of order $1/\log\epsilon^{-1}$. Of course, in practice we must compute a finite sum, and the next lemma estimates the two parts of the associated truncation error.

Lemma 2 *Let $h > 0$, $0 < \theta < 1$ and $0 < t \le 1$. Then the function f defined in (23) satisfies*

$$\frac{h}{\Gamma(\beta)} \sum_{M=-\infty}^{-M-1} f(nh) \le C_2\exp(-\beta e^{Mh}) \quad \textit{for} \quad Mh \ge \begin{cases} \log(\beta^{-1}-1), & 0<\beta<1/2,\\ 0, & \beta \ge 1/2, \end{cases} \tag{25}$$

and

$$\frac{h}{\Gamma(\beta)} \sum_{n=N+1}^{\infty} f(nh) \le \frac{C_3}{t^{\beta}} \exp\bigl(-\theta t e^{Nh-1}\bigr) \quad \textit{for} \quad Nh \ge 1 + \log(\beta t^{-1}). \tag{26}$$

When $0 < \beta \le 1$, the second estimate holds also with $\theta = 1$.

Proof If $n \le 0$, then $\varphi(nh, t) \ge -t + \beta e^{-nh}$ so

$$f(nh) \le g_1(-nh) \quad \text{where} \quad g_1(x) = \exp(t - \beta e^{x})(1+e^{x}).$$

The function $g_1(x)$ decreases for $x > \log(\beta^{-1} - 1)$ if $0 < \beta < 1/2$, and for all $x \geq 0$ if $\beta \geq 1/2$, so

$$h \sum_{n=-\infty}^{-M-1} f(nh) \leq h \sum_{n=M+1}^{\infty} g_1(nh) \leq \int_{Mh}^{\infty} g_1(x)\, dx \quad \text{for } M \text{ as in (25)},$$

and the substitution $p = e^x$ gives

$$\int_{Mh}^{\infty} g_1(x)\, dx = \int_{e^{Mh}}^{\infty} e^{t-\beta p}(1+p)\, \frac{dp}{p} \leq 2e^t \int_{e^{Mh}}^{\infty} e^{-\beta p}\, dp = \frac{2e^t}{\beta} \exp(-\beta e^{Mh}),$$

so the first estimate holds with $C_2 = 2e/\Gamma(\beta + 1)$.

If $n \geq 0$ we have $\varphi(nh, t) \geq t \exp(nh - 1) - \beta nh$ and $1 + e^{-nh} \leq 2$, so

$$f(nh) \leq g_2(nh) \quad \text{where} \quad g_2(x) = 2 \exp(-te^{x-1} + \beta x).$$

The function $g_2(x)$ decreases for $x > 1 + \log(\beta t^{-1})$, so

$$h \sum_{n=N+1}^{\infty} f(nh) \leq \int_{Nh}^{\infty} g_2(x)\, dx \quad \text{for } N \text{ as in (26)},$$

and the substitution $p = e^x$ gives

$$\int_{Nh}^{\infty} g_2(x)\, dx \leq 2 \int_{e^{Nh}}^{\infty} e^{-te^{-1}p} p^{\beta}\, \frac{dp}{p} = 2\left(\frac{e}{t}\right)^{\beta} \int_{te^{Nh-1}}^{\infty} e^{-p} p^{\beta-1}\, dp.$$

Since $te^{Nh-1} \geq \beta$, if $0 < \beta \leq 1$ then the integral on the right is bounded above by $\beta^{\beta-1} \exp(-te^{Nh-1})$. If $\beta > 1$, then $p^{\beta-1} e^{-(1-\theta)p}$ is bounded for $p > 0$ so

$$\int_{te^{Nh-1}}^{\infty} e^{-p} p^{\beta-1}\, dp = \int_{te^{Nh-1}}^{\infty} e^{-\theta p} (p^{\beta-1} e^{-(1-\theta)p})\, dp \leq C \exp(-\theta te^{Nh-1}),$$

completing the proof. □

It is now a simple matter to see that the number of terms $L = M + 1 + N$ needed to ensure a relative accuracy ϵ for $\delta \leq t \leq 1$ is of order $(\log \epsilon^{-1}) \log(\delta^{-1} \log \epsilon^{-1})$.

Theorem 5 *Let a_n and w_n be defined by (12). For $0 < \delta \leq 1$ and for a sufficiently small $\epsilon > 0$, if*

$$\frac{1}{h} \geq \frac{1}{2\pi r} \log \frac{3C_1}{\epsilon}, \quad M \geq \frac{1}{h} \log\left(\frac{1}{\beta} \log \frac{3C_2}{\epsilon}\right), \quad N \geq 1 + \frac{1}{h} \log\left(\frac{1}{\theta\delta} \log \frac{3C_3}{\epsilon}\right),$$

then

$$|\rho_M^N(t)| \le \epsilon \quad \textit{for } \delta \le t \le 1.$$

Proof The inequalities for h, M and N imply that each of $C_1 e^{-2\pi r/h}$, $C_2 \exp(-\beta e^{Mh})$ and $C_3 \exp(-\theta t e^{Nh-1})$ is bounded above by $\epsilon t^{-\beta}/3$, so the error estimate is a consequence of Theorem 4, Lemma 2 and the triangle inequality. Note that the restrictions on M and N in (25) and (26) will be satisfied for ϵ sufficiently small. □

Although the error bounds above require $t \in [\delta, 1]$, a simple rescaling allows us to treat a general compact subinterval $[\delta, T]$. If $\check{a}_n = a_n/T$ and $\check{w}_n = w_n/T^\beta$, then

$$\frac{1}{t^\beta} = \frac{1}{T^\beta}\frac{1}{(t/T)^\beta} \approx \frac{1}{\Gamma(\beta)}\sum_{n=-M}^{N} \check{w}_n e^{-\check{a}_n t}$$

for $\delta \le t/T \le 1$, or in other words for $\delta \cdot T \le t \le T$. Moreover, the relative error $\check{\rho}_M^N(t) = \rho_M^N(t/T)$ is unchanged by the rescaling.

Example 2 We took the same values for β, δ, T, ϵ, ϵ_{RD} and ϵ_{RT} as in Example 1. Since the constant C_1 of Theorem 4 is difficult to estimate, we again used (16) to choose $h = 0.47962$. Likewise, the constant C_3 in Lemma 2 is difficult to estimate, so we chose $N = \lceil h^{-1} x_{\delta/T} \rceil = 40$. However, knowing $C_2 = 2e/\Gamma(\beta+1)$ we easily determined that $C_2 \exp(-\beta e^{Mh}) \le \epsilon_{\mathrm{RT}}$ for $M = 8$. The exponents and weights (12) were computed for the interval $[\delta/T, 1]$, and then rescaled as above to create an approximation for the interval $[\delta, T]$ with $M + 1 + N = 49$ terms and a relative error whose magnitude is at worst 2.2×10^{-8}.

The behaviour of the relative error $\rho_M^N(t)$, shown in Fig. 5, suggests a modified strategy: construct the approximation for $[\delta, 10T]$ but use it only on $[\delta, T]$. We found that doing so required $N = 45$, that is, 5 additional terms, but resulted in a nearly uniform amplitude for the relative error of about 0.97×10^{-8}. Finally, after applying Prony's method with $L = 17$ and $K = 6$ we were able to reduce the number of terms from $M + 1 + N = 54$ to 43 without increasing the relative error.

To compare these results with those of Li [5], let $0 < \alpha < 1$ and let $k(t) = t^{\alpha-1}/\Gamma(\alpha)$ denote the kernel for the fractional integral of order α. Taking $\beta = 1 - \alpha$ we compute the weights w_l and exponents a_l as above and define

$$k_M^N(t) = \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)}\sum_{n=-M}^{N} w_n e^{-a_n t} \quad \text{for } \delta \le t \le T.$$

The fast algorithm evaluates

$$(\mathcal{K}_M^N U)^n = \int_0^{t_{n-1}} k_M^N(t-s)\tilde{U}(s)\,ds + \int_{t_{n-1}}^{t_n} k(t_n - s)\tilde{U}(s)\,ds$$

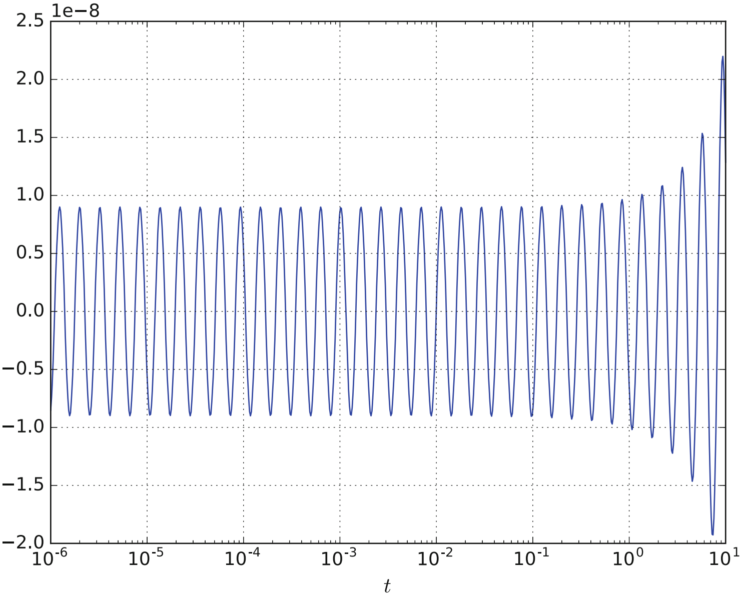


Fig. 5 The relative error for the initial approximation from Example 2

and our bound $|\rho_M^N(t)| \le \epsilon$ implies that $|k_M^N(t) - k(t)| \le \epsilon t^{\alpha-1}/\Gamma(\alpha)$ for $\delta \le t \le T$, so

$$\bigl|(\mathscr{K}_M^N U)^n - (\mathscr{K}\tilde{U})(t_n)\bigr| \le \epsilon \int_0^{t_{n-1}} \frac{(t_n - s)^{\alpha-1}}{\Gamma(\alpha)} \,|\tilde{U}(s)|\,ds \le \frac{\epsilon t_n^\alpha}{\Gamma(\alpha+1)} \max_{1\le j\le n} |U^j|,$$

provided $\Delta t_n \ge \delta$ and $t_n \le T$. Similarly, the method of Li yields $(\mathscr{K}_Q U)^n$ but with a bound for the *absolute* error in (14), so that $|k_Q(t) - k(t)| \le \epsilon'$ for $\delta' \le t < \infty$. Thus,

$$\bigl|(\mathscr{K}_Q U)^n - (\mathscr{K}\tilde{U})(t_n)\bigr| \le \epsilon' \frac{\sin \pi\alpha}{\pi} \int_0^{t_{n-1}} |\tilde{U}(s)|\,ds \le \epsilon' t_n \frac{\sin \pi\alpha}{\pi} \max_{1\le j\le n} |U^j|,$$

provided $\Delta t_n \ge \delta$. Li [5, Fig. 3 (d)] required about $Q = 250$ points to achieve an (absolute) error $\epsilon' \le 10^{-6}$ for $t \ge \delta' = 10^{-4}$ when $\alpha = 1/4$ (corresponding to $\beta = 1-\alpha = 3/4$). In Examples 1 and 2, our methods give a smaller error $\epsilon \le 10^{-8}$ using only $M + 1 + N = 43$ terms with a less restrictive lower bound for the time step, $\delta = 10^{-6}$. Against these advantages, the method of Li permits arbitrarily large t_n.

5 Conclusion

Comparing Examples 1 and 2, we see that, for comparable accuracy, the approximation based on the second substitution results in far fewer terms because we are able to use a much smaller choice of M. However, after applying Prony's method both approximations are about equally efficient. If Prony's method is not used, then the second approximation is clearly superior. Another consideration is that the first approximation has more explicit error bounds so we can, a priori, more easily determine suitable choices of h, M and N to achieve a desired accuracy.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1965)
2. Alpert, B., Greengard, L., Hagstrom, T.: Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation. SIAM J. Numer. Anal. **37**, 1138–1164 (2000)
3. Beylkin, G., Monzón, L.: Approximation by exponential sums revisited. Appl. Comput. Harmon. Anal. **28**, 131–149 (2010)
4. Jiang, S., Zhang, J., Zhang, Q., Zhang, Z.: Fast evaluation of the Caputo fractional derivative and its applications to fractional diffusion equations. Commun. Comput. Phys. **21**(3), 650–678 (2017)
5. Li, J.R.: A fast time stepping method for evaluating fractional integrals. SIAM J. Sci. Comput. **31**, 4696–4714 (2010)
6. McNamee, J., Stenger, F., Whitney, E.L.: Whittaker's cardinal function in retrospect. Math. Comput. **25**, 141–154 (1971)
7. Xu, K., Jiang, S.: A bootstrap method for sum-of-poles approximations. J. Sci. Comput. **55**, 16–39 (2013)

Approximate Quadrature Measures on Data-Defined Spaces

Hrushikesh N. Mhaskar

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract An important question in the theory of approximate integration is to study the conditions on the nodes $x_{k,n}$ and weights $w_{k,n}$ that allow an estimate of the form

$$\sup_{f\in\mathscr{B}_\gamma}\left|\sum_k w_{k,n}f(x_{k,n})-\int_{\mathbb{X}}fd\mu^*\right|\le cn^{-\gamma},\qquad n=1,2,\cdots,$$

where $\mathbb{X}$ is often a manifold with its volume measure μ^*, and $\mathscr{B}_\gamma$ is the unit ball of a suitably defined smoothness class, parametrized by γ. In this paper, we study this question in the context of a quasi-metric, locally compact, measure space $\mathbb{X}$ with a probability measure μ^*. We show that quadrature formulas exact for integrating the so called diffusion polynomials of degree $< n$ satisfy such estimates. Without requiring exactness, such formulas can be obtained as a solutions of some kernel-based optimization problem. We discuss the connection with the question of optimal covering radius. Our results generalize in some sense many recent results in this direction.

1 Introduction

The theory of approximate integration of a function based on finitely many samples of the function is a very old subject. Usually, one requires the necessary quadrature formula to be exact for some finite dimensional space. For example, we mention the following theorem, called Tchakaloff's theorem [28, Exercise 2.5.8, p. 100].

H. N. Mhaskar (✉)
Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, USA
e-mail: hrushikesh.mhaskar@cgu.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_41

(For simplicity of exposition, the notation used in the introduction may not be the same as in the rest of this paper.)

Theorem 1 *Let $\mathbb{X}$ be a compact topological space, $\{\phi_j\}_{j=0}^{N-1}$ be continuous real valued functions on $\mathbb{X}$, and μ^* be a probability measure on $\mathbb{X}$ (i.e., μ^* is a positive Borel measure with $\mu^*(\mathbb{X}) = 1$). Then there exist $N + 1$ points $x_1, \cdots, x_{N+1}$, and non-negative numbers $w_1, \cdots, w_{N+1}$ such that*

$$\sum_{k=1}^{N+1} w_k = 1, \qquad \sum_{k=1}^{N+1} w_k\phi_j(x_k) = \int_{\mathbb{X}} \phi_j(x)d\mu^*(x), \qquad j = 0, \cdots, N-1. \tag{1}$$

It is very easy to see that under the conditions of Theorem 1, if $f : \mathbb{X} \to \mathbb{R}$ is continuous, and $V_N = \mathsf{span}\{1, \phi_0, \cdots, \phi_{N-1}\}$ then

$$\left|\int_{\mathbb{X}} f(x)d\mu^*(x) - \sum_{k=1}^{N+1} w_k f(x_k)\right| \le 2 \min_{P\in V_N} \max_{x\in\mathbb{X}} |f(x) - P(x)|. \tag{2}$$

A great deal of modern research is concerned with various variations of this theme, interesting from the point of view of computation. For example, can we ensure all the weights w_k to be equal by a judicious choice of the points x_k, or can we obtain estimates similar to (2) with essentially arbitrary points x_k, with or without requiring that the weights be positive, or can we obtain better rates of convergence for subspaces (e.g., suitably defined Bessel potential spaces) of the space of continuous functions than that guaranteed by (2)? Of course, this research typically requires $\mathbb{X}$ to have some additional structure.

The current paper is motivated by the spate of research within the last couple of years, in particular, by the results in [5–9]. The focus in [6, 7] is the case when $\mathbb{X}$ is the unit sphere $\mathbb{S}^q$ embedded in $\mathbb{R}^{q+1}$, and the weights are all equal. A celebrated result by Bondarenko et al. in [4] shows that for every large enough n, there exist $\mathcal{O}(n^q)$ points on $\mathbb{S}^q$ such that equal weight quadrature formulas based at these points are exact for integrating spherical polynomials of degree $< n$. In these papers (see also [20]), it is shown that for such formulas an estimate of the following form holds:

$$\left|\int_{\mathbb{S}^q} f(x)d\mu^*(x) - \frac{1}{n}\sum_k f(x_k)\right| \le \frac{c}{n^s}\|f\|_{s,p,q}, \tag{3}$$

where $\|\cdot\|_{s,p,q}$ is a suitably defined Bessel potential subspace of $L^p(\mu^*)$ and the smoothness index s satisfies $s > q/p$, so that functions in this subspace are actually continuous. More generally, the systems of points x_k for which an estimate of the form (3) holds is called an approximate QMC design. Various constructions and properties of such designs are studied. The papers [5, 8] study certain analogous questions in the context of a smooth, compact, Riemannian manifold. The case of a Grassmanian manifold is studied in [9], with numerical illustrations.

Our paper is also motivated by machine learning considerations, where one is given a data set of the form $\{(x_j, y_j)\}$, sampled from some unknown probability distribution μ. The objective is to approximate $f(x) = \mathbb{E}(y|x)$. In this context, $\mathscr{P} = \{x_j\}$ is usually referred to as a *point cloud*, and is considered to be sampled from the marginal distribution μ^*. Thus, in contrast to the research described above, the points $\{x_j\}$ in this context are *scattered*; i.e., one does not have a choice of stipulating their locations in advance.

Typically, the points $\{x_j\}$ are in a high dimensional *ambient space*, and the classical approximation results are inadequate for practical applications. A very powerful relatively recent idea to work with such problems is the notion of a data-defined manifold. Thus, we assume that the points $\{x_j\}$ lie on a low dimensional manifold embedded in the ambient space. This manifold itself is not known, but some relation on the set $\mathscr{P}$ is assumed to be known, giving rise to a graph structure with vertices on the manifold. Various quantities such as the eigenvalues and eigenfunctions of the Laplace-Beltrami (or a more general elliptic partial differential) operator on the manifold can be approximated well by the corresponding objects for the so called graph Laplacian that can be computed directly from the points $\{x_j\}$ themselves (e.g., [1–3, 23, 29, 30]). It is shown in [21] that a local coordinate chart on such data-defined manifolds can be obtained in terms of the heat kernel on the manifold.

In our theoretical investigations, we will not consider the statistical problem of machine learning, but assume that the marginal distribution μ^* is known. Since the heat kernel can be approximated well using the eigen-decomposition of the graph Laplacian [10], we find it convenient and essential to formulate all our assumptions in this theory only in terms of the measure μ^* on the manifold and the heat kernel. In particular, we do not consider the question of estimating the eigenvalues and eigenfunctions of this kernel, but assume that they are given.

In [14, 15], we have studied the existence of quadrature formulas exact for certain eigenspaces in this context. They play a critical role in approximation theory based on these eigenspaces; e.g., [13, 24, 25]. Although our proofs of the existence of quadrature formulas based on scattered data so far require the notion of gradient on a manifold, the approximation theory itself has been developed in a more general context of locally compact, quasi-metric, measure spaces.

In this paper, we will prove certain results analogous to those in [6, 7] in the context of locally compact, quasi-metric, measure spaces. In order to do so, we need to generalize the notion of an approximate QMC design to include non-equal weights, satisfying certain regularity conditions. We will show that an estimate of the form (3) holds if and only if it holds for what we call diffusion polynomials of certain degree. Conversely, if this estimate holds with non-negative weights, then the assumption of regularity is automatically satisfied. We will also point out the connection between such quadratures and the so called covering radius of the points on which they are based. Our results include their counterparts in [6, 7], except that we deal with slightly larger smoothness classes. We will discuss a construction of the approximate quadratures that yield a bound of the form (3), without referring to the eigen-decomposition itself.

We describe our general set up in Sect. 2, and discuss the main results in Sect. 3. The proofs are given in Sect. 5. Section 4 reviews some preparatory results required in the proofs.

2 The Set-Up

In this section, we describe our general set up. In Sect. 2.1, we introduce the notion of a data-defined space. In Sect. 2.2, we review some measure theoretic concepts. The smoothness classes in which we study the errors in approximate integration are defined in Sect. 2.3.

2.1 The Quasi-Metric Measure Space

Let $\mathbb{X}$ be a non-empty set. A *quasi-metric* on $\mathbb{X}$ is a function $\rho : \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ that satisfies the following properties: For all $x, y, z \in \mathbb{X}$,

1. $\rho(x,y) \ge 0$,
2. $\rho(x,y) = 0$ if and only if $x = y$,
3. $\rho(x,y) = \rho(y,x)$,
4. there exists a constant $\kappa_1 \ge 1$ such that

$$\rho(x,y) \le \kappa_1\{\rho(x,z) + \rho(z,y)\}, \qquad x, y, z \in \mathbb{X}. \tag{4}$$

For example, the geodesic distance on a Riemannian manifold $\mathbb{X}$ is a quasi-metric.

The quasi-metric ρ gives rise to a topology on $\mathbb{X}$, with

$$\{y \in \mathbb{X} : \rho(x,y) < r\}, \qquad x \in \mathbb{X},\ r > 0.$$

being a basis for the topology. In the sequel, we will write

$$\mathbb{B}(x,r) = \{y \in \mathbb{X} : \rho(x,y) \le r\},\ \Delta(x,r) = \mathbb{X}\setminus\mathbb{B}(x,r), \qquad x \in \mathbb{X},\ r > 0.$$

In remainder of this paper, let μ^* be a fixed probability measure on $\mathbb{X}$. We fix a non-decreasing sequence $\{\lambda_k\}_{k=0}^{\infty}$ of nonnegative numbers such that $\lambda_0 = 0$, and $\lambda_k \uparrow \infty$ as $k \to \infty$. Also, we fix a system of continuous, bounded, and integrable functions $\{\phi_k\}_{k=0}^{\infty}$, orthonormal with respect to μ^*; namely, for all nonnegative integers j, k,

$$\int_{\mathbb{X}} \phi_k(x)\phi_j(x)d\mu^*(x) = \begin{cases} 1, \text{ if } j = k, \\ 0, \text{ otherwise.}\end{cases} \tag{5}$$

We will assume that $\phi_0(x) = 1$ for all $x \in \mathbb{X}$.

For example, in the case of a compact Riemannian manifold $\mathbb{X}$, we may take the (normalized) volume measure on $\mathbb{X}$ to be μ^*, and take ϕ_k's to be the eigenfunctions of the Laplace–Beltrami operator on $\mathbb{X}$, corresponding to the eigenvalues $-\lambda_k^2$. If a different measure is assumed, then we may need to consider differential operators other than the Laplace–Beltrami operator. Also, it is sometimes not necessary to use the exact eigenvalues of such operators. For example, in the case when $\mathbb{X}$ is the unit sphere embedded in $\mathbb{R}^3$, the eigenvalues of the (negative) Laplace–Beltrami operator are given by $\sqrt{k(k+1)}$. The analysis is sometimes easier if we use k instead. In general, while the exact eigenvalues might be hard to compute, an asymptotic expression is often available. While these considerations motivate our definitions, we observe that we are considering a very general scenario with quasi-metric measure spaces, where differential operators are not defined. Nevertheless, we will refer to each ϕ_k as an *eigenfunction* corresponding to the *eigenvalue* λ_k, even though they are not necessarily obtained from an eigen-decomposition of any predefined differential or integral operator.

In our context, the role of polynomials will be played by diffusion polynomials, which are finite linear combinations of $\{\phi_j\}$. In particular, an element of

$$\Pi_n := \mathsf{span}\{\phi_j : \lambda_j < n\}$$

will be called a diffusion polynomial of degree $< n$.

For reasons explained in the introduction, we will formulate our assumptions in terms of a formal heat kernel. The *heat kernel* on $\mathbb{X}$ is defined formally by

$$K_t(x,y) = \sum_{k=0}^{\infty} \exp(-\lambda_k^2 t)\phi_k(x)\phi_k(y), \qquad x,y \in \mathbb{X},\ t > 0. \tag{6}$$

Although K_t satisfies the semigroup property, and in light of the fact that $\lambda_0 = 0$, $\phi_0(x) \equiv 1$, we have formally

$$\int_{\mathbb{X}} K_t(x,y)d\mu^*(y) = 1, \qquad x \in \mathbb{X}, \tag{7}$$

yet K_t may not be the heat kernel in the classical sense. In particular, we need not assume K_t to be nonnegative.

Definition 1 The system $\Xi = (\mathbb{X}, \rho, \mu^*, \{\lambda_k\}_{k=0}^{\infty}, \{\phi_k\}_{k=0}^{\infty}))$ is called a **data-defined space** if each of the following conditions are satisfied.

1. For each $x \in \mathbb{X}$ and $r > 0$, the ball $\mathbb{B}(x,r)$ is compact.
2. There exist $q > 0$ and $\kappa_2 > 0$ such that the following power growth bound condition holds:

$$\mu^*(\mathbb{B}(x,r)) = \mu^*(\{y \in \mathbb{X} : \rho(x,y) < r\}) \le \kappa_2 r^q, \qquad x \in \mathbb{X},\ r > 0. \tag{8}$$

3. The series defining $K_t(x, y)$ converges for every $t \in (0, 1]$ and $x, y \in \mathbb{X}$. Further, with q as above, there exist $\kappa_3, \kappa_4 > 0$ such that the following Gaussian upper bound holds:

$$|K_t(x, y)| \le \kappa_3 t^{-q/2} \exp\left(-\kappa_4 \frac{\rho(x, y)^2}{t}\right), \qquad x, y \in \mathbb{X},\ 0 < t \le 1. \tag{9}$$

There is a great deal of discussion in the literature on the validity of the conditions in the above definition and their relationship with many other objects related to the quasi-metric space in question, (cf. for example, [11, 17–19]). In particular, it is shown in [11, Section 5.5] that all the conditions defining a data-defined space are satisfied in the case of any complete, connected Riemannian manifold with non-negative Ricci curvature. It is shown in [22] that our assumption on the heat kernel is valid in the case when $\mathbb{X}$ is a complete Riemannian manifold with bounded geometry, and $\{-\lambda_j^2\}$, respectively $\{\phi_j\}$, are eigenvalues, respectively eigenfunctions, for a uniformly elliptic second order differential operator satisfying certain technical conditions.

The bounds on the heat kernel are closely connected with the measures of the balls $\mathbb{B}(x, r)$. For example, using (9), Lemma 1 below, and the fact that

$$\int_{\mathbb{X}} |K_t(x, y)| d\mu^*(y) \ge \int_{\mathbb{X}} K_t(x, y) d\mu^*(y) = 1, \qquad x \in \mathbb{X},$$

it is not difficult to deduce as in [18] that

$$\mu^*(\mathbb{B}(x, r)) \ge c r^q, \qquad 0 < r \le 1. \tag{10}$$

In many of the examples cited above, the kernel K_t also satisfies a lower bound to match the upper bound in (9). In this case, Grigoryán [18] has also shown that (8) is satisfied for $0 < r < 1$.

We remark that the estimates (8) and (10) together imply that μ^* satisfies the homogeneity condition

$$\mu^*(\mathbb{B}(x, R)) \le c_1 (R/r)^q \mu^*(\mathbb{B}(x, r)), \qquad x \in \mathbb{X},\ r \in (0, 1],\ R > 0, \tag{11}$$

where $c_1 > 0$ is a suitable constant.

In the sequel, we assume that Ξ is a data-defined space, and make the following convention.

Constant Convention *In the sequel, the symbols $c, c_1, \cdots$ will denote positive constants depending only on $\mathbb{X}$, ρ, μ^*, $\kappa_1, \cdots, \kappa_5$, and other similar fixed quantities such as the parameters denoting the various spaces. They will not depend upon the systems $\{\phi_k\}$, $\{\lambda_k\}$ by themselves, except through the quantities mentioned above. On occasions when we need to have the constants depend upon additional variables, these will be listed explicitly. Their values may be different at different occurrences, even within a single formula. The notation $A \sim B$ will mean $c_1 A \le B \le c_2 A$.*

2.2 Measures

In this paper, it is necessary to consider a sequence of sets $\mathscr{C}_n = \{x_{1,n}, \cdots, x_{M_n,n}\}$, and the quadrature weights $w_{k,n}$, leading to sums of the form

$$\sum_{\substack{1\le k\le M_n\\ x_{k,n}\in B}} w_{k,n} f(x_{k,n}),$$

where $B \subseteq \mathbb{X}$. The precise locations of the points of $\mathscr{C}_n$, or the numbers $w_{k,n}$, or even the numbers M_n will play no role in our theoretical development. Therefore, we find it convenient to use a sequence of measures to abbreviate sums like the above. Accordingly, In this subsection, we review some measure theoretical notation and definitions.

If ν is a (signed) measure defined on a sigma algebra $\mathfrak{M}$ of $\mathbb{X}$, its total variation measure $|\nu|$ is defined by

$$|\nu|(B) = \sup \sum_{k=1}^{\infty} |\nu(U_k)|,$$

where the supremum is taken over all countable partitions $\{U_k\} \subseteq \mathfrak{M}$ of B. Here, the quantity $|\nu|(\mathbb{X})$ is called the total variation of ν. If ν is a signed measure, then its total variation is always finite. If ν is a positive measure, it is said to be of bounded variation if its total variation is finite. The measure ν is said to be complete if for any $B \in \mathfrak{M}$ with $|\nu|(B) = 0$ and any subset $A \subseteq B$, $A \in \mathfrak{M}$ and $|\nu|(A) = 0$. Since any measure can be extended to a complete measure by suitably enlarging the underlying sigma algebra, we will assume in the sequel that all the measures to be introduced in this paper are complete.

If $\mathscr{C} \subseteq \mathbb{X}$ is a finite set, the measure ν that associates with each $x \in \mathscr{C}$ the mass w_x, is defined by

$$\nu(B) = \sum_{x\in B} w_x.$$

for subsets $B \subseteq \mathbb{X}$. Obviously, the total variation of the measure ν is given by

$$|\nu|(B) = \sum_{x\in B} |w_x|, \qquad B \subseteq \mathbb{X}.$$

If $f : \mathscr{C} \to \mathbb{C}$, then for $B \subseteq \mathbb{X}$,

$$\int_B f d\nu = \sum_{x\in\mathscr{C}\cap B} w_x f(x).$$

Thus, in the example at the start of this subsection, if ν_n is the measure that associates the mass $w_{k,n}$ with $x_{k,n}$ for $k = 1, \cdots, M_n$, then we have a concise notation

$$\sum_{\substack{1 \le k \le M_n \\ x_{k,n} \in B}} w_{k,n} f(x_{k,n}) = \int_B f d\nu_n \left(= \int_B f(x) d\nu_n(x) \right).$$

In the sequel, we will assume that every measure introduced in this paper is a complete, sigma finite, Borel measure; i.e., the sigma algebra $\mathfrak{M}$ on which it is defined contains all Borel subsets of $\mathbb{X}$. In the rest of this paper, rather than stating that ν is defined on $\mathfrak{M}$, we will follow the usual convention of referring to members of $\mathfrak{M}$ as ν-measurable sets without mentioning the sigma algebra explicitly.

2.3 Smoothness Classes

If $B \subseteq \mathbb{X}$ is ν-measurable, and $f : B \to \mathbb{C}$ is a ν-measurable function, we will write

$$\|f\|_{\nu;B,p} := \begin{cases} \left\{ \int_B |f(x)|^p d|\nu|(x) \right\}^{1/p}, & \text{if } 1 \le p < \infty, \\ |\nu| - \operatorname*{ess\,sup}_{x \in B} |f(x)|, & \text{if } p = \infty. \end{cases}$$

We will write $L^p(\nu; B)$ to denote the class of all ν-measurable functions f for which $\|f\|_{\nu;B,p} < \infty$, where two functions are considered equal if they are equal $|\nu|$-almost everywhere. We will omit the mention of ν if $\nu = \mu^*$ and that of B if $B = \mathbb{X}$. Thus, $L^p = L^p(\mu^*; \mathbb{X})$. The L^p closure of the set of all diffusion polynomials will be denoted by X^p. For $1 \le p \le \infty$, we define $p' = p/(p-1)$ with the usual understanding that $1' = \infty$, $\infty' = 1$.

In the absence of a differentiability structure on $\mathbb{X}$, perhaps, the easiest way to define a Bessel potential space is the following. If $f_1 \in L^p, f_2 \in L^{p'}$ then

$$\langle f_1, f_2 \rangle := \int_{\mathbb{X}} f_1(x) f_2(x) d\mu^*(x).$$

In particular, we write

$$\hat{f}(k) = \langle f, \phi_k \rangle, \qquad k = 0, 1, \cdots.$$

For $r > 0$, the pseudo-differential operator Δ^r is defined formally by

$$\widehat{\Delta^r f}(k) = (\lambda_k + 1)^r \hat{f}(k), \qquad k = 0, 1, \cdots.$$

The class of all $f \in X^p$ for which there exists $\Delta^r f \in X^p$ with $\widehat{\Delta^r f}(k)$ as above is denoted by W^p_r. This definition is sometimes abbreviated in the form

$$W^p_r = \left\{ f \in X^p : \left\| \sum_k (\lambda_k + 1)^r \hat{f}(k) \phi_k \right\|_p < \infty \right\}.$$

However, since the series expansion need not converge in the L^p norm, we prefer the distributional definition as we have given.

While the papers [5–9] all deal with the spaces which we have denoted by W^p_r, we find it easier to consider a larger class, H^p_γ, defined as follows. If $f \in X^p$, $r > 0$, we define a K-functional for $\delta > 0$ by

$$\omega_r(p; f, \delta) := \inf\{\|f - f_1\|_p + \delta^r \|\Delta^r f_1\|_p \ : \ f_1 \in W^p_r\}. \tag{12}$$

If $\gamma > 0$, we choose $r > \gamma$, and define the smoothness class H^p_γ to be the class of all $f \in X^p$ such that

$$\|f\|_{H^p_\gamma} := \|f\|_p + \sup_{\delta \in (0,1]} \frac{\omega_r(p; f, \delta)}{\delta^\gamma} < \infty. \tag{13}$$

For example, if $\mathbb{X} = \mathbb{R}/(2\pi\mathbb{Z})$, μ^* is the arc measure on $\mathbb{X}$, $\{\phi_k\}$'s are the trigonometric monomials $\{1, \cos(k\circ), \sin(k\circ)\}_{k=1}^\infty$, and the eigenvalue λ_k corresponding to $\cos(k\circ)$, $\sin(k\circ)$ is $|k|$, then the class W^∞_2 is the class of all twice continuously differentiable functions, while the class H^∞_2 includes $f(x) = |\sin x|$. The importance of the spaces H^p_γ is well known in approximation theory [12]. We now describe the connection with approximation theory in our context.

If $f \in L^p$, $W \subseteq L^p$, we define

$$\mathsf{dist}(p; f, W) := \inf_{P \in W} \|f - P\|_p.$$

The following theorem is shown in [24, Theorem 2.1] (where a different notation is used).

Proposition 1 *Let $f \in X^p$. Then*

$$\|f\|_{H^p_\gamma} \sim \|f\|_p + \sup_{n > 0} n^\gamma \mathsf{dist}(p; f, \Pi_n). \tag{14}$$

In particular, different values of $r > \gamma$ give rise to the same smoothness class with equivalent norms (cf. [12]). We note that $W^p_r \subset H^p_r$ for every $r > 0$.

3 Main Results

In this paper, we wish to state our theorems without the requirement that the quadrature formulas have positive weights, let alone equal weights. A substitute for this requirement is the notion of regularity (sometimes called continuity) condition. The space of all signed (or positive), complete, sigma finite, Borel measures on $\mathbb{X}$ will be denoted by $\mathscr{M}$.

Definition 2 Let $d > 0$. A measure $\nu \in \mathscr{M}$ will be called ***d*-regular** if

$$|\nu|(\mathbb{B}(x,d)) \le cd^q, \qquad x \in \mathbb{X}. \tag{15}$$

The infimum of all constants c which work in (15) will be denoted by $|\!|\!|\nu|\!|\!|_{R,d}$, and the class of all d-regular measures will be denoted by $\mathscr{R}_d$.

For example, μ^* itself is in $\mathscr{R}_d$ with $|\!|\!|\mu^*|\!|\!|_{R,d} \le \kappa_2$ for *every* $d > 0$ (cf. (8)). If $\mathscr{C} \subset \mathbb{X}$, we define the mesh norm $\delta(\mathscr{C})$ (also known as fill distance, covering radius, density content, etc.) and minimal separation $\eta(\mathscr{C})$ by

$$\delta(\mathscr{C}) = \sup_{x\in\mathbb{X}} \inf_{y\in\mathscr{C}} \rho(x,y), \qquad \eta(\mathscr{C}) = \inf_{x,y\in\mathscr{C},\ x\neq y} \rho(x,y). \tag{16}$$

It is easy to verify that if $\mathscr{C}$ is finite, the measure that associates the mass $\eta(\mathscr{C})^q$ with each point of $\mathscr{C}$ is $\eta(\mathscr{C})$-regular [25, Lemma 5.3].

Definition 3 Let $n \ge 1$. A measure $\nu \in \mathscr{M}$ is called a **quadrature measure of order** n if

$$\int_{\mathbb{X}} P d\nu = \int_{\mathbb{X}} P d\mu^*, \qquad P \in \Pi_n. \tag{17}$$

An **MZ (Marcinkiewicz-Zygmund) quadrature measure** of order n is a quadrature measure ν of order n for which $|\!|\!|\nu|\!|\!|_{R,1/n} < \infty$.

Our notion of approximate quadrature measures is formulated in the following definition.

Definition 4 Let $\aleph = \{\nu_n\}_{n=1}^\infty \subset \mathscr{M}$, $\gamma > 0$, $1 \le p \le \infty$. We say that $\aleph$ is a sequence of **approximate quadrature measures of class** $\mathscr{A}(\gamma, p)$ if each of the following conditions hold.

1.

$$\sup_{n\ge 1} |\nu_n|(\mathbb{X}) < \infty. \tag{18}$$

2.

$$\sup_{n\ge 1} |\!|\!|\nu_n|\!|\!|_{R,1/n} < \infty. \tag{19}$$

3. For $n \ge 1$,

$$\left|\int_{\mathbb{X}} Pd\mu^* - \int_{\mathbb{X}} Pd\nu_n\right| \le A\frac{\|P\|_{H^p_\gamma}}{n^\gamma}, \qquad P \in \Pi_n, \tag{20}$$

for a positive constant A independent of P, but possibly dependent on $\aleph$ in addition to the other fixed parameters.

With an abuse of terminology, we will often say that ν is an approximate quadrature measure of order n (and write $\nu \in \mathscr{A}(\gamma, p, n)$) to mean tacitly that it is a member of a sequence $\aleph$ of approximate quadrature measures for which (20) holds.

Clearly, if each ν_n is a quadrature measure of order n, then (20) is satisfied for every $\gamma > 0$ and $1 \le p \le \infty$. In the case of a compact Riemannian manifold satisfying some additional conditions, the existence of quadrature measures based on scattered data that satisfy the other two conditions in the above definition are discussed in [14, 15]. In particular, we have shown in [15] that under certain additional conditions, a sequence $\aleph$, where each ν_n is a positive quadrature measure of order n necessarily satisfies the first two conditions in Definition 4. In Theorem 5 below, we will give the analogue of this result in the present context.

First, we wish to state a theorem reconciling the notion of approximate quadrature measures with the usual notion of worst case error estimates.

Theorem 2 *Let $n \ge 1$, $1 \le p \le \infty$, $\gamma > q/p$, ν be a $1/n$-regular measure satisfying $|\nu|(\mathbb{X}) < \infty$ and (20). Then for every $f \in H^p_\gamma$,*

$$\left|\int_{\mathbb{X}} fd\mu^* - \int_{\mathbb{X}} fd\nu\right| \le c\left(A + \|\!|\nu|\!\|_{R,1/n}^{1/p}(|\nu|(\mathbb{X}))^{1/p'}\right)\frac{\|f\|_{H^p_\gamma}}{n^\gamma}. \tag{21}$$

For example, if $\mathscr{C} \subset \mathbb{X}$ is a finite set with $\eta(\mathscr{C}) \sim |\mathscr{C}|^{-1/q}$, and ν is the measure that associates the mass $|\mathscr{C}|^{-1}$ with each point of $\mathscr{C}$, then our notion of approximate quadrature measures generalizes the notion of approximate QMC designs in [6–8]. Since an MZ quadrature measure of order n on a compact Riemannian manifold is in $\mathscr{A}(\gamma, p, n)$, Theorem 2 generalizes essentially [20, Theorem 5] (for the spaces denoted there by $B^\gamma_{p,\infty}$, which are our H^p_γ) as well as [5, Asssertions (B), (C)] (except that we consider the larger smoothness class than defined directly with the Bessel potentials). We note finally that together with Proposition 3, Theorem 2 implies that if $\nu \in \mathscr{A}(\gamma, p, n)$ then $\nu \in \mathscr{A}(\gamma, p, \alpha n)$ for any positive $\alpha > 0$ (although the various constants will then depend upon α).

Next, we demonstrate in Theorem 3 below that a sequence of approximate quadrature measures can be constructed as solutions of certain optimization problems under certain additional conditions. These optimization problems involve certain kernels of the form $G(x, y) = \sum_k b(\lambda_k)\phi_k(x)\phi_k(y)$, where (intuitively) $b(\lambda_k) \sim (\lambda_k + 1)^{-\beta}$ for some β. In the case of integrating functions in L^2, only the order of magnitude estimates on the coefficients $b(\lambda_k)$ play a role. In the other spaces, this is not sufficient because of an absence of the Parseval identity. On the other hand, restricting ourselves to Bessel potentials is not always an option in the case of the data-defined spaces; there is generally no closed form formula for these. A middle ground is provided by the following definition [25, Definition 2.3].

Definition 5 Let $\beta \in \mathbb{R}$. A function $b : \mathbb{R} \to \mathbb{R}$ will be called a mask of type β if b is an even, S times continuously differentiable function such that for $t > 0$, $b(t) = (1+t)^{-\beta}F_b(\log t)$ for some $F_b : \mathbb{R} \to \mathbb{R}$ such that $|F_b^{(k)}(t)| \le c(b)$, $t \in \mathbb{R}$, $k = 0, 1, \cdots, S$, and $F_b(t) \ge c_1(b)$, $t \in \mathbb{R}$. A function $G : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ will be called a kernel of type β if it admits a formal expansion $G(x, y) = \sum_{j=0}^{\infty} b(\lambda_j)\phi_j(x)\phi_j(y)$ for some mask b of type $\beta > 0$. If we wish to specify the connection between G and b, we will write $G(b; x, y)$ in place of G.

The definition of a mask of type β can be relaxed somewhat, for example, the various bounds on F_b and its derivatives may only be assumed for sufficiently large values of $|t|$ rather than for all $t \in \mathbb{R}$. If this is the case, one can construct a new kernel by adding a suitable diffusion polynomial (of a fixed degree) to G, as is customary in the theory of radial basis functions, and obtain a kernel whose mask satisfies the definition given above. This does not add any new feature to our theory. Therefore, we assume the more restrictive definition as given above.

Theorem 3 *Let $1 \le p \le \infty$, $\beta > q/p$, G be a kernel of type β in the sense of Definition 5. For a measure ν, we denote*

$$M_p(\nu) = \left\| \int_{\mathbb{X}} G(x, \circ)d\nu(x) - \int_{\mathbb{X}} G(x, \circ)d\mu^*(x) \right\|_{p'}. \tag{22}$$

Let $n > 0$, K be any compact subset of measures such that $\sup_{\nu\in K} |\nu|(\mathbb{X}) \le c$ *and* $\sup_{\nu\in K} |\!|\!|\nu|\!|\!|_{R,2^{-n}} \le c$. *If there exists a quadrature measure ν^* of order 2^n in K, and $\nu^\# \in K$ satisfies*

$$M_p(\nu^\#) \le c \inf_{\nu\in K} M_p(\nu). \tag{23}$$

Then $\nu^\#$ satisfies (20) for every γ, $0 < \gamma < \beta$, and in particular, $\nu^\# \in \mathscr{A}(\gamma, p, 2^n)$ for each such γ.

The main purpose of Theorem 3 is to suggest a way to construct an approximate quadrature measure $\nu^\#$ by solving a minimization problem involving different compact sets as appropriate for the applications. We illustrate by a few examples.

In some applications, the interest is in the choice of the points, stipulating the quadrature weights. For example, in the context of the sphere, the existence of points yielding equal weights quadrature is now known [4]. So, one may stipulate equal weights and seek an explicit construction for the points to yield equal weights approximate quadrature measures as in [6]. In this case, K can be chosen to be the set of all equal weight measures supported at points on $\mathbb{S}^q$, the compactness of this set following from that of the tensor product of the spheres. In the context of machine learning, the points cannot be chosen and the interest is in finding the weights of an approximate quadrature measure. The existence of positive quadrature formulas (necessarily satisfying the required regularity conditions) are known in the case of a manifold, subject to certain conditions on the points and the manifold in question [14, 15]. In this case, the set K can be taken to be that of all positive unit measures supported at these points. In general, if $\mathbb{X}$ is compact, Tchakaloff's theorem shows that one could seek to obtain approximate quadrature measures computationally by minimizing the quantity $M_p(\nu)$ over all positive unit measures ν supported on a number of points equal to the dimension of Π_n for each n.

We make some remarks regarding the computational aspects. The problem of finding exact quadrature weights is the problem of solving an under-determined system of equations involving the eigenfunctions. Since these eigenfunctions are themselves known only approximately from the data, it is desirable to work directly with a kernel. In the case of L^2, the minimization problem to find non-negative weights is the problem of minimizing

$$\sum_{j,\ell} w_j w_\ell G^*(x_j, x_\ell), \qquad x_j, x_\ell \in \mathscr{C} \subset \mathscr{P}$$

over all non-negative w_j with $\sum_j w_j = 1$, where

$$G^*(x,y) = \sum_{k=1}^{\infty} b(\lambda_k)^2 \phi_k(x)\phi_k(y).$$

Thus, the optimization problem involves only the training data. In the context of semi-supervised learning, a large point cloud $\mathscr{P}$ is given, but the labels are available only at a much smaller subset $\mathscr{C} \subset \mathscr{P}$. In this case, we need to seek approximate quadrature measures supported only on $\mathscr{C}$, but may use the entire set $\mathscr{P}$ to compute these. Thus, in order to apply Theorem 3, we may choose K to be the set of all measures with total variation ≤ 1, supported on $\mathscr{C}$, and estimate the necessary norm expressions using the entire point cloud $\mathscr{P}$. We observe that we have not stipulated a precise solution of an optimization problem, only a solution in the sense of (23). In the context of data-defined spaces, the data-based kernels themselves are only approximations of the actual kernels, and hence, Theorem 3 provides a theoretical justification for using algorithms to find sub-optimal solutions to the minimization problem in order to find approximate quadrature formulas.

We end this discussion by observing a proposition used in the proof of Theorem 3.

Proposition 2 *Let $1 \le p \le \infty$, $\beta > q/p$, G be a kernel of type β in the sense of Definition 5. If $n > 0$, $\nu^\# \in \mathscr{M}$, and $M_p(\nu^\#) \le \tilde{A}2^{-n\beta}$, then $\nu^\#$ satisfies (20) with 2^n replacing n and $A = c\tilde{A}$.*

Next, we consider the density of the supports of the measures in a sequence of approximate quadrature measures. It is observed in [7, 8] that there is a close connection between approximate QMC designs and the (asymptotically) optimal covering radii of the supports of these designs. The definition in these papers is given in terms of the number of points in the support. Typically, for a QMC design of order n, this number is $\sim n^q$. Therefore, it is easy to interpret this definition in terms of the mesh norm of the support of a QMC design of order n being $\sim 1/n$. Our definition of an approximate quadrature measure sequence does not require the measures involved to be finitely supported. Therefore, the correct analogue of this definition seems to be the assertion that every ball of radius $\sim 1/n$ should intersect the support of an approximate quadrature measure of order n. The following Theorem 4 gives a sharper version of this sentiment.

Theorem 4 *Let $\gamma > 0$, $1 \le p \le \infty$ and $\aleph = \{\nu_n\}$ be a sequence in $\mathscr{A}(\gamma, p)$. Let $\tilde{p} = 1 + q/(\gamma p')$. Then there exists a constant C_1 such that for $n \ge c$,*

$$|\nu_n|(\mathbb{B}(x, C_1/n^{1/\tilde{p}})) \ge c_1 n^{-q/\tilde{p}}, \qquad x \in \mathbb{X}. \tag{24}$$

In particular, if $\aleph$ is a sequence of approximate quadrature measures of class $\mathscr{A}(\gamma, 1)$, then

$$|\nu_n|(\mathbb{B}(x, C_1/n)) \sim c_1 n^{-q}, \qquad x \in \mathbb{X}. \tag{25}$$

The condition (25) ensures that the support of the measure $\nu \in \mathscr{A}(\gamma, 1, n)$ must contain at least cn^q points. In several papers, including [5], this fact was used to show the existence of a function in H^p_γ for which there holds a lower bound corresponding to the upper bound in (20). The construction of the "bad function" in these papers involves the notion of infinitely differentiable functions and their pointwise defined derivatives. Since we are not assuming any differentiability structure on $\mathbb{X}$, so the notion of a C^∞ function in the sense of derivatives is not possible in this context.

Finally, we note in this connection that the papers [5–7] deal exclusively with non-negative weights. The definition of approximate QMC designs in these papers does not require a regularity condition as we have done. We show under some extra conditions that if $\aleph$ is a sequence of positive measures such that for each $n \ge 1$, ν_n satisfies (20) with $p = 1$ and some $\gamma > 0$, then $\aleph \in \mathscr{A}(\gamma, 1)$.

For this purpose, we need to overcome a technical hurdle. In the case of the sphere, the product of two spherical polynomials of degree $< n$ is another spherical polynomial of degree $< 2n$. Although a similar fact is valid in many other manifolds, and has been proved in [15, 16] in the context of eigenfunctions of very general elliptic differential operators on certain manifolds, we need to make an explicit assumption in the context of the present paper, where we do not assume any differentiability structure.

Product Assumption *For $A, N > 0$, let*

$$\epsilon_{A,N} := \sup_{\lambda_j, \lambda_k \le N} \mathsf{dist}(\infty; \phi_j\phi_k, \Pi_{AN}). \tag{26}$$

We assume that there exists $A^* \geq 2$ *with the following property: for* ***every*** $R > 0$, $\lim_{N\to\infty} N^R \epsilon_{A^*,N} = 0$.

In the sequel, for any $H : \mathbb{R} \to \mathbb{R}$, we define formally

$$\Phi_N(H; x, y) := \sum_{j=0}^{\infty} H(\lambda_j/N)\phi_j(x)\phi_j(y), \qquad x, y \in \mathbb{X},\ N > 0. \tag{27}$$

In the remainder of this paper, we will fix an infinitely differentiable, even function $h : \mathbb{R} \to \mathbb{R}$ such that $h(t) = 1$ if $|t| \leq 1/2$, $h(t) = 0$ if $|t| \geq 1$, and h is non-increasing on $[1/2, 1]$. The mention of this function will be usually omitted from the notation; e.g., we write $\Phi_n(x, y)$ in place of $\Phi_n(h; x, y)$.

Theorem 5 *Let* $n > 0$, $1 \leq p \leq \infty$, ν *be a positive measure satisfying* (20) *for some* $\gamma > 0$. *We assume that the product assumption holds, and that in addition the following inequality holds: there exists* $\beta > 0$ *such that*

$$\min_{y \in \mathbb{B}(x, \beta/m)} |\Phi_m(x, y)| \geq c m^q, \qquad x \in \mathbb{X},\ m \geq 1. \tag{28}$$

Then

$$\nu(\mathbb{B}(x, 1/n)) \leq c n^{-q/p}, \qquad x \in \mathbb{X}. \tag{29}$$

In particular, if $p = 1$ *then* $|\nu|(\mathbb{X}) \leq c$, $\nu \in \mathscr{R}_{1/n}$, *and* $|||\nu|||_{R,1/n} \leq c$.

The condition (28) is proved in [15, Lemma 7.3] in the case of compact Riemannian manifolds satisfying a gradient condition on the heat kernel (in particular, the spaces considered in the above cited papers).

4 Preparatory Results

In this section, we collect together some known results. We will supply the proofs for the sake of completeness when they are not too complicated.

4.1 Results on Measures

The following proposition (cf. [15, Proposition 5.6]) reconciles different notions of regularity condition on measures defined in our papers.

Proposition 3 *Let* $d \in (0, 1]$, $\nu \in \mathscr{M}$.

(a) *If* ν *is d-regular, then for each* $r > 0$ *and* $x \in \mathbb{X}$,

$$|\nu|(\mathbb{B}(x, r)) \leq c|||\nu|||_{R,d}\, \mu^*(\mathbb{B}(x, c(r + d))) \leq c_1 |||\nu|||_{R,d}(r + d)^q. \tag{30}$$

Conversely, if for some $A > 0$, $|\nu|(\mathbb{B}(x,r)) \le A(r+d)^q$ or each $r > 0$ and $x \in \mathbb{X}$, then ν is d-regular, and $|\!|\!|\nu|\!|\!|_{R,d} \le 2^q A$.

(b) *For each $\alpha > 0$,*

$$|\!|\!|\nu|\!|\!|_{R,\alpha d} \le c_1(1+1/\alpha)^q |\!|\!|\nu|\!|\!|_{R,d} \le c_1^2(1+1/\alpha)^q(\alpha+1)^q |\!|\!|\nu|\!|\!|_{R,\alpha d}, \tag{31}$$

where c_1 is the constant appearing in (30)*.*

If $K \subseteq \mathbb{X}$ is a compact subset and $\epsilon > 0$, we will say that a subset $\mathscr{C} \subseteq K$ is ϵ-separated if $\rho(x,y) \ge \epsilon$ for every $x, y \in \mathscr{C}$, $x \ne y$. Since K is compact, there exists a finite, maximal ϵ-separated subset $\{x_1, \cdots, x_M\}$ of K. If $x \in K \setminus \cup_{k=1}^M \mathbb{B}(x_k, \epsilon)$, then $\{x, x_1, \cdots, x_M\}$ is a strictly larger ϵ-separated subset of K. So, $K \subseteq \cup_{k=1}^M \mathbb{B}(x_k, \epsilon)$. Moreover, with κ_1 as in (4), the balls $\mathbb{B}(x_k, \epsilon/(3\kappa_1))$ are mutually disjoint.

Proof of Proposition 3 In the proof of part (a) only, let $\lambda > |\!|\!|\nu|\!|\!|_{R,d}$, $r > 0$, $x \in \mathbb{X}$, and let $\{y_1, \cdots, y_N\}$ be a maximal $2d/3$-separated subset of $\mathbb{B}(x, r+2d/3)$. Then $\mathbb{B}(x,r) \subseteq \mathbb{B}(x, r+2d/3) \subseteq \cup_{j=1}^N \mathbb{B}(y_j, 2d/3)$. So,

$$\begin{aligned}|\nu|(\mathbb{B}(x,r)) \le |\nu|(\mathbb{B}(x, r+2d/3)) &\le \sum_{j=1}^N |\nu|(\mathbb{B}(y_j, 2d/3)) \\ &\le \sum_{j=1}^N |\nu|(\mathbb{B}(y_j, d)) \le \lambda N d^q.\end{aligned}$$

The balls $\mathbb{B}(y_j, d/(3\kappa_1))$ are mutually disjoint, and $\cup_{j=1}^N \mathbb{B}(y_j, d/(3\kappa_1)) \subseteq \mathbb{B}(x, c(r+d))$. In view of (10), $d^q \le c\mu^*(\mathbb{B}(y_j, d/(3\kappa_1)))$ for each j. So,

$$\begin{aligned}|\nu|(\mathbb{B}(x,r)) &\le \lambda N d^q \le c\lambda \sum_{j=1}^N \mu^*(\mathbb{B}(y_j, d/(3\kappa_1))) = c\lambda\mu^*(\cup_{j=1}^N \mathbb{B}(y_j, d/(3\kappa_1))) \\ &\le c\lambda\mu^*(\mathbb{B}(x, c(r+d))).\end{aligned}$$

Since $\lambda > |\!|\!|\nu|\!|\!|_{R,d}$ was arbitrary, this leads to the first inequality in (30). The second inequality follows from (8). The converse statement is obvious. This completes the proof of part (a).

Using (30) with αd in place of r, we see that

$$|\nu|(\mathbb{B}(x, \alpha d)) \le c_1(\alpha+1)^q d^q |\!|\!|\nu|\!|\!|_{R,d} = c_1(1+1/\alpha)^q (\alpha d)^q |\!|\!|\nu|\!|\!|_{R,d}.$$

This implies the first inequality in (31). The second inequality follows from the first, applied with $1/\alpha$ in place of α. □

Next, we prove a lemma (cf. [25, Proposition 5.1]) which captures many details of the proofs in Sect. 4.2.

Lemma 1 *Let $\nu \in \mathscr{R}_d$, $N > 0$. If $g_1 : [0,\infty) \to [0,\infty)$ is a nonincreasing function, then for any $N > 0$, $r > 0$, $x \in \mathbb{X}$,*

$$N^q \int_{\Delta(x,r)} g_1(N\rho(x,y))d|\nu|(y) \le c\frac{2^q(1+(d/r)^q)q}{1-2^{-q}}|||\nu|||_{R,d} \times \int_{rN/2}^{\infty} g_1(u)u^{q-1}du. \tag{32}$$

Proof By replacing ν by $|\nu|/|||\nu|||_{R,d}$, we may assume that ν is positive, and $|||\nu|||_{R,d} = 1$. Moreover, for $r > 0$, $\nu(\mathbb{B}(x,r)) \le c(1+(d/r)^q)r^q$. In this proof only, we will write $\mathbb{A}(x,t) = \{y \in \mathbb{X} \ : \ t < \rho(x,y) \le 2t\}$. We note that $\nu(\mathbb{A}(x,t)) \le c2^q(1+(d/r)^q)t^q$, $t \ge r$, and

$$\int_{2^{R-1}}^{2^R} u^{q-1}du = \frac{1-2^{-q}}{q}2^{Rq}.$$

Since g_1 is nonincreasing, we have

$$\begin{aligned}
\int_{\Delta(x,r)} g_1(N\rho(x,y))d\nu(y) &= \sum_{R=0}^{\infty}\int_{\mathbb{A}(x,2^Rr)} g_1(N\rho(x,y))d\nu(y)\\
&\le \sum_{R=0}^{\infty} g_1(2^RrN)\nu(\mathbb{A}(x,2^Rr)) \le c2^q(1+(d/r)^q)\sum_{R=0}^{\infty} g_1(2^RrN)(2^Rr)^q\\
&\le c\frac{2^q(1+(d/r)^q)q}{1-2^{-q}}r^q\sum_{R=0}^{\infty}\int_{2^{R-1}}^{2^R} g_1(urN)u^{q-1}du\\
&= c\frac{2^q(1+(d/r)^q)q}{1-2^{-q}}r^q\int_{1/2}^{\infty} g_1(urN)u^{q-1}du\\
&= c\frac{2^q(1+(d/r)^q)q}{1-2^{-q}}N^{-q}\int_{rN/2}^{\infty} g_1(v)v^{q-1}dv.
\end{aligned}$$

This proves (32). □

4.2 Results on Kernels

In our theory, a fundamental role is played by the kernels defined formally in (27):

$$\Phi_N(H;x,y) := \sum_{j=0}^{\infty} H(\lambda_j/N)\phi_j(x)\phi_j(y), \qquad x,y \in \mathbb{X},\ N > 0. \tag{33}$$

To describe the properties of this kernel, we introduce the notation

$$|||H|||_S := \max_{0\le k\le S}\max_{x\in\mathbb{R}} |H^{(k)}(x)|.$$

A basic and important property of these kernels is given in the following theorem.

Theorem 6 *Let $S > q$ be an integer, $H : \mathbb{R} \to \mathbb{R}$ be an even, S times continuously differentiable, compactly supported function. Then for every $x, y \in \mathbb{X}$, $N > 0$,*

$$|\Phi_N(H; x, y)| \le \frac{cN^q |||H|||_S}{\max(1, (N\rho(x, y))^S)}. \tag{34}$$

Theorem 6 is proved in [24], and more recently in much greater generality in [27, Theorem 4.3]. In [24], Theorem 6 was proved under the conditions that the so called finite speed of wave propagation holds, and the following *spectral bounds* hold for the so called Christoffel (or spectral) function (defined by the sum expression in (35) below):

$$\sum_{\lambda_j<N} |\phi_j(x)|^2 \le cN^q, \qquad x \in \mathbb{X},\ N > 0. \tag{35}$$

We have proved in [14, Theorem 4.1] that (9) with $y \ne x$ is equivalent to the finite speed of wave propagation. We have also shown in [14, Proposition 4.1] and [25, Lemma 5.2] that (9) with $y = x$ is equivalent to (35).

The following proposition follows easily from Lemma 1 and Theorem 6.

Proposition 4 *Let S, H be as in Theorem 6, $d > 0$, $\nu \in \mathscr{R}_d$, and $x \in \mathbb{X}$.*

(a) *If $r \ge 1/N$, then*

$$\int_{\Delta(x,r)} |\Phi_N(H; x, y)|d|\nu|(y) \le c(1 + (dN)^q)(rN)^{-S+q} |||\nu|||_{R,d} |||H|||_S. \tag{36}$$

(b) *We have*

$$\int_{\mathbb{X}} |\Phi_N(H; x, y)|d|\nu|(y) \le c(1 + (dN)^q)|||\nu|||_{R,d} |||H|||_S, \tag{37}$$

$$\|\Phi_N(H; x, \circ)\|_{\nu;\mathbb{X},p} \le cN^{q/p'}(1 + (dN)^q)^{1/p} |||\nu|||_{R,d}^{1/p} |||H|||_S, \tag{38}$$

and

$$\left\|\int_{\mathbb{X}} |\Phi_N(H; \circ, y)|d|\nu|(y)\right\|_p \le c(1 + (dN)^q)^{1/p'} |||\nu|||_{R,d}^{1/p'} \times (|\nu|(\mathbb{X}))^{1/p} |||H|||_S. \tag{39}$$

Proof Without loss of generality, we assume that ν is a positive measure and assume also the normalizations $|\!|\!|\nu|\!|\!|_{R,d} = |\!|\!||H|\!|\!|_S = 1$. Let $x \in \mathbb{X}$, $N > 0$. For $r \geq 1/N$, $d/r \leq dN$. In view of (34) and (32), we have for $x \in \mathbb{X}$:

$$\begin{aligned}\int_{\Delta(x,r)} |\Phi_N(H; x, y)| d\nu(y) &\leq cN^q \int_{\Delta(x,r)} (N\rho(x, y))^{-S} d\nu(y)\\ &\leq c(1 + (dN)^q) \int_{rN/2}^{\infty} v^{-S+q-1} dv\\ &\leq c(1 + (dN)^q)(rN)^{-S+q}.\end{aligned}$$

This proves (36).

Using (36) with $r = 1/N$, we obtain that

$$\int_{\Delta(x,1/N)} |\Phi_N(H; x, y)| d\nu(y) \leq c(1 + (dN)^q). \tag{40}$$

We observe that in view of (34), and the fact that $\nu(\mathbb{B}(x, 1/N)) \leq c(1/N + d)^q \leq cN^{-q}(1 + (dN)^q)$,

$$\int_{\mathbb{B}(x,1/N)} |\Phi_N(H; x, y)| d\nu(y) \leq cN^q \nu(\mathbb{B}(x, 1/N)) \leq c(1 + (dN)^q).$$

Together with (40), this leads to (37).

The estimate (38) follows from (34) in the case $p = \infty$, and from (37) in the case $p = 1$. For $1 < p < \infty$, it follows from the convexity inequality

$$\| F\|_{\nu;\mathbb{X},p} \leq \| F\|_{\nu;\mathbb{X},\infty}^{1/p'} \| F\|_{\nu;\mathbb{X},1}^{1/p}. \tag{41}$$

The estimate (39) is the same as (37) in the case when $p = \infty$. In addition, using (37) with μ^* in place of ν, $1/N$ in place of d, we obtain

$$\int_{\mathbb{X}} |\Phi_N(H; x, y)| d\mu^*(x) = \int_{\mathbb{X}} |\Phi_N(H; y, x)| d\mu^*(x) \leq c.$$

Therefore,

$$\int_{\mathbb{X}} \int_{\mathbb{X}} |\Phi_N(H; x, y)| d|\nu|(y) d\mu^*(x) = \int_{\mathbb{X}} \int_{\mathbb{X}} |\Phi_N(H; x, y)| d\mu^*(x) d|\nu|(y) \leq c|\nu|(\mathbb{X}).$$

This proves (39) in the case when $p = 1$. The estimate in the general case follows from the cases $p = 1, \infty$ and (41). □

Next, we study some operators based on these kernels. If ν is any measure on $\mathbb{X}$ and $f \in L^p$, we may define formally

$$\sigma_N(H;\nu;f,x) := \int_{\mathbb{X}} f(y)\Phi_N(H;x,y)d\nu(y). \tag{42}$$

The following is an immediate corollary of Proposition 4, used with μ^* in place of ν, $d = 0$.

Corollary 1 *We have*

$$\sup_{x\in\mathbb{X}} \int_{\mathbb{X}} |\Phi_N(H;x,y)|d\mu^*(y) \le c|||H|||_S, \tag{43}$$

and for every $1 \le p \le \infty$ *and* $f \in L^p$,

$$\|\sigma_N(H;\mu^*;f)\|_p \le c|||H|||_S\|f\|_p. \tag{44}$$

We recall that $h : \mathbb{R} \to \mathbb{R}$ denotes a fixed, infinitely differentiable, and even function, nonincreasing on $[0,\infty)$, such that $h(t) = 1$ if $|t| \le 1/2$ and $h(t) = 0$ if $|t| \ge 1$. We omit the mention of h from the notation, and all constants $c, c_1, \cdots$ may depend upon h. As before, we will omit the mention of ν if $\nu = \mu^*$ and that of H if $H = h$. Thus, $\Phi_N(x,y) = \Phi_N(h;x,y)$, and similarly $\sigma_N(f,x) = \sigma_N(h;\mu^*;f,x)$, $\sigma_N(\nu;f,x) = \sigma_N(h;\nu;f,x)$. The slight inconsistency is resolved by the fact that we use μ^*, ν, $\tilde{\nu}$ etc. to denote measures and h, g, b, H, etc. to denote functions. We do not consider this to be a sufficiently important issue to complicate our notations.

The following proposition gives the approximation properties of the kernels, and summarizes some important inequalities in approximation theory in this context. Different parts of this proposition are proved in [24, 25].

Proposition 5 *Let* $1 \le p \le \infty$, $N > 0$, $r > 0$.

(a) *For* $f \in L^p$,

$$\mathsf{dist}(p;f,\Pi_N) \le \|f - \sigma_N(f)\|_p \le c\mathsf{dist}(p;f,\Pi_{N/2}). \tag{45}$$

(b) *If* $f \in W^p_r$, *then*

$$\mathsf{dist}(p;f,\Pi_N) \le \|f - \sigma_N(f)\|_p \le cN^{-r}\|\Delta^r f\|_p. \tag{46}$$

(c) *For* $P \in \Pi_N$,

$$\|\Delta^r P\|_p \le cN^r\|P\|_p. \tag{47}$$

(d) *For* $f \in L^p$,

$$\omega_r(p;f,1/N) \le \|f - \sigma_N(f)\|_p + N^{-r}\|\Delta^r\sigma_N(f)\|_p \le c\omega_r(p;f,1/N). \tag{48}$$

Proof If $P \in \Pi_{N/2}$ is chosen so that $\|f - P\|_p \le 2\mathsf{dist}(p; f, \Pi_{N/2})$, then (44) implies that

$$\|f - \sigma_N(f)\|_p = \|f - P - \sigma_N(f - P)\|_p \le c\|f - P\|_p \le c\mathsf{dist}(p; f, \Pi_{N/2}).$$

This proves part (a).

The parts (b) and (c) are proved in [24, Theorem 6.1].

Next, let f_1 be chosen so that $\|f - f_1\|_p + N^{-r}\|\Delta^r f_1\|_p \le 2\omega_r(p; f, 1/N)$. Then using (44), (46), and (47), we deduce that

$$\begin{aligned}
&\|f - \sigma_N(f)\|_p + N^{-r}\|\Delta^r\sigma_N(f)\|_p \\
&\le \|f - f_1 - \sigma_N(f - f_1)\|_p + \|f_1 - \sigma_N(f_1)\|_p \\
&\qquad\qquad + N^{-r}\left(\|\Delta^r\sigma_N(f - f_1)\|_p + \|\Delta^r\sigma_N(f_1)\|_p\right) \\
&\le c\{\|f - f_1\|_p + N^{-r}\|\Delta^r f_1\|_p + \|\sigma_N(f - f_1)\|_p + N^{-r}\|\sigma_N(\Delta^r f_1)\|_p\} \\
&\le c\{\|f - f_1\|_p + N^{-r}\|\Delta^r f_1\|_p\} \le c\omega_r(p; f, 1/N).
\end{aligned}$$

This proves (48). □

We note next a corollary of this proposition.

Corollary 2 *Let $r > \gamma > 0$, $\delta \in (0, 1]$, $1 \le p \le \infty$, $f \in X^p$, and $n \ge 1$. Then*

$$\omega_r(p; \sigma_n(f), \delta) \le c\omega_r(p; f, \delta) \qquad \|\sigma_n(f)\|_{H^p_\gamma} \le c\|f\|_{H^p_\gamma}. \tag{49}$$

Proof Let $N \ge 1$ be chosen such that $1/(2N) < \delta \le 1/N$. A comparison of the Fourier coefficients shows that

$$\sigma_N(\sigma_n(f)) = \sigma_n(\sigma_N(f)), \qquad \Delta^r(\sigma_N(\sigma_n(f))) = \sigma_n(\Delta^r(\sigma_N(f))).$$

Consequently, using (48), we conclude that

$$\begin{aligned}
\omega_r(p; \sigma_n(f), \delta) &\le \omega_r(p; \sigma_n(f), 1/N) \\
&\le \|\sigma_n(f) - \sigma_N(\sigma_n(f))\|_p + \frac{1}{N^r}\|\Delta^r(\sigma_N(\sigma_n(f)))\|_p \\
&= \|\sigma_n(f) - \sigma_n(\sigma_N(f))\|_p + \frac{1}{N^r}\|\sigma_n(\Delta^r(\sigma_N(f)))\|_p \\
&\le c\{\|f - \sigma_N(f)\|_p + \frac{1}{N^r}\|\Delta^r(\sigma_N(f))\|_p\} \\
&\le c\omega_r(p; f, 1/N) \le c_1\omega_r(p; f, 1/(2N)) \le c_1\omega_r(p; f, \delta).
\end{aligned}$$

This proves the first inequality in (49). The second inequality is now immediate from the definitions. □

Next, we state another fundamental result, that characterizes the space H^p_γ in terms of a series expansion of the functions. In the sequel, we will write for $f \in X^1 \cap X^\infty$, $x \in \mathbb{X}$,

$$\tau_j(f,x) = \begin{cases} \sigma_1(f,x), & \text{if } j = 0, \\ \sigma_{2^j}(f,x) - \sigma_{2^{j-1}}(f,x), & \text{if } j = 1, 2, \cdots . \end{cases}$$

The following lemma summarizes some relevant properties of these operators.

Lemma 2 *Let* $1 \le p \le \infty, f \in X^p$.

(a) *We have*

$$f = \sum_{j=0}^{\infty} \tau_j(f), \tag{50}$$

with convergence in the sense of L^p.

(b) *For each* $j = 2, 3, \cdots$,

$$\|\tau_j(f)\|_p \le c\mathsf{dist}(p; f, \Pi_{2^{j-2}}) \le c \sum_{k=j-1}^{\infty} \|\tau_k(f)\|_p. \tag{51}$$

In particular, if $\gamma > 0$ *and* $f \in H^p_\gamma$, *then*

$$\|f\|_p + \sup_{j\ge 0} 2^{j\gamma} \|\tau_j(f)\|_p \sim \|f\|_{H^p_\gamma}. \tag{52}$$

(c) *If* $j \ge 2$, $d > 0$, *and* $\nu \in \mathscr{R}_d$ *then*

$$\|\tau_j(f)\|_{\nu;1} \le (1 + (2^j d)^q)^{1/p} |\!|\!|\nu|\!|\!|_{R,d}^{1/p} (|\nu|(\mathbb{X}))^{1/p'} \|\tau_j(f)\|_p. \tag{53}$$

Proof Part (a) is an immediate consequence of (45). Since $\Pi_{2^{j-2}} \subset \Pi_{2^{j-1}}$, (45) implies

$$\|\tau_j(f)\|_p \le \|f - \sigma_{2^j}(f)\|_p + \|f - \sigma_{2^{j-1}}(f)\|_p \le c\mathsf{dist}(p; f, \Pi_{2^{j-2}}).$$

This proves the first estimate in (51). The second follows from (50). The estimate (52) can be derived easily using (51) and Proposition 1. This completes the proof of part (b).

Next, we prove part (c). In this proof only, let

$$\tilde{G}(t) = h(t/2) - h(4t), \qquad g(t) = h(t) - h(2t).$$

Then g is supported on $[1/4, 1]$, while

$$\tilde{G}(t) = \begin{cases} 0, & \text{if } 0 \le t \le 1/8, \\ 1, & \text{if } 1/4 \le t \le 1, \\ 0, & \text{if } t \ge 2. \end{cases}$$

Therefore, it is easy to verify that $\tilde{G}(t)g(t) = g(t)$ for all t, $\tau_j(f) = \sigma_{2^j}(g;f)$, and hence, for all $f \in L^1$, $x \in \mathbb{X}$,

$$\tau_j(f, x) = \int_{\mathbb{X}} \tau_j(f, y)\Phi_{2^j}(\tilde{G}; x, y)d\mu^*(y).$$

Using Hölder inequality followed by (39) with p' in place of p and $\tilde{G}$ in place of H, we obtain that

$$\begin{aligned}\int_{\mathbb{X}} |\tau_j(f, x)|d|\nu|(x) &\le \int_{\mathbb{X}}\int_{\mathbb{X}} |\Phi_{2^j}(\tilde{G}; x, y)|d|\nu|(x)||\tau_j(f, y)|d\mu^*(y) \\ &\le \left\| \int_{\mathbb{X}} |\Phi_{2^j}(\tilde{G}; x, \circ)|d|\nu|(x) \right\|_{p'} \|\tau_j(f)\|_p \\ &\le c\{(1 + (d2^j)^q)\}^{1/p} |\!|\!|\nu|\!|\!|_{R,d}^{1/p} (|\nu|(\mathbb{X}))^{1/p'} \|\tau_j(f)\|_p.\end{aligned}$$

This proves (53). □

We will use the following corollary of this lemma in our proofs.

Corollary 3 *If $n \ge 1$, $0 < \gamma < r$, $P \in \Pi_n$, then*

$$\sup_{\delta\in(0,1]} \frac{\omega_r(p; P, \delta)}{\delta^r}. \sim \|\Delta^r P\|_p \le cn^{r-\gamma}\|P\|_{H^p_\gamma}. \tag{54}$$

Further,

$$\|P\|_{H^p_\gamma} \le c\{\|P\|_p + \|\Delta^\gamma P\|_p\} \le cn^\gamma \|P\|_p. \tag{55}$$

Proof In view of the fact that $\sigma_{2n}(P) = P$, we conclude from (48) used with $2n$ in place of N that

$$\frac{1}{(2n)^r}\|\Delta^r P\|_p = \|P - \sigma_{2n}(P)\|_p + \frac{1}{(2n)^r}\|\Delta^r \sigma_{2n}(P)\|_p \le c\omega_r(p; P, 1/(2n)).$$

This shows that

$$\|\Delta^r P\|_p \le c \sup_{\delta\in(0,1]} \frac{\omega_r(p; P, \delta)}{\delta^r}.$$

The estimate in (54) in the other direction follows from the definition of $\omega_r(p; P, \delta)$.

Let m be an integer, $2^m \le n < 2^{m+1}$. Since the expansion for P as given in (50) is only a finite sum, we see that

$$\|\Delta^r P\|_p = \left\|\sum_j \Delta^r \tau_j(P)\right\|_p = \left\|\sum_{j=0}^{m+2} \Delta^r \tau_j(P)\right\|_p \le \sum_{j=0}^{m+2} \|\Delta^r \tau_j(P)\|_p.$$

Hence, using (47) and (52), we deduce that

$$\|\Delta^r P\|_p \le c \sum_{j=0}^{m+2} 2^{j(r-\gamma)} 2^{j\gamma} \|\tau_j(P)\|_p \le c 2^{m(r-\gamma)} \|P\|_{H^p_\gamma}.$$

This implies the last estimate in (54).

If $N \ge n$ then $\mathsf{dist}(p, P, \Pi_N) = 0$. If $N < n$, then the estimate (46) yields

$$\mathsf{dist}(p, P, \Pi_N) \le c N^{-\gamma} \|\Delta^\gamma P\|.$$

Hence, Proposition 1 shows that

$$\|P\|_{H^p_\gamma} \sim \|P\|_p + \sup_{N \ge 1} N^\gamma \mathsf{dist}(p, P, \Pi_N) \le c\{\|P\|_p + \|\Delta^\gamma P\|\}.$$

This proves the first estimate in (55); the second follows from (47). □

Next, we recall yet another preparatory lemma. The following lemma is proved in [26, Lemma 5.4]. (In this lemma, the statement (57) is stated only for $p = \infty$, but the statement below follows since μ^* is a probability measure.)

Lemma 3 *Let $N \ge 1$, $P \in \Pi_N$, $0 < p_1 \le p_2 \le \infty$. Then*

$$\|P\|_{p_2} \le c N^{q(1/p_1 - 1/p_2)} \|P\|_{p_1}. \tag{56}$$

Further, let the product assumption hold, $P_1, P_2 \in \Pi_N$, $1 \le p, p_1, p_2 \le \infty$, and $R > 0$ be arbitrary. Then there exists $Q \in \Pi_{A^*N}$ such that

$$\|P_1 P_2 - Q\|_p \le c(R) N^{-R} \|P_1\|_{p_1} \|P_2\|_{p_2}. \tag{57}$$

The following embedding theorem is a simple consequence of the results stated so far.

Lemma 4

(a) *Let $1 \le p_1 < p_2 \le \infty$, $\gamma > q(1/p_1 - 1/p_2)$, $f \in H^{p_1}_\gamma$. Then*

$$\|f\|_{H^{p_2}_{\gamma - q(1/p_1 - 1/p_2)}} \le c \|f\|_{H^{p_1}_\gamma}. \tag{58}$$

(b) *Let* $1 \le p < \infty$, $\gamma > q/p$, *and* $f \in H^p_\gamma$. *Then* $f \in H^\infty_{\gamma - q/p}$ *(*$f \in X^\infty$ *in particular), and*

$$\|f\|_{H^\infty_{\gamma-q/p}} \le c\|f\|_{H^p_\gamma}. \tag{59}$$

Proof In this proof only, we will write $\alpha = q(1/p_1 - 1/p_2)$. Let $n \ge 0, f \in H^{p_1}_\gamma$, and $r > \gamma$. Without loss of generality, we may assume that $\|f\|_{H^{p_1}_\gamma} = 1$. In view of (48),

$$\frac{1}{2^{nr}}\|\Delta^r \sigma_{2^n}(f)\|_{p_1} \le c2^{-n\gamma}.$$

Since $\Delta^r \sigma_{2^n}(f) \in \Pi_{2^n}$, Lemma 3 shows that

$$\frac{1}{2^{nr}}\|\Delta^r \sigma_{2^n}(f)\|_{p_2} \le \frac{2^{n\alpha}}{2^{nr}}\|\Delta^r \sigma_{2^n}(f)\|_{p_1} \le c2^{-n(\gamma-\alpha)}. \tag{60}$$

Further, since each $\tau_j(f) \in \Pi_{2^j}$, we deduce from (56), (52), and the fact that $\gamma > \alpha$, that

$$\sum_{j=n+1}^{\infty} \|\tau_j(f)\|_{p_2} \le c \sum_{j=n+1}^{\infty} 2^{j\alpha}\|\tau_j(f)\|_{p_1} \le c \sum_{j=n+1}^{\infty} 2^{-j(\gamma-\alpha)} = c2^{-n(\gamma-\alpha)}. \tag{61}$$

Consequently, the series

$$\sigma_{2^n}(f) + \sum_{j=n+1}^{\infty} \tau_j(f)$$

converges in L^{p_2}, necessarily to f. Therefore, $f \in X^{p_2}$. Further, (61) shows that

$$\|f - \sigma_{2^n}(f)\|_{p_2} \le c2^{-n(\gamma-\alpha)}.$$

Together with (60) we have thus shown that

$$\|f - \sigma_{2^n}(f)\|_{p_2} + \frac{1}{2^{nr}}\|\Delta^r \sigma_{2^n}(f)\|_{p_2} \le c2^{-n(\gamma-\alpha)}.$$

In view of (48), this proves (58).

Part (b) is special case of part (a). □

5 Proofs of the Main Results

We start with the proof of Theorem 2. This proof mimics that of [20, Theorem 5]. However, while this theorem was proved in the case of the sphere for (exact) quadrature measures, the following theorem assumes only approximate quadratures and is, of course, valid for generic data-defined spaces.

Proof of Theorem 2 Without loss of generality, we may assume in this proof that $\|f\|_{H^p_\gamma} = 1$. Let $m \ge 0$ be an integer such that $2^m \le n < 2^{m+1}$. Proposition 3(b) shows that

$$|\!|\!|\nu|\!|\!|_{R,2^{-m}} \sim |\!|\!|\nu|\!|\!|_{R,1/n} \sim |\!|\!|\nu|\!|\!|_{R,2^{-m-1}} \le c.$$

Since $\gamma > q/p$, Lemma 4 shows that $f \in X^\infty$, so that Lemma 2(a) leads to

$$f = \sigma_{2^m}(f) + \sum_{j=m+1}^{\infty} \tau_j(f),$$

where the series converges uniformly. Hence, using (53) with $d = 2^{-m}$ and (52), we obtain

$$\begin{aligned}\left|\int_{\mathbb{X}} fd\nu - \int_{\mathbb{X}} \sigma_{2^m}(f)d\nu\right| &\le \sum_{j=m+1}^{\infty} \left|\int_{\mathbb{X}} \tau_j(f)d\nu\right| \le \sum_{j=m+1}^{\infty} \|\tau_j(f)\|_{\nu;1} \\ &\le c|\!|\!|\nu|\!|\!|_{R,d}^{1/p}(|\nu|(\mathbb{X}))^{1/p'} \sum_{j=m+1}^{\infty} 2^{(j-m)q/p}\|\tau_j(f)\|_p \\ &\le c2^{-mq/p}|\!|\!|\nu|\!|\!|_{R,d}^{1/p}(|\nu|(\mathbb{X}))^{1/p'} \sum_{j=m+1}^{\infty} 2^{-j(\gamma-q/p)} \\ &= c2^{-m\gamma}|\!|\!|\nu|\!|\!|_{R,d}^{1/p}(|\nu|(\mathbb{X}))^{1/p'}. \end{aligned} \tag{62}$$

In view of (20), we obtain using Corollary 2 that

$$\left|\int_{\mathbb{X}} \sigma_{2^m}(f)d\mu^* - \int_{\mathbb{X}} \sigma_{2^m}(f)d\nu\right| \le \frac{A}{2^{m\gamma}}\|\sigma_{2^m}(f)\|_{H^p_\gamma} \le c\frac{A}{2^{m\gamma}}\|f\|_{H^p_\gamma} = c\frac{A}{2^{m\gamma}}.$$

Using this observation and (62), we deduce that

$$\begin{aligned}\left|\int_{\mathbb{X}} fd\mu^* - \int_{\mathbb{X}} fd\nu\right| &= \left|\int_{\mathbb{X}} \sigma_{2^m}(f)d\mu^* - \int_{\mathbb{X}} fd\nu\right| \\ &\le \left|\int_{\mathbb{X}} \sigma_{2^m}(f)d\mu^* - \int_{\mathbb{X}} \sigma_{2^m}(f)d\nu\right| + \left|\int_{\mathbb{X}} fd\nu - \int_{\mathbb{X}} \sigma_{2^m}(f)d\nu\right| \\ &\le c\left(A + |\!|\!|\nu|\!|\!|_{R,d}^{1/p}(|\nu|(\mathbb{X}))^{1/p'}\right)2^{-m\gamma}.\end{aligned}$$

This proves Theorem 2. □

In order to prove Theorem 3, we first summarize in Proposition 6 below some properties of the kernels G introduced in Definition 5, including the existence of such a kernel. This proposition is proved in [25, Proposition 5.2]; we state it with p' in [25, Proposition 5.2] replaced by p per the requirement of our proof. Although the set up there is stated as that of a compact smooth manifold without boundary, the proofs are verbatim the same for data-defined spaces.

Let b be a mask of type $\beta \in \mathbb{R}$. In the sequel, if $N > 0$, we will write $b_N(t) = b(Nt)$.

Proposition 6 *Let $1 \le p \le \infty$, $\beta > q/p$, G be a kernel of type β.*

(a) *For every $y \in \mathbb{X}$, there exists $\psi_y := G(\circ, y) \in X^{p'}$ such that $\langle \psi_y, \phi_k \rangle = b(\lambda_k)\phi_k(y)$, $k = 0, 1, \cdots$. We have*

$$\sup_{y\in\mathbb{X}} \|G(\circ, y)\|_{p'} \le c. \tag{63}$$

(b) *Let $n \ge 1$ be an integer, $\nu \in \mathscr{R}_{2^{-n}}$, and for $F \in L^1(\nu) \cap L^\infty(\nu)$, $m \ge n$,*

$$U_m(F, x) := \int_{y\in\mathbb{X}} \{G(x, y) - \Phi_{2^m}(hb_{2^m}; x, y)\}F(y)d\nu(y).$$

Then

$$\|U_m(F)\|_{p'} \le c2^{-m\beta}2^{q(m-n)/p}\|\nu\|_{R,2^{-n}}\|F\|_{\nu;\mathbb{X},p'}. \tag{64}$$

It is convenient to prove Proposition 2 before proving Theorem 3.

Proof of Proposition 2 Let $P \in \Pi_{2^n}$, and we define

$$\mathscr{D}_G(P)(x) = \sum_j \frac{\hat{P}(j)}{b(\lambda_j)}\phi_j(x), \qquad x \in \mathbb{X}.$$

Then it is easy to verify that

$$P(x) = \int_{\mathbb{X}} G(x, y)\mathscr{D}_G(P)(y)d\mu^*(y). \tag{65}$$

Using Fubini's theorem and the condition that $M_p(\nu^\#) \le \tilde{A}2^{-n\beta}$, we deduce that

$$\begin{aligned}
&\left|\int_{\mathbb{X}} P(x)d\nu^\#(x) - \int_{\mathbb{X}} P(x)d\mu^*(x)\right| \\
&= \left|\int_{\mathbb{X}} \mathscr{D}_G(P)(y)\left\{\int_{\mathbb{X}} G(x, y)d\nu^\#(x) - \int_{\mathbb{X}} G(x, y)d\mu^*(x)\right\} d\mu^*(y)\right| \\
&\le \|\mathscr{D}_G(P)\|_p M_p(\nu^\#) \le \tilde{A}2^{-n\beta}\|\mathscr{D}_G(P)\|_p.
\end{aligned} \tag{66}$$

Let $0 < \gamma < r < \beta$. We have proved in [25, Lemma 5.4(b)] that

$$\|\mathscr{D}_G(P)\|_p \le c2^{n(\beta-r)}\|\Delta^r P\|_p.$$

Hence, Corollary 3 implies that

$$\|\mathscr{D}_G(P)\|_p \le c2^{n(\beta-\gamma)}\|P\|_{H^p_\gamma}.$$

Using (66), we now conclude that

$$\left|\int_{\mathbb{X}} P(x)d\nu^\#(x) - \int_{\mathbb{X}} P(x)d\mu^*(x)\right| \le \tilde{A}2^{-n\beta}\|\mathscr{D}_G(P)\|_p \le c\tilde{A}2^{-n\beta}2^{n(\beta-\gamma)}\|P\|_{H^p_\gamma}.$$

Thus, $\nu^\#$ satisfies (20). □

Proof of Theorem 3 We note that

$$\int_{\mathbb{X}} \{G(x,y) - \Phi_{2^n}(hb_{2^n}; x, y)\}\, d\mu^*(x) = 0, \qquad y \in \mathbb{X}.$$

Since $\nu^* \in K$, ν^* is a quadrature measure of order 2^n, and $\Phi_{2^n}(hb_{2^n}; x, \circ) \in \Pi_{2^n}$, we obtain that

$$\begin{aligned} M_p(\nu^\#) &\le c \inf_{\nu\in K} M_p(\nu) \le cM_p(\nu^*) \\ &= c\left\|\int_{\mathbb{X}} \{G(x,\circ) - \Phi_{2^n}(hb_{2^n}; x, \circ)\}\, d\nu^*(x)\right. \\ &\qquad\qquad \left. - \int_{\mathbb{X}} \{G(x,\circ) - \Phi_{2^n}(hb_{2^n}; x, \circ)\}\, d\mu^*(x)\right\|_{p'} \\ &= c\left\|\int_{\mathbb{X}} \{G(x,\circ) - \Phi_{2^n}(hb_{2^n}; x, \circ)\}\, d\nu^*(x)\right\|_{p'}. \end{aligned} \tag{67}$$

We now use Proposition 6(b) with $F \equiv 1$, $m = n$, to conclude that

$$M_p(\nu^\#) \le c2^{-n\beta}. \tag{68}$$

Thus, $\nu^\#$ satisfies the conditions in Proposition 2, and hence, (20). □

The main idea in the proof of Theorem 4 below is to show that the localization of the kernels Φ_N imply via Proposition 4 that the integral of $\Phi_N(x, \cdot)$ on $\mathbb{X}$ is concentrated on a ball of radius $\sim 1/N$ around x.

Proof of Theorem 4 Let $n \ge 1$, $\nu = \nu_n \in \aleph$, and $x \in \mathbb{X}$. In this proof only, let $\alpha \in (0,1)$ be fixed (to be chosen later), N be defined by

$$N^{\gamma+q/p'} = (\alpha n)^\gamma; \quad \text{i.e.; } N = (\alpha n)^{1/\tilde{p}}. \tag{69}$$

We consider the polynomial $P = \Phi_N(x, \circ)$ and note that $P \in \Pi_N \subseteq \Pi_n$. Since ν satisfies (20),

$$\left|1 - \int_{\mathbb{X}} P(y)d\nu(y)\right| = \left|\int_{\mathbb{X}} P(y)d\mu^*(y) - \int_{\mathbb{X}} P(y)d\nu(y)\right| \le \frac{c}{n^\gamma}\|P\|_{H^p_\gamma}. \tag{70}$$

In view of (55) in Corollary 3, and Proposition 1, we deduce using the definition (69) that

$$\|P\|_{H^p_\gamma} \le cN^\gamma\|P\|_p \le cN^{\gamma+q/p'}\|P\|_1 \le cN^{\gamma+q/p'} = c\alpha^\gamma n^\gamma.$$

Therefore, (70) leads to

$$\left|1 - \int_{\mathbb{X}} P(y)d\nu(y)\right| \le c\alpha^\gamma.$$

We now choose α to be sufficiently small to ensure that

$$\left|1 - \int_{\mathbb{X}} \Phi_N(x,y)d\nu(y)\right| = \left|1 - \int_{\mathbb{X}} P(y)d\nu(y)\right| \le 1/2, \qquad n \ge 1. \tag{71}$$

Next, we use (36) with h in place of H, $d = 1/n$, and $r = \lambda/N$ for sufficiently large $\lambda \ge 1$ to be chosen later. Recalling that $N \le n$, this yields

$$\begin{aligned}\left|\int_{\Delta(x,\lambda/N)} P(y)d\nu(y)\right| &\le \int_{\Delta(x,\lambda/N)} |\Phi_N(x,y)|d|\nu|(y)\\ &\le c(1+(N/n)^q)(\lambda)^{-S+q} \le c\lambda^{-S+q},\end{aligned} \tag{72}$$

where we recall our convention that $|||\nu|||_{R,1/n}$ is assumed to be bounded independently of n. We now choose λ to be large enough so that

$$\int_{\Delta(x,\lambda/N)} |\Phi_N(x,y)|d|\nu|(y) \le 1/4. \tag{73}$$

Together with (71), (72), this leads to

$$1/4 \le \int_{\mathbb{B}(x,\lambda/N)} \Phi_N(x,y)d\nu(y) \le 7/4, \qquad n \ge 1. \tag{74}$$

Since $|\Phi_N(x,y)| \le cN^q$ (cf. (34)), we deduce that

$$\begin{aligned}1/4 &\le \int_{\mathbb{B}(x,\lambda/N)} \Phi_N(x,y)d\nu(y) \le \int_{\mathbb{B}(x,\lambda/N)} |\Phi_N(x,y)|d|\nu|(y)\\ &\le cN^q|\nu|(\mathbb{B}(x,\lambda/N)).\end{aligned}$$

This implies (24). □

Finally, the proof of Theorem 5 mimics that of [15, Theorem 5.8] (see also [15, Theorem 5.5(a)] to see the connection with regular measures). Unlike in that proof, we use only the approximate quadrature measures rather than exact quadrature measures.

Proof of Theorem 5 Let $x \in \mathbb{X}$. With A^* defined as in the product assumption and β as in (28), let $\tilde{A} = \max(A^*, 1/\beta)$. In view of (28) and the fact that ν is a positive measure, we obtain that

$$n^{2q}\nu(\mathbb{B}(x,1/n)) \le c\int_{\mathbb{B}(x,1/n)} |\Phi_{n/\tilde{A}}(x,y)|^2 d\nu(y) \le \int_{\mathbb{X}} |\Phi_{n/\tilde{A}}(x,y)|^2 d\nu(y). \qquad (75)$$

Let $R \ge 1$. In view of (57) in Lemma 3, there exists $Q \in \Pi_n$ such that

$$\|\Phi_{n/\tilde{A}}(x,\circ)^2 - Q\|_\infty \le cn^{-R}\|\Phi_{n/\tilde{A}}(x,\circ)\|_1^2 \le cn^{-R}. \qquad (76)$$

Since ν satisfies (20), we obtain first that

$$\left|\int_{\mathbb{X}} \phi_0 d\mu^* - \int_{\mathbb{X}} \phi_0 d\nu\right| \le c,$$

so that $|\nu|(\mathbb{X}) \le c$, and then conclude using (76) that

$$\begin{aligned}&\left|\int_{\mathbb{X}} |\Phi_{n/\tilde{A}}(x,y)|^2 d\mu^*(y) - \int_{\mathbb{X}} |\Phi_{n/\tilde{A}}(x,y)|^2 d\nu(y)\right|\\ &\le \int_{\mathbb{X}} ||\Phi_{n/\tilde{A}}(x,y)|^2 - Q(y)|d\mu^*(y) + \int_{\mathbb{X}} ||\Phi_{n/\tilde{A}}(x,y)|^2 - Q(y)|d\nu(y)\\ &\qquad + \left|\int_{\mathbb{X}} Qd\mu^* - \int_{\mathbb{X}} Qd\nu\right| \le \frac{c}{n^R} + c\frac{\|Q\|_{H^p_\gamma}}{n^\gamma}. \end{aligned} \qquad (77)$$

Since $Q \in \Pi_n$, Lemma 3, Corollary 3, and (76) lead to

$$\begin{aligned}\frac{\|Q\|_{H^p_\gamma}}{n^\gamma} &\le c\|Q\|_p \le cn^{q/p'}\|Q\|_1 \le cn^{q/p'}\left\{\|\Phi_{n/\tilde{A}}(x,\circ)^2 - Q\|_\infty + \|\Phi_{n/\tilde{A}}(x,\circ)^2\|_1\right\}\\ &\le cn^{q/p'}\{n^{-R} + \|\Phi_{n/\tilde{A}}(x,\circ)^2\|_1\}.\end{aligned} \qquad (78)$$

In view of (38) in Proposition 4, used with μ^* in place of ν, $d = 0, p = 2$, we see that

$$\|\Phi_{n/\tilde{A}}(x,\circ)^2\|_1 = \|\Phi_{n/\tilde{A}}(x,\circ)\|_2^2 \le cn^q.$$

Therefore, (78) and (77) imply that

$$\int_{\mathbb{X}} |\Phi_{n/\tilde{A}}(x,y)|^2 d\nu(y) \le cn^{q+q/p'}.$$

Together with (75), this leads to (29). □

Acknowledgements The research of this author is supported in part by ARO Grant W911NF-15-1-0385. We thank the referees for their careful reading and helpful comments.

References

1. Belkin, M., Niyogi, P.: Semi-supervised learning on Riemannian manifolds. Mach. Learn. **56**(1–3), 209–239 (2004)
2. Belkin, M., Niyogi, P.: Convergence of Laplacian eigenmaps. Adv. Neural Inf. Proces. Syst. **19**, 129 (2007)
3. Belkin, M., Niyogi, P.: Towards a theoretical foundation for Laplacian-based manifold methods. J. Comput. Syst. Sci. **74**(8), 1289–1308 (2008)
4. Bondarenko, A., Radchenko, D., Viazovska, M.: Optimal asymptotic bounds for spherical designs. Ann. Math. **178**(2), 443–452 (2013)
5. Brandolini, L., Choirat, C., Colzani, L., Gigante, G., Seri, R., Travaglini, G.: Quadrature rules and distribution of points on manifolds. Ann. Sc. Norm. Super. Pisa Cl. Sci **13**, 889–923 (2014)
6. Brauchart, J., Saff, E., Sloan, I., Womersley, R.: QMC designs: optimal order quasi monte carlo integration schemes on the sphere. Math. Comput. **83**(290), 2821–2851 (2014)
7. Brauchart, J.S., Dick, J., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces. J. Math. Anal. Appl. **431**(2), 782–811 (2015)
8. Breger, A., Ehler, M., Graef, M.: Points on manifolds with asymptotically optimal covering radius. Preprint at arXiv:1607.06899 (2016)
9. Breger, A., Ehler, M., Graef, M.: Quasi monte carlo integration and kernel-based function approximation on Grassmannians. Preprint at arXiv:1605.09165 (2016)
10. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006)
11. Davies, E.B.: L^p spectral theory of higher-order elliptic differential operators. Bull. Lond. Math. Soc. **29**(05), 513–546 (1997)
12. DeVore, R.A., Lorentz, G.G.: Constructive Approximation, vol. 303. Springer Science & Business Media, Berlin (1993)
13. Ehler, M., Filbir, F., Mhaskar, H.N.: Locally learning biomedical data using diffusion frames. J. Comput. Biol. **19**(11), 1251–1264 (2012)
14. Filbir, F., Mhaskar, H.N.: A quadrature formula for diffusion polynomials corresponding to a generalized heat kernel. J. Fourier Anal. Appl. **16**(5), 629–657 (2010)
15. Filbir, F., Mhaskar, H.N.: Marcinkiewicz–Zygmund measures on manifolds. J. Complex. **27**(6), 568–596 (2011)

16. Geller, D., Pesenson, I.Z.: Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds. J. Geom. Anal. **21**(2), 334–371 (2011)
17. Grigor'yan, A.: Estimates of heat kernels on Riemannian manifolds. Lond. Math. Soc. Lect. Note Ser. **273**, 140–225 (1999)
18. Grigor'yan, A.: Heat kernels on metric measure spaces with regular volume growth. In: Handbook of Geometric Analysis, no. 2. Adv. Lect. Math. (ALM), vol. 13, pp. 1–60. Int. Press, Somerville (2010)
19. Grigor'yan, A.: Heat kernels on weighted manifolds and applications. Contemp. Math **398**, 93–191 (2006)
20. Hesse, K., Mhaskar, H.N., Sloan, I.H.: Quadrature in Besov spaces on the Euclidean sphere. J. Complex. **23**(4), 528–552 (2007)
21. Jones, P.W., Maggioni, M., Schul, R.: Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. Proc. Natl. Acad. Sci. **105**(6), 1803–1808 (2008)
22. Kordyukov, Y.A.: L^p-theory of elliptic differential operators on manifolds of bounded geometry. Acta Appl. Math. **23**(3), 223–260 (1991)
23. Lafon, S.S.: Diffusion maps and geometric harmonics. Ph.D. thesis, Yale University (2004)
24. Maggioni, M., Mhaskar, H.N.: Diffusion polynomial frames on metric measure spaces. Appl. Comput. Harmon. Anal. **24**(3), 329–353 (2008)
25. Mhaskar, H.N.: Eignets for function approximation on manifolds. Appl. Comput. Harmon. Anal. **29**(1), 63–87 (2010)
26. Mhaskar, H.N.: A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. Neural Netw. **24**(4), 345–359 (2011)
27. Mhaskar, H.N.: A unified framework for harmonic analysis of functions on directed graphs and changing data. Appl. Comput. Harmon. Anal. Published online June 28 (2016)
28. Rivlin, T.J.: The Chebyshev Polynomials. Wiley, New York (1974)
29. Rosasco, L., Belkin, M., Vito, E.D.: On learning with integral operators. J. Mach. Learn. Res. **11**, 905–934 (2010)
30. Singer, A.: From graph to manifold Laplacian: the convergence rate. Appl. Comput. Harmon. Anal. **21**(1), 128–134 (2006)

Tractability of Multivariate Problems for Standard and Linear Information in the Worst Case Setting: Part II

Erich Novak and Henryk Woźniakowski

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We study QPT (quasi-polynomial tractability) in the worst case setting for linear tensor product problems defined over Hilbert spaces. We assume that the domain space is a reproducing kernel Hilbert space so that function values are well defined. We prove QPT for algorithms that use only function values under the three assumptions:

1. the minimal errors for the univariate case decay polynomially fast to zero,
2. the largest singular value for the univariate case is simple and
3. the eigenfunction corresponding to the largest singular value is a multiple of the function value at some point.

The first two assumptions are necessary for QPT. The third assumption is necessary for QPT for some Hilbert spaces.

1 Introduction

In Part I [6] we presented a lower error bound for approximating linear multivariate operators defined over Hilbert spaces with algorithms that use function values. In this Part II we study upper bounds and algorithms for the same problem. We want to understand the intrinsic difficulty of approximation of d-variate problems when

E. Novak (✉)
Jena University, Math Institute, Jena, Germany
e-mail: erich.novak@uni-jena.de

H. Woźniakowski
Institute of Applied Mathematics and Mechanics, University of Warsaw, Warsaw, Poland

Department of Computer Science, Columbia University, New York, USA
e-mail: henryk@cs.columbia.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_42

d is large. Algorithms that approximate d-variate problems may use finitely many functionals from the class $\Lambda^{\rm all}$ of information or from the standard class $\Lambda^{\rm std}$ of information. The class $\Lambda^{\rm all}$ consists of arbitrary linear functionals, whereas the class $\Lambda^{\rm std}$ consists of only function values.

We wish to approximate a d-variate problem in the worst case setting to within an error threshold $\varepsilon \in (0,1)$. The intrinsic difficulty is measured by the information complexity which is defined as the minimal number of linear functionals from the class $\Lambda \in \{\Lambda^{\rm all}, \Lambda^{\rm std}\}$ which is needed to find an ε-approximation, see (2) for the precise definition.

Tractability deals with how the information complexity depends on d and on ε^{-1}, see [3–5]. In particular, we would like to know when the information complexity is exponential in d, the so-called curse of dimensionality, and when we have a specific dependence on d which is not exponential. There are various ways of measuring the lack of exponential dependence and that leads to different notions of tractability. In particular, we have polynomial tractability (PT) when the information complexity is polynomial in d and ε^{-1}, and quasi-polynomial tractability (QPT) if the information complexity is at most proportional to

$$\exp\big(t\,(1+\ln\varepsilon^{-1})(1+\ln d)\big) = \big(e\,\varepsilon^{-1}\big)^{t(1+\ln d)}$$

for some non-negative t independent of d and ε. This means that the exponent of ε^{-1} may depend weakly on d through $\ln\, d$.

In this paper we study QPT for linear (unweighted) tensor product problems, $\mathbb{S} = \{S_d\}$ with $S_d = S_1^{\otimes d}$ and a compact linear non-zero $S_1 : F_1 \to G_1$ for Hilbert spaces F_1 and G_1. Since we want to use function values we need to assume that F_1 is a reproducing kernel Hilbert space of univariate functions defined on a non-empty $D \subseteq \mathbb{R}$. For simplicity we consider real valued functions. By

$$K_1 : D_1 \times D_1 \to \mathbb{R}$$

we denote the reproducing kernel of F_1. Then $S^{\otimes d} : F_1^{\otimes d} \to G_1^{\otimes d}$ and $F_1^{\otimes d}$ is a reproducing kernel Hilbert space of d-variate functions defined on $D \times D \times \cdots \times D$ (d times) with the reproducing kernel

$$K_d(x,t) = \prod_{j=1}^{d} K_1(x_j,t_j) \qquad \text{for all} \quad x_j, t_j \in D_1.$$

Obviously, tractability may depend on which class $\Lambda^{\rm std}$ or $\Lambda^{\rm all}$ is used. Tractability results for $\Lambda^{\rm std}$ cannot be better than for $\Lambda^{\rm all}$. The main question is when they are more or less the same. In particular, it is known when QPT holds for $\Lambda^{\rm all}$. Namely, let $\{\lambda_j, \eta_j\}$ be the ordered sequence of eigenvalues λ_j and orthonormal eigenfunctions η_j of $S_1^*S_1 : F_1 \to F_1$. Here $S_1^* : G_1 \to F_1$ is the adjoint operator of S_1. Let

$$\text{decay}_\lambda := \sup\{\, r \ge 0 \;:\; \lim_{j\to\infty} j^r \lambda_j = 0\,\}$$

denote the polynomial decay of the eigenvalues λ_j. Since S_1 is assumed to be non-zero and compact, we have $\lambda_1 > 0$, $\lim_j \lambda_j = 0$, and decay_λ is well defined. However, it may happen that $\mathrm{decay}_\lambda = 0$.

It is known, see [1], that

$$\mathbb{S} \text{ is QPT for } \Lambda^{\mathrm{all}} \quad \text{iff} \quad \lambda_2 < \lambda_1 \text{ and } \mathrm{decay}_\lambda > 0.$$

Furthermore, if $\lambda_2 > 0$ then $\mathbb{S}$ is not PT for Λ^{all} (and for Λ^{std}). On the other hand, if $\lambda_2 = \lambda_1 > 0$ then $\mathbb{S}$ suffers from the curse of dimensionality for the class Λ^{all} (and for Λ^{std}).

We now discuss QPT for Λ^{std}. To motivate the need for the assumption on the eigenfunction η_1 corresponding to the largest eigenvalue λ_1, we cite a result from Part I, see [6], for the Sobolev space

$$F_1 \text{ with the reproducing kernel } K_1^*(x,t) = 1 + \min(x,t) \text{ for } x,t \in [0,1].$$

Then $\mathbb{S}$ suffers from the curse of dimensionality if

$$\eta_1 \neq \pm\, [K_1^*(t,t)]^{-1/2}\, K_1^*(\cdot,t) = \pm\,(1+t)^{-1/2}\,(1+\min(\cdot,t)) \quad \text{for all} \quad t \in [0,1]. \tag{1}$$

Furthermore, for the approximation problem, $S_1 f = \mathrm{APP}_1 f = f \in G_1 = L_2([0,1])$, the assumption (1) holds, $\lambda_2 < \lambda_1$ and $\mathrm{decay}_\lambda = 2$. Therefore for $\mathrm{APP} = \{\mathrm{APP}_1^{\otimes d}\}$ we have

$$\text{Curse for } \Lambda^{\mathrm{std}} \quad \text{and} \quad \text{QPT for } \Lambda^{\mathrm{all}}.$$

In this paper we prove that the assumption (1) is essential for the curse and QPT can hold for the class Λ^{std} if (1) is not satisfied.

This will be shown by establishing a result for general linear non-zero tensor product problems for which F_1 is an arbitrary reproducing kernel Hilbert space with the reproducing kernel $K_1 : D_1 \times D_1 \to \mathbb{R}$. For the class Λ^{std}, the role of the sequence $\lambda = \{\lambda_j\}$ is replaced by the sequence $e = \{e_n(S_1)\}$ of the minimal worst case errors of algorithms that use at most n function values. First of all, note that

$$\lim_n e_n(S_1) = 0.$$

Indeed, this holds for S_1 being a continuous linear functional, see [4, p. 79] and for a compact linear operator S_1 and for all positive ε it is enough to approximate sufficiently well finitely many linear functionals. We define the polynomial decay of the minimal errors $e_n(S_1)$ as for the eigenvalues by

$$\mathrm{decay}_e := \sup\{\, r \geq 0 \,:\, \lim_{n\to\infty} n^r e_n(S_1) = 0 \,\}.$$

The main result of this paper is the following theorem.

Theorem 1 *Let $\mathbb{S}$ be a non-zero linear tensor product with a compact linear S_1 for which*

- $\lambda_2 < \lambda_1$,
- $\mathrm{decay}_e > 0$,
- $\eta_1 = \pm\, K_1(t,t)^{-1/2}\, K_1(\cdot,t)$ *for some* $t \in D_1$.

Then $\mathbb{S}$ is QPT for the class $\Lambda^{\rm std}$.

We now comment on the assumptions of this theorem. The first assumption is the same as for the class $\Lambda^{\rm all}$. As already said, for $\lambda_2 = \lambda_1 > 0$ we have the curse of dimensionality for $\Lambda^{\rm std}$. The second assumption is necessary for QPT and the class $\Lambda^{\rm std}$. Indeed, if $\mathrm{decay}_e = 0$ then even the univariate case cannot be solved polynomially in ε^{-1}. This assumption corresponds to the assumption $\mathrm{decay}_\lambda > 0$ for the class $\Lambda^{\rm all}$. For many problems we have $\mathrm{decay}_e = \mathrm{decay}_\lambda$. However, there are problems for which $\mathrm{decay}_\lambda = 1$, $\mathrm{decay}_e = 0$, and $e_n(S_1)$ can go to zero arbitrarily slowly, i.e., like $1/\ln(\ln(\cdots\ln(n)))$, where the number of ln can be arbitrarily large, see [2] which is also reported in [5] pp. 292–304. In this case, i.e., when $\mathrm{decay}_\lambda > 0$ and $\mathrm{decay}_e = 0$, we have QPT for $\Lambda^{\rm all}$ and no QPT for $\Lambda^{\rm std}$.

We now discuss the last assumption which states that the eigenfunction η_1 corresponding to the largest eigenvalue λ_1 is of a very special form. First of all, note that the scaling which is used in (1) and here is needed to guarantee that $\|\eta_1\| = 1$. This implies that $K_1(t,t) > 0$. For $\eta_1 = \pm\, K_1(t,t)^{-1/2}\, K_1(\cdot,t)$ we have

$$\langle f, \eta_1\rangle_{F_1} = \pm\, K_1(t,t)^{-1/2}\ \langle f, K_1(\cdot,t)\rangle_{F_1} = \pm\, K_1(t,t)^{-1/2} f(t).$$

This means that the inner product $\langle f, \eta_1\rangle_{F_1}$ now can be computed exactly by one function value. Apparently, this important property allows us to achieve QPT for the class $\Lambda^{\rm std}$. If this last assumption is not satisfied then we may decrease F_1 slightly by a rank 1 modification to obtain QPT for the modified problem, see Sect. 6.

Theorem 1 will be proved constructively by presenting an algorithm $A_{d,\varepsilon}$ that computes an ε-approximation and uses at most $\mathcal{O}\left(\exp\left(t\,(1+\ln\varepsilon^{-1})(1+\ln d)\right)\right)$ function values for some t independent of d and ε^{-1}. The algorithm $A_{d,\varepsilon}$ is a modification of the Smolyak (sparse grid) algorithm applied to components of the operators S_d, see [7, 9] and Chapter 15 of [4] as well as Chapter 27 of [5].

It seems interesting to apply Theorem 1 to the space F_1 with the reproducing kernel K_1^* which was used before. Combining the results of Part I with Theorem 1 we obtain the following corollary.

Corollary 1 *Consider the spaces with K_1^* as above. Then $\mathbb{S}$ is QPT for the class $\Lambda^{\rm std}$ iff*

- $\lambda_2 < \lambda_1$,
- $\mathrm{decay}_e > 0$,
- $\eta_1 = \pm\,(1+t)^{-1/2}\,(1+\min(\cdot,t))$ *for some* $t \in D_1$.

2 Preliminaries

Let $S : F \to G$ be a continuous linear non-zero operator, where F is a reproducing kernel Hilbert space of real functions f defined over a common non-empty domain $D \subset \mathbb{R}^k$ for some positive integer k, and G is a Hilbert space. We approximate S by algorithms A_n that use at most n function values, i.e., we use the class Λ^{std}. Without loss of generality we may assume that A is linear, see e.g., [3, 8]. That is,

$$A_n f = \sum_{j=1}^{n} f(t_j)\, g_j$$

for some $t_j \in D$ and $g_j \in S(F) \subseteq G$. The worst case error of A_n is defined as

$$e(A_n) = \sup_{\|f\|_F \le 1} \|Sf - A_n f\|_G = \|S - A_n\|_{F \to G}.$$

For $n = 0$, we take $A_n = 0$ and then we obtain the initial error which is

$$e(0) = e_0(S) = \|S\|_{F \to G}.$$

Since S is non-zero, the initial error is positive.

We are ready to define the information complexity for the class Λ^{std} and for the so-called normalized error criterion. It is defined as the minimal number of function values which are needed to reduce the initial error by a factor $\varepsilon \in (0, 1)$. That is,

$$n(\varepsilon, S) = \min\{\, n \,:\, \exists A_n \text{ such that } e(A_n) \le \varepsilon\, e_0(S)\}. \tag{2}$$

Assume now that we have a sequence

$$\mathbb{S} = \{S_d\}_{d=1}^{\infty}$$

of continuous linear non-zero operators $S_d : F_d \to G_d$, where F_d is a reproducing kernel Hilbert space of real functions defined over a non-empty $D_d \subset \mathbb{R}^d$ and G_d is a Hilbert space. In this case, we want to verify how the information complexity $n(\varepsilon, S_d)$ depends on ε^{-1} and d. We say that $\mathbb{S}$ is quasi-polynomially tractable (QPT) for the class Λ^{std} iff there are non-negative numbers C and t such that

$$n(\varepsilon, S_d) \le C \exp\big(t\,(1 + \ln\, \varepsilon^{-1})(1 + \ln\, d)\big) \quad \text{for all } \ \varepsilon \in (0, 1),\ d \in \mathbb{N}.$$

More about other tractability concepts can be found in [3–5].

3 Linear Tensor Products

We obtain a linear tensor product problem if the spaces $F = F_d$ and $G = G_d$ as well as $S = S_d$ are given by tensor products of d copies of F_1 and G_1 as well as a continuous linear non-zero operator $S_1 : F_1 \to G_1$, respectively, where F_1 is a reproducing kernel Hilbert space of real univariate functions defined over a non-empty $D_1 \subset \mathbb{R}$ and G_1 is a Hilbert space. To simplify the notation we assume that F_1 is of infinite dimension. Then F_d is an infinite dimensional space of d-variate real functions defined on $D_d = D_1 \times D_1 \times \cdots \times D_1$ (d times).

We assume that S_1 is compact. Then all S_d are also compact. Let (λ_j, η_j) be the eigenpairs of $W_1 = S_1^* S_1 : F_1 \to F_1$ with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \quad \text{and} \quad \langle \eta_i, \eta_j \rangle_{F_1} = \delta_{i,j}.$$

Clearly, $\|S_1\|_{F_1 \to G_1} = \sqrt{\lambda_1}$. Since S is non-zero, $\lambda_1 > 0$. We have$f \in F_1$ iff

$$f = \sum_{j=1}^{\infty} \langle f, \eta_j \rangle_{F_1} \eta_j \quad \text{with} \quad \|f\|_{F_1}^2 = \sum_{j=1}^{\infty} \langle f, \eta_j \rangle_{F_1}^2 < \infty.$$

Then

$$S_1 f = \sum_{j=1}^{\infty} \langle f, \eta_j \rangle_{F_1} S_1 \eta_j, \tag{3}$$

where

$$\langle S_1 \eta_i, S_1 \eta_j \rangle_{G_1} = \langle \eta_i, W_1 \eta_j \rangle_{F_1} = \lambda_j \delta_{i,j}.$$

This means that the sequence $\{S_1 \eta_j\}$ is orthogonal in G_1 and

$$\|S_1 f\|_{G_1}^2 = \sum_{j=1}^{\infty} \langle f, \eta_j \rangle_{F_1}^2 \lambda_j.$$

For $d \geq 2$, the eigenpairs (λ_j, η_j) of $W_d = S_d^* S_d : F_d \to F_d$ are given in terms of the eigenpairs (λ_j, η_j) of the univariate operator $W_1 = S_1^* S_1 : F_1 \to F_1$. We have

$$\{\lambda_{d,j}\}_{j=1}^{\infty} = \{\lambda_{j_1} \lambda_{j_2} \cdots \lambda_{j_d}\}_{j_1, j_2, \ldots, j_d = 1}^{\infty}.$$

Similarly, the eigenfunctions of W_d are of product form

$$\{\eta_{d,j}\}_{j=1}^{\infty} = \{\eta_{j_1} \otimes \eta_{j_2} \otimes \cdots \otimes \eta_{j_d}\}_{j_1, j_2, \ldots, j_d = 1}^{\infty},$$

where

$$[\eta_{j_1} \otimes \eta_{j_2} \otimes \cdots \otimes \eta_{j_d}](x) = \prod_{k=1}^{d} \eta_{j_k}(x_k) \quad \text{for all } \ x = [x_1, \dots, x_d] \in D_d$$

and

$$\big\langle \eta_{i_1} \otimes \eta_{i_2} \otimes \cdots \otimes \eta_{i_d}, \eta_{j_1} \otimes \eta_{j_2} \otimes \cdots \otimes \eta_{j_d} \big\rangle_{F_d} = \delta_{i_1,j_1}\delta_{i_2,j_2}\cdots\delta_{i_d,j_d}.$$

Then $\|S_d\|_{F_d \to G_d} = \|W_d\|_{F_d \to F_d}^{1/2} = \lambda_1^{d/2}$. Hence, the initial error is $e_0(S_d) = \lambda_1^{d/2}$.

We have $f \in F_d$ iff

$$f = \sum_{(j_1,j_2,\dots,j_d) \in \mathbb{N}^d} \big\langle f, \eta_{j_1} \otimes \cdots \otimes \eta_{j_d} \big\rangle_{F_d} \eta_{j_1} \otimes \cdots \otimes \eta_{j_d}$$

with

$$\|f\|_{F_d}^2 = \sum_{(j_1,j_2,\dots,j_d) \in \mathbb{N}^d} \big\langle f, \eta_{j_1} \otimes \cdots \otimes \eta_{j_d} \big\rangle_{F_d}^2 < \infty.$$

In particular, for $x = (x_1, x_2, \dots, x_d) \in D_d$ we have

$$f(x) = \sum_{(j_1,j_2,\dots,j_d) \in \mathbb{N}^d} \big\langle f, \eta_{j_1} \otimes \cdots \otimes \eta_{j_d} \big\rangle_{F_d} \eta_{j_1}(x_1) \cdots \eta_{j_d}(x_d).$$

4 Decomposition of Linear Tensor Products

In this section we assume, as in Theorem 1, that

$$\eta_1 = \pm K_1(t,t)^{-1/2} K_1(\cdot, t) \quad \text{for some } \ t \in D_1.$$

Then for $j \geq 2$ we obtain

$$0 = \big\langle \eta_1, \eta_j \big\rangle_{F_1} = K_1(t,t)^{-1/2}\, \eta_j(t).$$

Hence, $\eta_j(t) = 0$ for all $j \geq 2$. This implies that

$$f(t, \dots, t) = \big\langle f, \eta_1^{\otimes d} \big\rangle_{F_d} \eta_1^d(t),$$

and for any $k = 1, 2, \ldots, d-1$ and any vector $x = (t, \ldots, t, x_{k+1}, \ldots, x_d)$ we have

$$f(x) = \sum_{(j_{k+1},\ldots,j_d)\in\mathbb{N}^{d-k}} \left\langle f, \eta_1^{\otimes k} \otimes \eta_{j_{k+1}} \cdots \otimes \eta_{j_d} \right\rangle_{F_d} \\ [\eta_1(t)]^k \eta_{j_{k+1}}(x_{k+1}) \cdots \eta_{j_d}(x_d). \tag{4}$$

We start the decomposition of S_d from the univariate case, $d = 1$. From (3) we have

$$S_1 = V_1 + V_2$$

with

$$V_1 f = \langle f, \eta_1 \rangle_{F_1} S_1 \eta_1 = \pm K_1(t,t)^{-1/2} f(t)\, S_1 \eta_1,$$
$$V_2 f = \sum_{j=2}^{\infty} \left\langle f, \eta_j \right\rangle_{F_1} S_1 \eta_j$$

for all $f \in F_1$. Clearly,

$$\|V_1\|_{F_1\to G_1} = \|S_1\eta_1\|_{G_1} = \sqrt{\lambda_1} \quad \text{and} \quad \|V_2\|_{F_1\to G_1} = \|S_1\eta_2\|_{G_1} = \sqrt{\lambda_2}.$$

We stress that we can compute $V_1 f$ exactly by using one function value.

For $d \geq 2$, we obtain

$$S_d = (V_1 + V_2)^{\otimes d} = \sum_{(j_1,j_2,\ldots,j_d)\in\{1,2\}^d} V_{j_1} \otimes V_{j_2} \otimes \cdots \otimes V_{j_d}.$$

For $j = (j_1, j_2, \ldots, j_d) \in \{1,2\}^d$ we define

$$|j|_2 = |\{j_i \mid j_i = 2\}|$$

as the number of indices equal to 2. Clearly,

$$\|V_{j_1} \otimes V_{j_2} \otimes \cdots \otimes V_{j_d}\|_{F_d\to G_d} = \|V_1\|_{F_1\to G_1}^{d-|j|_2} \|V_2\|_{F_1\to G_1}^{|j|_2} = \lambda_1^{(d-|j|_2)/2} \lambda_2^{|j|_2/2}.$$

5 Algorithms for Linear Tensor Products

We now derive an algorithm for linear tensor products for which the assumptions of Theorem 1 hold and we conclude QPT for the class Λ^{std} from an estimate of the worst case error of this algorithm.

To simplify the notation we assume that $\lambda_1 = 1$. This can be done without loss of generality since otherwise we can replace S_1 by $\lambda_1^{-1/2} S_1$.

For $\lambda_1 = 1$ and due to the first assumption in Theorem 1, we have

$$\|V_1\|_{F_1 \to G_1} = 1 \quad \text{and} \quad \|V_2\|_{F_1 \to G_1} = \lambda_2^{1/2} < 1.$$

Consider first $V_{2,d} = V_2^{\otimes d}$ with an exponentially small norm since

$$\|V_{2,d}\|_{F_d \to G_d} = \lambda_2^{d/2}.$$

From the assumptions $\mathrm{decay}_e > 0$ and $\lambda_2 < 1$, it was concluded in [9], see in particular Lemma 1 and Theorem 2 of this paper, that for all $d \in \mathbb{N}$ there is a Smolyak/sparse grid algorithm

$$A_{d,n} f = \sum_{m=1}^{n} f(t_{d,n,m})\, g_{d,n,m} \quad \text{for all} \quad f \in F_d$$

for some $t_{d,n,m} \in D_d$ and $g_{d,n,m} = g_{d,n,m,1} \otimes \cdots \otimes g_{d,n,m,d}$ with $g_{d,n,m,\ell} \in V_1(F_1) \subseteq G_1$, such that

$$e(A_{d,n}) = \|V_{2,d} - A_{d,n}\|_{F_d \to G_d} \le \alpha\, n^{-r} \quad \text{for all } d, n \in \mathbb{N} \tag{5}$$

for some positive α and r. We stress that α and r are independent of d and n.

From the third assumption of Theorem 1 we know that

$$\eta_1 = \delta\, K_1(t,t)^{-1/2} K_1(\cdot, t), \quad \text{where} \quad \delta \in \{-1, 1\}.$$

For an integer $k \in [0, d]$, consider $V_1^{\otimes (d-k)} \otimes V_2^{\otimes k}$. For $k = 0$ we drop the second factor and for $k = d$ we drop the first factor so that $V_1^{\otimes d} \otimes V_2^{\otimes 0} = V_1^{\otimes d}$ and $V_1^{\otimes 0} \otimes V_2^{\otimes d} = V_2^{\otimes d}$.

For $k = 0$, we approximate $V_1^{\otimes d}$ by the algorithm

$$A_{d,n,0} f = \frac{\delta^d}{K_1(t,t)^{d/2}} f(t, t, \ldots, t)\, (S_1 \eta_1)^{\otimes d} \quad \text{for all} \quad f \in F_d.$$

Clearly, the error of this approximation is zero since $A_{d,n,0} = V_1^{\otimes d}$ and $A_{d,n,0}$ uses one function value.

For $k = d$, we approximate $V_2^{\otimes d}$ by the algorithm $A_{d,n}$ with error at most $\alpha\, n^{-r}$.

For $k = 1, 2, \ldots, d-1$, we approximate $V_1^{\otimes (d-k)} \otimes V_2^{\otimes k}$ by the algorithm

$$A_{d,n,k} f = \frac{\delta^{d-k}}{[K_1(t,t)]^{(d-k)/2}} \sum_{m=1}^{n} f(t, \ldots, t, t_{k,n,m}) (S_1 \eta_1)^{\otimes (d-k)} \otimes g_{k,n,m}$$

for all $f \in F_d$. We now show that

$$A_{d,n,k} = V_1^{\otimes\,(d-k)} \,\otimes\, A_{k,n}. \tag{6}$$

Indeed, we know that $V_1 \eta_j = 0$ for all $j \geq 2$. Then

$$\begin{aligned}
&(V_1^{\otimes\,(d-k)} \,\otimes\, A_{k,n})f \\
&= (V_1^{\otimes\,(d-k)} \,\otimes\, A_{k,n}) \sum_{(j_1,\dots,j_d)\in\mathbb{N}^d} \big\langle f, \eta_{j_1} \otimes \cdots \otimes \eta_{j_d} \big\rangle_{F_d} \eta_{j_1} \otimes \cdots \otimes \eta_{j_d} \\
&= \sum_{(j_1,\dots,j_d)\in\mathbb{N}^d} \big\langle f, \otimes_{\ell=1}^d \eta_{j_\ell} \big\rangle_{F_d} (V_1 \eta_{j_1}) \otimes \cdots \otimes (V_1 \eta_{j_{d-k}}) \otimes A_{k,n}(\eta_{j_{d-k+1}} \otimes \cdots \otimes \eta_{j_d}) \\
&= \alpha_k \sum_{(j_{d-k+1},\dots,j_d)\in\mathbb{N}^k} \Big\langle f, \eta_1^{\otimes\,(d-k)} \otimes_{\ell=1}^k \eta_{j_{d-k+\ell}} \Big\rangle_{F_d} (S_1 \eta_1)^{\otimes\,(d-k)} \otimes \\
&\qquad\qquad \sum_{m=1}^n \big(\otimes_{\ell=1}^k \eta_{j_{d-k+\ell}}\big)\,(t_{k,n,m}) g_{k,n,m} \\
&= \frac{\delta^{d-k}}{[K_1(t,t)]^{(d-k)/2}} \sum_{m=1}^n h_m (S_1 \eta_1)^{\otimes\,(d-k)} \,\otimes\, g_{k,n,m},
\end{aligned}$$

where $\alpha_k = \delta^{d-k} K_1(t,t)^{(d-k)/2} \eta_1(t)^{d-k}$ and

$$\begin{aligned}
h_m = \sum_{(j_{d-k+1},\dots,j_d)\in\mathbb{N}^k} &\Big\langle f, \eta_1^{\otimes\,(d-k)} \otimes \eta_{j_{d-k+1}} \otimes \cdots \otimes \eta_{j_d} \Big\rangle_{F_d} \\
&\cdot \eta_1(t)^{d-k} \big(\eta_{j_{d-k+1}} \otimes \cdots \otimes \eta_{j_d}\big)\,(t_{k,n,j}).
\end{aligned}$$

From (4) we conclude that

$$h_j = f(t, \dots, t, t_{k,n,j})$$

and

$$V_1^{\otimes\,(d-k)} \,\otimes\, A_{k,n} = A_{d,n,k},$$

as claimed. From this, we see that

$$V_1^{\otimes\,(d-k)} \,\otimes\, V_2^{\otimes\,k} - A_{d,n,k} = V_1^{\otimes\,(d-k)} \,\otimes\, (V_2^{\otimes\,k} - A_{k,n})$$

and

$$e(A_{d,n,k}) = \|V_2^{\otimes\,k} - A_{k,n}\|_{F_k \to G_k} \leq \alpha\, n^{-r}.$$

We now explain how we approximate $V_{j_1} \otimes \cdots \otimes V_{j_d}$ for an arbitrary

$$j = (j_1, \ldots, j_d) \in \{1, 2\}^d.$$

The idea is the same as before, i.e., for the indices $j_\ell = 1$ we approximate V_{j_ℓ} by itself, and for the rest of the indices, which are equal to 2, we apply the Smolyak/sparse grid algorithm for proper parameters. More precisely, let $k = |j|_2$. The cases $k = 0$ and $k = d$ have been already considered. Assume then that $k \in [1, d-1] := \{1, 2, \ldots, d-1\}$. Let $\ell_i \in [1, d]$ be the *ith* occurrence of 2 in the vector j, i.e., $1 \le \ell_1 < \ell_2 < \cdots < \ell_k \le d$, and $j_{\ell_1} = j_{\ell_2} = \cdots = j_{\ell_k} = 2$.

Define the algorithm

$$A_{d,n,j} f = \frac{\delta^{d-k}}{K_1(t,t)^{(d-k)/2}} \sum_{m=1}^{n} f(y_{d,n,j,m})\, h_{d,n,j,m,1} \otimes \cdots \otimes h_{d,n,j,m,d},$$

where the vector $y_{d,n,j,m} = (y_{d,n,j,m,1}, \ldots, y_{d,n,j,m,d})$ is given by

$$y_{d,n,j,m,\ell} = \begin{cases} t & \text{if } j_\ell = 1, \\ t_{k,n,m,i} & \text{if } j_\ell = 2 \text{ and } \ell = \ell_i, \end{cases}$$

and

$$h_{d,n,j,m.\ell} = \begin{cases} S_1 \eta_1 & \text{if } j_\ell = 1, \\ g_{k,n,m,i} & \text{if } j_\ell = 2 \text{ and } \ell = \ell_i \end{cases}$$

for $\ell = 1, 2, \ldots, d$.

The error of the algorithm $A_{d,n,j}$ is the same as the error of the algorithm $A_{d,n,|j|_2}$ since for (unweighted tensor) products the permutation of indices does not matter.

Hence, for all $j \in \{1, 2\}$, the algorithm $A_{d,n,j}$ uses at most n function values and

$$e(A_{d,n,j}) \le \alpha\, n^{-r}, \tag{7}$$

and this holds for all d.

We now define an algorithm which approximates S_d with error at most $\varepsilon \in (0,1)$. The idea of this algorithm is based on approximation of all $V_{j_1} \otimes \cdots \otimes V_{j_d}$ whose norm is $\|V_2\|^{|j|_2} = \lambda_2^{|j|_2/2}$. If $\lambda_2^{|j|_2/2} \le \varepsilon/2$ we approximate $V_{j_1} \otimes \cdots \otimes V_{j_d}$ by zero otherwise by the algorithm $A_{d,n,j}$ for specially chosen n. More precisely, let

$$k = \min\left(d, \left\lceil \frac{2 \ln \frac{2}{\varepsilon}}{\ln \frac{1}{\lambda_2}} \right\rceil \right).$$

Define the algorithm

$$A_{d,n,\varepsilon} = \sum_{j\in\{1,2\}^d} A_{d,n,\varepsilon,j} \tag{8}$$

with

$$A_{d,n,\varepsilon,j} = \begin{cases} 0 & \text{if } |j|_2 > k, \\ A_{d,n,j} & \text{if } |j|_2 \le k. \end{cases}$$

Note that non-zero terms in (8) correspond to $|j|_2 \le k$ and each of them uses at most n function values. Therefore the algorithm $A_{d,n,\varepsilon}$ uses at most

$$\mathrm{card}(A_{d,n,\varepsilon}) \le n \sum_{\ell=0}^{k} \binom{d}{\ell}$$

function values.

We now analyze the error of $A_{d,n,\varepsilon}$. We have

$$S_d - A_{d,n,\varepsilon} = \sum_{j\in\{1,2\}^d,\ |j|_2\le k} \big(V_{j_1}\otimes\cdots\otimes V_{j_d} - A_{d,n,\varepsilon,j}\big) + \sum_{j\in\{1,2\}^d,\ |j|_2> k} V_{j_1}\otimes\cdots\otimes V_{j_d}.$$

Note that the second operator in the sum above is zero if $k = d$. For $k < d$ the terms of the second operator are orthogonal and therefore it has norm at most $\lambda_2^{k/2} \le \varepsilon/2$ by the definition of k.

From (7) we conclude

$$\|S_d - A_{d,n,\varepsilon}\|_{F_d\to G_d} \le \alpha\, n^{-r} \sum_{\ell=0}^{k} \binom{d}{\ell} \ + \ \varepsilon/2.$$

We now consider two cases $k \le d/2$ and $k > d/2$. We opt for simplicity at the expense of some error overestimates which are still enough to establish QPT.

- Case $k \le d/2$.

  Then the binomial coefficients $\binom{d}{\ell}$ are increasing and

$$\sum_{\ell=0}^{k} \binom{d}{\ell} \le (k+1)\binom{d}{k} \le (k+1)\,\frac{d^k}{k!} \le 2d^k.$$

If we take n such that

$$\frac{2\alpha\, d^k}{n^r} \le \varepsilon/2 \tag{9}$$

then

$$e(A_{d,n,\varepsilon}) = \|S_d - A_{d,n,\varepsilon}\|_{F_d \to G_d} \le \varepsilon.$$

Since $k \le 1 + 2\ln(2\varepsilon^{-1})/\ln(\lambda_2^{-1})$, we have

$$d^k \le \alpha_1 (1+\varepsilon^{-1})^{\alpha_2\,(1+\ln\, d)}$$

for some α_1 and α_2 independent of d and ε^{-1}. Therefore

$$n = \mathcal{O}\left(\exp\left(\mathcal{O}((1+\ln\,\varepsilon^{-1})(1+\ln\,d))\right)\right)$$

satisfies (9). Furthermore, the cardinality of $A_{d,n,\varepsilon}$ is bounded by

$$2d^k\, n = \mathcal{O}\left(\exp\left(\mathcal{O}((1+\ln\,\varepsilon^{-1})(1+\ln\,d))\right)\right).$$

- Case $k > d/2$.
  We now have $d \le 2k \le 2(1 + 2\ln(2\varepsilon^{-1})/\ln(\lambda_2^{-1})) = \mathcal{O}(1+\ln\,\varepsilon^{-1})$. We estimate $\sum_{\ell=0}^{k}\binom{d}{\ell}$ by $2^d = \exp(\mathcal{O}(1+\ln\,\varepsilon^{-1}))$. Then $2\alpha\, 2^d\, n^{-r} \le \varepsilon/2$ for

  $$n = \mathcal{O}\left(\exp\left(\mathcal{O}(1+\ln\,\varepsilon^{-1})\right)\right).$$

  Hence

  $$e(A_{d,n,\varepsilon}) \le \varepsilon$$

  and the cardinality of $A_{d,n,\varepsilon}$ is bounded by

  $$2^d\, n = \mathcal{O}\left(\exp\left(\mathcal{O}(1+\ln\,\varepsilon^{-1})\right)\right).$$

In both cases, $k \le d/2$ and $k > d/2$, we show that the error of the algorithm $A_{d,n,\varepsilon}$ is at most ε and the number of function values used by this algorithm is at most

$$\alpha_3 \exp\left(\alpha_4\,(1+\ln\,\varepsilon^{-1})(1+\ln\,d)\right)$$

for some α_3 and α_4 independent of ε^{-1} and d. This shows that the problem $\mathbb{S} = \{S_d\}$ is QPT. This also proves Theorem 1.

6 Final Comments

Let us assume, as in Theorem 1, that $\mathbb{S}$ is a non-zero linear tensor product problem with a compact linear S_1 for which

- $\lambda_2 < \lambda_1$,
- $\mathrm{decay}_e > 0$,

but the last condition is not fulfilled, i.e.,

$$\eta_1 \neq \pm K_1(t,t)^{-1/2} K_1(\cdot,t) \quad \text{for all} \quad t \in D.$$

Then, as we have seen, we cannot in general conclude QPT for the class Λ^{std}.

We can ask whether we can modify the problem somehow, by decreasing the class F_1, in order to obtain QPT for the modified (smaller) spaces. It turns out that this is possible. For notational convenience we assume again that $\lambda_1 = 1$.

Since η_1 is non-zero, there exists a point $t^* \in D$ such that $\eta_1(t^*) \neq 0$. Define

$$\widetilde{F}_1 = \{f \in F_1 \mid \langle f, \eta_1 \rangle_{F_1} = [\eta_1(t^*)]^{-1} f(t^*)\}.$$

Note that $\eta_1 \in \widetilde{F}_1$ and $\widetilde{F}_1$ is a linear subspace of F_1. Let

$$\widetilde{f} = \eta_1 - \frac{K_1(\cdot,t^*)}{\eta_1(t^*)}.$$

Clearly, $\widetilde{f} \in F_1$ and $\widetilde{f} \neq 0$. Then $\widetilde{F}_1$ can be rewritten as

$$\widetilde{F}_1 = \{f \in F_1 \mid \langle f, \widetilde{f} \rangle_{F_1} = 0\}.$$

It is easy to verify that the reproducing kernel $\widetilde{K}_1$ of $\widetilde{F}$ is

$$\widetilde{K}_1(x,y) = K_1(x,y) - \frac{\widetilde{f}(x)\widetilde{f}(y)}{\|\widetilde{f}\|^2} \quad \text{for all} \quad x,y \in D.$$

Furthermore, it is also easy to check that

$$\eta_1 = \widetilde{K}_1(t^*,t^*)^{-1/2}\,\widetilde{K}_1(\cdot,t^*).$$

The operator $\widetilde{S}_1 = S_1|_{\widetilde{F}_1}$, which is the restriction of S_1 to the subspace $\widetilde{F}_1$ satisfies all assumptions of Theorem 1. Indeed, let $\widetilde{\lambda}_n$ be the ordered eigenvalues of

$$\widetilde{W}_1 = \widetilde{S}_1^*\widetilde{S}_1 : \widetilde{F}_1 \to \widetilde{F}_1.$$

Since $\eta_1 \in \widetilde{F}_1$ we have $\widetilde{\lambda}_1 = \lambda_1 = 1$, whereas $\widetilde{\lambda}_n \le \lambda_n$ for all $n \ge 2$ since $\widetilde{F}_1 \subseteq F_1$. Therefore $\widetilde{\lambda}_2 < \widetilde{\lambda}_1$. Similarly, for both classes $\Lambda^{\rm all}$ and $\Lambda^{\rm std}$, the minimal worst case errors for $\widetilde{S}_1$ are no larger than the minimal worst case errors for S_1. Hence, applying Theorem 1 for $\widetilde{S}_1^{\otimes d}$, we conclude QPT for the modified problem.

References

1. Gnewuch, M., Woźniakowski, H.: Quasi-polynomial tractability. J. Complex. **27**, 312–330 (2011)
2. Hinrichs, A., Novak, E., Vybiral, J.: Linear information versus function evaluations for L_2-approximation. J. Approx. Theory **153**, 97–107 (2008)
3. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume I: Linear Information. European Mathematical Society Publishing House, Zürich (2008)
4. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. European Mathematical Society Publishing House, Zürich (2010)
5. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume III: Standard Information for Operators. European Mathematical Society Publishing House, Zürich (2012)
6. Novak, E., Woźniakowski, H.: Tractability of multivariate problems for standard and linear information in the worst case setting: Part I. J. Approx. Theory **207**, 177–192 (2016)
7. Smolyak, S.A.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Akad. Nauk SSSR **4**, 240–243 (1963), in Russian
8. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic Press, London (1988)
9. Wasilkowski, G.W., Woźniakowski, H.: Weighted tensor-product algorithms for linear multivariate problems. J. Complex. **15**, 402–447 (1999)

The Analysis of Vertex Modified Lattice Rules in a Non-periodic Sobolev Space

Dirk Nuyens and Ronald Cools

Dedicated to Ian H. Sloan's beautiful contributions to the existence and construction of lattice rules, on the occasion of his 80th birthday.

Abstract In a series of papers, in 1993, 1994 and 1996, Ian Sloan together with Harald Niederreiter introduced a modification of lattice rules for non-periodic functions, called "vertex modified lattice rules", and a particular breed called "optimal vertex modified lattice rules", see Numerical Integration IV (Birkhäuser 1993) pp. 253–265, J Comput Appl Math 51(1):57–70, 1994, and Comput Math Model 23(8–9):69–77, 1996. These are like standard lattice rules but they distribute the point at the origin to all corners of the unit cube, either by equally distributing the weight and so obtaining a multi-variate variant of the trapezoidal rule, or by choosing weights such that multilinear functions are integrated exactly. In the 1994 paper, Niederreiter and Sloan concentrate explicitly on Fibonacci lattice rules, which are a particular good choice of 2-dimensional lattice rules. Error bounds in this series of papers were given related to the star discrepancy.

In this paper we pose the problem in terms of the so-called unanchored Sobolev space, which is a reproducing kernel Hilbert space often studied nowadays in which functions have L_2-integrable mixed first derivatives. It is known constructively that randomly shifted lattice rules, as well as deterministic tent-transformed lattice rules and deterministic fully symmetrized lattice rules can achieve close to $O(N^{-1})$ convergence in this space, see Sloan et al. (Math Comput 71(240):1609–1640, 2002) and Dick et al. (Numer Math 126(2):259–291, 2014) respectively, where possible $\log^s(N)$ terms are taken care of by weighted function spaces.

We derive a break down of the worst-case error of vertex modified lattice rules in the unanchored Sobolev space in terms of the worst-case error in a Korobov space, a multilinear space and some additional "mixture term". For the 1-dimensional case

D. Nuyens (✉) · R. Cools
Department of Computer Science, KU Leuven, Leuven, Belgium
e-mail: dirk.nuyens@cs.kuleuven.be; ronald.cools@cs.kuleuven.be

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_43

this worst-case error is obvious and gives an explicit expression for the trapezoidal rule. In the 2-dimensional case this mixture term also takes on an explicit form for which we derive upper and lower bounds. For this case we prove that there exist lattice rules with a nice worst-case error bound with the additional mixture term of the form $N^{-1}\log^2(N)$.

1 Introduction

We study the numerical approximation of an s-dimensional integral over the unit cube

$$I(f) := \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

A (rank-1) *lattice rule* with N points in s dimensions is an equal weight cubature rule

$$Q(f;\boldsymbol{z},N) := \frac{1}{N}\sum_{k=0}^{N-1} f\left(\left\{\frac{\boldsymbol{z}k}{N}\right\}\right), \tag{1}$$

where $\boldsymbol{z} \in \mathbb{Z}^s$ is the *generating vector* of which the components are most often chosen to be relatively prime to N, and the curly braces $\{\cdot\}$ mean to take the fractional part componentwise. Clearly, as this is an equal weight rule, the constant function is integrated exactly. The classical theory, see [6, 14], is mostly concerned with periodic functions and then uses the fact that f can be expressed in an absolutely converging Fourier series to study the error. See also [11] for a recent overview of this "spectral" error analysis and its application to lattice rules. In this paper we only consider real-valued integrand functions.

In a series of papers [7–9] Niederreiter and Sloan introduced vertex modified lattice rules, and, more general, vertex modified quasi-Monte Carlo rules, to also cope with non-periodic functions. In this paper we revisit these vertex modified lattice rules using the technology of reproducing kernel Hilbert spaces, more precisely the unanchored Sobolev space of smoothness 1. The inner product for the one-dimensional unanchored Sobolev space is defined by

$$\langle f,g\rangle_{\mathrm{usob1},1,\gamma_1} := \int_0^1 f(x)\,\mathrm{d}x \int_0^1 g(x)\,\mathrm{d}x + \frac{1}{\gamma_1}\int_0^1 f'(x)\,g'(x)\,\mathrm{d}x, \tag{2}$$

where, more generally, γ_j is a "product weight" associated with dimension j, which is used to model the importance of different dimensions, see, e.g., [15]. In the multivariate case we take the tensor product such that the norm is defined by

$$\|f\|^2_{\mathrm{usob1},s,\boldsymbol{\gamma}} := \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}^{-1} \int_{[0,1]^{|\mathfrak{u}|}} \left(\int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}_{-\mathfrak{u}} \right)^2 \mathrm{d}\boldsymbol{x}_{\mathfrak{u}}$$

$$= \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}^{-1} \left\| \int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}_{-\mathfrak{u}} \right\|^2_{L_2}, \qquad (3)$$

with $\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j$. We use the short hand notation $\{1 : s\} = \{1, \ldots, s\}$ and thus in (3) $\mathfrak{u}$ ranges over all subsets of $\{1, \ldots, s\}$, and $-\mathfrak{u}$ is the complement with respect to the full set, $-\mathfrak{u} = \{1 : s\} \setminus \mathfrak{u}$. Note that (3) is a sum of L_2-norms of mixed first derivatives for all variables in $\mathfrak{u}$ where all other variables are averaged out.

2 Vertex Modified Lattice Rules

The *vertex modified lattice rule* proposed in [7] is given by

$$Q^{\mathrm{vm}}(f; \boldsymbol{z}, N, \boldsymbol{w}) = \sum_{\boldsymbol{a}\in\{0,1\}^s} w(\boldsymbol{a}) f(\boldsymbol{a}) + \frac{1}{N} \sum_{k=1}^{N-1} f\left(\left\{ \frac{\boldsymbol{z}k}{N} \right\} \right), \qquad (4)$$

with well chosen vertex weights $w(\boldsymbol{a})$ such that the constant function is still integrated exactly. It is assumed that $\gcd(z_j, N) = 1$, for all $j = 1, \ldots, s$, such that only the lattice point for $k = 0$ is on the edge of the domain $[0, 1]^s$, and this is why the second sum only ranges over $k = 1, \ldots, N-1$, i.e., the interior points. We note that typically N equals the number of function evaluations. This is not true anymore for vertex modified lattice rules. We define M to be the total number of function evaluations, and this is given by

$$M = 2^s + N - 1. \qquad (5)$$

The 2^s term makes us focus on the low-dimensional cases only, and we derive explicit results for $s = 2$ later. The vertex modified rule can then be represented as a standard cubature rule of the form

$$Q(f; \{(w_k, \boldsymbol{x}_k)\}_{k=1}^M) = Q(f) = \sum_{k=1}^{M} w_k f(\boldsymbol{x}_k), \qquad (6)$$

with appropriate choices for the pairs $(w_k, \boldsymbol{x}_k)$. For the vertex modified rules we only need to specify the weights at the vertices of the unit cube, all other remain unchanged from the standard lattice rule and are $1/N$.

Two particular choices for the weights $w(\boldsymbol{a})$ have been proposed [7–9]. The first one has constant weights $w(\boldsymbol{a}) \equiv 1/(2^s N)$ which mimics the trapezoidal rule in each one-dimensional projection:

$$T(f; z, N) := Q^{\mathrm{vm}}(f; z, N, \frac{1}{2^s N}) = \frac{1}{2^s N} \sum_{\boldsymbol{a} \in \{0,1\}^s} f(\boldsymbol{a}) + \frac{1}{N} \sum_{k=1}^{N-1} f\left(\left\{\frac{zk}{N}\right\}\right).$$

A second particular choice of weights $w^*(\boldsymbol{a})$ leads to the so-called *optimal vertex modified lattice rule* [7]:

$$Q^*(f; z, N) := Q^{\mathrm{vm}}(f; z, N, \boldsymbol{w}^*) = \sum_{\boldsymbol{a} \in \{0,1\}^s} w^*(\boldsymbol{a}) f(\boldsymbol{a}) + \frac{1}{N} \sum_{k=1}^{N-1} f\left(\left\{\frac{zk}{N}\right\}\right).$$

This rule integrates all multilinear polynomials exactly, i.e.,

$$Q^*(f; z, N) = Q^{\mathrm{vm}}(f; z, N, \boldsymbol{w}^*) = I(f) \quad \text{for all} \quad f(\boldsymbol{x}) = \prod_{j=1}^{s} x_j^{k_j} \quad \text{with } k_j \in \{0, 1\}.$$

There is no need to solve a linear system of equations to find the weights $w^*(\boldsymbol{a})$. The following result from [7] shows they can be determined explicitly.

Proposition 1 *For every $\boldsymbol{a} \in \{0,1\}^s$ define $\mathfrak{u}$ to be the support of $\boldsymbol{a}$, i.e., $\mathfrak{u} = \mathfrak{u}(\boldsymbol{a}) = \{1 \le j \le s : a_j \neq 0\}$. Then the weight $w^*(\boldsymbol{a})$ is given by*

$$w^*(\boldsymbol{a}) = w^*_{\mathfrak{u}} = \frac{1}{2^s} - \frac{1}{N} \sum_{k=1}^{N-1} \ell_{\mathfrak{u}}\left(\left\{\frac{zk}{N}\right\}\right) \quad \textit{where} \quad \ell_{\mathfrak{u}}(\boldsymbol{x}) := \prod_{j \in \mathfrak{u}} x_j \prod_{j \in \{1:s\} \setminus \mathfrak{u}} (1 - x_j).$$

Proof The idea is to use a kind of a Lagrange basis which is 0 in all vertex points $\boldsymbol{a} \in \{0,1\}^s$ except in one. For this purpose, consider the basis, for $\mathfrak{u} \subseteq \{1 : s\}$,

$$\ell_{\mathfrak{u}}(\boldsymbol{x}) = \prod_{j \in \mathfrak{u}} x_j \prod_{j \in \{1:s\} \setminus \mathfrak{u}} (1 - x_j)$$

such that $\ell_{\mathfrak{u}}(\boldsymbol{a}) = \mathbb{1}_{\mathfrak{u}(\boldsymbol{a}) = \mathfrak{u}}$. Demanding that $Q(\ell_{\mathfrak{u}}) = I(\ell_{\mathfrak{u}})$ for some $\mathfrak{u} \subseteq \{1 : s\}$, gives

$$\sum_{\boldsymbol{a} \in \{0,1\}^s} w^*(\boldsymbol{a})\, \ell_{\mathfrak{u}}(\boldsymbol{a}) + \frac{1}{N} \sum_{k=1}^{N-1} \ell_{\mathfrak{u}}\left(\left\{\frac{zk}{N}\right\}\right) = \int_{[0,1]^s} \ell_{\mathfrak{u}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

from where the result follows. □

3 Reproducing Kernel Hilbert Spaces

In this section we collect some well known results. For more details the reader is referred to, e.g., [2, 3, 5, 10, 11].

The reproducing kernel $K : [0,1] \times [0,1] \to \mathbb{R}$ of a one-dimensional reproducing kernel Hilbert space $\mathscr{H}(K)$ is a symmetric, positive definite function which has the reproducing property

$$f(y) = \langle f, K(\cdot, y) \rangle_K \quad \text{for all } f \in \mathscr{H}(K) \text{ and } y \in [0,1].$$

The induced norm in the space will be denoted by $\|f\|_K = \sqrt{\langle f, f \rangle_K}$. If the space has a countable basis $\{\varphi_h\}_h$ which is orthonormal with respect to the inner product of the space, then, by virtue of Mercer's theorem, the kernel is given by

$$K(x,y) = \sum_h \varphi_h(x)\, \overline{\varphi_h(y)}.$$

For the multivariate case we consider the tensor product space and the kernel is then given by

$$K_s(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} K(x_j, y_j).$$

We define the *worst-case error* of integration using a cubature rule Q to be

$$\mathrm{wce}(Q; K) := \sup_{\substack{f \in \mathscr{H}(K) \\ \|f\|_K \le 1}} |Q(f) - I(f)|.$$

For a general cubature formula $Q(f) = \sum_{k=1}^{M} w_k f(\boldsymbol{x}_k)$ the squared worst-case error can be written as, see, e.g., [5],

$$\mathrm{wce}(Q;K)^2 = \int_{[0,1]^{2s}} K(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} - 2 \sum_{k=1}^{M} w_k \int_{[0,1]^s} K(\boldsymbol{x}_k, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} + \sum_{k,\ell=1}^{M} w_k w_\ell\, K(\boldsymbol{x}_k, \boldsymbol{x}_\ell). \tag{7}$$

For all kernels in the remainder of the text we have that $\int_0^1 \int_0^1 K(x,y)\, \mathrm{d}x \mathrm{d}y = 1$ and $\int_0^1 K(x,y)\, \mathrm{d}y = 1$ for all $x \in [0,1]$ and this also holds for the multivariate kernel due to the product structure.

3.1 The Korobov Space

A well known example is the Korobov space which consists of periodic functions which can be expanded in an absolutely converging Fourier series. We refer the reader to the general references in the beginning of this section for further information on the Korobov space. Denote the Fourier coefficients by

$$\hat{f}(\boldsymbol{h}) := \int_{[0,1]^s} f(\boldsymbol{x}) \, \exp(-2\pi \mathrm{i}\, \boldsymbol{h} \cdot \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \boldsymbol{h} \in \mathbb{Z}^s.$$

In the one-dimensional case, if we assume an algebraic decay of $h^{-\alpha}$, $\alpha > 1/2$, by means of

$$\|f\|^2_{\mathrm{kor}\alpha,1,\gamma_1} := |\hat{f}(0)|^2 + \sum_{0 \neq h \in \mathbb{Z}} |\hat{f}(h)|^2 \, \gamma_1^{-1} |h|^{2\alpha} < \infty,$$

then the reproducing kernel is given by

$$K^{\mathrm{kor}\alpha}_{1,\gamma_1}(x,y) := 1 + \gamma_1 \sum_{0 \neq h \in \mathbb{Z}} \frac{\exp(2\pi \mathrm{i}\, h(x-y))}{|h|^{2\alpha}}.$$

We now specifically concentrate on the case $\alpha = 1$ as this will be of use throughout the paper. For $\alpha = 1$ the reproducing kernel for the s-variate case can be written as

$$K^{\mathrm{kor1}}_{s,\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{s} \left(1 + 2\pi^2 \gamma_j B_2(\{x_j - y_j\})\right) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \prod_{j \in \mathfrak{u}} 2\pi^2 \gamma_j B_2(\{x_j - y_j\}),$$

where $B_2(t) = t^2 - t + \frac{1}{6} = \frac{1}{2\pi^2} \sum_{0 \neq h \in \mathbb{Z}} \frac{\exp(2\pi \mathrm{i}\, ht)}{h^2}$, for $0 \leq t \leq 1$, is the 2nd degree Bernoulli polynomial and $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^s$ is a set of product weights which are normally used to model dimension importance. Here we will not make use of the weights $\boldsymbol{\gamma}$, except for scaling, such that the worst-case error of one space shows up in the worst-case error expression of another space.

For a general cubature rule $Q(f) = \sum_{k=1}^M w_k f(\boldsymbol{x}_k)$, with $\sum_{k=1}^M w_k = 1$, using (7) one obtains

$$\mathrm{wce}(Q; K^{\mathrm{kor1}}_{s,\boldsymbol{\gamma}})^2 = \sum_{k,\ell=1}^{M} w_k w_\ell \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \prod_{j \in \mathfrak{u}} 2\pi^2 \gamma_j B_2(\{x_{k,j} - x_{\ell,j}\}). \tag{8}$$

In case $Q(f) = Q(f; \boldsymbol{z}, N)$ is a lattice rule then the difference of two points is also a point of the point set and therefore the squared worst-case error formula simplifies to

$$\mathrm{wce}(Q(\cdot; \boldsymbol{z}, N); K^{\mathrm{kor1}}_{s,\boldsymbol{\gamma}})^2 = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \prod_{j \in \mathfrak{u}} 2\pi^2 \gamma_j B_2(x_{k,j}).$$

We remark that, apart from the higher cost, using a vertex modified lattice rule in the Korobov space makes no difference to the worst-case error,

$$\mathrm{wce}(Q^{\mathrm{vm}}(\cdot;\boldsymbol{z},N,\boldsymbol{w});K_{s,\boldsymbol{\gamma}}^{\mathrm{kor}\alpha}) = \mathrm{wce}(Q(\cdot;\boldsymbol{z},N);K_{s,\boldsymbol{\gamma}}^{\mathrm{kor}\alpha}), \tag{9}$$

since $K_{s,\boldsymbol{\gamma}}^{\mathrm{kor}\alpha}(\boldsymbol{a},\boldsymbol{0}) = K_{s,\boldsymbol{\gamma}}^{\mathrm{kor}\alpha}(\boldsymbol{0},\boldsymbol{0})$ for all $\boldsymbol{a} \in \{0,1\}^s$ and the weights $w(\boldsymbol{a})$ are such that they sum to $1/N$ due to the constraint of integrating the constant function exactly.

3.2 The Space of Multilinear Functions

Define the following multilinear functions, for $\mathfrak{u} \subseteq \{1:s\}$,

$$g_{\mathfrak{u}}(\boldsymbol{x}) := \prod_{j\in\mathfrak{u}} \sqrt{12}\,(x_j - \tfrac{1}{2}) = \prod_{j\in\mathfrak{u}} \sqrt{12}\,B_1(x_j),$$

so $g_{\emptyset}(\boldsymbol{x}) = 1, g_{\{1\}}(\boldsymbol{x}) = \sqrt{12}\,(x_1-\frac{1}{2})$ and so on, where $B_1(t) = t-\frac{1}{2}$ is the 1st degree Bernoulli polynomial. These functions form an orthonormal basis $\{g_{\mathfrak{u}}\}_{\mathfrak{u}\subseteq\{1:s\}}$ with respect to the standard L_2 inner product and we can thus construct a reproducing kernel for this finite dimensional space:

$$K_{s,\boldsymbol{\gamma}}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{y}) := \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}\, g_{\mathfrak{u}}(\boldsymbol{x})\, g_{\mathfrak{u}}(\boldsymbol{y}) = 1 + \sum_{\emptyset\neq\subseteq\{1:s\}} \prod_{j\in\mathfrak{u}} 12\,\gamma_j\, B_1(x_j)\, B_1(y_j),$$

where we introduced standard product weights. The worst-case error for a general cubature rule $Q(f) = \sum_{k=1}^{M} w_k f(\boldsymbol{x}_k)$, for which $\sum_{k=1}^{M} w_k = 1$, is given by

$$\mathrm{wce}(Q;K_{s,\boldsymbol{\gamma}}^{\mathrm{lin}})^2 = \sum_{k,\ell=1}^{M} w_k w_\ell \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:s\}} \prod_{j\in\mathfrak{u}} 12\,\gamma_j\,(x_{k,j}-\tfrac{1}{2})\,(x_{\ell,j}-\tfrac{1}{2}). \tag{10}$$

We remark that this space is not such an interesting space on its own. The one-point rule which samples at the point $(\frac{1}{2},\ldots,\frac{1}{2})$ has worst-case error equal to zero in this space, as can be seen immediately from (10). The worst-case error in this multilinear space will show up as part of the worst-case error in the Sobolev space that we will discuss next. Also note that, by construction, the optimal vertex modified lattice rule has

$$\mathrm{wce}(Q^*;K_{s,\boldsymbol{\gamma}}^{\mathrm{lin}}) = 0.$$

Naturally for $s = 1$ also $\mathrm{wce}(T;K_{1,\boldsymbol{\gamma}}^{\mathrm{lin}}) = 0$.

3.3 *The Unanchored Sobolev Space of Smoothness* 1

The reproducing kernel of the unanchored Sobolev space of smoothness 1 is given by

$$K_{s,\boldsymbol{\gamma}}^{\mathrm{usob1}}(\boldsymbol{x},\boldsymbol{y}) := \prod_{j=1}^{s} \left(1 + \gamma_j B_1(x_j) B_1(y_j) + \gamma_j \frac{B_2(\{x_j - y_j\})}{2}\right),$$

and the norm by (3). (The inner product is built as the tensor product based on the one-dimensional inner product (2).) We note that for functions from the Korobov space with $\alpha = 1$

$$\|f\|_{\mathrm{usob1},s,\boldsymbol{\gamma}} = \|f\|_{\mathrm{kor1},s,\boldsymbol{\gamma}/(2\pi)^2} \qquad \text{for all} \quad f \in \mathscr{H}(K_{s,\boldsymbol{\gamma}}^{\mathrm{kor1}}),$$

where $\boldsymbol{\gamma}/(2\pi)^2$ means all weights are rescaled by a factor of $1/(2\pi)^2$, which can easily be seen from the one-dimensional case using (2) and the Fourier series of f, see also [3].

Lattice rules were studied in the unanchored Sobolev space in [3] using the tent-transform and were shown to achieve $O(N^{-1})$ convergence rate without the need for random shifting as was previously known. A second approach in that paper used full symmetrisation of the point set (reflection around $\frac{1}{2}$ for each combination of dimensions; this is the generalization of the 1-point rule at $\frac{1}{2}$ for the multilinear space as discussed above, making sure all multilinear functions are integrated exactly). In a way we can look at vertex modified lattice rules Q^{vm} as being only the symmetrisation of the node $\mathbf{0}$ but with different weights. Using equal weights leads to the rule $T(\cdot; \boldsymbol{z}, N)$ which is the full symmetrisation of the point $\mathbf{0}$ (but does not necessarily integrate the multilinear functions exactly). For the rule $Q^{*}(\cdot; \boldsymbol{z}, N)$ the weights are chosen in a more intrinsic way such that they integrate multilinear functions exactly and we will concentrate our analysis on this rule.

4 Error Analysis

4.1 *Decomposing the Error for the Unanchored Sobolev Space*

We study the worst-case error of using a vertex modified lattice rule in the unanchored Sobolev space. First note

$$\begin{aligned} K_{s,\boldsymbol{\gamma}}^{\mathrm{usob1}}(\boldsymbol{x},\boldsymbol{y}) = 1 &+ \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \prod_{j \in \mathfrak{u}} \gamma_j B_1(x_j) B_1(y_j) + \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \prod_{j \in \mathfrak{u}} \gamma_j \frac{B_2(\{x_j - y_j\})}{2} \\ &+ \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \sum_{\emptyset \neq \mathfrak{v} \subset \mathfrak{u}} \prod_{j \in \mathfrak{u}} \gamma_j B_1(x_j) B_1(y_j) \prod_{j' \in \mathfrak{v}} \gamma_{j'} \frac{B_2(\{x_{j'} - y_{j'}\})}{2}. \end{aligned} \qquad (11)$$

From this the following break down of the worst-case error can be obtained.

Proposition 2 *The squared worst-case error for a general cubature rule* $Q(f) = \sum_{k=1}^{M} w_k f(\boldsymbol{x}_k)$, *with* $\sum_{k=1}^{M} w_k = 1$, *in the unanchored Sobolev space of smoothness* 1 *is given by*

$$\mathrm{wce}(Q; K^{\mathrm{usob1}}_{s,\boldsymbol{\gamma}})^2 = \mathrm{wce}(Q; K^{\mathrm{lin}}_{s,\boldsymbol{\gamma}/12})^2 + \mathrm{wce}(Q; K^{\mathrm{kor1}}_{s,\boldsymbol{\gamma}/(2\pi)^2})^2$$

$$+ \sum_{k,\ell=1}^{M} w_k w_\ell \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \sum_{\emptyset \neq \mathfrak{v} \subset \mathfrak{u}} \prod_{j \in \mathfrak{u} \setminus \mathfrak{v}} \gamma_j B_1(x_{k,j}) B_1(x_{\ell,j}) \prod_{j' \in \mathfrak{v}} \gamma_{j'} \frac{B_2(\{x_{k,j'} - x_{\ell,j'}\})}{2}.$$

Proof This can be found by direct calculation using (11) in (7) and comparing terms with the worst-case errors in the Korobov space (8) and the multilinear space (10). □

This means our worst-case error is constituted of the worst-case error in the multilinear space (with the weights scaled by $1/12$) and the worst-case error in the Korobov space of smoothness 1 (with the weights rescaled by $1/(2\pi)^2$) plus a "mixture term". For the optimal modified lattice rule Q^* the error in the multilinear space is zero. Additionally, the worst-case error in the Korobov space does not change for a vertex modified lattice rule as it just distributes the weight of the point $\mathbf{0}$ to the other vertices, but such that the sum of all vertex weights is still $1/N$, see (9).

Obviously, in only one dimension, the mixture term is not present as we cannot take both $\mathfrak{u}$ and $\mathfrak{v}$ non-empty, and then the worst-case error in the Sobolev space of smoothness 1 equals the worst-case error of the respective lattice rule in the Korobov space of smoothness 1 (with rescaled weights) when multilinear functions are integrated exactly. In two dimensions the mixture term can be rewritten into a nice form as we show in the next proposition which gives the worst-case errors for $s = 1$ and $s = 2$.

Proposition 3 *For* $s = 1$ *with any* $Q(f) = \sum_{k=1}^{M} w_k f(x_k)$, *where* $\sum_{k=1}^{M} w_k = 1$,

$$\mathrm{wce}(Q; K^{\mathrm{usob1}}_{1,\boldsymbol{\gamma}})^2 = \mathrm{wce}(Q; K^{\mathrm{lin}}_{1,\boldsymbol{\gamma}/12})^2 + \mathrm{wce}(Q; K^{\mathrm{kor1}}_{1,\boldsymbol{\gamma}/(2\pi)^2})^2.$$

Specifically the one-dimensional trapezoidal rule, $T(f) = \frac{1}{N}\sum_{k=1}^{N-1} f(k/N) + (f(0) + f(1))/(2N)$, *which is equal to the optimal vertex modified rule for* $s = 1$, *gives*

$$\mathrm{wce}(T; K^{\mathrm{usob1}}_{1,\boldsymbol{\gamma}}) = \mathrm{wce}(Q^*; K^{\mathrm{usob1}}_{1,\boldsymbol{\gamma}}) = \sqrt{\frac{\gamma_1}{12}}\,\frac{1}{N}.$$

For $s = 2$ we have for an optimal vertex modified lattice rule $Q^*(\cdot; \boldsymbol{z}, N)$, with $\gcd(z_1, N) = 1$ *and* $\gcd(z_2, N) = 1$,

$$\mathrm{wce}(Q^*; K^{\mathrm{usob1}}_{2,\boldsymbol{\gamma}})^2 = \mathrm{wce}(Q^*; K^{\mathrm{kor1}}_{2,\boldsymbol{\gamma}/(2\pi)^2})^2 + \frac{\gamma_1\gamma_2}{8\,\pi^2 N^2} \sum_{j\in\{1,2\}} \sum_{\substack{h\geq 1 \\ h\not\equiv 0 \pmod N}} \frac{\cot^2(\pi h w_j/N)}{h^2}, \tag{12}$$

where we have set $w_1 \equiv z_1^{-1} z_2 \pmod N$ *and* $w_2 \equiv z_2^{-1} z_1 \pmod N$, *such that* $w_2 \equiv w_1^{-1} \pmod N$. *Furthermore*

$$\mathrm{wce}(Q^*; K^{\mathrm{usob1}}_{2,\boldsymbol{\gamma}})^2 > \mathrm{wce}(Q^*; K^{\mathrm{kor1}}_{2,\boldsymbol{\gamma}/(2\pi)^2})^2 + \frac{\gamma_1\gamma_2}{8\,\pi^2 N^2} \sum_{j\in\{1,2\}} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{h^2}$$

$$\mathrm{wce}(Q^*; K^{\mathrm{usob1}}_{2,\boldsymbol{\gamma}})^2 < \mathrm{wce}(Q^*; K^{\mathrm{kor1}}_{2,\boldsymbol{\gamma}/(2\pi)^2})^2 + \frac{\gamma_1\gamma_2}{48\,N^2} \sum_{j\in\{1,2\}} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{h^2}.$$

Proof For $s = 1$ and the trapezoidal rule we see from Proposition 2 that we only need to consider the error for the space $K^{\mathrm{kor1}}_{1,\gamma_1/(2\pi)^2}$ since $T = Q^*$ for $s = 1$. So we need to look at the twofold quadrature of $B_2(\{x - y\})$. Since this function is periodic the trapezoidal rule T reduces to the standard lattice rule (1) such that

$$\mathrm{wce}(T; K^{\mathrm{kor1}}_{1,\boldsymbol{\gamma}/(2\pi)^2})^2 = \frac{\gamma_1}{N^2} \sum_{k,\ell=0}^{N-1} \frac{B_2((k-\ell \bmod N)/N)}{2} = \frac{\gamma_1}{N} \sum_{k=0}^{N-1} \frac{B_2(k/N)}{2} = \frac{\gamma_1}{12\,N^2}.$$

For $s = 2$ and a general cubature rule $Q(f) = \sum_{k=1}^{M} w_k f(\boldsymbol{x}_k)$ there are two 2-dimensional mixture terms in Proposition 2: for $j = 1, j' = 2$ and $j = 2, j' = 1$ we have

$$\sum_{k,\ell=1}^{M} w_k w_\ell\, \gamma_j B_1(x_{k,j}) B_1(x_{\ell,j})\, \gamma_{j'}\, \frac{B_2(\{x_{k,j'} - x_{\ell,j'}\})}{2} \tag{13}$$

$$= \frac{\gamma_j\,\gamma_{j'}}{(2\pi)^2} \sum_{k=1}^{M} w_k B_1(x_{k,j}) \sum_{\ell=1}^{M} w_\ell B_1(x_{\ell,j}) \sum_{0\neq h\in\mathbb{Z}} \frac{\exp(2\pi\mathrm{i}\, h(x_{k,j'} - x_{\ell,j'}))}{h^2}$$

$$= \frac{\gamma_j\,\gamma_{j'}}{(2\pi)^2} \sum_{0\neq h\in\mathbb{Z}} \frac{1}{h^2} \left[\sum_{k=1}^{M} w_k B_1(x_{k,j}) \exp(2\pi\mathrm{i}\, h x_{k,j'})\right] \left[\sum_{\ell=1}^{M} w_\ell B_1(x_{\ell,j}) \exp(-2\pi\mathrm{i}\, h x_{\ell,j'})\right]$$

$$= \frac{\gamma_j\,\gamma_{j'}}{(2\pi)^2} \sum_{0\neq h\in\mathbb{Z}} \frac{1}{h^2} \left|\sum_{k=1}^{M} w_k B_1(x_{k,j}) \exp(2\pi\mathrm{i}\, h x_{k,j'})\right|^2,$$

where we used the Fourier expansion of B_2 as given in Sect. 3.1. We now focus on the 2-dimensional cubature sum inside the modulus. For the optimal vertex modified lattice rule Q^* this cubature sum gives

$$\begin{aligned}
&\sum_{k=1}^{M} w_k\, B_1(x_{k,j})\, \exp(2\pi \mathrm{i}\, h x_{k,j'}) \\
&= \sum_{\boldsymbol{a}\in\{0,1\}^2} w^*(\boldsymbol{a})\, B_1(a_j)\, \exp(2\pi \mathrm{i}\, h a_{j'}) + \frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\left\{\frac{z_j k}{N}\right\}\right) \exp(2\pi \mathrm{i}\, h z_{j'} k/N).
\end{aligned}$$

In the first part the exponential disappears as $\exp(2\pi \mathrm{i}\, h a_{j'}) = 1$ for all $\boldsymbol{a} \in \{0,1\}^2$. Furthermore the whole sum over $\boldsymbol{a} \in \{0,1\}^2$ vanishes as, using $\gcd(z_j, N) = 1$,

$$\begin{aligned}
&Q^*(B_1(x_j); z, N) = 0 \\
&\qquad = \sum_{\boldsymbol{a}\in\{0,1\}^2} w^*(\boldsymbol{a})\, B_1(a_j) + \frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\left\{\frac{z_j k}{N}\right\}\right) = \sum_{\boldsymbol{a}\in\{0,1\}^2} w^*(\boldsymbol{a})\, B_1(a_j),
\end{aligned}$$

where the equality to zero follows from the exactness for multilinear functions and the sum over k vanishes due to symmetry. Thus, using Q^* and making use of the forthcoming Lemma 1 and the fact that $\gcd(z_{j'}, N) = 1$, we find, for $w_j = z_j^{-1} z_{j'} \bmod N$, with z_j^{-1} the multiplicative inverse of z_j modulo N,

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\left\{\frac{z_j k}{N}\right\}\right) \exp(2\pi \mathrm{i}\, h z_{j'} k/N) \\
&\qquad = \begin{cases} 0 & \text{when } h w_j \equiv 0 \pmod{N}, \\ -\mathrm{i}\cot(\pi h w_j/N)/(2N) & \text{otherwise.} \end{cases}
\end{aligned}$$

It thus follows that, for $Q = Q^*$, each mixture term takes the form

$$\begin{aligned}
&\frac{\gamma_j\,\gamma_{j'}}{(2\pi)^2} \sum_{0\neq h\in\mathbb{Z}} \frac{1}{h^2} \left| \sum_{k=1}^{M} w_k B_1(x_{k,j}) \exp(2\pi \mathrm{i}\, h x_{k,j'}) \right|^2 \\
&\qquad = \frac{\gamma_j\,\gamma_{j'}}{(4\pi)^2 N^2} \sum_{\substack{0\neq h\in\mathbb{Z} \\ h w_j \not\equiv 0 \pmod{N}}} \frac{\cot^2(\pi h w_j/N)}{h^2}.
\end{aligned}$$

Making use of $\gcd(w_j, N) = 1$ and using the sign-symmetry on the sum we obtain

$$\begin{aligned}
\frac{2\,\gamma_j\,\gamma_{j'}}{(4\pi)^2 N^2} \sum_{\substack{h\geq 1 \\ hw_j \not\equiv 0 \pmod N}} \frac{\cot^2(\pi h w_j/N)}{h^2} &= \frac{\gamma_j\,\gamma_{j'}}{8\,\pi^2 N^2} \sum_{\substack{h\geq 1 \\ h \not\equiv 0 \pmod N}} \frac{\cot^2(\pi h w_j/N)}{h^2} \\
&= \frac{\gamma_j\,\gamma_{j'}}{8\,\pi^2 N^2} \sum_{\ell\geq 0} \sum_{h=1}^{N-1} \frac{\cot^2(\pi(\ell N + h) w_j/N)}{(\ell N + h)^2} \\
&= \frac{\gamma_j\,\gamma_{j'}}{8\,\pi^2 N^2} \sum_{\ell\geq 0} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{(\ell N + h)^2} \\
&= \frac{\gamma_j\,\gamma_{j'}}{8\,\pi^2 N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{h^2} \sum_{\ell\geq 0} \frac{1}{(\ell N/h + 1)^2} \\
&< \frac{\gamma_j\,\gamma_{j'}}{8\,\pi^2 N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{h^2} \sum_{\ell\geq 1} \frac{1}{\ell^2} \\
&= \frac{\gamma_j\,\gamma_{j'}}{48\,N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w_j/N)}{h^2}.
\end{aligned}$$

For the upper bound we have set $h = N - 1$ in the sum over $\ell \geq 0$ and then used $N/(N-1) > 1$ and $\sum_{\ell\geq 1} \ell^{-2} = \pi^2/6$. The lower bound is easily derived from the same line by considering the case $\ell = 0$ only. □

It is a little bit unfortunate that the $\cot^2$-sum for both w_1 and w_2 appears in (12). We strongly believe that the infinite sum over h is the same for w_1 and w_2, and this is equivalent to obtaining the same value for (13). If this is true than also in the upper and lower bound we just remain with twice either of the sums. We verified the equality on (13) numerically for all $N \leq 4001$ and $z \in \{1, \ldots, N-1\}$ with $\gcd(z, N) = 1$ and could not find a counter example. Moreover in Corollary 1, forthcoming, we show equality to always hold in case of Fibonacci lattice rules. Therefore we make the following conjecture.

Conjecture 1 Given integers z and N, with $\gcd(z, N) = 1$, we have

$$\begin{aligned}
&\sum_{k,\ell=1}^{N-1} B_1(k/N)\, B_2((z(k-\ell) \bmod N)/N)\, B_1(\ell/N) \\
&\qquad = \sum_{k,\ell=1}^{N-1} B_1(k/N)\, B_2((z^{-1}(k-\ell) \bmod N)/N)\, B_1(\ell/N),
\end{aligned}$$

where z^{-1} is the multiplicative inverse of z modulo N.

The following lemma was used in the proof of Proposition 3 for the cubature sum of the linear Bernoulli polynomial in dimension j with a single exponential function in dimension j', taking $\theta = hz_{j'}$. The lemma is also valid for a product of exponential functions which in the case of lattice rules would give $\theta = \boldsymbol{h}_{\mathfrak{u}} \cdot \boldsymbol{z}_{\mathfrak{u}}$ for some $\mathfrak{u} \subset \{1 : s\}$.

Lemma 1 *For $\theta \in \mathbb{Z}$ and $\gcd(z_j, N) = 1$, denote by z_j^{-1} the multiplicative inverse of z_j modulo N, then*

$$\frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\left\{\frac{z_j k}{N}\right\}\right)\exp(2\pi \mathrm{i}\,\theta\,k/N) = \begin{cases} 0, & \text{if } \theta \equiv 0 \pmod N, \\ \dfrac{-\mathrm{i}}{2N}\cot(\pi z_j^{-1}\theta/N), & \text{otherwise.} \end{cases}$$

Proof With $a = \exp(2\pi \mathrm{i}\, z_j^{-1}\theta/N)$ and $z_j^{-1}\theta \not\equiv 0 \pmod N$ we have

$$\frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\frac{k}{N}\right) a^k = -\frac{1}{2N}\sum_{k=1}^{N-1} a^k + \frac{1}{N}\sum_{k=1}^{N-1}\frac{k}{N}a^k,$$

where $\sum_{k=1}^{N-1} a^k = -1$ as $a^N = 1$. Now using

$$\sum_{k=1}^{N-1}\frac{k}{N}(f(k+1) - f(k)) = -\frac{1}{N}\sum_{k=1}^{N-1} f(k) + \frac{N-1}{N}f(N),$$

and, for $a \neq 1$,

$$a^k = \frac{a^{k+1}}{a-1} - \frac{a^k}{a-1}$$

we find

$$\sum_{k=1}^{N-1}\frac{k}{N}a^k = -\frac{1}{N}\frac{1}{a-1}\sum_{k=1}^{N-1} a^k + \frac{N-1}{N}\frac{a^N}{a-1},$$

where again $a^N = 1$ and $\sum_{k=1}^{N-1} a^k = -1$. Thus

$$\frac{1}{N}\sum_{k=1}^{N-1} B_1\left(\frac{k}{N}\right) a^k = \frac{1}{2N} + \frac{1}{N}\frac{1}{a-1}.$$

The proof is then completed by taking $t = \pi z_j^{-1}\theta/N$ in the identity $-\mathrm{i}\cot(t) = 1 + 2/(\exp(2\mathrm{i}\,t) - 1)$. □

4.2 Upper and Lower Bound

In Proposition 3 we already obtained an upper and a lower bound on $\mathrm{wce}(Q^*; K_{2,\boldsymbol{\gamma}}^{\mathrm{usob1}})^2$, but they were in terms of the sum

$$\frac{1}{N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{h^2}, \tag{14}$$

with $\gcd(w, N) = 1$. In fact also the sum with w^{-1}, the multiplicative inverse of w modulo N, should be considered if Conjecture 1 is false. If the conjecture would be false then this can be fixed in the end by assuming N to be large enough (see the remark after Proposition 4). Note that the sum is 1-periodic in $t = w/N$ as well as having the symmetry $\cot^2(\pi t) = \cot^2(\pi(1-t)) = \cot^2(-\pi t)$.

Below we will use the series

$$H_N(a) := \sum_{h=1}^{N} \frac{1}{h^a}, \tag{15}$$

where we consider $a \geq 1$. This is known as the harmonic number of N of order a. If we set $N = \infty$ we get the Riemann zeta function

$$\zeta(a) := \sum_{h=1}^{\infty} \frac{1}{h^a}, \tag{16}$$

which is finite for $a > 1$. Since $\zeta(1) = \infty$ we can look at how $H_N(1)$ increases. For $N \geq 3$ we have

$$H_N(1) \leq \frac{11}{6 \log(3)} \log(N). \tag{17}$$

The above elementary bound follows from the definition of the Euler–Mascheroni constant $\lim_{N\to\infty} H_N(1) - \log(N) \approx 0.5772$, which converges monotonically from above. Solving $H_3(1) = c \log(3)$ results in (17) for $N \geq 3$. We will also make use of the following identity

$$\frac{1}{N-1} \sum_{w=1}^{N-1} \cot^2(\pi w/N) = \frac{N-2}{3}, \tag{18}$$

which can be seen by the closed form solution of the Dedekind sum $S(z, N)$ with $z = 1$.

The standard approach to show existence of a good generating vector is to prove a good upper bound for the average over all possible generating vectors. We first show a general lower bound and then an upper bound for the average choice of generating vector on the above $\cot^2$-sum.

Lemma 2 *For $N \geq 3$ and any choice of w such that $\gcd(w, N) = 1$, the following lower bound holds:*

$$\frac{1}{N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{h^2} > \frac{1}{6N^2}.$$

Proof We have $\{hw \bmod N : h \in \{1, \ldots, N-1\}\} = \{1, \ldots, N-1\}$ since $\gcd(w, N) = 1$, thus

$$\sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{h^2} > \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{(N-1)^2} = \sum_{h=1}^{N-1} \frac{\cot^2(\pi h/N)}{(N-1)^2} = \frac{N-2}{3(N-1)},$$

where we used (18). □

The previous lemma shows that we cannot expect the worst-case error to be better behaving than $1/N$ which is not a surprise as this is the expected convergence for 1D. We now check what happens if we uniformly pick a w from $\{1, \ldots, N-1\}$ for prime $N \geq 3$. Surprisingly this can be calculated exactly.

Lemma 3 *For a prime $N \geq 3$, the average over $w \in \{1, \ldots, N-1\}$ of the* $\cot^2$*-sum* (14) *is given by*

$$\frac{1}{N-1} \sum_{w=1}^{N-1} \frac{1}{N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{h^2} = \frac{N-2}{3N^2} H_{N-1}(2) \quad \leq \quad \frac{\pi^2}{18N}.$$

Proof Since N is prime we have $\gcd(h, N) = 1$ and thus $\{hw \bmod N : w \in \{1, \ldots, N-1\}\} = \{1, \ldots, N-1\}$. Therefore

$$\begin{aligned} \frac{1}{N-1} \sum_{w=1}^{N-1} \frac{1}{N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h w/N)}{h^2} &= \frac{1}{N^2} \sum_{h=1}^{N-1} \frac{1}{h^2} \frac{1}{N-1} \sum_{w=1}^{N-1} \cot^2(\pi w/N) \\ &= \frac{N-2}{3N^2} \sum_{h=1}^{N-1} \frac{1}{h^2} \leq \frac{\zeta(2)}{3N} = \frac{\pi^2}{18N}, \end{aligned}$$

where we used (18). □

Unfortunately the above result only allows us to say that the expected worst-case error is only as good as the Monte Carlo rate of $N^{-1/2}$ (since the sum (14) appears in the squared worst-case error). To get a better bound we need another approach. If we pick the w which gives the best possible value for the *square root* of the sum (14) then this will also be the best value for the sum directly. Furthermore, using the following inequality, often called “Jensen’s” inequality, we have

$$\left(\frac{1}{N^2}\sum_{h=1}^{N-1}\frac{\cot^2(\pi hw/N)}{h^2}\right)^{1/2} \le \frac{1}{N}\sum_{h=1}^{N-1}\frac{|\cot(\pi hw/N)|}{h}. \tag{19}$$

We now use a popular trick in proving existence: the value for the best choice w^* to minimize (either side of) (19) will be at least as small as the average over all possible choices of w, thus

$$\frac{1}{N}\sum_{h=1}^{N-1}\frac{|\cot(\pi hw^*/N)|}{h} \le \frac{1}{N-1}\sum_{w=1}^{N-1}\frac{1}{N}\sum_{h=1}^{N-1}\frac{|\cot(\pi hw/N)|}{h}.$$

The next lemma will give an upper bound for the right hand side above. The argument that the best choice will be at least as good as the average is used numerous times in Ian Sloan’s work and is also used inductively in component-by-component algorithms to construct lattice rules achieving nearly the optimal convergence order, see, e.g., [2, 16, 17].

Lemma 4 *For a prime $N \ge 3$, the average over $w \in \{1, \ldots, N-1\}$ of the $|\cot|$-sum in* (19) *is given by*

$$\frac{1}{N-1}\sum_{w=1}^{N-1}\frac{1}{N}\sum_{h=1}^{N-1}\frac{|\cot(\pi hw/N)|}{h} \le \frac{H_{N-1}(1)}{N}\frac{6}{\pi}\log(N).$$

Proof Similar as in the proof of Lemma 3 we use the fact that the multiplicative inverse of h exists and we can thus just look at the sum over w. For $N \ge 3$

$$\begin{aligned}\frac{1}{N-1}\sum_{w=1}^{N-1}|\cot(\pi w/N)| &= \frac{2}{N-1}\left[\cot(\pi/N) + \sum_{w=2}^{(N-1)/2}\cot(\pi w/N)\right]\\ &\le \frac{2}{N-1}\left[\cot(\pi/N) + \int_{1/N}^{(N-1)/(2N)}\cot(\pi t)\,N\,\mathrm{d}t\right]\\ &= \frac{2}{N-1}\left[\cot(\pi/N) + \frac{N}{\pi}\left(-\log(2\sin(\pi/(2N)))\right)\right]\\ &< \frac{2.2}{\pi}(1+\log(4N/3))\\ &< \frac{3}{\pi}\log(3N),\end{aligned}$$

where we used $2/\sin(\pi/(2N)) \le 4N/3$ for $N \ge 3$, with equality for $N = 3$, and some elementary bounds. □

We can now combine the previous results in estimating an upper bound for the worst-case error of a good choice of w for the optimal vertex modified rule Q^* for $s = 2$. From Proposition 3, again using Jensen's inequality by taking square-roots on both sides, we obtain

$$\mathrm{wce}(Q^*; K_{2,\boldsymbol{\gamma}}^{\mathrm{usob1}}) < \mathrm{wce}(Q^*; K_{2,\boldsymbol{\gamma}/(2\pi)^2}^{\mathrm{kor1}}) + \frac{\sqrt{\gamma_1\gamma_2}}{\sqrt{48}\,N} \sum_{j\in\{1,2\}} \sum_{h=1}^{N-1} \frac{|\cot(\pi h w_j/N)|}{h} \tag{20}$$

where the sum over j could be replaced by $\sqrt{2}$ if Conjecture 1 is true.

Piecing everything together we obtain the following result.

Proposition 4 *Given a sufficiently large prime N, then there exist $w \in \{1, \ldots, N-1\}$ such that the optimal vertex modified rule Q^*, with generating vector $\boldsymbol{z} = (1, w)$, has worst-case error in the unanchored Sobolev space for $s = 2$ of*

$$\mathrm{wce}(Q^*; K_{2,\boldsymbol{\gamma}}^{\mathrm{usob1}}) < \mathrm{wce}(Q^*; K_{2,\boldsymbol{\gamma}/(2\pi)^2}^{\mathrm{kor1}}) + \frac{11\sqrt{2}\,\gamma_1\gamma_2}{\pi\sqrt{48}\log 3} \frac{\log^2(N)}{N}.$$

If Conjecture 1 is true then sufficiently large can be replaced by a prime $N \geq 3$.

Proof From Proposition 3 we obtain (20) and combine this with Eq. (17) and Lemma 4. □

It is well known that there exist lattice rules for the Korobov space of order 1 which have convergence $N^{-1+\delta}$ for $\delta > 0$, see, e.g., [2, 14]. The question of finding a good optimal vertex modified rule for the unanchored Sobolev space now boils down to having N large enough such that the set of good w for the Korobov space and the set of good w for the $|\cot|$-sum overlap. This is done by showing there exist at least $N/2$ good choices that satisfy twice the average and then necessarily these two sets overlap. We will not digress here. See, e.g., [1] for such a technique. Similarly, if the conjecture is not true, then the same technique can be applied by taking N large enough such that all three good sets overlap and one obtains the desired convergence.

4.3 Fibonacci Lattice Rules

In [8], Niederreiter and Sloan turn to Fibonacci lattice rules as it is well known they perform best possible in view of many different quality criteria for numerical integration in two dimensions, see, e.g., [6]. The Fibonacci numbers can be defined recursively by $F_0 = 0$, $F_1 = 1$ and $F_k = F_{k-1} + F_{k-2}$ for $k \geq 2$. A Fibonacci lattice rule then takes the number of points a Fibonacci number $N = F_k$ and the generating vector $\boldsymbol{z} = (1, F_{k-1})$, for $k \geq 3$.

We can now show that Conjecture 1 is true for the explicit case of Fibonacci lattice rules.

Lemma 5 *For $z = F_{k-1}$ or F_{k-2} and $N = F_k$, $k \geq 3$, we have* $\gcd(z, N) = 1$, *and*

$$\sum_{k,\ell=1}^{N-1} B_1(k/N)\, B_2((z(k-\ell) \bmod N)/N)\, B_1(\ell/N) = \sum_{k,\ell=1}^{N-1} B_1(k/N)\, B_2((z^{-1}(k-\ell) \bmod N)/N)\, B_1(\ell/N),$$

where z^{-1} is the multiplicative inverse of z modulo N.

Proof It is known that $F_{k-1}^{-1} \equiv \pm F_{k-1} \pmod{F_k}$ with a plus sign for k even and a minus sign for k odd. The result follows from the symmetry $B_2(t) = B_2(1-t)$ for $0 \leq t \leq 1$. □

Combining Lemma 5 with Proposition 3 gives then an exact expression for the worst-case error in case of Fibonacci lattice rules. Note that N does not need to be prime for this proof.

Corollary 1 *For Q_k^* an optimal vertex modified lattice rule based on a Fibonacci lattice rule with generator $(1, F_{k-1})$ modulo F_k, $k \geq 4$, we have*

$$\mathrm{wce}(Q_k^*; K_{2,\boldsymbol{\gamma}/(2\pi)^2}^{\mathrm{usob1}})^2 = \mathrm{wce}(Q_k^*; K_{2,\boldsymbol{\gamma}}^{\mathrm{kor1}})^2 + \frac{\gamma_1 \gamma_2}{4\pi^2 N^2} \sum_{\substack{h \geq 1 \\ h \not\equiv 0 \pmod{N}}} \frac{\cot^2(\pi h w/N)}{h^2}$$

where $w = F_{k-1}$ and $N = F_k$.

5 Numerics and a Convolution Algorithm

In this section we restrict ourselves to N prime. Similar in spirit as [11, 12] it is possible to evaluate the sum

$$S_N(z/N) := \frac{1}{N^2} \sum_{h=1}^{N-1} \frac{\cot^2(\pi h z/N)}{h^2}$$

for all $z \in \{1, \ldots, N-1\}$ simultaneously by a (fast) convolution algorithm. Take a generator for the cyclic group $\mathbb{Z}_N^\times := \{1, \ldots, N-1\} = \langle g \rangle$ and represent $z = \langle g^\beta \rangle$

Table 1 Optimal choices of generating vector $(1,z)$ for a selection of prime N for the unanchored Sobolev space of order 1

N	z	$\mathrm{wce}^2(Q^*, K^{\mathrm{usob1}}_{2,\mathbf{1}})$	$= \mathrm{wce}^2(Q^*, K^{\mathrm{kor1}}_{2,\mathbf{1}/(2\pi)^2})$	$+$ mixingterm
17	5	$2.16 \cdot 10^{-3}$	$1.92 \cdot 10^{-3}$	$2.39 \cdot 10^{-4}$
37	11	$5.33 \cdot 10^{-4}$	$4.57 \cdot 10^{-4}$	$7.63 \cdot 10^{-5}$
67	18	$1.73 \cdot 10^{-4}$	$1.46 \cdot 10^{-4}$	$2.66 \cdot 10^{-5}$
131	76	$4.67 \cdot 10^{-5}$	$3.92 \cdot 10^{-5}$	$7.47 \cdot 10^{-6}$
257	76	$1.37 \cdot 10^{-5}$	$1.12 \cdot 10^{-5}$	$2.47 \cdot 10^{-6}$
521	377	$3.48 \cdot 10^{-6}$	$2.83 \cdot 10^{-6}$	$6.48 \cdot 10^{-7}$
1031	743	$9.75 \cdot 10^{-7}$	$7.81 \cdot 10^{-7}$	$1.94 \cdot 10^{-7}$
2053	794	$2.70 \cdot 10^{-7}$	$2.13 \cdot 10^{-7}$	$5.70 \cdot 10^{-8}$
4099	2511	$7.06 \cdot 10^{-8}$	$5.53 \cdot 10^{-8}$	$1.53 \cdot 10^{-8}$
8209	3392	$1.88 \cdot 10^{-8}$	$1.46 \cdot 10^{-8}$	$4.19 \cdot 10^{-9}$
16,411	6031	$4.82 \cdot 10^{-9}$	$3.73 \cdot 10^{-9}$	$1.09 \cdot 10^{-9}$
32,771	20,324	$1.26 \cdot 10^{-9}$	$9.71 \cdot 10^{-10}$	$2.91 \cdot 10^{-10}$
65,537	25,016	$3.34 \cdot 10^{-10}$	$2.55 \cdot 10^{-10}$	$7.90 \cdot 10^{-11}$
131,101	80,386	$8.97 \cdot 10^{-11}$	$6.79 \cdot 10^{-11}$	$2.18 \cdot 10^{-11}$
262,147	159,921	$2.30 \cdot 10^{-11}$	$1.74 \cdot 10^{-11}$	$5.64 \cdot 10^{-12}$

and $h = \langle g^{-\gamma} \rangle$, where $\langle \cdot \rangle$ denotes calculation modulo N. Then consider for all $0 \le \beta \le N-2$

$$S_N(\langle g^{\beta} \rangle) = \frac{1}{N^2} \sum_{\gamma=0}^{N-2} \frac{\cot^2(\pi \langle g^{\beta-\gamma} \rangle / N)}{\langle g^{-\gamma} \rangle^2}.$$

This is the cyclic convolution of two length $N-1$ vectors and can be calculated by an FFT algorithm. In Table 1 we show the best choice of z obtained by this method and the associated squared worst-case errors. Instead of using h^2 in the denominator of S_N we actually used a generalized zeta function $\zeta(2, h/N)/N^2$ which is the exact value of the infinite sum in Proposition 3. These results are plotted in Fig. 1. Several reference lines have been superimposed with different powers of $\log(N)$. We note that for this range of N the $\log^2(N)/N$, see Proposition 4, seems to be an overestimate for the square root of the mixing term. On the other hand, from the figure we see that the total error for this range of N behaves like $\log^{1/2}(N)/N$ for all practical purposes. It is interesting to compare this behavior with the results in [4, 18] which shows a different algorithm for modifying two-dimensional quasi-Monte Carlo point sets to the non-periodic setting (with $M = 5N-2$, while here we have $M = N+3$ for 2D) where an upper bound of $\log^{1/2}(N)/N$ is shown (which is also proven to be the lower bound there).

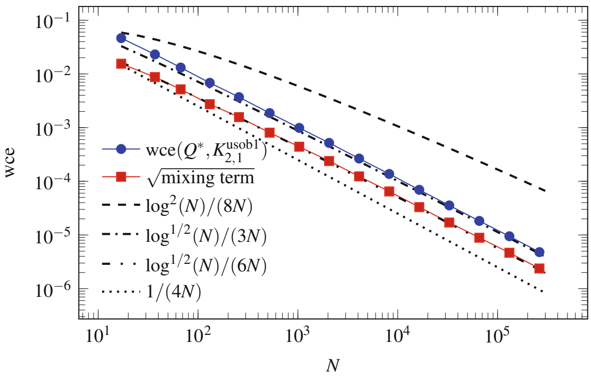


Fig. 1 Plot of optimal worst-case error and square root of mixture term from Table 1

6 Conclusion

In this paper we revisited (optimal) vertex modified lattice rules [7–9] introduced by Niederreiter and Sloan, and studied their error in the unanchored Sobolev space which is one of the typical reproducing kernel Hilbert spaces used to study lattice rules nowadays. The analysis makes use of a breakdown of the squared worst-case error into the squared worst-case error in a multilinear space, the Korobov space and an additional "mixture" term where combinations of basis functions from those two previous spaces appear. For $s = 2$ we showed that there exist optimal vertex modified lattice rules for which the square root of the mixture term converges like $N^{-1} \log^2(N)$. Because of the 2^s cost of evaluating the integrand on all vertices of the unit cube, it does not look very interesting to extend the analysis to an arbitrary number of dimensions. Although we restricted our detailed analysis to the case $s = 2$, we remark that a similar breakdown was achieved in terms of the L_2 discrepancy in [13], which shows that the cost of 2^s vertices still pays off for $s < 12$ in their numerical tests. Such tests would also be useful for the analysis in this paper, as would a component-by-component algorithm for $s > 2$. Finally, a comparison with the bounds in [4, 18] suggests the power of the $\log(N)$ term could be improved, as is hinted at by our numerical results. These are suggestions for future work.

Acknowledgements We thank Jens Oettershagen for useful comments and pointers to [4, 18]. We also thank the Taiwanese National Center for Theoretical Sciences (NCTS)—Mathematics Division, and the National Taiwan University (NTU)—Department of Mathematics, where part of this work was carried out. We thank the referees for their helpful comments and acknowledge financial support from the KU Leuven research fund (OT:3E130287 and C3:3E150478).

References

1. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing lattice rules based on weighted degree of exactness and worst case error. Computing **87**(1–2), 63–89 (2010)
2. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
3. Dick, J., Nuyens, D., Pillichshammer, F.: Lattice rules for nonperiodic smooth integrands. Numer. Math. **126**(2), 259–291 (2014)
4. Dũng, D., Ullrich, T.: Lower bounds for the integration error for multivariate functions with mixed smoothness and optimal Fibonacci cubature for functions on the square. Math. Nachr. **288**, 743–762 (2015)
5. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**(221), 299–322 (1998)
6. Niederreiter, H.: Random number generation and quasi-Monte Carlo methods. In: Regional Conference Series in Applied Mathematics, vol. 63, SIAM, Philadelphia (1992)
7. Niederreiter, H., Sloan, I.H.: Quasi-Monte Carlo methods with modified vertex weights. In: Brass, H., Hämmerlin, G. (eds.) Numerical Integration IV (Oberwolfach, 1992), International Series of Numerical Mathematics, pp. 253–265. Birkhäuser, Basel (1993)
8. Niederreiter, H., Sloan, I.H.: Integration of nonperiodic functions of two variables by Fibonacci lattice rules. J. Comput. Appl. Math. **51**(1), 57–70 (1994)
9. Niederreiter, H., Sloan, I.H.: Variants of the Koksma–Hlawka inequality for vertex-modified Quasi-Monte Carlo integration rules. Comput. Math. Model. **23**(8–9), 69–77 (1996)
10. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems — Volume I: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society Publishing House, Zürich (2008)
11. Nuyens, D.: The construction of good lattice rules and polynomial lattice rules. In: Kritzer, P., Niederreiter, H., Pillichshammer, F., Winterhof, A. (eds.) Radon Series on Computational and Applied Mathematics, pp. 223–256. De Gruyter, Berlin (2014)
12. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. Math. Comput. **75**(254), 903–920 (2006)
13. Reddy, M.V., Joe, S.: An average discrepancy for optimal vertex-modified number-theoretic rules. Adv. Comput. Math. **12**(1), 59–69 (2000)
14. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford Science Publications, Oxford (1994)
15. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complex. **14**(1), 1–33 (1998)
16. Sloan, I.H., Reztsov, A.V.: Component-by-component construction of good lattice rules. Math. Comput. **71**(237), 263–273 (2002)
17. Sloan, I.H., Kuo, F.Y., Joe, S.: On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces. Math. Comput. **71**(240), 1609–1640 (2002)
18. Ullrich, T.: Optimal cubature in Besov spaces with dominating mixed smoothness on the unit square. J. Complex. **30**, 72–94 (2014)

Matching Schur Complement Approximations for Certain Saddle-Point Systems

John W. Pearson and Andy Wathen

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The solution of many practical problems described by mathematical models requires approximation methods that give rise to linear(ized) systems of equations, solving which will determine the desired approximation. This short contribution describes a particularly effective solution approach for a certain class of so-called saddle-point linear systems which arises in different contexts.

1 Introduction

Iterative methods are now widely used in various applications for the solution of linear(ized) systems of equations. A key aspect is preconditioning [38]. Without appropriate preconditioners, convergence can be unacceptably slow, whereas an effective preconditioner can *enable* the solution of matrix systems of vast dimension, and thus allow large scale computational modelling.

There continues important work on algebraic preconditioners—preconditioners which require only the entries of a (sparse) matrix for construction; triangular factorization remains an important paradigm, and algebraic multigrid techniques are finding ever wider application. However, it is now keenly realised that preconditioners which exploit matrix structures often have considerable utility. In particular, state-of-the-art preconditioners for so-called saddle-point systems [5] have found application in many areas [6, 7, 13, 15, 17, 20].

J. W. Pearson
University of Edinburgh, Edinburgh, UK
e-mail: j.pearson@ed.ac.uk

A. Wathen (✉)
Oxford University, Oxford, UK
e-mail: wathen@maths.ox.ac.uk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_44

In this short contribution, we examine matrices with a particular type of saddle-point structure, namely

$$\underbrace{\begin{bmatrix} \alpha^2 A & 0 & B^T \\ 0 & \gamma^2 A & -A \\ B & -A & 0 \end{bmatrix}}_{\mathscr{A}_3} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \\ \mathbf{h} \end{bmatrix}, \qquad (1)$$

where $A, B \in \mathbb{R}^{n\times n}$, with A being symmetric and invertible, and with α, γ non-zero (real) parameters. We survey applications which give rise to equations of this form in Sect. 4 below, and believe the methodology presented could be applied to areas of research other than those specifically mentioned.

In fact, by simple block elimination, it is easily seen that (1) is equivalent to the 2×2 block system

$$\underbrace{\begin{bmatrix} \alpha^2 A & B^T \\ B & -\gamma^{-2} A \end{bmatrix}}_{\mathscr{A}_2} \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{h} + \gamma^{-2}\mathbf{g} \end{bmatrix},$$

with $\mathbf{v} = \gamma^{-2}\mathbf{w} + \gamma^{-2}A^{-1}\mathbf{g}$.

Further block elimination leads to the equivalent "1×1 block system"—the Schur complement system

$$\underbrace{\left(\gamma^{-2}A + \alpha^{-2}BA^{-1}B^T\right)}_{S} \mathbf{w} = \alpha^{-2}BA^{-1}\mathbf{f} - \mathbf{h} - \gamma^{-2}\mathbf{g}.$$

In this case $\mathbf{u} = \alpha^{-2}A^{-1}\mathbf{f} - \alpha^{-2}A^{-1}B^T\mathbf{w}$, and $\mathbf{v}$ can be recovered as above.

As an alternative, one may decompose the 2×2 block system to write

$$\underbrace{\left(\alpha^2 A + \gamma^2 B^T A^{-1} B\right)}_{S_1} \mathbf{u} = \mathbf{f} + \gamma^2 B^T A^{-1}\mathbf{h} + B^T A^{-1}\mathbf{g},$$

and then recover $\mathbf{w} = \gamma^2 A^{-1}B\mathbf{u} - \gamma^2 A^{-1}\mathbf{h} - A^{-1}\mathbf{g}$ and $\mathbf{v}$ as above.

The equivalence of these 3×3, 2×2 and 1×1 block systems is well known—see, for example, [16]—and, via the result of [19, 21], the solution of any of them crucially depends on having a good approximation for the Schur complement matrix $S = \gamma^{-2}A + \alpha^{-2}BA^{-1}B^T$ or S_1. Approximations $\widehat{S}$ for which the eigenvalues of $\widehat{S}^{-1}S$ do not depend on the parameters α, γ, or on any implicit parameters (such as mesh size) which arise in A, B, are particularly valuable since they lead to iterative solvers which converge in a number of iterations independent of all such parameters, as we shall demonstrate.

Given the 3 × 3 block system, use of a block diagonal preconditioner

$$\mathscr{P}_3 = \begin{bmatrix} \alpha^2 A & 0 & 0 \\ 0 & \gamma^2 A & 0 \\ 0 & 0 & S \end{bmatrix}$$

allows the solution of (1) in exactly 3 iterations using the Krylov subspace iteration method MINRES [21]. Similarly, given the 2 × 2 block system, use of

$$\mathscr{P}_2 = \begin{bmatrix} \alpha^2 A & 0 \\ 0 & S \end{bmatrix}$$

and MINRES [22] again is guaranteed to yield the solution for any right hand side vector in 3 iterations. In either case, replacing S with a $\widehat{S}$ for which the eigenvalues of $\widehat{S}^{-1}S$ do not depend on any problem parameters yields solvers based on MINRES which require not 3, but just a few more iteration steps, and still a number independent of the parameters α, γ and the problem dimension. For the 1 × 1 system, the Conjugate Gradient method can also be employed effectively with $\widehat{S}$ as a preconditioner. These guarantees will be described below, but we first describe in generality a 'matching Schur complement approximation' for which the required parameter-independent eigenvalues are guaranteed.

2 Matching Schur Complement Approximation

By simple calculation it is seen that

$$S := \gamma^{-2}A + \alpha^{-2}BA^{-1}B^T = \widehat{S} - \alpha^{-1}\gamma^{-1}(B + B^T),$$

where $\widehat{S} := \left(\gamma^{-1}A + \alpha^{-1}B\right) A^{-1} \left(\gamma^{-1}A + \alpha^{-1}B\right)^T$. The original motivation for this choice of approximation, $\widehat{S}$, arose in the context of PDE-constrained optimization [26], where it was argued that the approximation allows one to match all terms except for $\alpha^{-1}\gamma^{-1}(B+B^T)$.[1] Previous suggestions had more significant 'unmatched' terms [31].

Theorem 1 *If A is positive definite then all eigenvalues of $\widehat{S}^{-1}S$ are real, and are greater than or equal to $\frac{1}{2}$. If further the symmetric part of B is positive or negative semi-definite, then the eigenvalues of $\widehat{S}^{-1}S$ all lie in the real interval $[\frac{1}{2}, 1]$.*

[1] In more detail, the multiplication of the terms $\left(\gamma^{-1}A\right) A^{-1} \left(\gamma^{-1}A\right)$ within $\widehat{S}$ allows one to capture the first term of the exact Schur complement, $\gamma^{-2}A$. In a similar way, the multiplication of the terms $\left(\alpha^{-1}B\right) A^{-1} \left(\alpha^{-1}B\right)^T$ within $\widehat{S}$ leads to the second term of S, that is $\alpha^{-2}BA^{-1}B^T$.

Proof The desired eigenvalues are bounded by the extreme values of the generalized Rayleigh quotient

$$R := \frac{\gamma^{-2}\mathbf{x}^T A\mathbf{x} + \alpha^{-2}\mathbf{x}^T BA^{-1}B^T\mathbf{x}}{\gamma^{-2}\mathbf{x}^T A\mathbf{x} + \alpha^{-2}\mathbf{x}^T BA^{-1}B^T\mathbf{x} + \alpha^{-1}\gamma^{-1}\mathbf{x}^T(B + B^T)\mathbf{x}}.$$

Since A is symmetric and positive definite, we can write $\mathbf{a} := \gamma^{-1}A^{1/2}\mathbf{x}$, $\mathbf{b} := \alpha^{-1}A^{-1/2}B^T\mathbf{x}$ so that

$$R = \frac{\mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b} + \mathbf{a}^T\mathbf{b} + \mathbf{b}^T\mathbf{a}} = \frac{1}{2} + \frac{1}{2}\frac{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})}{(\mathbf{a}+\mathbf{b})^T(\mathbf{a}+\mathbf{b})},$$

which evidently implies that $R \geq \frac{1}{2}$ whatever the properties of B. Further, since we are at liberty to choose the signs of α and γ, if $B + B^T$ is semi-definite then $\mathbf{a}^T\mathbf{b} + \mathbf{b}^T\mathbf{a} \geq \mathbf{0}$ with appropriate choice of signs, so that the denominator in R is clearly greater than or equal to the numerator. This gives the result. □

Some comments are in order. The multiplicative form of $\widehat{S}$ means that application of its inverse requires the solution of two systems with coefficient matrix $\gamma^{-1}A + \alpha^{-1}B$, and multiplication with A. In Sect. 4, we describe situations where these computations are relatively straightforward using, for example, multigrid technology. That the eigenvalue spectrum is so tightly confined is somewhat remarkable, but very helpful, in particular in the context of Krylov subspace iterative methods.

Furthermore, one may use a similar analysis[2] to show that the eigenvalues of $\widehat{S}_1^{-1}S_1$ are also contained in $[\frac{1}{2}, 1]$, where

$$\widehat{S}_1 := (\alpha A + \gamma B)^T A^{-1} (\alpha A + \gamma B).$$

Both results are useful, depending on which arrangement of the saddle-point system we examine.

3 Predicted Convergence Rate of the Krylov Subspace Method

We now wish to analyze the convergence rate we can expect from an iterative method combined with our choice of preconditioner, focusing on the 3 × 3 block matrix $\mathscr{A}_3$ with a suitable preconditioner, applied within the MINRES algorithm.

[2] The analysis reads the same as presented for Theorem 1, except with $\mathbf{a} := \alpha A^{1/2}\mathbf{x}, \mathbf{b} := \gamma A^{-1/2}B\mathbf{x}$.

3.1 Eigenvalue Bounds for Preconditioned System $\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3$

Let us first consider eigenvalue bounds for $\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3$, where

$$\widehat{\mathscr{P}}_3 := \begin{bmatrix} \alpha^2\widehat{A} & 0 & 0 \\ 0 & \gamma^2\widehat{A} & 0 \\ 0 & 0 & \widehat{S} \end{bmatrix},$$

in other words where the $(1,1)$ block of the preconditioner is suitably approximated using a matrix $\widehat{A}$, and a matching strategy is used to approximate the Schur complement of $\mathscr{A}_3$.

The starting point of our analysis is the following fundamental result of Rusten and Winther [32]:

Theorem 2 *Consider the saddle-point matrix*

$$\mathscr{A}_{\Phi,\Psi} = \begin{bmatrix} \Phi & \Psi^T \\ \Psi & 0 \end{bmatrix},$$

where Φ is symmetric positive definite, and Ψ has full rank. Let $\mu_{\max}$ and $\mu_{\min}$ denote the largest and smallest eigenvalues of Φ, and let $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular values of Ψ. Then the spectrum of $\mathscr{A}_{\Phi,\Psi}$ satisfies

$$\lambda\left(\mathscr{A}_{\Phi,\Psi}\right) \in \left[\frac{1}{2}\left(\mu_{\min} - \sqrt{\mu_{\min}^2 + 4\sigma_{\max}^2}\right), \frac{1}{2}\left(\mu_{\max} - \sqrt{\mu_{\max}^2 + 4\sigma_{\min}^2}\right)\right]$$
$$\cup \left[\mu_{\min}, \frac{1}{2}\left(\mu_{\max} + \sqrt{\mu_{\max}^2 + 4\sigma_{\max}^2}\right)\right].$$

We suppose that the positive definite approximation $\widehat{A}$ is such that the eigenvalues of $\widehat{A}^{-1}A$ are contained in $[1-\zeta, 1+\eta]$, for some (preferably small) constants $\zeta \in [0,1)$, $\eta \geq 0$. Within $\widehat{\mathscr{P}}_3$, the Schur complement approximation is obtained using our matching strategy, and we assume for now that it is applied exactly.

Note that the eigenvalues of the preconditioned matrix $\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3$ are the same as those of the following (similar) matrix:

$$\widehat{\mathscr{P}}_3^{-1/2}\mathscr{A}_3\widehat{\mathscr{P}}_3^{-1/2} = \begin{bmatrix} \alpha^2\widehat{A} & 0 & 0 \\ 0 & \gamma^2\widehat{A} & 0 \\ 0 & 0 & \widehat{S} \end{bmatrix}^{-1/2} \begin{bmatrix} \alpha^2 A & 0 & B^T \\ 0 & \gamma^2 A & -A \\ B & -A & 0 \end{bmatrix} \begin{bmatrix} \alpha^2\widehat{A} & 0 & 0 \\ 0 & \gamma^2\widehat{A} & 0 \\ 0 & 0 & \widehat{S} \end{bmatrix}^{-1/2}$$
$$= \begin{bmatrix} \widehat{A}^{-1/2}A\widehat{A}^{-1/2} & 0 & \alpha^{-1}\widehat{A}^{-1/2}B^T\widehat{S}^{-1/2} \\ 0 & \widehat{A}^{-1/2}A\widehat{A}^{-1/2} & -\gamma^{-1}\widehat{A}^{-1/2}A\widehat{S}^{-1/2} \\ \alpha^{-1}\widehat{S}^{-1/2}B\widehat{A}^{-1/2} & -\gamma^{-1}\widehat{S}^{-1/2}A\widehat{A}^{-1/2} & 0 \end{bmatrix}.$$

Thus consider the eigenvalues of $\widehat{\mathscr{P}}_3^{-1/2}\mathscr{A}_3\widehat{\mathscr{P}}_3^{-1/2}$. In the setting of Theorem 2,

$$\Phi = \begin{bmatrix} \widehat{A}^{-1/2}A\widehat{A}^{-1/2} & 0 \\ 0 & \widehat{A}^{-1/2}A\widehat{A}^{-1/2} \end{bmatrix}, \quad \Psi = \left[\alpha^{-1}\widehat{S}^{-1/2}B\widehat{A}^{-1/2} \; -\gamma^{-1}\widehat{S}^{-1/2}A\widehat{A}^{-1/2}\right].$$

First, observing that the matrix $\widehat{A}^{-1/2}A\widehat{A}^{-1/2}$ is similar to $\widehat{A}^{-1}A$ gives straightforwardly that

$$\mu_{\min} = 1-\zeta, \qquad \mu_{\max} = 1+\eta,$$

again using the notation of Theorem 2.

To find values for $\sigma_{\min}$ and $\sigma_{\max}$, we then need to look for the singular values of Ψ, which are equal to the square root of the eigenvalues of

$$\Psi\Psi^T = \alpha^{-2}\widehat{S}^{-1/2}B\widehat{A}^{-1}B^T\widehat{S}^{-1/2} + \gamma^{-2}\widehat{S}^{-1/2}A\widehat{A}^{-1}A\widehat{S}^{-1/2}. \tag{2}$$

The matrix (2) is similar to

$$\widehat{S}^{-1}\left(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A\right),$$

and so its eigenvalues may be bounded by the extreme values of the Rayleigh quotient

$$\begin{aligned} &\frac{\mathbf{x}^T(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A)\mathbf{x}}{\mathbf{x}^T\widehat{S}\mathbf{x}} \\ &\quad = \underbrace{\frac{\mathbf{x}^T(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A)\mathbf{x}}{\mathbf{x}^TS\mathbf{x}}}_{R_1} \cdot \underbrace{\frac{\mathbf{x}^TS\mathbf{x}}{\mathbf{x}^T\widehat{S}\mathbf{x}}}_{R_2}. \end{aligned} \tag{3}$$

Note that

$$R_1 = \frac{\mathbf{x}^T(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A)\mathbf{x}}{\mathbf{x}^T(\alpha^{-2}BA^{-1}B^T + \gamma^{-2}AA^{-1}A)\mathbf{x}} = \frac{\mathbf{x}^TC\widehat{A}_2^{-1}C^T\mathbf{x}}{\mathbf{x}^TCA_2^{-1}C^T\mathbf{x}},$$

where $A_2 := \mathrm{blkdiag}(A,A)$, $\widehat{A}_2 := \mathrm{blkdiag}(\widehat{A},\widehat{A})$, and

$$C = \left[\alpha^{-1}B \; \gamma^{-1}A\right]$$

is full rank by assumption. Thus, with $\mathbf{y} = A_2^{-1/2}C^T\mathbf{x}$, we have

$$R_1 = \frac{\mathbf{y}^TA_2^{1/2}\widehat{A}_2^{-1}A_2^{1/2}\mathbf{y}}{\mathbf{y}^T\mathbf{y}},$$

from which it readily follows that $R_1 \in [1-\zeta, 1+\eta]$. We also know from Theorem 1 that $R_2 \in [\frac{1}{2}, 1]$. Putting these pieces together, we see that

$$\sigma_{\min} \geq \sqrt{\frac{1-\zeta}{2}}, \qquad \sigma_{\max} \leq \sqrt{1+\eta}.$$

Applying Theorem 2 along with our bounds for $\mu_{\min}$, $\mu_{\max}$, $\sigma_{\min}$ and $\sigma_{\max}$ gives us the following result:

Lemma 1 *The eigenvalues of the preconditioned system $\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3$ are contained in*

$$\left[\frac{1}{2}\left(1-\zeta-\sqrt{(1-\zeta)^2+4(1+\eta)}\right), \frac{1}{2}\left(1+\eta-\sqrt{(1+\eta)^2+2(1-\zeta)}\right)\right]$$
$$\cup \left[1-\zeta, \frac{1}{2}\left(1+\eta+\sqrt{5+6\eta+\eta^2}\right)\right],$$

where $\zeta \in [0,1)$ and $\eta \geq 0$ are constants such that the bounds $\lambda(\widehat{A}^{-1}A) \in [1-\zeta, 1+\eta]$ are exactly attained.

Note that in the case $\zeta = 0 = \eta$, which corresponds to the situation where the only approximation in the preconditioner $\widehat{\mathscr{P}}_3$ is the matching approximation $\widehat{S}$ for the exact Schur complement S, we have

$$\lambda(\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3) \in \left[\frac{1}{2}(1-\sqrt{5}), \frac{1}{2}(1-\sqrt{3})\right] \cup \left[1, \frac{1}{2}(1+\sqrt{5})\right].$$

3.2 Convergence Rate of MINRES

It is possible to exploit the result of Lemma 1 to guarantee a resulting convergence rate of the MINRES algorithm with preconditioner $\widehat{\mathscr{P}}_3$. To do this, we make use of the following theorem [13, Theorem 4.14]:

Theorem 3 *After k steps of the preconditioned* MINRES *method, applied to a system with matrix $\mathscr{A}$ and preconditioner $\mathscr{P}$, the residual $\mathbf{r}^{(k)}$ satisfies*

$$\frac{\|\mathbf{r}^{(k)}\|_{\mathscr{P}^{-1}}}{\|\mathbf{r}^{(0)}\|_{\mathscr{P}^{-1}}} \leq 2\left(\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)^{\lfloor \frac{k}{2} \rfloor},$$

where $a, b, c, d > 0$ are such that $a-b = d-c$, and

$$\lambda(\mathscr{P}^{-1}\mathscr{A}) \in [-a,-b] \cup [c,d].$$

Therefore, considering our preconditioner $\widehat{\mathscr{P}}_3$ for the matrix $\mathscr{A}_3$, we may write

$$\begin{aligned}
a &= \frac{1}{2}\left(-1+\zeta+\sqrt{(1-\zeta)^2+4(1+\eta)}\right), \quad c = 1-\zeta, \\
b &= \frac{1}{2}\left(-1-\eta+\sqrt{(1+\eta)^2+2(1-\zeta)}\right), \\
d &= \frac{1}{2}\left(1+\eta+\sqrt{5+6\eta+\eta^2}\right).
\end{aligned} \tag{4}$$

Clearly, in this setting, the condition $a-b=d-c$ is not satisfied. In fact it may be readily shown that $a-b<d-c$, as

$$\begin{aligned}
b+d &= \frac{1}{2}\left(\sqrt{(1+\eta)^2+2(1-\zeta)}+\sqrt{(1+\eta)^2+4(1+\eta)}\right) \\
&\geq \frac{1}{2}\left(\sqrt{(1-\zeta)^2+2(1-\zeta)}+\sqrt{(1-\zeta)^2+4(1+\eta)}\right) \\
&> \frac{1}{2}\left(1-\zeta+\sqrt{(1-\zeta)^2+4(1+\eta)}\right) \\
&= c+a.
\end{aligned}$$

However, it may clearly be stated that

$$\lambda(\widehat{\mathscr{P}}_3^{-1}\mathscr{A}_3) \in [-b+c-d,-b] \cup [c,d],$$

as the left interval has been stretched and includes the original interval $[-a,-b]$. We may use this to state the following result:

Lemma 2 *After k steps of the preconditioned* MINRES *method, applied to the* 3×3 *block system $\mathscr{A}_3$ and preconditioned by $\widehat{\mathscr{P}}_3$, the residual $\mathbf{r}^{(k)}$ will satisfy*

$$\frac{\|\mathbf{r}^{(k)}\|_{\widehat{\mathscr{P}}_3^{-1}}}{\|\mathbf{r}^{(0)}\|_{\widehat{\mathscr{P}}_3^{-1}}} \leq 2\left(\frac{\sqrt{d(d-c+b)}-\sqrt{bc}}{\sqrt{d(d-c+b)}+\sqrt{bc}}\right)^{\lfloor\frac{k}{2}\rfloor},$$

where b,c,d are the quantities stated in (4).

This result illustrates that the matching strategy outlined in the previous section is able to achieve rapid and robust convergence for the class of matrix systems under consideration, since the convergence bound in Lemma 2 is independent of α,γ and the dimensions of $A,\mathscr{A}$, provided only that ζ,η have such independence.

3.3 Approximate Application of $\widehat{S}$

An important question is whether such a strategy can be readily applied if the Schur complement approximation $\widehat{S}$ is applied inexactly. In more detail, the matrices $L := \gamma^{-1}A + \alpha^{-1}B$ and L^T may not be straightforward to invert, so one may wish to instead approximate the Schur complement by

$$\widetilde{S} := \widehat{L}A^{-1}\widehat{L}^T,$$

where $\widehat{L}$ is some suitable (cheap) approximation of L. This then fits into the preconditioner

$$\widetilde{\mathscr{P}}_3 := \begin{bmatrix} \alpha^2\widehat{A} & 0 & 0 \\ 0 & \gamma^2\widehat{A} & 0 \\ 0 & 0 & \widetilde{S} \end{bmatrix}.$$

To analyse the performance of this preconditioner, we need to consider the eigenvalues of $\widetilde{\mathscr{P}}_3^{-1/2}\mathscr{A}_3\widetilde{\mathscr{P}}_3^{-1/2}$. Then, in the notation of Theorem 2,

$$\Phi = \begin{bmatrix} \widehat{A}^{-1/2}A\widehat{A}^{-1/2} & 0 \\ 0 & \widehat{A}^{-1/2}A\widehat{A}^{-1/2} \end{bmatrix}, \quad \Psi = \begin{bmatrix} \alpha^{-1}\widetilde{S}^{-1/2}B\widehat{A}^{-1/2} & -\gamma^{-1}\widetilde{S}^{-1/2}A\widehat{A}^{-1/2} \end{bmatrix}.$$

The quantities $\mu_{\min}$ and $\mu_{\max}$ are therefore identical to the values when $\widehat{S}$ is applied exactly (i.e. $\widetilde{S} = \widehat{S}$) within the preconditioner.

To find suitable values for $\sigma_{\min}$ and $\sigma_{\max}$, we may apply a similar working as above, and consider the Rayleigh quotient

$$\frac{\mathbf{x}^T(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A)\mathbf{x}}{\mathbf{x}^T\widetilde{S}\mathbf{x}} = \underbrace{\frac{\mathbf{x}^T(\alpha^{-2}B\widehat{A}^{-1}B^T + \gamma^{-2}A\widehat{A}^{-1}A)\mathbf{x}}{\mathbf{x}^TS\mathbf{x}}}_{R_1} \cdot \underbrace{\frac{\mathbf{x}^TS\mathbf{x}}{\mathbf{x}^T\widehat{S}\mathbf{x}}}_{R_2} \cdot \underbrace{\frac{\mathbf{x}^T\widehat{S}\mathbf{x}}{\mathbf{x}^T\widetilde{S}\mathbf{x}}}_{R_3}.$$

As for (3), we may write that $R_1 \in [1-\zeta, 1+\eta]$ and $R_2 \in [\frac{1}{2}, 1]$. We now wish to know what can be said about the quantity R_3. A useful observation, in particular if the matrix A is well conditioned, is that

$$\begin{aligned} R_3 = \frac{\mathbf{x}^TLA^{-1}L^T\mathbf{x}}{\mathbf{x}^T\widehat{L}A^{-1}\widehat{L}^T\mathbf{x}} &= \frac{\mathbf{x}^TLA^{-1}L^T\mathbf{x}}{\mathbf{x}^TLL^T\mathbf{x}} \cdot \frac{\mathbf{x}^TLL^T\mathbf{x}}{\mathbf{x}^T\widehat{L}\widehat{L}^T\mathbf{x}} \cdot \frac{\mathbf{x}^T\widehat{L}\widehat{L}^T\mathbf{x}}{\mathbf{x}^T\widehat{L}A^{-1}\widehat{L}^T\mathbf{x}} \\ &= \frac{\mathbf{y}^T\mathbf{y}}{\mathbf{y}^TA\mathbf{y}} \cdot \frac{\mathbf{x}^TLL^T\mathbf{x}}{\mathbf{x}^T\widehat{L}\widehat{L}^T\mathbf{x}} \cdot \frac{\mathbf{z}^TA\mathbf{z}}{\mathbf{z}^T\mathbf{z}}, \end{aligned}$$

where $\mathbf{y} = A^{-1/2}L^T\mathbf{x}$ and $\mathbf{z} = A^{-1/2}\widehat{L}^T\mathbf{x}$. It is clear that $\mathbf{y}^T\mathbf{y}/\mathbf{y}^TA\mathbf{y} \in [\frac{1}{\lambda_{\max}(A)}, \frac{1}{\lambda_{\min}(A)}]$ and $\mathbf{z}^TA\mathbf{z}/\mathbf{z}^T\mathbf{z} \in [\lambda_{\min}(A), \lambda_{\max}(A)]$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively.

The remaining quantity is that of $\mathbf{x}^TLL^T\mathbf{x}/\mathbf{x}^T\widehat{L}\widehat{L}^T\mathbf{x}$, and the question becomes: if $\widehat{L}$ is a good approximation of L, does this imply that $\widehat{L}\widehat{L}^T$ is a good approximation of LL^T? In general this is in fact not the case; however, as observed by Braess and Peisker in [11], if one takes $\widehat{L}$ to be m steps of a convergent iterative process applied to a symmetric L, one may state that

$$\frac{\mathbf{x}^TL^2\mathbf{x}}{\mathbf{x}^T\widehat{L}\widehat{L}^T\mathbf{x}} \in \left[(1-\omega_m)^2, (1+\omega_m)^2\right].$$

Here ω_m relates to the rate of convergence of the iterative method for L, and satisfies $\omega_m \to 0$ as $m \to \infty$. Similar observations can possibly be applied to nonsymmetric matrices L, as LL^T itself is clearly symmetric.

Using this property, we may bound the constants $\sigma_{\min}$ and $\sigma_{\max}$ as follows:

$$\sigma_{\min} \geq \sqrt{\frac{1-\zeta}{2\kappa(A)}}\,(1-\omega_m), \qquad \sigma_{\max} \leq \sqrt{(1+\eta)\kappa(A)}\,(1+\omega_m),$$

where $\kappa(A)$ denotes the condition number of A. Inserting the bounds for $\mu_{\min}$, $\mu_{\max}$, $\sigma_{\min}$ and $\sigma_{\max}$ into the result of Theorem 2 tells us that $\lambda(\widetilde{\mathscr{P}}_3^{-1}\mathscr{A}) \in [-\widetilde{a}, -\widetilde{b}] \cup [c, \widetilde{d}]$, where

$$\widetilde{a} = \frac{1}{2}\left(-1+\zeta+\sqrt{(1-\zeta)^2+4(1+\eta)(1+\omega_m)^2\kappa(A)}\right),$$

$$\widetilde{b} = \frac{1}{2}\left(-1-\eta+\sqrt{(1+\eta)^2+\frac{2(1-\zeta)}{\kappa(A)}(1-\omega_m)^2}\right),$$

$$\widetilde{d} = \frac{1}{2}\left(1+\eta+\sqrt{(1+\eta)^2+4(1+\eta)(1+\omega_m)^2\kappa(A)}\right).$$

We note that these bounds for the eigenvalues of $\widetilde{\mathscr{P}}_3^{-1}\mathscr{A}$ are weak, sometimes extremely so, as when the approximations of L become increasingly accurate (i.e. $\omega_m \to 0$), the values of $\widetilde{a}, \widetilde{b}, \widetilde{d}$ should tend to those of a, b, d in (4). However, in the above expressions, the factors of $\kappa(A)$ remain when inserting $\omega_m = 0$. Therefore, if $\kappa(A)$ is well conditioned, as for many problems in PDE-constrained optimization for example, the theoretical guarantee of the effectiveness of $\widetilde{\mathscr{P}}_3$ is obtained straightforwardly. If this is not the case, this highlights the necessity of a potent scheme to approximate the (inverse action of) L and L^T appropriately. In practice, a number of cycles of a tailored multigrid scheme is often found to perform this function, for instance.

We highlight that the theoretical issues surrounding the approximation of matrices of the form LL^T is not restricted to the matching strategy presented in this paper, and arises when using many different preconditioners for saddle-point systems, due to the structure of the Schur complement of the saddle-point system itself.

3.4 Comments on 2×2 and 1×1 Block Cases

The analysis presented in this section has focused on solving matrix systems of the structure $\mathscr{A}_3$ using preconditioner $\widehat{\mathscr{P}}_3$ within the MINRES algorithm. Of course, it is perfectly legitimate to reduce the system to the form $\mathscr{A}_2$, and solve this using MINRES with preconditioner

$$\widehat{\mathscr{P}}_2 := \begin{bmatrix} \alpha^2 \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix}.$$

Literature such as [33] considers eigenvalue bounds for saddle-point systems with non-zero $(2,2)$ block, which arise when considering the matrix of importance in this case:

$$\widehat{\mathscr{P}}_2^{-1/2} \mathscr{A}_2 \widehat{\mathscr{P}}_2^{-1/2} = \begin{bmatrix} \widehat{A}^{-1/2} A \widehat{A}^{-1/2} & \alpha^{-1} \widehat{A}^{-1/2} B^T \widehat{S}^{-1/2} \\ \alpha^{-1} \widehat{S}^{-1/2} B \widehat{A}^{-1/2} & -\gamma^{-2} \widehat{S}^{-1/2} A \widehat{S}^{-1/2} \end{bmatrix}.$$

The analysis in this case of 2×2 blocks is more standard and is summarised, for example, in Chapter 4 of [13], or [28]. It can be applied to demonstrate that a similar MINRES convergence rate to the 3×3 case is to be expected for such 2×2 systems when the same approximations are employed within the preconditioner. The matching strategy is also, therefore, an effective approach for systems of the form $\mathscr{A}_2$.

It is also possible to consider the Schur complement (1×1 block) system itself, and apply preconditioned Conjugate Gradients with our matching strategy. In this case the potency of the iterative method will depend directly on the effectiveness of the matching strategy, which we have ascertained to guarantee compact eigenvalue bounds. However, we emphasize that such a solver will require a matrix-vector multiplication with S or S_1, and therefore an exact representation of A^{-1} will generally be required. Such a method should therefore only be applied if A has a simple structure, for instance if it is a diagonal matrix.

4 Applications of Matching Approach

In this section, we wish to briefly survey applications in which the matching strategy discussed in this paper has been applied.

PDE-Constrained Optimization The class of problems for which the authors originally derived this approach was that of PDE-constrained optimization problems of the following form:

$$\begin{aligned} \min_{y,c} \ & \frac{1}{2}\,\| y - \widehat{y} \|_{L_2(\Omega)}^2 + \frac{\beta}{2}\,\|c\|_{L_2(\Omega)}^2 \\ \text{s.t.} \quad & \mathscr{L} y = c, \quad \text{in } \Omega, \\ & y = h, \quad \text{on } \partial\Omega. \end{aligned}$$

Here, y and c denote *state* and *control variables* which we wish to find, with $\widehat{y}$ a given *desired state* and $\beta > 0$ a *regularization parameter*. The constraints of the optimization problem are derived from a PDE operator $\mathscr{L}$ on a domain Ω, and given Dirichlet boundary conditions h on the boundary $\partial\Omega$ of the domain. Other boundary conditions are possible, though boundary control problems have a slightly different form [18].

If the PDE operator $\mathscr{L} = -\nabla^2 + \mathbf{w} \cdot \nabla$, where $\mathbf{w}$ is some given wind vector, then the problem under consideration is that of *convection–diffusion control*, and upon discretization of this problem the matrix system to be solved is [26, 27]

$$\begin{bmatrix} M & 0 & \bar{K}^T \\ 0 & \beta M & -M \\ \bar{K} & -M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \\ \mathbf{h} \end{bmatrix}, \tag{5}$$

with $\mathbf{y}$, $\mathbf{c}$ and $\mathbf{p}$ the discretized state, control and *adjoint variables*, and $\mathbf{f}$ and $\mathbf{h}$ including terms arising from the desired state and boundary conditions. Here, M is a finite element mass matrix which is symmetric positive definite, and $\bar{K}$ is a finite element matrix relating to $\mathscr{L}$ which has the property that $\bar{K} + \bar{K}^T$ is positive semidefinite. If $\mathbf{w} = \mathbf{0}$, then the control problem reduces to that of *Poisson control*, and $\bar{K} = \bar{K}^T = K$ is a stiffness matrix. For either problem, the system (5) is of the form $\mathscr{A}_3$, with $A = M$, $B = \bar{K}$, $\alpha = 1$, $\gamma = \sqrt{\beta}$, and the theory of this paper can be applied (see [26, 27]). Such problems have the additional convenient property that the mass matrix M is well conditioned.

This theory has been extended to a range of other PDE-constrained optimization problems of different structure, for example to time-independent and time-dependent fluid flow control problems [23, 24, 36], reaction-diffusion control problems from chemical reactions and pattern formation processes in mathematical biology [25, 37], and active set Newton methods for problems with additional bound constraints [29]. Further, the papers [3, 4] examine PDE-constrained optimization problems with uncertain inputs using this strategy, low-rank methods are derived for a class of time-dependent problems in [35], and optimization problems with fractional differential equation constraints are studied in [12] (where a result of the type shown in Theorem 1 is proved using the Binomial Theorem).

Complex Valued Linear Systems Matrix systems of similar type to the 2 × 2 block form $\mathscr{A}_2$ are discussed in [1] in the context of complex valued linear systems. In more detail, consider the solution of the complex matrix system $C\mathbf{z} = \mathbf{d}$, where $C = A + iB$, $\mathbf{z} = \mathbf{u} + i\mathbf{w}$ and $\mathbf{d} = \mathbf{f} + i\mathbf{h}$. Therefore $(A + iB)(\mathbf{u} + i\mathbf{w}) = \mathbf{f} + i\mathbf{h}$, whereupon comparing real and imaginary parts gives the matrix system

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{h} \end{bmatrix}. \tag{6}$$

It is clear that, if A and B are symmetric, one may rearrange the system (6) to a symmetric matrix of form $\mathscr{A}_2$, with $\alpha = \gamma = 1$.

In [1], the authors derive a preconditioner for the system (6) based on the matching strategy. Further, in [2], preconditioned modified Hermitian and skew-Hermitian splitting (PMHSS) iteration methods for 2 × 2 block linear systems are considered using the same methodology.

Cahn–Hilliard Models Another major application area of the approach we have outlined is that of the numerical solution of Cahn–Hilliard models describing phase separation. For instance, in [9] the authors consider the H^{-1}–gradient flow of the Ginzburg–Landau energy

$$E(u) := \int_{\Omega} \frac{\delta\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon}\psi(u) \,\mathrm{d}\Omega,$$

with $\delta, \varepsilon > 0$, and an *obstacle potential* given by $\psi(u) = \frac{1}{2}(1 - u^2) + I_{[-1,1]}(u)$ with an indicator function I (though there are other possible choices for this potential).

Upon discretizing the resulting PDEs, the authors are required to solve matrix systems of the form

$$\begin{bmatrix} -H & M \\ M & \tau K \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{h} \end{bmatrix},$$

where M, K are defined as above, H is a symmetric matrix which involves the sum of a stiffness matrix and terms involving $\psi'(u)$, and $\tau > 0$ is the time-step used within the numerical method. Although this system is not precisely of the form $\mathscr{A}_2$, the authors were able to use convenient properties of H to obtain good numerical results for certain restrictions of the time-step (i.e. $\tau < \varepsilon^2$). In [8] some theoretical guarantees are provided for similar preconditioners for image inpainting problems.

We note that many scientists have applied the matching strategy to Cahn–Hilliard models. In [10] preconditioners for large scale binary Cahn–Hilliard models are considered, matrix systems arising from the evolution of diblock copolymer melts are tackled in [14], and solvers for the phase field crystal equation, which is itself of Cahn–Hilliard type, are constructed in [30].

We note that the fields categorised above do not represent an exhaustive list of applications for the approach presented in this paper. For instance, see [34] for a discussion of preconditioners for discontinuous Galerkin time-stepping methods, and many other papers discussing optimal control problems and Cahn–Hilliard equations, for other recent developments of this method.

5 Concluding Remarks

We have considered block preconditioners for a particular class of saddle-point matrices which arise in various applications. Specifically, we have demonstrated the efficacy of an approach which employs a ‘matching strategy’ for the approximation of a Schur complement. The use of the resulting preconditioners is shown to enable the iterative solution of corresponding systems of equations in a number of iterations independent of parameters in the problem and of the dimension of the relevant matrices. This is therefore a highly effective solution approach for such systems of equations.

Acknowledgements Ian Sloan has been a great friend and mentor over many years. Through his mathematical insight, he has derived really useful approximations in different contexts throughout his long career. We wish to express our deep appreciation for his leadership in our field and to acknowledge his encouragement of all those who seek to contribute to it.

The authors are grateful to Shev MacNamara and an anonymous referee for their interesting and helpful comments which have left us with further avenues to consider. JWP was funded for this research by the Engineering and Physical Sciences Research Council (EPSRC) Fellowship EP/M018857/1.

References

1. Axelsson, O., Neytcheva, M., Ahmad, B.: A comparison of iterative methods to solve complex valued linear algebraic systems. Numer. Algorithms **66**, 811–841 (2014)
2. Bai, Z.-Z., Benzi, M., Chen, F., Wang, Z.-Q.: Preconditioned MHSS iteration methods for a class of block two-by-two linear systems with applications to distributed control problems. IMA J. Numer. Anal. **33**, 343–369 (2013)
3. Benner, P., Dolgov, S., Onwunta, A., Stoll, M.: Low-rank solvers for unsteady Stokes–Brinkman optimal control problem with random data. Comput. Methods Appl. Mech. Eng. **304**, 26–54 (2016)
4. Benner, P., Onwunta, A., Stoll, M.: Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs. SIAM J. Matrix Anal. Appl. **37**(2), 491–518 (2016)
5. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
6. Benzi, M., Haber, E., Taralli, L.: Multilevel algorithms for large-scale interior point methods. SIAM J. Sci. Comput. **31**(6), 4152–4175 (2009)
7. Biros, G., Ghattas, O.: Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part I: The Krylov–Schur solver. SIAM J. Sci. Comput. **27**(2), 687–713 (2005)

8. Bosch, J., Kay, D., Stoll, M., Wathen, A.J.: Fast solvers for Cahn–Hilliard inpainting. SIAM J. Imaging Sci. **7**(1), 67–97 (2014)
9. Bosch, J., Stoll, M., Benner, P.: Fast solution of Cahn–Hilliard variational inequalities using implicit time discretization and finite elements. J. Comput. Phys. **262**, 38–57 (2014)
10. Boyanova, P., Do-Quang, M., Neytcheva, M: Efficient preconditioners for large scale binary Cahn–Hilliard models. Comput. Methods Appl. Math. **12**(1), 1–22 (2012)
11. Braess, D., Peisker, D.: On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning. IMA J. Numer. Anal. **6**, 393–404 (1986)
12. Dolgov, S., Pearson, J.W., Savostyanov, D.V., Stoll, M.: Fast tensor product solvers for optimization problems with fractional differential equations as constraints. Appl. Math. Comput. **273**, 604–623 (2016)
13. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics. Oxford University Press, New York (2014)
14. Farrell, P.E., Pearson, J.W.: A preconditioner for the Ohta–Kawasaki equation. SIAM J. Matrix Anal. Appl. **38**(1), 217–225 (2017)
15. Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. Academic Press, London (1982)
16. Greif, C., Moulding, E., Orban, D.: Bounds on eigenvalues of matrices arising from interior-point methods. SIAM J. Optim. **24**(1), 49–83 (2014)
17. Haber, E., Ascher, U.M.: Preconditioned all-at-once methods for large, sparse parameter estimation problems. Inverse Prob. **17**(6), 1847–1864 (2001)
18. Heidel, G., Wathen, A.J.: Preconditioning for boundary control problems in incompressible fluid dynamics. Numer. Linear Alg. Appl. (2017, submitted)
19. Ipsen, I.C.F.: A note on preconditioning nonsymmetric matrices. SIAM J. Sci. Comput. **23**(3), 1050–1051 (2001)
20. Le Gia, Q.T., Sloan, I.H., Wathen, A.J.: Stability and preconditioning for a hybrid approximation on the sphere. Numer. Math. **118**(4), 695–711 (2011)
21. Murphy, M.F., Golub, G.H., Wathen, A.J.: A note on preconditioning for indefinite linear systems. SIAM J. Sci. Comput. **21**(6), 1969–1972 (2000)
22. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12**, 617–629 (1975)
23. Pearson, J.W.: On the development of parameter-robust preconditioners and commutator arguments for solving Stokes control problems. Electron. Trans. Numer. Anal. **44**, 53–72 (2015)
24. Pearson, J.W.: Preconditioned iterative methods for Navier–Stokes control problems. J. Comput. Phys. **292**, 194–207 (2015)
25. Pearson, J.W., Stoll, M.: Fast iterative solution of reaction–diffusion control problems arising from chemical processes. SIAM J. Sci. Comput. **35**(5), B987–B1009 (2013)
26. Pearson, J.W., Wathen, A.J.: A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. Numer. Linear Algebra Appl. **19**(5), 816–829 (2012)
27. Pearson, J.W., Wathen, A.J.: Fast iterative solvers for convection–diffusion control problems. Electron. Trans. Numer. Anal. **40**, 294–310 (2013)
28. Pestana, J., Wathen, A.J.: Natural preconditioning and iterative methods for saddle point systems. SIAM Rev. **57**(1), 71–91 (2015)
29. Porcelli, M., Simoncini, V., Tani, M.: Preconditioning of active-set Newton methods for PDE-constrained optimal control problems. SIAM J. Sci. Comput. **37**(5), S472–S502 (2014)
30. Praetorius, S., Voigt, M.: Development and analysis of a block-preconditioner for the phase-field crystal equation. SIAM J. Sci. Comput. **37**(3), B425–B451 (2015)
31. Rees, T., Dollar, H.S., Wathen, A.J.: Optimal solvers for PDE-constrained optimization. SIAM J. Sci. Comput. **32**(1), 271–298 (2010)
32. Rusten, T., Winther, R.: A preconditioned iterative method for saddle point problems. SIAM J. Matrix Anal. Appl. **13**(3), 887–904 (1992)
33. Silvester, D., Wathen, A.: Fast iterative solution of stabilised Stokes systems. Part II: using general block preconditioners. SIAM J. Numer. Anal. **31**(5), 1352–1367 (1994)

34. Smears, I.: Robust and efficient preconditioners for the discontinuous Galerkin time-stepping method. IMA J. Numer. Anal. **37**(4), 1961–1985 (2017)
35. Stoll, M., Breiten, T.: A low-rank in time approach to PDE-constrained optimization. SIAM J. Sci. Comput. **37**(1), B1–B29 (2015)
36. Stoll, M., Wathen, A.: All-at-once solution of time-dependent Stokes control. J. Comput. Phys. **232**(1), 498–515 (2013)
37. Stoll, M., Pearson, J.W., Maini, P.K.: Fast solvers for optimal control problems from pattern formation. J. Comput. Phys. **304**, 27–45 (2016)
38. Wathen, A.J.: Preconditioning. Acta Numer. **24**, 329–376 (2015)

Regularized Quadrature Methods for Fredholm Integral Equations of the First Kind

Sergei V. Pereverzev, Evgeniya V. Semenova, and Pavlo Tkachenko

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract Although quadrature methods for solving ill-posed integral equations of the first kind were introduced just after the publication of the classical papers on the regularization by A.N. Tikhonov and D.L. Phillips, there are still no known results on the convergence rate of such discretization. At the same time, some problems appearing in practice, such as Magnetic Particle Imaging (MPI), allow one only a discretization corresponding to a quadrature method. In the present paper we study the convergence rate of quadrature methods under general regularization scheme in the Reproducing Kernel Hilbert Space setting.

1 Introduction

The so-called direct or discretization methods for solving Fredholm integral equations can be conventionally subdivided into three groups, namely: (1) degenerate-kernel methods, (2) projection methods, including collocation methods, Galerkin methods, least squares methods, and others, and (3) the Nyström or quadrature methods. Note that the well-known Sloan iteration [5, 27] belongs to the first named group. It can be built on the top of all projection methods, but was initially used with a Galerkin approximation.

For Fredholm equations of the second kind, there exist fairly complete results on the analysis and optimization of methods from all the above mentioned groups. Just to mention a few references, we refer to the books [1, 11, 23, 26].

S. V. Pereverzev · P. Tkachenko
Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria
e-mail: sergei.pereverzyev@oeaw.ac.at; pavlo.tkachenko@oeaw.ac.at

E. V. Semenova (✉)
Institute of Mathematics National Academy of Sciences of Ukraine, Kyiv, Ukraine

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_45

As to Fredholm equations of the first kind with a smooth kernel function, due to the fact that they are usually ill-posed, direct methods for solving them are usually combined with a regularization procedure. Of course, for unperturbed equations a regularization is less crucial [21], but in general it is unavoidable. Regularized degenerate-kernel methods were studied extensively. A few selected references are [7, 12, 25]. There are also studies on regularized collocation methods [10, 18, 24].

At the same time, quadrature methods have not been investigated enough for ill-posed Fredholm equations of the first kind. In spite of the fact that a quadrature method is the first direct method suggested for such equations [28], to the best of our knowledge, no results about convergence rates of regularized quadrature methods are known in the literature up to now. Our paper sheds a light on this issue.

The distinguishing feature of a quadrature method for a Fredholm integral equation of the first kind

$$\int_{\Omega} s(t,x)c(x)d\Omega(x) = u(t), \qquad t \in \omega \subset \mathbb{R}^{d_2},\ x \in \Omega \subset \mathbb{R}^{d_1} \tag{1}$$

is that it uses another type of information than Galerkin type or collocation methods. Namely, for a Galerkin type method we should be given Fourier coefficients of the kernel $s(t,x)$ and the right-hand side u. A collocation method uses the values of the right-hand side u at the collocation points $\{t_i\}_{i=1}^{N} \in \omega$ and the information about the kernel $s(t,x)$ in the form

$$\int_{\Omega} s(t_i,x)s(t_j,x)d\Omega(x).$$

However such kind of information is not available for some problems. For example, the information acquisition of Magnetic Particle Imaging (MPI) technology [4, 22] allows an access only to a discretized form of the corresponding Eq. (1) with respect to x in such a way that one should deal with the following system

$$\sum_{j=1}^{M} w_j s(t,x^j)c(x^j) = u(t), \tag{2}$$

where w_j are some positive weights and $\{x^j\} \subset \Omega$ is a system of knots that can be formed by the so-called Lissajous nodes, for example [4]. Therefore, a quadrature method naturally appears for such problems and further investigation of the properties of this discretization strategy is required.

It is intended that the discrete solution of (2) approximates the vector of values of the solution of (1) at knots. The following simple argument kindly provided by an anonymous referee shows the difficulty with a straightforward application of quadrature methods by inverting (2).

Write a system corresponding to (2) as

$$SW\mathbf{c} = \mathbf{u}, \tag{3}$$

where W is a diagonal matrix composed of the quadrature weights, S is a matrix with elements $s_{ij} = s(t_i, x^j)$, $\mathbf{c} = (c(x^1), \ldots, c(x^M))$, $\mathbf{u} = (u(t_1), \ldots, u(t_N))$.

Assume that the system (3) is nonsingular for some quadrature rule, say the midpoint rectangular rule. Then S is also nonsingular. To see the difficulty with an unregularized quadrature approximation, assume that a second quadrature method is used, say Simpson's method, using the same quadrature node points $\{x^j\}_{j=1}^M$ and the same points $\{t_i\}_{i=1}^N$. Denote the corresponding approximation by

$$SW^*\mathbf{c}^* = \mathbf{u}$$

noting that the matrix S does not change. The invertibility of S implies

$$W^*\mathbf{c}^* = W\mathbf{c}$$

Then, if the approximation $\mathbf{c}$ converges to u, it is not possible that $\mathbf{c}^*$ can converge to u as the weights in W^* are different from those in W, but instead

$$W^{-1}W^*\mathbf{c}^* \to u.$$

Thus, only one quadrature method, at most, can lead to a convergent schema, an absurd conclusion; and this indicates the difficulty of a purely quadrature-based approach and the need for an analysis of a regularized algorithm.

The paper organized as follows: In the next section we introduce some basic assumptions and definitions. Then, in Sect. 3 we consider quadrature methods for the discretization of Fredholm integral equations of the first kind and investigate some of their characteristics. In Sect. 4 we discuss the application of the general regularization scheme for dealing with the ill-posedness of the problem and estimate the rate of convergence of the proposed method. Finally, in the last section the algorithms and some numerical illustrations are presented.

2 Preliminaries

Let $L_2(\Omega)$, $\Omega \subset \mathbb{R}^{d_1}$, and $L_2(\omega)$, $\omega \subset \mathbb{R}^{d_2}$ be the Hilbert spaces of square summable functions on Ω and ω equipped with the standard inner products with respect to the measures $d\Omega(x)$ and $d\omega(t)$. Moreover we also consider the space $C(\Omega)$ of continuous functions on Ω.

The discretization (2) presupposes that the integral operator of (1) acts from the space allowing the evaluation of functions at the points of Ω. It is known that Reproducing Kernel Hilbert Spaces (RKHS) are natural spaces with such property [19, 20]. It is also known that any RKHS, say H_K, can be generated by the corresponding reproducing kernel $K = K(x, y)$, $x, y \in \Omega$, which is a symmetric and positive-definite function. Moreover, H_K is equipped with an inner product $\langle \cdot, \cdot \rangle_{H_K}$ such that for any $f \in H_K$ we have

$$f(x) = \langle f, K_x \rangle_{H_K}, \tag{4}$$

where $K_x = K_x(\cdot) = K(x, \cdot)$.

Let's consider a compact integral operator $S_\Omega : H_K \to L_2(\omega)$ given by

$$S_\Omega c(t) := \int_\Omega s(t, x)c(x)d\Omega(x), \qquad t \in \omega. \tag{5}$$

Further we impose additional assumptions on the space H_K and the kernel $s(t, x)$.

Assumption 1 *Let H_K be compactly embedded in $L_2(\Omega)$ and the kernel $K(x, y)$ be such that*

$$K(x, y) := \sum_l \beta_l T_l(x)T_l(y),$$

where for any l $\beta_l > 0$, $T_l(x) \in C(\Omega)$, and $\{T_l(x)\}_{l=1}^{\infty}$ is a linearly independent system of functions.

Assumption 2 *Let W^τ be the so-called space with a given rate of convergence $\tau = \tau(N)$ for the system $\{T_l(x)\}_{l=1}^{\infty}$ (see details about such spaces in [2]), i.e. W^τ is a normed space embedded in $C(\Omega)$ such that for any $f \in W^\tau$ it holds true*

$$\min_{u \in span\{T_l\}_{l=1}^N} \|f - u\|_{C(\Omega)} \le \tau(N)\|f\|_{W^\tau}. \tag{6}$$

Remark 1 To illustrate Assumption 2, let us consider the space $W^r_\infty(\Omega)$, $\Omega = [-1, 1]$, of functions on Ω having absolutely continuous derivatives of order up to $(r - 1)$ and $\|f^{(r)}\|_{L_\infty} \le \infty$,

$$\|f\|_{W^r_\infty(\Omega)} = \sum_{l=0}^{r} \left\|f^{(l)}\right\|_{L_\infty}$$

Consider the system $T_l(x) = \cos(l \arccos x)$, $x \in [-1, 1]$, of Chebyshev polynomials of the first kind, that are extensively used in the context of MPI-technology with Lissajous acquisition points [4]. Then from [3] it follows that for any $f \in W^r_\infty(\Omega)$ the condition (6) holds with $\tau(N) = O(N^{-r})$. Thus, $W^r_\infty(\Omega)$ can be seen as W^τ with $\tau(N) = cN^{-r}$.

Assumption 3 *Let the sequence $\{\beta_l\}$ be such that for any N*

$$\sum_{l=N+1}^{\infty} \beta_l \|T_l\|^2_{C(\Omega)} \leq \tau^2(N). \tag{7}$$

Remark 2 In the context of Remark 1 it is enough to assume that $\beta_l = O(l^{-\gamma})$ for $\gamma > 2r + 1$.

3 Discretization by a Quadrature Rule

In this section we introduce quadrature methods for discretizing the operator S_Ω and show that under our assumptions the discretization error is of order $O(\tau(N))$.

Consider a quadrature rule $Q_{w,M}$ such that for any $g(x) \in H_K$

$$Q_{w,M}(g) = \sum_{j=1}^{M} w_j g(x^j), \tag{8}$$

where $\{x^j\}_{j=1}^M \subset \Omega$ and $\{w_j\}_{j=1}^M \subset \mathbb{R}^+$ are the systems of quadrature knots and weights respectively.

Assumption 4 *Let for any natural N there exists $M = M(N)$ such that for $l_1, l_2 = 1, 2, \ldots, N$ it holds*

$$Q_{w,M}(T_{l_1} T_{l_2}) = \int_\Omega T_{l_1}(x) T_{l_2}(x) d\Omega(x).$$

Remark 3 In the context of Remark 1 the Gaussian quadrature formula $Q_{w,M}$ meets Assumption 4 with $M = N+1$ for $T_{l_i} = \cos(l_i \arccos(x)), i = 1, 2, x_j = \cos(\frac{2j-1}{2M}\pi)$ and $w_j = \frac{\pi}{M}$. However, in general the set $\{T_{l_1}(x)T_{l_2}(x)\}$ consists of N^2 different functions, and therefore M could be of order $O(N^2)$.

Now we apply the quadrature rule $Q_{w,M}$ for the discretization of the operator S_Ω. According to our notation we have

$$S_{w,M}c = (S_{w,M}c)(t) := Q_{w,M}(s(t,\cdot)c(\cdot)) = \sum_{j=1}^{M} w_j s(t, x^j) c(x^j), \tag{9}$$

where $S_{w,M} : H_K \to L_2(\omega)$.

Assumption 5 *Let $\{e_k\}_{k=1}^{\infty}$ be an orthonormal basis of $L_2(\omega)$. Assume that*

(i) for any ν the function $S_\nu(\cdot) = \int_\omega e_\nu(t)s(t,\cdot)d\omega(t)$ belongs to W^τ;
(ii) there exists some constant c_1 such that $(\sum_\nu \|S_\nu\|^2_{W^\tau})^{1/2} \le c_1$;
(iii) for any fixed $t \in \omega$ the function $g(y) = \int_\Omega s(t,x)K_y(x)d\Omega(x)$ belongs to H_K.

Remark 4 We illustrate Assumption 5 (conditions (ii)) in terms of the space $W^\tau = W^r_\infty(\Omega)$ for $\omega = \Omega = [-1,1]$. Suppose that the kernel $s(t,x)$ has bounded continuous 1-periodic mixed partial derivatives $\frac{\partial^{l+1}}{\partial x^l \partial t}s(t,x)$, $l = 1,\dots,r$, and the orthonormal basis $\{e_k\}_{k=1}^{\infty}$ is the system of trigonometric functions. From the integration by parts formula it follows that for any 1-periodic continuously differentiable function g

$$\left|\int_\omega g(t)e_\nu(t)d\omega(t)\right| \le c\nu^{-1}\max_{t\in\Omega}\left|g'(t)\right|,$$

where c is come constant that does not depend on ν. Using the above inequality and the definition of the norm in $W^r_\infty(\Omega)$ we have

$$\begin{aligned}\|S_\nu(x)\|^2_{W^r_\infty(\Omega)} &= \left(\sum_{l=0}^{r}\max_{x\in\Omega}\left|\int_\omega \frac{\partial^l s(t,x)}{\partial x^l}e_\nu(t)d\omega(t)\right|\right)^2 \\ &\le \left(c\nu^{-1}\sum_{l=0}^{r}\max_{t,x\in[-1,1]}\left|\frac{\partial^{l+1}s(t,x)}{\partial x^l\partial t}\right|\right)^2 \\ &\le \left(c\nu^{-1}r\max_{l\in(1,\dots,r)}\max_{t,x\in[-1,1]}\left|\frac{\partial^{l+1}s(t,x)}{\partial x^l\partial t}\right|\right)^2.\end{aligned}$$

By summing over all ν we finally obtain

$$\begin{aligned}\sqrt{\sum_\nu \|S_\nu\|^2_{W^\tau}} &\le cr\sqrt{\sum_\nu \nu^{-2}}\max_{l\in(1,..,r)}\max_{t,x\in[-1,1]}\left|\frac{\partial^{l+1}s(t,x)}{\partial x^l\partial t}\right| \\ &= cr\max_{l\in(1,..,r)}\max_{t,x\in[-1,1]}\left|\frac{\partial^{l+1}s(t,x)}{\partial x^l\partial t}\right| < \infty.\end{aligned}$$

Thus the condition (ii) fulfills.

We need the following lemmas to estimate the accuracy of the quadrature approximation (9) for operator S_Ω.

Lemma 1 *Let $f(x) \in W^\tau$. If Assumptions 2, 4 are satisfied then for $M = M(N)$ and any $l < N$ it holds*

$$\left|\int_\Omega f(x)T_l(x)d\Omega(x) - Q_{w,M}(fT_l)\right| \le 2\mu_\Omega \tau(N)\|f\|_{W^\tau}\|T_l\|_{C(\Omega)},$$

where $\mu_\Omega = \int_\Omega d\Omega(x)$.

Proof According to Assumption 2 there exist a function $u \in span\{T_l\}_{l=1}^N$ such that (6) is satisfied. Then taking into account Assumption 4 we obtain

$$\begin{aligned}&\left|\int_\Omega f(x)T_l(x)d\Omega(x) - Q_{w,M}(fT_l)\right|\\ &= \left|\int_\Omega (f(x)-u(x))T_l(x)d\Omega(x) - Q_{w,M}((f-u)T_l)\right|\\ &\le 2\mu_\Omega\|f-u\|_{C(\Omega)}\|T_l\|_{C(\Omega)} \le 2\mu_\Omega\tau(N)\|f\|_{W^\tau}\|T_l\|_{C(\Omega)},\end{aligned}$$

which proves the statement. □

Lemma 2 *Let $f \in W^\tau$ and Assumptions 1–4 be satisfied. Then*

$$\left\|\int_\Omega f(x)K_y(x)d\Omega(x) - Q_{w,M}(fK_y)\right\|_{H_K} \le 2\mu_\Omega\sqrt{1+\tau^2(0)}\|f\|_{W^\tau}\tau(N).$$

Proof Using the Assumption 1 and (4) we get

$$\begin{aligned}&\left\|\int_\Omega f(x)K_y(x)d\Omega(x) - \sum_{j=1}^M w_j f(x^j)K_y(x^j)\right\|_{H_K}^2\\ &= \Bigg\langle \int_\Omega f(x)K_y(x)d\Omega(x) - \sum_{j=1}^M w_j f(x^j)K_y(x^j),\\ &\qquad \int_\Omega f(\widetilde{x})K_y(\widetilde{x})d\Omega(\widetilde{x}) - \sum_{i=1}^M w_i f(x^i)K_y(x^i)\Bigg\rangle_{H_K}\\ &= \int_\Omega f(\widetilde{x})\int_\Omega f(x)\left\langle K_y(\widetilde{x}), K_y(x)\right\rangle_{H_K} d\Omega(x)d\Omega(\widetilde{x})\\ &\quad - \int_\Omega f(x)\sum_{i=1}^M w_i f(x^i)\left\langle K_y(x), K_y(x^i)\right\rangle_{H_K} d\Omega(x)\\ &\quad - \int_\Omega f(\widetilde{x})\sum_{j=1}^M w_j f(x^j)\left\langle K_y(\widetilde{x}), K_y(x^j)\right\rangle_{H_K} d\Omega(\widetilde{x})\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{M} w_j f(x^j) \sum_{i=1}^{M} w_i f(x^i) \left\langle K_y(x^j), K_y(x^i)\right\rangle_{H_K} \\
&= \int_{\Omega} f(\widetilde{x}) \int_{\Omega} f(x) K(x,\widetilde{x}) d\Omega(x) d\Omega(\widetilde{x}) - \int_{\Omega} f(x) \sum_{i=1}^{M} w_i f(x^i) K(x, x^i) d\Omega(x) \\
&\quad - \int_{\Omega} f(\widetilde{x}) \sum_{j=1}^{M} w_j K(\widetilde{x}, x^j) d\Omega(\widetilde{x}) + \sum_{j=1}^{M} w_j f(x^j) \sum_{i=1}^{M} w_i f(x^i) K(x^j, x^i) \\
&= \sum_{l=1}^{\infty} \beta_l \Bigg[\int_{\Omega} f(\widetilde{x}) T_l(\widetilde{x}) d\Omega(\widetilde{x}) \int_{\Omega} f(x) T_l(x) d\Omega(x) \\
&\qquad - \int_{\Omega} f(x) T_l(x) d\Omega(x) \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \\
&\quad - \int_{\Omega} f(\widetilde{x}) T_l(\widetilde{x}) d\Omega(\widetilde{x}) \sum_{j=1}^{M} w_j f(x^j) T_l(x^j) + \sum_{j=1}^{M} w_j f(x^j) T_l(x^j) \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \Bigg] \\
&= \sum_{l=1}^{\infty} \beta_l \left[\int_{\Omega} f(x) T_l(x) d\Omega(x) - \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \right]^2 .
\end{aligned}
$$

Taking into account Lemma 1 and Assumption 3 we obtain

$$
\begin{aligned}
&\sum_{l=1}^{\infty} \beta_l \left[\int_{\Omega} f(x) T_l(x) d\Omega(x) - \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \right]^2 \\
&= \sum_{l=1}^{N} \beta_l \left[\int_{\Omega} f(x) T_l(x) d\Omega(x) - \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \right]^2 \\
&\quad + \sum_{l=N+1}^{\infty} \beta_l \left[\int_{\Omega} f(x) T_l(x) d\Omega(x) - \sum_{i=1}^{M} w_i f(x^i) T_l(x^i) \right]^2 \\
&\le \sum_{l=1}^{N} 4\beta_l \mu_{\Omega}^2 \tau^2(N) \|f\|_{W^\tau}^2 \|T_l\|_{C(\Omega)}^2 + \sum_{l=N+1}^{\infty} 4\beta_l \mu_{\Omega}^2 \|f\|_{W^\tau}^2 \|T_l\|_{C(\Omega)}^2 \\
&\le 4\mu_{\Omega}^2 \tau^2(N) \|f\|_{W^\tau}^2 (1 + \tau^2(0))
\end{aligned}
$$

that completes the proof. □

Theorem 1 *Let Assumptions 1–5 be satisfied. Then it holds*

$$\|S_\Omega - S_{w,M}\|_{H_K \to L_2(\omega)} \le c_2 \tau(N), \tag{10}$$

where $c_2 = 2\mu_\Omega c_1 \sqrt{1 + \tau^2(0)}$.

Proof For the orthonormal basis $\{e_k\}_{k=1}^\infty \in L_2(\omega)$ the Parseval identity asserts that for $S_\Omega c - S_{w,M} c \in L_2(\omega)$

$$\|S_\Omega c - S_{w,M} c\|_{L_2(\omega)} = \sqrt{\sum_\nu \langle S_\Omega c - S_{w,M} c, e_\nu \rangle^2_{L_2(\omega)}}. \tag{11}$$

Thus, to prove the theorem it is necessary to bound the corresponding inner product in (11).

Since $c(x)$ can be represented by using (4) we have

$$S_\Omega c(t) - S_{w,M} c(t) = \int_\Omega s(t,x) c(x) d\Omega(x) - \sum_{j=1}^M w_j s(t, x^j) c(x^j)$$

$$= \int_\Omega s(t,x) \langle c, K_x \rangle_{H_K} d\Omega(x) - \sum_{j=1}^M w_j s(t, x^j) \langle c, K_{x^j} \rangle_{H_K}$$

$$= \left\langle c, \int_\Omega s(t,x) K_x d\Omega(x) - \sum_{j=1}^M w_j s(t, x^j) K_{x^j} \right\rangle_{H_K}.$$

Then using Cauchy-Schwarz inequality and Lemma 2 we get

$$\langle S_\Omega c - S_{w,M} c, e_\nu \rangle = \int_\omega e_\nu(t) \left\langle c, \int_\Omega s(t,x) K_x d\Omega(x) - \sum_{j=1}^M w_j s(t, x^j) K_{x^j} \right\rangle_{H_K} d\omega(t)$$

$$= \left\langle c, \int_\Omega S_\nu(x) K_x d\Omega(x) - \sum_{j=1}^M w_j K_{x^j} S_\nu(x^j) \right\rangle_{H_K}$$

$$\le \|c\|_{H_K} \left\| \int_\Omega S_\nu(x) K_x d\Omega(x) - \sum_{j=1}^M w_j K_{x^j} S_\nu(x^j) \right\|_{H_K}$$

$$\le \|c\|_{H_K} 2\mu_\Omega \sqrt{1 + \tau^2(0)} \|S_\nu(x)\|_{W^\tau} \tau(N).$$

Substituting this bound into (11) and taking into account Assumption 5 we can finally obtain

$$\begin{aligned}\|S_\Omega c - S_{w,M}c\|_{L_2(\omega)} &\le 2\mu_\Omega\sqrt{1+\tau^2(0)}\tau(N)\|c\|_{H_K}\sqrt{\sum_\nu \|S_\nu(x)\|_{W^\tau}}\\ &\le 2\mu_\Omega c_1\sqrt{1+\tau^2(0)}\|c\|_{H_K}\tau(N). \qquad \square\end{aligned}$$

4 Regularization

As we already mentioned in Introduction, Fredholm integral equations of the first kind with a smooth kernel function are usually ill-posed, and therefore regularization is needed for their stable solution [6, 8, 9].

Consider now a noisy version of Eq. (1) that can be written as

$$S_\Omega c = u^\delta, \tag{12}$$

where u^δ is such that $\|u - u^\delta\|_{L_2(\omega)} \le \delta$. We assume that (12) for $\delta = 0$ has solutions and denote by $c^\dagger$ the so-called Moore-Penrose solution. From [13] we know that there is always a continuous, strictly increasing function $\phi, \phi(0) = 0$, called index function such that $c^\dagger \in H_K$ belongs to the range of the operator $\phi(S_\Omega^* S_\Omega)$. By S_Ω^* we denote the adjoint of S_Ω.

Furthermore we assume that the index function ϕ is operator monotone.

Definition 1 The function ϕ is operator monotone on $(0, a)$ if for any pair of self-adjoint operators A, B with spectrum in $(0, a)$, where $a = \max\{\|A\|, \|B\|\}$, we have $\phi(A) \le \phi(B)$ whenever $A \le B$.

It is known [14] that if ϕ is an operator monotone function on $(0, a)$ then for any pair of self-adjoint operators A, B, $\|A\|, \|B\| \le b$, $b < a$ there exists a constant d_1 such that

$$\|\phi(A) - \phi(B)\| \le d_1\phi(\|A - B\|). \tag{13}$$

Moreover from [17] we know that the operator monotone functions satisfy the inequality

$$\phi(t)/t \le T\phi(s)/s, \quad \text{whenever} \quad 0 < s < t < a, \tag{14}$$

where T is some constant. Summarizing our discussion above we impose the following assumption.

Assumption 6 *Let* $c^\dagger = \phi(S_\Omega^* S_\Omega)v$, *where* $\|v\|_{H_K} \leq 1$, *and* ϕ *is an operator monotone function on the interval* $(0, \|S_\Omega^* S_\Omega\|]$.

For regularization of Eq. (12) we consider general regularization scheme combined with the discretization according to (9), namely the approximate solution $c_{\alpha,\delta}^N$ is calculated as $c_{\alpha,\delta}^N := g_\alpha(S_{w,M}^* S_{w,M}) S_{w,M}^* u^\delta$, where $\{g_\alpha(\lambda)\}, 0 < \lambda \leq \|S_\Omega^* S_\Omega\|$ is a parametric family of bounded functions with α being a regularization parameter.

Of course, not every family can be used as a regularization.

Definition 2 A family $\{g_\alpha\}$ is called a regularization, if there are constants $\chi_{-1}, \chi_{1/2}, \chi_0$ for which

$$\sup_{0<\lambda\leq\|S_\Omega^* S_\Omega\|} |1 - \lambda g_\alpha(\lambda)| \leq \chi_0$$

$$\sup_{0<\lambda\leq\|S_\Omega^* S_\Omega\|} \sqrt{\lambda}\,|g_\alpha(\lambda)| \leq \frac{\chi_{1/2}}{\sqrt{\alpha}}$$

$$\sup_{0<\lambda\leq\|S_\Omega^* S_\Omega\|} |g_\alpha(\lambda)| \leq \frac{\chi_{-1}}{\alpha}.$$

Moreover, due to (14) the following properties of $\{g_\alpha(\lambda)\}, 0 < \lambda \leq \|S_\Omega^* S_\Omega\|$ can be derived [16]

$$\sup_{0<\lambda\leq\|S_\Omega^* S_\Omega\|} |1 - \lambda g_\alpha(\lambda)|\phi(\lambda) \leq \chi\phi(\alpha), \tag{15}$$

where the number χ does not depend on α and ϕ.

Remark 5 For Tiknonov-Phillips regularization, which will be used in the next section

$$g_\alpha(\lambda) = \frac{1}{\alpha + \lambda},$$

and the above mentioned conditions are satisfied with $\chi_{-1} = \chi_0 = \chi = 1, \chi_{1/2} = 1/2$.

Next we prove the following lemma.

Lemma 3 *Under Assumptions 1–6 it holds*

$$\|c^\dagger - c_{\alpha,\delta}^N\|_{H_K} \leq \chi\phi(\alpha) + \chi_0 d_1 \phi(c_2\tau(N)) + \frac{\chi_{1/2}\delta}{\sqrt{\alpha}} + \frac{\chi_{1/2} c_2 \tau(N) \|c^\dagger\|_{H_K}}{\sqrt{\alpha}}.$$

Proof The statement of the lemma follows from [15, formula (4)], but for completeness we present the proof here as well.

Taking into account the above properties of the regularization family $g_\alpha(\lambda)$ we have

$$\begin{aligned}\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} &\le \|(I - g_\alpha(S^*_{w,m}S_{w,m})S^*_{w,m}S_{w,m})\phi(S^*_{w,m}S_{w,m})\|_{H_K\to H_K}\\ &+ \|(I - g_\alpha(S^*_{w,m}S_{w,m})S^*_{w,m}S_{w,m})(\phi(S^*_{w,m}S_{w,m}) - \phi(S^*_\Omega S_\Omega)\|_{H_K\to H_K}\\ &+ \|g_\alpha(S^*_{w,m}S_{w,m})S^*_{w,m}(u - u^\delta)\|_{H_K}\\ &+ \|g_\alpha(S^*_{w,m}S_{w,m})S^*_{w,m}(S_\Omega - S_{w,m})\|_{H_K\to H_K}\|c^\dagger\|_{H_K}\\ &\le \chi\phi(\alpha) + \chi_0\|\phi(S^*_{w,m}S_{w,m}) - \phi(S^*_\Omega S_\Omega)\|_{H_K\to H_K} + \chi_{1/2}\frac{\delta}{\sqrt{\alpha}}\\ &+ \chi_{1/2}\|c^\dagger\|_{H_K}\frac{\|S_{w,m} - S_\Omega\|_{H_K\to L_2(\omega)}}{\sqrt{\alpha}}.\end{aligned}$$

Using the property (13) and Theorem 1 we complete the proof:

$$\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} \le \chi\phi(\alpha) + \chi_0 d_1\phi(c_2\tau(N)) + \chi_{1/2}\frac{\delta}{\sqrt{\alpha}} + \chi_{1/2}\|c^\dagger\|_{H_K}\frac{c_2\tau(N)}{\sqrt{\alpha}}. \qquad \square$$

Theorem 2 *Let N be the smallest positive integer such that*

$$c_2\tau(N) \le \begin{cases} \alpha, & \phi(t) \ge \sqrt{t} \\ \delta, & \phi(t) < \sqrt{t} \end{cases}.$$

Then for $\alpha = \theta^{-1}(\delta)$, *where* $\theta(t) = \sqrt{t}\phi(t), \quad t \in [0, \|S_\Omega\|^2]$, *we have*

$$\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} \le c\phi(\theta^{-1}(\delta)),$$

with some constant c that does not depend on δ, N, M.

Proof If $\phi(\lambda) \ge \sqrt{\lambda}$, $t \in [0, \|S_\Omega\|^2]$, then from Lemma 3 we conclude that

$$\begin{aligned}\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} &\le \chi\phi(\alpha) + \chi_0 d_1\phi(\alpha) + \frac{\chi_{1/2}\delta}{\sqrt{\alpha}} + \chi_{1/2}\sqrt{\alpha}\|c^\dagger\|_{H_K}\\ &\le (\chi + \chi_0 d_1 + \chi_{1/2}\|c^\dagger\|_{H_K})\phi(\alpha) + \frac{\chi_{1/2}\delta}{\sqrt{\alpha}}.\end{aligned}$$

Thus from the definition of the function $\theta(t)$ it holds that $\delta/\sqrt{\theta^{-1}(\delta)} = \phi(\theta^{-1}(\delta))$ and we obtain

$$\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} \le c\phi(\theta^{-1}(\delta)).$$

For the case $\phi(\lambda) < \sqrt{\lambda}$, $t \in [0, \|S_\Omega\|^2]$, from Lemma 3 the following estimation holds true

$$\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} \leq \chi\phi(\alpha) + \chi_0 d_1 \phi(\delta) + \frac{\chi_{1/2}\delta}{\sqrt{\alpha}} + \frac{\chi_{1/2}\delta\|c^\dagger\|_{H_K}}{\sqrt{\alpha}}.$$

Taking into account that $\phi(\delta) < \sqrt{\delta}$ and $\alpha = \theta^{-1}(\delta) > \delta$ we finally obtain

$$\|c^\dagger - c^N_{\alpha,\delta}\|_{H_K} \leq c\phi(\theta^{-1}(\delta)). \qquad \square$$

Remark 6 It is known [16] that the error bound of order $O(\phi(\theta^{-1}(\delta)))$ is optimal for $c^\dagger \in \text{Range}(\phi(S^*_\Omega S_\Omega))$, i.e. it can't be improved for the class of solutions under consideration.

5 Algorithms and Numerical Illustrations

In this section we compare two discretization strategies for Fredholm integral equation of the first kind: the regularized collocation method studied in [18] and the regularized quadrature approximation described in the present paper. For reader convenience we describe both algorithm in ready-to-use form below. Note that the presented descriptions correspond to the Tikhonov-Phillips regularization scheme, namely $g_\alpha(\lambda) = 1/(\alpha + \lambda)$.

5.1 Regularized Collocation Method

According to [18], the solution $c^N_{\alpha,\delta}$ can be represented as

$$c^N_{\alpha,\delta} = \sum_{j=1}^{M} c_j w_j s(t_j, \cdot),$$

where the vector $\mathbf{c} \in \mathbb{R}^M$ of the coefficients c_j can be found from the system

$$\alpha\mathbf{c} + A\mathbf{c} = \mathbf{u}^\delta,$$

with $A = MW$,

$$W = diag(w_1, \dots, w_M), \;\; M = \left[m_{ij}\right], \;\; m_{ij} = \int_\Omega s(t_j, t)s(t_i, t)d\Omega(t),$$
$$\mathbf{u}^\delta = (u^\delta(t_1), \dots, u^\delta(t_M)).$$

5.2 Regularized Quadrature Method

Recall that for a regularization of Eq. (12) we use Tikhonov-Phillips method combined with the discretization according to (9), namely

$$\alpha c + S^*_{w,M} S_{w,M} c = S^*_{w,M} u^\delta, \tag{16}$$

where for any $b(t) \in L_2(\omega)$

$$(S^*_{w,M} b)(\cdot) = \sum_{j=1}^{M} w_j K(\cdot, x^j) \int_\omega s(t, x^j) b(t) d\omega(t). \tag{17}$$

Note that the solution of (16) belongs to Range$(S^*_{w,M})$. It means that due to (17) the element $c^N_{\alpha,\delta}$ can be represented as

$$c^N_{\delta,\alpha} = \sum_{k=1}^{M} c_k K(\cdot, x^k).$$

Thus, the solution of Eq. (16) is derived from the system of linear equations with respect to $c_k, k = 1, .., M$, namely the values c_k can be found from the system

$$\alpha c_k + \sum_{p=1}^{M} c_p s_{k,p} = u^\delta_k, \quad k = 1..M,$$

where $u^\delta_k = \int_\omega s(t, x^k) u^\delta(t) d\omega(t)$ and

$$s_{k,p} = \sum_{\mu=1}^{M} w_k w_\mu K(x^\mu, x^p) \int_\omega s(t, x^\mu) s(t, x^k) d\omega(t).$$

5.3 Numerical Comparison

In our numerical tests we put both algorithms side-by-side for integral equations (1) defined on $\omega = \Omega = [0, 1]$ with kernels

$$s(t, x) = \sum_{l=1}^{D} \sum_{m=1}^{D} d_{lm} \frac{\cos(2\pi l t) \cos(2\pi m x)}{(2\pi l)^p (2\pi m)^q},$$

and the exact solutions

$$c^{\dagger}(t) = \sum_{k=1}^{D} c_k^{\dagger} \frac{\cos(2\pi kt)}{(2\pi k)^{\nu}},$$

where $d_{lm}, l, m = 1, \ldots, D$ and $c_k^{\dagger}, k = 1, \ldots, D$, are uniformly distributed random numbers from $(0, 1)$. In our experiments $D = 100, p = 1, q = \{1, 3\}$, and $\nu = 2$.

We choose such values of the parameters p and q to demonstrate the advantage in the sense of the accuracy of the quadrature method over the collocation in the case when kernels $s(t, x)$ are smoother with respect to integration variables.

It is clear that for such kernels and solutions the right-hand sides $u(t)$ can be calculated explicitly as well:

$$u(t) = \sum_{l=1}^{D} \sum_{m=1}^{D} d_{lm} \frac{\cos(2\pi lt) c_k^{\dagger}}{2(2\pi l)^p (2\pi m)^{q+\nu}}.$$

Then we generate the noisy data u^{δ} as follows

$$u^{\delta}(t) = \sum_{l=1}^{D} \cos(2\pi lt) \sum_{m=1}^{D} \left(d_{lm} \frac{c_k^{\dagger}}{2(2\pi l)^p (2\pi m)^{q+\nu}} + \delta \xi_l \right),$$

where δ is the noise intensity and $\xi_l, l = 1, \ldots, D$ are uniformly distributed random numbers from $(-1, 1)$.

Both considered methods are tested with $t_i = x^i = \frac{i}{M}, i = 1, 2, \ldots, M, M = 100$. Moreover, we consider a quadrature rule with equal weights.

Note that for the realization of the quadrature method one should also define the RKHS H_K, and for simplicity we consider $K(x, x') = \sum_{j=1}^{M} l_j(x) l_j(x')$, where $l_j, j = 1, 2, \ldots, M$, are the fundamental interpolation functions associated with the knots $x^j, j = 1, 2, \ldots, M$. In this case $K(x^i, x^j) = \delta_{ij}$.

We employed Tikhonov-Phillips regularization with the optimal choice of α from the set $\alpha_i = 10^{-20} \cdot 1.08^{i-1}, i = 1, 2 \ldots, 1000$.

Tables 1 and 2 report the mean relative error over 10 simulations as described above. Figures 1 and 2 show a result of a particular simulation. The presented results demonstrate the advantage of a quadrature method. Moreover, in view of Remarks 1 and 4 one may conclude that for $q = 1$ the convergence rate $\tau = \tau(N)$ is worse than

Table 1 Average errors of collocation and quadrature methods over 10 simulations of the noisy data, $p = 1, q = 3, \nu = 2$

	Mean relative error				
	$\delta = 10^{-5}$	$\delta = 10^{-6}$	$\delta = 10^{-7}$	$\delta = 10^{-8}$	$\delta = 10^{-9}$
Collocation method	0.4615	0.1764	0.1542	0.1547	0.1547
Quadrature method	0.4021	0.1749	0.1040	0.0722	0.0380

Table 2 Average errors of collocation and quadrature methods over 10 simulations of the noisy data, $p = 1, q = 1, \nu = 2$

	Mean relative error				
	$\delta = 10^{-5}$	$\delta = 10^{-6}$	$\delta = 10^{-7}$	$\delta = 10^{-8}$	$\delta = 10^{-9}$
Collocation method	0.2744	0.2262	0.2262	0.2261	0.2261
Quadrature method	0.3197	0.1553	0.0863	0.0787	0.0783

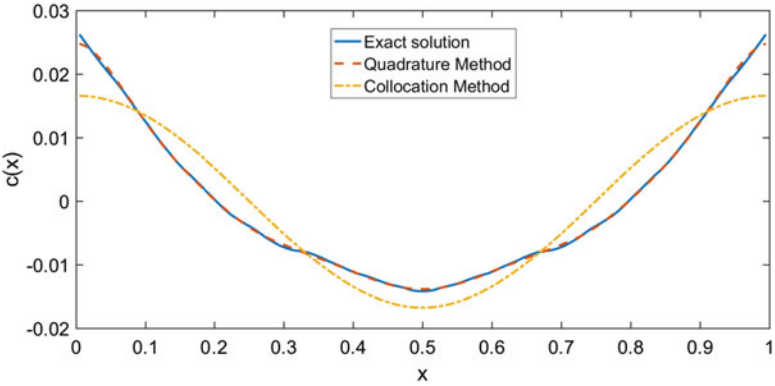


Fig. 1 Reconstruction of $c^\dagger$ (exact solution) by the collocation and quadrature methods, $\delta = 10^{-9}, p = 1, q = 3, \nu = 2$

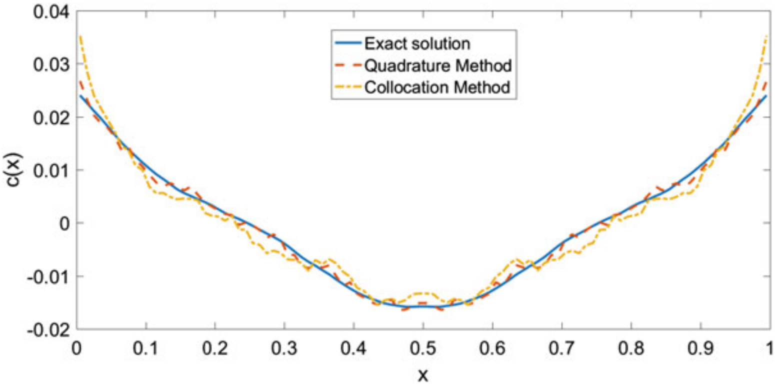


Fig. 2 Reconstruction of $c^\dagger$ (exact solution) by the collocation and quadrature methods, $\delta = 10^{-9}, p = 1, q = 1, \nu = 2$

for $q = 3$. Therefore, in view of Theorem 1 for sufficiently small noise δ and $q = 1$ one should expect larger error than for $q = 3$. The comparison of Tables 1 and 2, as well as Figs. 1 and 2, confirms the above expectation.

Acknowledgements This work was done while the second author was visiting Johann Radon Institute within the EU-Horizon 2020 MSC-RISE project AMMODIT.

The authors affiliated with Johann Radon Institute gratefully acknowledge the support of the Austrian Science Fund (FWF): project I1669.

References

1. Atkinson, K.E.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge (1997)
2. Cohen, A., DeVore, R., Kerkyacharian, G., Picard, D.: Maximal spaces with given rate of convergence for thresholding algorithms. Appl. Comput. Harmon. Anal. **11**, 167–191 (2001)
3. Dzyadyk, V.K., Shevchuk, I.A.: Theory of Uniform Approximation of Functions by Polynomials. Walter de Gruyter, Berlin (2008)
4. Erb, W., Kaethner, C., Ahlborg, M., Buzug, T.M.: Bivariate Lagrange interpolation at the node points of non-degenerate Lissajous curves. Numer. Math. **133**, 685–705 (2016)
5. Graham, I.G., Atkinson, K.E.: On the Sloan iteration applied to integral equations of the first kind. IMA J. Numer. Anal. **13**, 29–41 (1993)
6. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind. Research Notes in Mathematics. Pitman, London (1984)
7. Groetsch, C.W.: Convergence analysis of a regularized degenerate kernel methods for Fredholm integral equations of the first kind. Integr. Equ. Oper. Theory **13**, 63–75 (1990)
8. Groetsch, C.W.: Stable Approximate Evaluation of Unbounded Operators. Lecture Notes in Mathematics. Springer, Berlin (2007)
9. Kirsch, A.: An Introduction to the Mathematical Theory of Inverse Problems. Springer, Berlin (2011)
10. Krebs, J., Louis, A.K., Wendland, H.: Sobolev error estimates and a priori parameter selection for semi-discrete Tikhonov regularization. J. Inverse Ill-Posed Prob. **17**, 845–869 (2009)
11. Kress, R.: Linear Integral Equations. Springer, Berlin (2014)
12. Maas, P., Pereverzev, S.V., Ramlau R., Solodky, S.G.: An adaptive discretization for Tikhonov-Phillips regularization with a posteriori parameter selection. Numer. Math. **87**, 485–502 (2001)
13. Mathé, P., Hofmann, B.: How general are general source conditions? Inverse Prob. **24**, 1–5 (2008)
14. Mathé, P., Pereverzev, S.V.: Moduli of continuity for operator valued functions. Numer. Funct. Anal. Optim. **23**, 623–631 (2002)
15. Mathé, P., Pereverzev, S.V.: Discretization strategy for ill-posed problems in variable Hilbert scales. Inverse Prob. **19**, 1263–1277 (2003)
16. Mathé, P., Pereverzev, S.V.: Geometry of linear ill-posed problems in variable Hilbert scales. Inverse Prob. **19**, 789–803 (2003)
17. Mathé, P., Pereverzev, Sergei V.: Regularization of some linear ill-posed problems with discretized random noisy data. Math. Comput. **75**, 1913–1929 (2006)
18. Nair, M.T., Pereverzev, S.: Regularized collocation method for Fredholm integral equations of the first kind. J. Complex. **23**, 454–467 (2007)
19. Nashed, M., Wahba, G.: Regularization and approximation of linear operator equations in reproducing kernel spaces. Bull. Am. Math. Soc. **80**, 1213–1218 (1974)
20. Nashed, M., Wahba, G.: Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operator equations. SIAM J. Math. Anal. **5**, 974–987 (1974)
21. Nédélec, J.: Curved finite element methods for the solution of singular integral equations on surfaces in $\mathbb{R}^3$. Comput. Methods Appl. Mech. Eng. **8**, 61–80 (1976)
22. Panagiotopoulos, N., Duschka, R., Ahlborg, M., Bringout, G., Debbeler, C., Graeser, M., Kaethner, C., Lüdtke-Buzug, K., Medimagh, H., Stelzner, J., Buzug, T.M., Barkhausen, J., Vogt, F. M., Haegele, J.: Magnetic particle imaging – current developments and future directions. Int. J. Nanomedicine **10**, 3097–3114 (2015)

23. Pereverzev, S.V.: Optimization of Methods for Approximate Solution of Operator Equations. Nova Science Publishers, New York (1996)
24. Pereverzev, S.V., Solodky, S.G., Volynets, E.A.: The balancing principle in solving semi-discrete inverse problems in Sobolev scales by Tikhonov method. Appl. Anal. **91**, 435–446 (2012)
25. Plato, R., Vainikko, G.: On the regularization of projection methods for solving ill-posed problems. Numer. Math. **57**, 63–79 (1990)
26. Reinhard, H.-J.: Analysis of Approximation Methods for Differential and Integral Equations. Springer, New York (1985)
27. Sloan, I.H.: Error analysis for a class of degenerate kernel methods. Numer. Math. **25**, 231–238 (1976)
28. Twomey, S.: On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. J. Assoc. Comput. Mach. **10**, 97–101 (1963)

On Linear Versus Nonlinear Approximation in the Average Case Setting

Leszek Plaskota

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We compare the average errors of n-term linear and nonlinear approximations assuming that the coefficients in an orthogonal expansion of the approximated element are scaled i.i.d. random variables. We show that generally the n-term nonlinear approximation can be even exponentially better than the n-term linear approximation. On the other hand, if the scaling parameters decay no faster than polynomially then the average errors of nonlinear approximations do not converge to zero faster than those of linear approximations, as $n \to +\infty$. The main motivation and application is the approximation of Gaussian processes. In this particular case, the nonlinear approximation is, roughly, no more than n times better than its linear counterpart.

1 Introduction

Let F be a linear space over the reals. For a given countable *dictionary*

$$\mathscr{D} := \left\{ \xi_j \,|\, j = 1, 2, \ldots \right\} \subset F,$$

the n-term *linear approximation* relies on approximating elements of F by elements of the linear subspace

$$V_n := \operatorname{span}\{\xi_1, \ldots, \xi_n\} = \left\{ \sum_{j=1}^{n} c_j \xi_j \,\middle|\, c_j \in \mathbb{R} \right\}$$

L. Plaskota (✉)
Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
e-mail: leszekp@mimuw.edu.pl

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_46

with respect to a given norm $\|\cdot\|$. In the n-term *nonlinear approximation*, the approximations are in the nonlinear manifold

$$\widetilde{V}_n := \Big\{ \sum_{j=1}^{n} c_j \xi_{i_j} \,\Big|\, c_j \in \mathbb{R},\ i_1 < i_2 < \cdots < i_n \Big\}.$$

The nonlinear approximation is an important practical tool because of its possible applications to, e.g., data storage or compressed sensing. Therefore it is important to know whether it offers better results than linear approximation. This topic has usually been studied assuming that the approximated elements are of deterministic nature. We only mention [3, 5–8, 12, 14] as examples. Much less is known on nonlinear approximation in non-deterministic cases where the approximated elements are assumed to be random elements or processes, see, e.g., [1, 2, 4, 9, 17].

In this paper, we consider the average error of approximation with respect to a probability measure μ defined on F. Denoting by

$$\mathrm{dist}(f, V_n) := \inf_{v_n \in V_n} \|f - v_n\| \qquad \text{and} \qquad \mathrm{dist}(f, \widetilde{V}_n) := \inf_{v_n \in \widetilde{V}_n} \|f - v_n\|$$

the minimal errors of linear and nonlinear approximations for each individual $f \in F$, respectively, we are interested in the average errors

$$e_\mu^{\mathrm{ave}}(n) = \left(\int_F \mathrm{dist}(f, V_n)^2 \, \mu(\mathrm{d}f) \right)^{1/2} \text{ and } \widetilde{e}_\mu^{\mathrm{ave}}(n) = \left(\int_F \mathrm{dist}(f, \widetilde{V}_n)^2 \, \mu(\mathrm{d}f) \right)^{1/2}.$$

The author's motivation to study relations between $e_\mu^{\mathrm{ave}}(n)$ and $\widetilde{e}_\mu^{\mathrm{ave}}(n)$ comes from the average case approximation based on partial information only, which is studied in *information-based complexity* theory [15]. Here F is a separable Banach space, μ is a zero mean Gaussian measure, and the error is taken with respect to a Hilbert norm. In this case, the optimal choice of the dictionary leads to the situation where one approximates $f \in F$ from its coefficients in an orthonormal expansion, where the coefficients are scaled normally distributed random variables. This corresponds to the Karhunen-Loéve decomposition in the theory of random processes. See Sect. 2 for details.

In order to obtain some general results, we consider in Sect. 3 a more general situation where the coefficients in the orthogonal expansion are arbitrary scaled i.i.d.'s. Our general results show that the n-term nonlinear approximation can be exponentially (and no more than exponentially) better than the n-term linear approximation. On the other hand, if the scaling parameters decay at most polynomially then the orders of linear and nonlinear approximations are the same. In the particular Gaussian case we have an additional result that the nonlinear approximation is, roughly, no more than n times better than its linear counterpart, see Corollary 2 of Sect. 4. Finally, we compare the average case errors for Gaussian measures with the corresponding worst case errors over the unit ball of F, see Remark 2.

2 Motivation: Approximation of Gaussian Processes

Let F be a real separable Banach space with a norm $\|\cdot\|_F$, equipped with a Gaussian measure μ defined on the corresponding σ-field of Borel sets. Let the mean element of μ be zero and its covariance operator $C_\mu : F^* \to F$, i.e.,

$$\int_F L(f)\,\mu(\mathrm{d}f) = 0 \qquad \forall L \in F^*$$

and

$$\int_F L_1(f)L_2(f)\,\mu(\mathrm{d}f) = L_1(C_\mu L_2), \qquad \forall L_1, L_2 \in F^*.$$

(See e.g. [10, 16] for more about Gaussian measures in Banach spaces.) Suppose we want to approximate elements $f \in F$ with error measured in another norm $\|\cdot\|$. We assume that the only *a priori* knowledge of f is that it is distributed according to μ. We first collect some information about f and then combine this information to obtain an approximation $A_n(f)$. Specifically, we assume that the norm $\|\cdot\|$ is weaker than the original norm $\|\cdot\|_F$ and is generated by an inner product $\langle\cdot,\cdot\rangle$. Let H be the completion of F to a Hilbert space with respect to $\langle\cdot,\cdot\rangle$, so that $(F, \|\cdot\|_F)$ is continously embedded in $(H, \|\cdot\|)$. The allowed approximations to $f \in F$ are of the form

$$A_n(f) = \phi_n(L_1 f, L_2 f, \dots, L_n f)$$

where L_j, $1 \le j \le n$, are some continuous linear functionals on F, and $\phi_n : \mathbb{R}^n \to H$ is an arbitrary measurable mapping. The average error of approximation is defined as

$$\mathrm{err}_\mu^{\mathrm{ave}}(A_n) := \left(\int_F \|f - A_n(f)\|^2 \mu(\mathrm{d}f)\right)^{1/2}.$$

In this setting, for fixed n, the optimal approximation A_n^* (i.e., the one that uses n functional evaluations and minimizes the average error) is as follows, see e.g. [11, 13, 15]. Define $W_\mu : H \to H$ as

$$W_\mu f = C_\mu L_f \quad \text{where} \quad L_f = \langle\cdot, f\rangle.$$

W_μ is a self-adjoint, nonnegative definite operator with finite trace. Let ξ_j^*, $j \ge 1$, be a complete and orthonormal in H system of eigenelements of W_μ, and λ_j be the corresponding eigenvalues,

$$W_\mu \xi_j^* = \lambda_j \xi_j^*, \qquad j \ge 1, \tag{1}$$

where

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \quad \text{and} \quad \sum_{j=1}^{+\infty} \lambda_j < +\infty.$$

Then each $f \in F$ admits the orthonormal expansion

$$f = \sum_{j=1}^{+\infty} c_j \xi_j^* \quad \text{with} \quad c_j = \langle f, \xi_j^* \rangle = \lambda_j^{1/2} y_j, \tag{2}$$

where y_j are i.i.d. standard normal variables. The optimal approximation is given by the partial sum

$$A_n^*(f) = \sum_{j=1}^{n} \langle f, \xi_j^* \rangle \xi_j^*,$$

and

$$\mathrm{err}_\mu^{\mathrm{ave}}(A_n^*) = \sqrt{\sum_{j=n+1}^{+\infty} \lambda_j}.$$

Thus $A_n^*(f)$ is just the H-orthogonal projection of f onto $V_n^* = \mathrm{span}(\xi_1^*, \xi_2^*, \ldots, \xi_n^*)$ and therefore is simultaneously the best n-term linear approximation, $\mathrm{err}_\mu^{\mathrm{ave}}(A_n^*) = e_\mu^{\mathrm{ave}}(n)$. Even more, the dictionary

$$\mathscr{D}^* = \{\xi_1^*, \xi_2^*, \xi_3^*, \ldots\}$$

is optimal in the sense that it offers the minimal average errors of the n-term linear approximation for all $n \geq 1$.

It follows that in order to have an approximation with error ε it is necessary and sufficient to perform

$$n(\varepsilon) = \min\left\{ n \,\middle|\, \sum_{j=n+1}^{+\infty} \lambda_j \leq \varepsilon^2 \right\}$$

functional evaluations. A natural question is whether for $f \in F$ it is necessary to store all the $n(\varepsilon)$ coefficients $\langle f, \xi_j^* \rangle$, $1 \leq j \leq n(\varepsilon)$, to represent the approximation without losing its quality. The answer is strictly related to the quality of nonlinear approximation as follows.

Denote by $\widetilde{A}_m^*(f)$ the optimal m-term nonlinear approximation for f with respect to the dictionary $\mathscr{D}^*$. It is given as

$$\widetilde{A}_m^*(f) = \sum_{j \in U(f)} \langle f, \xi_j^* \rangle \xi_j^*,$$

where $U(f) = \{i_1, i_2, \dots, i_m\}$ is the set of m indexes j for which the coefficients $|\langle f, \xi_j^* \rangle|$ are largest possible. It is important to notice that this approximation is generally *not* feasible since it requires m largest coefficients from the potentially infinite number of coefficients. However, it is completely feasible to apply the best m-term nonlinear approximation to $A^*(f)$ where we know that only the first n coefficients $\langle f, \xi_j^* \rangle$ are nonzero. That is, we let

$$\widetilde{A}_{n,m}(f) = \widetilde{A}_m^*\big(A_n^*(f)\big) = \sum_{j \in U_n(f)} \langle f, \xi_j^* \rangle \xi_j^*,$$

where $m \le n$ and $U_n(f)$ is the set of m indexes j amongst $\{1, 2, \dots, n\}$ for which $|\langle f, \xi_j^* \rangle|$ are largest possible. Since $f - A_n^*(f)$ and $A_n^*(f) - \widetilde{A}_{n,m}(f) \in V_n^*$ are H-orthogonal, we have

$$\begin{aligned} \|f - \widetilde{A}_{n,m}(f)\|^2 &= \|f - A_n^*(f)\|^2 + \|A_n^*(f) - \widetilde{A}_m^*\big(A_n^*(f)\big)\|^2 \\ &\le \|f - A_n^*(f)\|^2 + \|f - \widetilde{A}_m^*(f)\|^2. \end{aligned}$$

This immediately implies

$$\widetilde{e}_\mu^{\,\mathrm{ave}}(m) \le \mathrm{err}_\mu^{\mathrm{ave}}(\widetilde{A}_{n,m}) \le \sqrt{e_\mu^{\mathrm{ave}}(n)^2 + \widetilde{e}_\mu^{\,\mathrm{ave}}(m)^2}$$

and $\lim_{n \to +\infty} \mathrm{err}_\mu^{\mathrm{ave}}(\widetilde{A}_{n,m}) = \widetilde{e}_\mu^{\,\mathrm{ave}}(m)$.

We conclude that it is indeed possible to reduce the number of coefficients in the expansion of the approximation of f without losing the quality of approximation, if $\widetilde{e}_\mu^{\,\mathrm{ave}}(n)$ decreases faster than $e_\mu^{\mathrm{ave}}(n)$ as $n \to +\infty$. Hence the question now is when and how much is the nonlinear approximation better than linear approximation.

3 A General Setting

In order to study the relation between best n-term linear and nonlinear approximations we consider a more general setting than that of Sect. 2; namely, we assume that the y_j's in the expansion (2) are not necessarily normally distributed. Then the problem "linear versus nonlinear approximation" can be conveniently reformulated as follows.

Let Z be a real nonnegative random variable with expectation

$$\mathbb{E}Z = 1,$$

and let $\boldsymbol{x} = (x_1, x_2, \ldots)$ be an infinite sequence of independent and identically distributed copies of Z. Let $\boldsymbol{a} = (a_1, a_2, \ldots)$ be a non-increasing and summable sequence,

$$a_1 \geq a_2 \geq \cdots \geq 0,$$

so that

$$\mathbb{E}\left(\sum_{j=1}^{+\infty} a_j x_j \right) = \sum_{j=1}^{+\infty} a_j < +\infty. \tag{3}$$

Suppose we want to maximally reduce the expectation (3) by removing n components $a_j x_j$ from the sum. It is clear that then the best nonadaptive strategy (n-term linear approximation) is to remove the n first components, and the best adaptive strategy (n-term nonlinear approximation) is to remove the n largest components.[1] We correspondingly define the errors

$$e_n(Z, \boldsymbol{a}) = \mathbb{E}\left(\sum_{j=n+1}^{+\infty} a_j x_j \right) = \sum_{j=n+1}^{+\infty} a_j,$$

and

$$\widetilde{e}_n(Z, \boldsymbol{a}) = \mathbb{E}\left(\sum_{j \notin U(\boldsymbol{x})} a_j x_j \right),$$

where $U(\boldsymbol{x})$ is the set of n indexes for which $a_j x_j$ are largest possible.

Note that the adaptive strategy and its error are well defined since the condition (3) ensures that the sum $\sum_{j=1}^{+\infty} a_j x_j$ is finite almost surely. Moreover, both strategies are independent of $\boldsymbol{a}$.

We immediately observe that $e_n(Z, \boldsymbol{a})$ is independent of a_j for $j \leq n$, and that both, $e_n(Z, \boldsymbol{a})$ and $\widetilde{e}_n(Z, \boldsymbol{a})$, depend monotonically on $\boldsymbol{a}$. That is, if $\tilde{\boldsymbol{a}} \leq \boldsymbol{a}$ (coordinatewise) then $e_n(Z, \tilde{\boldsymbol{a}}) \leq e_n(Z, \boldsymbol{a})$ and $\widetilde{e}_n(Z, \tilde{\boldsymbol{a}}) \leq \widetilde{e}_n(Z, \boldsymbol{a})$.

We are interested in the ratio

$$R_n(Z, \boldsymbol{a}) = \frac{\widetilde{e}_n(Z, \boldsymbol{a})}{e_n(Z, \boldsymbol{a})}$$

(with convention $0/0 = 1$). Obviously, $0 < R_n(Z, \boldsymbol{a}) \leq 1$.

[1] The word ‘adaptive’ here means that the removed components depend on $\boldsymbol{x}$.

Remark 1 The assumption $\mathbb{E}Z = 1$ is only to simplify some formulas. Since the quantities $e_n(Z, \boldsymbol{a})$ and $\widetilde{e}_n(Z, \boldsymbol{a})$ are homogeneous with respect to multiplication of Z by a constant, for any Z and $c > 0$ we have $R_n(cZ, \boldsymbol{a}) = R_n(Z, \boldsymbol{a})$.

3.1 Worst Case Scenario

We first fix Z and n and ask for the 'worst' possible $\boldsymbol{a}$, i.e., we are interested in the quantity

$$r_n(Z) = \inf_{\boldsymbol{a}} R_n(Z, \boldsymbol{a}).$$

In other words, $r_n(Z)$ is the maximum gain from using adaptive strategy instead of nonadaptive strategy. Knowing it is especially important when $\boldsymbol{a}$ is unknown.

Theorem 1 *For any Z and $n \geq 0$ we have*

$$r_n(Z) = \widetilde{e}_n(Z, \boldsymbol{a}^*)$$

where $\boldsymbol{a}^* = (\underbrace{1, 1, \ldots, 1}_{n+1}, 0, 0, 0, \ldots)$.

Proof We first show that the theorem holds in case where only finitely many a_j's are positive. Let k be the largest integer for which $a_{n+k} > 0$. We can assume without loss of generality that $k \geq 1$ since otherwise $e_n(Z, \boldsymbol{a}) = \widetilde{e}_n(Z, \boldsymbol{a}) = 0$. We proceed by induction on k.

Let $k = 1$. Then, letting $\tilde{\boldsymbol{a}} = (\underbrace{a_{n+1}, \ldots, a_{n+1}}_{n+1}, 0, 0, \ldots)$, we have

$$\widetilde{e}_n(Z, \boldsymbol{a}) \geq \widetilde{e}_n(Z, \tilde{\boldsymbol{a}}) = a_{n+1}\widetilde{e}_n(Z, \boldsymbol{a}^*) = e_n(Z, \boldsymbol{a})\widetilde{e}_n(Z, \boldsymbol{a}^*).$$

Assume $k \geq 2$. Then we have almost surely that $U(\boldsymbol{x}) = \{1, 2, \ldots, n+k\} \setminus T(\boldsymbol{x})$ where $T(\boldsymbol{x})$ is the set of k indexes with smallest $a_j x_j$, $1 \leq j \leq n+k$. It can be decomposed as

$$T(\boldsymbol{x}) = T_0(\boldsymbol{x}) \cup \{t(\boldsymbol{x})\},$$

where $T_0(\boldsymbol{x})$ is the set of $(k-1)$ 'smallest' indexes from $\{1, 2, \ldots, n+k-1\}$, and $t(\boldsymbol{x})$ is the 'smallest' index from $\{1, 2, \ldots, n+k\} \setminus T_0(\boldsymbol{x})$. By inductive assumption applied to the sequence $\boldsymbol{a}' = (a_1, a_2, \ldots, a_{n+k-1}, 0, 0, \ldots)$ we have

$$\mathbb{E}\left(\sum_{j \in T_0(\boldsymbol{x})} a_j x_j\right) = \widetilde{e}_n(Z, \boldsymbol{a}') \geq \widetilde{e}_n(Z, \boldsymbol{a}^*)\, e_n(Z, \boldsymbol{a}') = \widetilde{e}_n(Z, \boldsymbol{a}^*)\left(\sum_{j=n+1}^{n+k-1} a_j\right). \tag{4}$$

Observe also that the expectation $\mathbb{E}\left(a_{t(\boldsymbol{x})}x_{t(\boldsymbol{x})}\right)$ is not smaller than $\widetilde{e}_n(Z,\tilde{\boldsymbol{a}})$ where $\tilde{\boldsymbol{a}} = (\underbrace{a_{n+k},\ldots,a_{n+k}}_{n+1},0,0,\ldots)$, since for such $\tilde{\boldsymbol{a}}$ almost surely $t(\boldsymbol{x})$ is the 'smallest' index from $\{1,2,\ldots,n+1\}$. This and case $k=1$ yield

$$\mathbb{E}\left(a_{t(\boldsymbol{x})}x_{t(\boldsymbol{x})}\right) \ge \widetilde{e}_n(Z,\tilde{\boldsymbol{a}}) \ge \widetilde{e}_n(Z,\boldsymbol{a}^*)\,e_n(Z,\tilde{\boldsymbol{a}}) = \widetilde{e}_n(Z,\boldsymbol{a}^*)\,a_{n+k}. \tag{5}$$

Taking together (4) and (5) we obtain

$$\begin{aligned}\widetilde{e}_n(Z,\boldsymbol{a}) &= \mathbb{E}\Bigg(\sum_{j\in T(\boldsymbol{x})} a_jx_j\Bigg) = \mathbb{E}\Bigg(\sum_{j\in T_0(\boldsymbol{x})} a_jx_j + a_{t(\boldsymbol{x})}x_{t(\boldsymbol{x})}\Bigg)\\ &= \mathbb{E}\Bigg(\sum_{j\in T_0(\boldsymbol{x})} a_jx_j\Bigg) + \mathbb{E}\Big(a_{t(\boldsymbol{x})}x_{t(\boldsymbol{x})}\Big)\\ &\ge \widetilde{e}_n(Z,\boldsymbol{a}^*)\left(\sum_{j=n+1}^{n+k-1} a_j + a_{n+k}\right)\\ &= \widetilde{e}_n(Z,\boldsymbol{a}^*)\,e_n(Z,\boldsymbol{a}).\end{aligned}$$

Consider now an arbitrary $\boldsymbol{a} = (a_1,a_2,\ldots)$. Let $\boldsymbol{a}^m = (a_1,\ldots,a_m,0,0,\ldots)$. Since

$$\widetilde{e}_n(Z,\boldsymbol{a}^m) \le \widetilde{e}_n(Z,\boldsymbol{a}) \le \widetilde{e}_n(Z,\boldsymbol{a}^m) + \sum_{j=m+1}^{+\infty} a_j,$$

we have $\widetilde{e}_n(Z,\boldsymbol{a}) = \lim_{m\to+\infty}\widetilde{e}_n(Z,\boldsymbol{a}^m)$. A similar equality holds for $e_n(Z,\boldsymbol{a})$. Hence

$$R_n(Z,\boldsymbol{a}) = \lim_{m\to+\infty} R_n(Z,\boldsymbol{a}^m) \le R_n(Z,\boldsymbol{a}^*),$$

as claimed. □

Now we want to see how small $r_n(Z)$ can be, i.e., how much the adaptive strategy helps. In particular, we want to know how $r_n(Z)$ depends on the cummulative distribution function

$$F_Z(t) = \mathrm{Prob}(Z \le t)$$

of the random variable Z. For this end, observe that $r_n(Z) = \mathbb{E}X$ where

$$X = \min_{1\le j\le n+1} x_j.$$

Since

$$\begin{aligned}\mathrm{Prob}\,(X \le t) &= 1 - \mathrm{Prob}\,(X > t)\\ &= 1 - \prod_{j=1}^{n+1} \mathrm{Prob}\,\big(x_j > t\big)\\ &= 1 - (1 - F_Z(t))^{n+1},\end{aligned}$$

we obtain the following result.

Corollary 1

$$r_n(Z) = \int_0^{+\infty} (1 - F_Z(t))^{n+1}\,\mathrm{d}t.$$

We immediately see that $r_n(Z) = 1$ for all n if $Z \equiv 1$, and in all other cases $r_n(Z)$ is a strictly decreasing function of n. It can decrease even at exponential rate which is the case, for instance, for the Bernoulli trials. Indeed, suppose $Z = 0$ or $Z = 2$ with probabilities $1/2$. Then $r_n(Z)$ equals 2 times the probability of success in $(n + 1)$ trials, which is 2^{-n}. We can generalize this example as follows. Suppose $0 < F_Z(0) < 1$. Then

$$r_n(Z) \le (1 - F_Z(0))^n \int_0^{+\infty} (1 - F_Z(t))\,\mathrm{d}t = (1 - F_Z(0))^n,$$

i.e., $r_n(Z)$ decreases exponentially fast.

Actually, $r_n(Z)$ cannot decrease faster than exponentially. Indeed, let $t_0 > 0$ be such that $F_Z(t_0) < 1$. (Such t_0 always exists since otherwise $Z \equiv 0$.) Then

$$r_n(Z) \ge \int_0^{t_0} (1 - F_Z(t_0))^{n+1}\,\mathrm{d}t = t_0(1 - F_Z(t_0))^{n+1},$$

as claimed.

On the other hand, if $F_Z(t_0) = 0$ for some $t_0 > 0$ then $r_n(Z) \ge t_0$ and $r_n(Z)$ does not converge to zero; that is, adaptive strategy does not help.

The discussion above shows that the really ‘interesting’ cases are those for which $F_Z(0) = 0$ and $F_Z(t) > 0$ for all $t > 0$. Since then the integral

$$\int_{t_0}^{+\infty} (1 - F_Z(t))^{n+1}\,\mathrm{d}t \le (1 - F_Z(t_0))^n \int_{t_0}^{+\infty} (1 - F_Z(t))\,\mathrm{d}t$$

converges exponentially fast to zero for any positive t_0, the convergence of $r_n(Z)$ depends only on $F_Z(t)$ in the neighborhood of 0.

For illustration, we consider two examples.

Example 1 Let

$$F_Z(t) = \begin{cases} (t/c)^{1/p} & 0 \le t \le c, \\ 1 & t > c. \end{cases}$$

Here $p > 0$ and $c = p + 1$ is the normalizing parameter. In other words, $Z = c\,|U|^p$ where U is uniform on $[-1, 1]$. In this case we can derive exact formula for $r_n(Z)$; namely, we have

$$r_n(Z) = \int_0^c \left(1 - (t/c)^{1/p}\right)^{n+1} \mathrm{d}t = c\,\psi(n, p) \tag{6}$$

where

$$\psi(n, p) = p \int_0^1 (1 - u)^{n+1} u^{p-1}\, \mathrm{d}u.$$

Integrating by parts we find that

$$\psi(n, p) = \left(\frac{n+1}{p+1}\right) \psi(n - 1, p + 1)$$

which, together with (6) and $\psi(0, p) = 1/(p + 1)$, yields the exact formula

$$r_n(Z) = \prod_{k=2}^{n+1} \frac{k}{p+k}. \tag{7}$$

We have that $r_n(Z) \asymp n^{-p}$ as $n \to +\infty$.[2] Indeed we take the logarithm in (7) and then estimate the resulting sum from below and from above by the integral of $\ln(x/(p + x))$ on the intervals $[1, n + 1]$ and $[3/2, n + 3/2]$, respectively. Using these estimates we obtain the desired result with the asymptotic constant between $\mathrm{e}^{-p}(p + 1)^{p+1}$ and $\mathrm{e}^{-p}(2/3)^{3/2}(p + 3/2)^{p+3/2}$.

Example 2 Suppose now that $Z = c\,|G|^p$, $p > 0$, where $G \sim \mathcal{N}(0, 1)$ is the standard Gaussian (normal) distribution on $\mathbb{R}$, and $c = (\mathrm{e}|G|^p)^{-1}$ is the normalizing parameter. That is,

$$F_Z(t) = \sqrt{\frac{2}{\pi}} \int_0^{(t/c)^{1/p}} \exp(-u^2/2)\, \mathrm{d}u.$$

In this case we have $F_Z(t) \approx \sqrt{2/\pi}\,(t/c)^{1/p} \asymp t^{1/p}$ as $t \to 0^+$ which, by Example 1, immediately yields $r_n(Z) \asymp n^{-p}$.

[2] For two positive sequences, we write $a_n \asymp b_n$ iff there are $0 < c < C < +\infty$ and n_0 such that $c \le a_n/b_n \le C$ holds for all $n \ge n_0$. We write $a_n \approx b_n$ iff $\lim_{n\to+\infty} a_n/b_n = 1$.

3.2 Asymptotics

For a given problem, i.e., for given $\boldsymbol{a}$, we typically ask how fast the error for adaptive strategy converges to zero as $n \to +\infty$ compared to the error of nonadaptive strategy. The results of the previous section do not provide a complete answer to this question since the 'worst' $\boldsymbol{a}$ depends on n. What we know right now is that $R_n(Z, \boldsymbol{a})$ cannot decay faster than exponentially, and this exponential decay is achieved, e.g., for the Bernoulli trials. We now investigate in more detail the behavior of $R_n(Z, \boldsymbol{a})$ for fixed $\boldsymbol{a}$ and $n \to +\infty$.

We first show the following simple, but useful fact.

Lemma 1 *Let $0 < p \le 1$. Suppose that Z takes only two values: $1/p$ (success) and 0 (failure), with probability of success p. Then*

$$\widetilde{e}_n(Z, \boldsymbol{a}) = \sum_{k=n+1}^{+\infty} a_k \, q_{n,k-1}$$

where $q_{n,k} = \sum_{i=n}^{k} \binom{k}{i} p^i (1-p)^{k-i}$ is the probability of at least n successes in k trials.

Proof Observe that the n largest components $a_k x_k$ are just the n first nonzero components. Letting A_i be the event that x_i is a success and there are exactly n successes in the first i trials, we have

$$\begin{aligned}\widetilde{e}_n(Z, \boldsymbol{a}) &= \sum_{i=n}^{+\infty} \mathbb{E}\left(\sum_{k=i+1}^{+\infty} a_k x_k \middle| A_i \right) \mathrm{Prob}(A_i) = \sum_{i=n}^{+\infty} \left(\sum_{k=i+1}^{+\infty} a_i \right) \mathrm{Prob}(A_i) \\ &= \sum_{k=n+1}^{+\infty} a_k \left(\sum_{i=n}^{k-1} \mathrm{Prob}(A_i) \right) = \sum_{k=n+1}^{+\infty} a_k q_{n,k-1},\end{aligned}$$

as claimed. □

Using the above lemma we can easily show that it is possible to achieve exponential decay of $R_n(Z, \boldsymbol{a})$ if a_k rapidly goes to zero.

Example 3 Let $a_1 = 1$ and $a_{k+1} = 2^{-k} a_k$, $k \ge 1$, so that

$$a_k = 2^{-\frac{(k-1)k}{2}}.$$

Consider the nonlinear approximation for Z as in Lemma 1 with $p = 1/2$. Then

$$\widetilde{e}_n(Z, \boldsymbol{a}) = p^n a_{n+1} + \sum_{k=n+2}^{+\infty} a_k \le \frac{a_{n+1}}{2^n} + a_{n+1} \left(\frac{1}{2^n} + \frac{1}{2^n 2^{n+1}} + \cdots \right) \le \frac{3 a_{n+1}}{2^n}.$$

Since we always have $e_n(Z,\boldsymbol{a}) \geq a_{n+1}$, then

$$R_n(Z,\boldsymbol{a}) \;\leq\; \frac{3}{2^n}.$$

This example has only theoretical interest. More important is the following result.

Theorem 2 *If the sequence $\boldsymbol{a} = (a_1, a_2, \ldots)$ decays no faster than polynomially then for any Z*

$$R_n(Z,\boldsymbol{a}) \;\asymp\; 1.$$

Hence adaptive schemes do not give a better convergence rate than nonadaptive schemes.

Proof Choose an arbitrary $t_0 > 0$ such that

$$p := \mathrm{Prob}(Z > t_0) > 0,$$

and consider the two-valued random variable Z' which takes t_0 with probability p and 0 with probability $1-p$. (Note that $\mathbb{E}(Z') = pt_0$ which need not be 1.) Since for the corresponding cummulative distribution functions is $F_{Z'}(t) \geq F_Z(t)$ for all t, we have

$$\widetilde{e}_n(Z,\boldsymbol{a}) \;\geq \widetilde{e}_n(Z',\boldsymbol{a}).$$

For $k \geq \lceil n/p \rceil$, the probability $q_{n,k}$ that there are at least n successes in k trials can be bounded from below by some $c > 0$ for all n sufficiently large, $n \geq n_0$ (where $c \approx 1/2$ as $n_0 \to +\infty$). Then, by Lemma 1 and polynomial decay of $a_1, a_2, a_3 \ldots,$

$$\widetilde{e}_n(Z',\boldsymbol{a}) \;\geq\; p\, t_0\, c \sum_{k=\lceil n/p \rceil}^{+\infty} a_k \;\asymp\; \sum_{k=n+1}^{+\infty} a_k \;=\; e_n(Z,\boldsymbol{a})$$

as claimed. □

4 Back to Gaussian Processes

We now relate the results of Sect. 3 to the average case approximation with respect to Gaussian measures of Sect. 2. Using the notation of Sect. 3, this corresponds to $Z = |G|^2$ with $G \sim \mathscr{N}(0,1)$, as in Example 2, and

$$\frac{\widetilde{e}_\mu^{\,\mathrm{ave}}(n)}{e_\mu^{\mathrm{ave}}(n)} \;=\; \sqrt{R_n(Z,\boldsymbol{a})},$$

where $\boldsymbol{a} = (\lambda_1, \lambda_2, \ldots)$ are given by (1). We have the following corollary.

Corollary 2 *Consider the average case linear and nonlinear approximations with respect to Gaussian measures μ.*

- *There exists $c > 0$ such that for any μ and any n*

$$\frac{\widetilde{e}^{\mathrm{ave}}_{\mu}(n)}{e^{\mathrm{ave}}_{\mu}(n)} \geq \frac{c}{n}.$$

- *If the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ associated with μ decay no faster than polynomially then*

$$\liminf_{n\to+\infty} \frac{\widetilde{e}^{\mathrm{ave}}_{\mu}(n)}{e^{\mathrm{ave}}_{\mu}(n)} > 0.$$

We stress that the polynomial decay of the eigenvalues λ_j is rather standard for Gaussian measures, such as the Wiener sheet measures. In these cases, nonlinear approximation does not give better convergence rate than linear approximation.

Remark 2 In this final remark, we compare the results of this paper with the corresponding results in the worst case setting, where the error of approximation is defined by the worst case error with respect to the unit ball of F (instead of the average error with respect to μ). We now assume that $(F, \|\cdot\|_F)$ is a Hilbert space that is compactly embedded in $(H, \|\cdot\|)$. The worst case error of an approximation A_n is defined as

$$\mathrm{err}^{\mathrm{wor}}(A_n) := \sup_{\|f\|_F\leq 1} \|f - A_n(f)\|.$$

In this setting, the best n-term linear approximation A_n^* is as follows, see, e.g., [15]. Let $W : H \to H$ be defined by the equation $\langle \cdot, f\rangle = \langle \cdot, Wf\rangle_F$. Then W is a self-adjoint, nonnegative definite and compact operator. Let $\xi_j^*, j \geq 1$, be a complete and orthonormal in H system of eigenelements of W, and

$$W\xi_j^* = \lambda_j \xi_j^*, \qquad j \geq 1,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \quad \lim_{j\to+\infty} \lambda_j = 0$. Then

$$A_n^*(f) = \sum_{j=1}^{n} \langle f, \xi_j^* \rangle \xi_j^*$$

and $\mathrm{err}^{\mathrm{wor}}(A_n^*) = \sqrt{\lambda_{n+1}}$. It is not difficult to see that the error of the best n-term nonlinear approximation $\widetilde{A}_n^*$ for the dictionary

$$\mathscr{D}^* = \{\xi_1^*, \xi_2^*, \xi_3^*, \ldots\}$$

equals

$$\mathrm{err}^{\mathrm{wor}}(\widetilde{A}_n^*) = \left(\sum_{j=1}^{n+1} \lambda_j^{-1}\right)^{-1/2},$$

and is achieved when $f = \sum_{i=1}^{+\infty} f_i \xi_i^*$ with

$$f_i = \left(\lambda_i \sum_{j=1}^{n+1} \lambda_j^{-1}\right)^{-1/2} \quad \text{for} \quad 1 \le i \le n+1,$$

and $f_i = 0$ for $i \ge n+1$. Hence the ratio between the errors of the best n-term nonlinear and linear approximations is

$$\frac{\widetilde{e}^{\mathrm{wor}}(n)}{e^{\mathrm{wor}}(n)} = \left(\sum_{j=1}^{n+1} \frac{\lambda_{n+1}}{\lambda_j}\right)^{-1/2}.$$

We easily see that this ratio is minimal when $\lambda_j = 1$ for all $1 \le j \le n+1$, and then it equals $(n+1)^{-1/2}$. We have the same order for polynomially decaying λ_j's, and only for very fast decaying λ_j's the nonlinear approximation does not help. Hence we have an opposite situation to that in the average case, where the faster the eigenvalues decay the more the nonlinear approximation helps.

References

1. Cioica, P.A., Dahlke, S., Döhring, N., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.L.: Adaptive wavelet methods for the stochastic Poisson equation. BIT Numer. Math. **52**, 589–614 (2012)
2. Cohen, A., D'Ales, J.-P.: Nonlinear approximation of random functions. SIAM J. Appl. Math. **57**, 518–540 (1997)
3. Cohen, A., Daubechies, I., Guleryuz, O.G., Orchard, M.T.: On the importance of combining wavelet-based nonlinear approximation with coding strategies. IEEE Trans. Inf. Theory **48**, 1895–1921 (2002)
4. Creutzig, J., Müller-Gronbach, T., Ritter, K.: Free-knot spline approximation of stochastic processes. J. Complexity **23**, 867–889 (2007)
5. DeVore, R.A.: Nonlinear approximation. Acta Numer. **8**, 51–150 (1998)
6. DeVore, R.A., Jawerth, B.: Optimal nonlinear approximation. Manuscripta Math. **63**, 469–478 (1992)
7. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
8. Kon, M.A., Plaskota, L.: Information complexity of neural networks. Neural Netw. **13**, 365–376 (2000)
9. Kon, M.A., Plaskota, L.: Information-based nonlinear approximation: an average case setting. J. Complexity **21**, 211–229 (2005)
10. Kuo, H.-H.: Gaussian Measures in Banach Spaces. Lecture Notes in Mathematics, vol. 463. Springer, New York (1975)

11. Plaskota, L.: Noisy Information and Computational Complexity. Cambridge University Press, Cambridge (1996)
12. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery, vol. 9, pp. 1–92. Walter de Gruyter, Berlin (2012)
13. Ritter, K.: Average-Case Analysis of Numerical Problems. Springer, Berlin (2000)
14. Temlyakov, V.N.: Nonlinear methods of approximation. Found. Comput. Math. **3**, 33–107 (2003)
15. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic Press, New York (1988)
16. Vakhania, N.N., Tarieladze, V.I., Chobanyan, S.A.: Probability Distributions on Banach Spaces. Kluwer, Dordrecht (1987)
17. Vybiral, J.: Average best *m*-term approximation. Constr. Approx. **36**, 83–115 (2012)

Integral Equations, Quasi-Monte Carlo Methods and Risk Modeling

Michael Preischl, Stefan Thonhauser, and Robert F. Tichy

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract We survey a QMC approach to integral equations and develop some new applications to risk modeling. In particular, a rigorous error bound derived from Koksma-Hlawka type inequalities is achieved for certain expectations related to the probability of ruin in Markovian models. The method is based on a new concept of isotropic discrepancy and its applications to numerical integration. The theoretical results are complemented by numerical examples and computations.

1 Introduction

During the last two decades quasi-Monte-Carlo methods (QMC-methods) have been applied to various problems in numerical analysis, statistical modeling and mathematical finance. In this paper we will give a brief survey on some of these developments and present new applications to more refined risk models involving discontinuous processes. Let us start with Fredholm integral equations of the second kind:

$$f(\mathbf{x}) = g(\mathbf{x}) + \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}, \tag{1}$$

where the kernel is given by $K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ with $k(\mathbf{x})$ having period 1 in each component of $\mathbf{x} = (x_1, \ldots, x_s)$. As it is quite common in applications of QMC-methods (see for example [13, 26, 37]) it is assumed that g and k belong to a weighted Korobov space. Of course, there exists a vast literature concerning the numerical solution of Fredholm equations, see for instance [7, 23] or [40]. In

M. Preischl (✉) · S. Thonhauser · R. F. Tichy
Institute of Analysis and Number Theory, Graz University of Technology, Graz, Austria
e-mail: preischl@math.tugraz.at; stefan.thonhauser@math.tugraz.at; tichy@tugraz.at

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_47

particular, we want to mention the work of I. Sloan in the late 1980's where he explored various quadrature rules for solving integral equations and applications to engineering problems ([35, 36] and [25]), which have also, after some modifications, been applied to Volterra type integral equations (see [10] or [11]). In [13] the authors approximate f using the Nyström method based on QMC rules.

For points $\mathbf{t}_1, \dots, \mathbf{t}_N$ in $[0, 1]^s$ the N-th approximation of f is given by

$$f_N(\mathbf{x}) := g(\mathbf{x}) + \frac{1}{N} \sum_{n=1}^{N} K(\mathbf{x}, \mathbf{t}_n) f_N(\mathbf{t}_n), \tag{2}$$

where the function values $f_N(\mathbf{t}_1), \dots, f_N(\mathbf{t}_N)$ are obtained by solving the linear system

$$f_N(\mathbf{t}_j) = g(\mathbf{t}_j) + \frac{1}{N} \sum_{n=1}^{N} K(\mathbf{t}_j, \mathbf{t}_n) f_N(\mathbf{t}_n), \; j = 1, \dots, N. \tag{3}$$

Under some mild conditions on K, N, and the integration points $\mathbf{t}_1, \dots, \mathbf{t}_N$, it is shown in [13] that there exists a unique solution of (3). Furthermore, the authors analyze the worst case error of this, so-called QMC-Nyström method. In addition, good lattice point sets $\mathbf{t}_1, \dots, \mathbf{t}_N$ are presented. Its convergence rate is best possible. A special focus of this important paper lies on the study of tractability and strong tractability of the QMC-Nyström method. For tractability theory in general we refer to the fundamental monograph of [31]. Using ideas of Hlawka [21] the third author of the present paper worked on iterative methods for solving Fredholm and Volterra equations, see also Hua-Wang [22].

The idea is to approximate the solution of integral equations by means of iterated (i.e. multi-dimensional) integrals. The convergence of this procedure follows from Banach's fixed point theorem and error estimates can be established following the proof of the Picard-Lindelöf approximation for ordinary differential equations. To be more precise, let us consider integration points $\mathbf{t}_1, \dots, \mathbf{t}_N \in [0, 1]^s$ with star discrepancy D_N^* defined as usual by

$$D_N^* = \sup_{J \subset [0,1]^s} \left| \frac{1}{N} \sharp\{n \le N : \mathbf{t}_n \in J\} - \lambda(J) \right|, \tag{4}$$

where the supremum is taken over all axis-aligned boxes J with one vertex in the origin and Lebesgue measure $\lambda(J)$. In [39] the following system of r integral equations has been considered for given functions g_j on $[0, 1]^{s+r}$ and h_j on $[0, 1]^s$:

$$\begin{aligned} f_j(\mathbf{x}) = \int_0^{x_1} \dots \int_0^{x_s} & g_j(\xi_1, \dots, \xi_s, f_1(\boldsymbol{\xi}), \dots, f_r(\boldsymbol{\xi})) d\xi_s \dots d\xi_1 \\ & + h_j(\mathbf{x}) \, , \; j = 1, \dots, r \end{aligned} \tag{5}$$

where we have used the notations $\mathbf{x} = (x_1, \ldots, x_s) \in [0,1]^s$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_s)$. Furthermore, we assume that the partial derivatives up to order s of the functions g_j and h_j, $j = 1, \ldots, r$, are bounded by some constants G and H, respectively. Then, for a given point set $\mathbf{t}_1, \ldots, \mathbf{t}_N$ in $[0,1]^s$ with discrepancy D_N^*, the solution $\mathbf{f} = (f_1, \ldots, f_r)$ of the system (5) can be approximated by the quantities $\mathbf{f}^{(k)} = (f_1^{(k)}, \ldots, f_r^{(k)})$, given recursively by

$$f_j^{(k+1)}(\mathbf{x}) = \frac{x_1 \cdots x_s}{N} \sum_{n=1}^{N} g_j(x_1 t_{1,n}, \ldots, x_s t_{s,n}, f_1^{(k)}(\mathbf{x} \cdot \mathbf{t}_n), \ldots, f_r^{(k)}(\mathbf{x} \cdot \mathbf{t}_n)); \tag{6}$$

here $\mathbf{x} \cdot \mathbf{t}_n$ stands for the inner product $x_1 t_{1,n} + \ldots + x_s t_{s,n}$, where $\mathbf{t}_n = (t_{1,n}, \ldots, t_{s,n})$. In [39] it is shown, that based on the classical Koksma-Hlawka inequality the worst case error, i.e., $\| \mathbf{f}^{(k)} - \mathbf{f} \|_\infty$ (sum of componentwise supremum norms) can be estimated in terms of the bounds G and H and the discrepancy D_N^* of the integration points. This method was also extended to integral equations with singularities, such as Abel's integral equation. The main focus of the present paper lies on applications in mathematical finance. In Albrecher and Kainhofer [3] the above method was used for the numerical solution of certain Cramér-Lundberg models in risk theory. However, it turned out that in these models certain discontinuities occur. This means, that one cannot assume bounds for the involved partial derivatives and simply apply the classical Koksma-Hlawka inequality. Moreover, the involved functions are indicator functions of simplices, thus not of bounded variation in the sense of Hardy and Krause, see Drmota and Tichy [14] and Kuipers and Niederreiter [24].

Albrecher and Kainhofer [3] considered a risk model with non-linear dividend barrier and made some assumptions to overcome the difficulties caused by discontinuities. For such applications it could help to use a different notion of variation for multivariate functions. Götz [18] proved a version of the Koksma-Hlawka inequality for general measures, Aistleitner and Dick [1] considered functions of bounded variation with respect to signed measures and Brandolini et al. [8, 9] replaced the integration domain $[0,1]^s$ by an arbitrary bounded Borel subset of $\mathbb{R}^s$ and proved the inequality for piecewise smooth integrands. Based on fundamental work of Harman [20], a new concept of variation was developed for a wide class of functions, see Pausinger and Svane [33] and Aistleitner et al. [2].

In the following we give a brief overview on concepts of multivariate variation and how they can be applied for error estimates in numerical integration. Let $f(\mathbf{x})$ be a function on $[0,1]^s$ and $\mathbf{a} = (a_1, \ldots, a_s) \leq \mathbf{b} = (b_1, \ldots, b_s)$ points in $[0,1]^s$, where $\leq$ denotes the natural componentwise partial order. Following the notation of Owen [32] and Aistleitner et al. [2] for a subset $u \subseteq \{1, \ldots, s\}$ we denote by $\mathbf{a}^u : \mathbf{b}^{-u}$ the point with ith coordinate equal to a_i if $i \in u$ and equal to b_i otherwise. Then for the box $R = [\mathbf{a}, \mathbf{b}]$ we introduce the s-dimensional difference operator

$$\Delta^{(d)}(f; R) = \Delta(f; R) = \sum_u (-1)^{|u|} f(\mathbf{a}^u : \mathbf{b}^{-u}),$$

where the summation is extended over all subsets $u \in \{1, \ldots, s\}$ with cardinality $|u|$ and complement $-u$. Next we define partitions of $[0,1]^s$ as they are used in the theory of multivariate Riemann integrals, which we call here *ladder*. A ladder $\mathscr{Y}$ in $[0,1]^s$ is the Cartesian product of one-dimensional partitions $0 = y_1^j < \ldots < y_{k_j}^j < 1$ (in any dimension $j = 1, \ldots, s$). Define the successor $(y_i^j)_+$ of y_i^j to be y_{i+1}^j if $i < k_j$ and $(y_{k_j}^j)_+ = 1$. For $\mathbf{y} = (y_{i_1}^1, \ldots, y_{i_s}^s) \in \mathscr{Y}$ we define the successor $\mathbf{y}_+ = ((y_{i_1}^1)_+, \ldots, (y_{i_s}^s)_+)$ and have

$$\Delta(f; [0,1]^s) = \sum_{\mathbf{y} \in \mathscr{Y}} \Delta(f; [\mathbf{y}, \mathbf{y}_+]).$$

Using the notation

$$V_{\mathscr{Y}}(f; [0,1]^s) = \sum_{\mathbf{y} \in \mathscr{Y}} \Delta(f; [\mathbf{y}, \mathbf{y}_+])$$

the Vitali variation of$f$ over $[0,1]^s$ is defined by

$$V(f; [0,1]^s) = \sup_{\mathscr{Y}} V_{\mathscr{Y}}(f; [0,1]^s). \tag{7}$$

Given a subset $u \subseteq \{1, \ldots, s\}$, let

$$\Delta_u(f; [\mathbf{a}, \mathbf{b}]) = \sum_{v \subseteq u} (-1)^{|v|} f(\mathbf{a}^v : \mathbf{b}^{-v})$$

and set $\mathbf{0} = (0, \ldots, 0), \mathbf{1} = (1, \ldots, 1) \in [0,1]^s$. For a ladder $\mathscr{Y}$ there is a corresponding ladder $\mathscr{Y}_u$ on the $|u|$-dimensional face of $[0,1]^s$ consisting of points of the form $\mathbf{x}^u : \mathbf{1}^{-u}$. Clearly,

$$\Delta_u(f; [0,1]^s) = \sum_{\mathbf{y} \in \mathscr{Y}_u} \Delta_u(f; [\mathbf{y}, \mathbf{y}_+]).$$

Using the notation

$$V_{\mathscr{Y}_u}(f; [0,1]^s) = \sum_{\mathbf{y} \in \mathscr{Y}_u} \Delta_u(f; [\mathbf{y}, \mathbf{y}_+])$$

for the variation over the ladder $\mathscr{Y}_u$ of the restriction of f to the face of $[0,1]^s$ specified by u, the Hardy-Krause variation is defined as

$$\mathscr{V}(f) = \mathscr{V}_{HK}(f; [0,1]^s) = \sum_{\emptyset \neq u \subseteq \{1, \ldots, s\}} \sup_{\mathscr{Y}_u} V_{\mathscr{Y}_u}(f; [0,1]^s).$$

Assuming that f is of bounded Hardy-Krause variation, the classical Koksma-Hlawka inequality reads as follows:

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) - \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x} \right| \leq \mathscr{V}(f) D_N^*, \tag{8}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is a finite point set in $[0, 1]^s$ with star discrepancy D_N^*. In the case $f : [0, 1]^s \to \mathbb{R}$ has continuous mixed partial derivatives up to order s the Vitali variation (7) is given by

$$\mathscr{V}(f; [0, 1]^s) = \int_{[0,1]^s} \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\mathbf{x}) \right| d\mathbf{x}. \tag{9}$$

Summing over all non-empty subsets $u \subseteq [0, 1]^s$ immediately yields an explicit formula for the Hardy-Krause variation in terms of integrals of partial derivatives, see Leobacher and Pillichshammer [27, Ch.3, p. 59]. In particular, the Hardy-Krause variation can be estimated from above by an absolute constant if we know global bounds on all partial derivatives up to order s.

In the remaining part of the introduction we briefly sketch a more general concept of multidimensional variation which was recently developed in [33]. Let $\mathscr{D}$ denote an arbitrary family of measurable subsets of $[0, 1]^s$ which contains the empty set $\emptyset$ and $[0, 1]^s$. Let $\mathscr{L}(\mathscr{D})$ denote the $\mathbb{R}$-vectorspace generated by the system of indicator functions $\mathbf{1}_A$ with $A \in \mathscr{D}$.

A set $A \subseteq [0, 1]^s$ is called an algebraic sum of sets in $\mathscr{D}$ if there exist $A_1, \ldots, A_m \in \mathscr{D}$ such that

$$\mathbf{1}_A = \sum_{i=1}^{n} \mathbf{1}_{A_i} - \sum_{i=n+1}^{m} \mathbf{1}_{A_i},$$

and $\mathscr{A}$ is defined to be the collection of algebraic sums of sets in $\mathscr{D}$. As in [33] we define the Harman complexity $h(A)$ of a non-empty set $A \in \mathscr{A}, A \neq [0, 1]^s$ as the minimal number m such there exist $A_1, \ldots, A_m$ with

$$\mathbf{1}_A = \sum_{i=1}^{n} \mathbf{1}_{A_i} - \sum_{i=n+1}^{m} \mathbf{1}_{A_i},$$

for some $1 \leq n \leq m$ and $A_i \in \mathscr{D}$ or $[0, 1]^s \setminus A_i \in \mathscr{D}$. Moreover, set $h([0, 1]^s) = h(\emptyset) = 0$ and for $f \in \mathscr{L}(\mathscr{D})$

$$V_{\mathscr{D}}^*(f) = \inf \left\{ \sum_{i=1}^{m} |\alpha_i| h_{\mathscr{D}(A_i)} : f = \sum_{i=1}^{m} \alpha_i \mathbf{1}_{A_i},\ \alpha_i \in \mathbb{R},\ A_i \in \mathscr{D} \right\}.$$

Furthermore, let $\mathscr{V}_\infty(\mathscr{D})$ denote the collection of all measurable, real-valued functions on $[0,1]^s$ which can be uniformly approximated by functions in $\mathscr{L}(\mathscr{D})$. Then the $\mathscr{D}$-variation of $f \in \mathscr{V}_\infty(\mathscr{D})$ is defined by

$$V_{\mathscr{D}}(f) = \inf\{ \liminf_{i\to\infty} V^*_{\mathscr{D}}(f_i) : f_i \in \mathscr{L}(\mathscr{D}), f = \lim_{i\to\infty} f_i \}, \tag{10}$$

and set $V_{\mathscr{D}}(f) = \infty$ if $f \notin \mathscr{V}_\infty(\mathscr{D})$. The space of functions of bounded $\mathscr{D}$-variation is denoted by $\mathscr{V}(\mathscr{D})$. Important classes of sets $\mathscr{D}$ are the class $\mathscr{K}$ of convex sets and the class $\mathscr{R}^*$ of axis aligned boxes containing $\mathbf{0}$ as a vertex. In Aistleitner et al. [2] it is shown that the Hardy-Krause variation $\mathscr{V}(f)$ coincides with $\mathscr{V}_{\mathscr{R}^*}(f)$. For various applications the $\mathscr{D}$-variation seems to be a more natural and suitable concept. A convincing example concerning an application to computational geometry is due to Pausinger and Edelsbrunner [15]. Pausinger and Svane [33] considered the variation $\mathscr{V}_{\mathscr{K}}(f)$ with respect to the class of convex sets. They proved the following version of the Koksma-Hlawka inequality:

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) - \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x} \right| \leq \mathscr{V}_{\mathscr{K}}(f) \tilde{D}_N,$$

where $\tilde{D}_N$ is the isotropic discrepancy of the point set $\mathbf{x}_1, \ldots, \mathbf{x}_N$, which is defined as follows

$$\tilde{D}_N = \sup_{C \in \mathscr{K}} \left| \frac{1}{N} \sharp\{n \leq N : \mathbf{x}_n \in C\} - \lambda(C) \right|.$$

Pausinger and Svane [33] have shown that twice continuously differentiable functions f admit finite $\mathscr{V}_{\mathscr{K}}(f)$, and in addition they gave a bound which will be useful in our context.

Our paper is structured as follows. In Sect. 2 we introduce specific Markovian models in risk theory where in a natural way integral equations occur. These equations are based on arguments from renewal theory and only in particular cases they can be solved analytically. In Sect. 3 we develop a QMC method for such equations. We give an error estimate based on Koksma-Hlawka type inequalities for such models. In Sect. 4 we compare our numerical results to exact solutions in specific instances.

2 Discounted Penalties in the Renewal Risk Model

2.1 Stochastic Modeling of Risks

In the following we assume a stochastic basis $(\Omega, \mathscr{F}, P)$ which is large enough to carry all the subsequently defined random variables. In risk theory the surplus process of an insurance portfolio is modeled by a stochastic process $X = (X_t)_{t\geq 0}$. In the classical risk model, going back to Lundberg [29], X takes the form

$$X_t = x + c\,t - \sum_{i=1}^{N_t} Y_i, \tag{11}$$

where the deterministic quantities $x \geq 0$ and $c \geq 0$ represent the initial capital and the premium rate. The stochastic ingredient $S_t = \sum_{i=1}^{N_t} Y_i$ is the cumulated claims process which is a compound Poisson process. The jump heights—or claim amounts—are $\{Y_i\}_{i\in\mathbb{N}}$ for which $Y_i \overset{iid}{\sim} F_Y$ with $F_Y(0) = 0$. The counting process $N = (N_t)_{t\geq 0}$ is a homogeneous Poisson process with intensity $\lambda > 0$. A crucial assumption in the classical model is the independence between $\{Y_i\}_{i\in\mathbb{N}}$ and N. A major topic in risk theory is the study of the ruin event. We introduce the time of ruin $\tau = \inf\{t \geq 0 \,|\, X_t < 0\}$, i.e., the first point in time at which the surplus becomes negative. In this setting τ is a stopping time with respect to the filtration generated by X, $\{\mathscr{F}_t^X\}_{t\geq 0}$ with $\mathscr{F}_t^X = \sigma\{X_s \,|\, 0 \leq s \leq t\}$. A first approach for quantifying the risk of X, is the study of the associated ruin probability

$$\psi(x) = P_x(X_t < 0 \ \text{ for some } \ t \geq 0) = P_x(\tau < \infty),$$

which is non-degenerate if $\mathbb{E}_x(X_1) > 0$, and satisfies the integral equation

$$\frac{c}{\lambda}\psi(x) = \int_x^\infty (1 - F_Y(y))dy + \int_0^x \psi(x-y)(1 - F_Y(y))dy.$$

In Gerber and Shiu [16, 17] so-called discounted penalty functions are introduced. This concept allows for an integral ruin evaluation and is based on a function $w : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ which links the deficit at ruin $|X_\tau|$ and the surplus prior to ruin $X_{\tau-} := \lim_{t\nearrow\tau} X_t$ via the function

$$V(x) = \mathbb{E}_x\left(e^{-\delta\tau} w(|X_\tau|, X_{\tau-})\,\mathbf{1}_{\{\tau<\infty\}}\right).$$

The time of ruin τ is included by means of a discounting factor $\delta > 0$ which gives more weight to an early ruin event. In this setting specific choices of w allow for an unified treatment of ruin related quantities. In the literature, this kind of expected, discounted penalty function is often called a Gerber-Shiu function.

Remark 1 When putting a focus on the study of $\psi(x)$, the condition $\mathbb{E}_x(X_1) > 0$ is crucial. It says that on average premiums exceed claim payments in one unit of time. Standard results, see Asmussen and Albrecher [6], show that under this condition $\lim_{t\to\infty} X_t = +\infty$ P-a.s. From an economic perspective the accumulation of an infinite surplus is unrealistic and risk models including shareholder participation via dividends are introduced in the literature. We refer to [6] for model extensions in this direction.

2.2 Markovian Risk Model

In the following we consider an insurance surplus process $X = (X_t)_{t \geq 0}$ of the form

$$X_t = x + \int_0^t c(X_{s-})ds - \sum_{i=1}^{N_t} Y_i.$$

The quantity $x \geq 0$ is called the initial capital, the cumulated claims are represented by $S_t = \sum_{i=1}^{N_t} Y_i$ and the state-dependent premium rate is $c(\cdot)$. The cumulated claims process $S = (S_t)_{t \geq 0}$ is given by a sequence $\{Y_i\}_{i \in \mathbb{N}}$ of positive, independently and identically distributed (iid) random variables and a counting process $N = (N_t)_{t \geq 0}$. For convenience we assume that the claims distribution admits a continuous density $f_Y : \mathbb{R}^+ \to \mathbb{R}^+$. In our setup we model the claim counting process $N = (N_t)_{t \geq 0}$ as a renewal counting process which is specified by the inter-jump times $\{W_i\}_{i \in \mathbb{N}}$ which are positive and iid random variables. Then, the time of the i-th jump is $T_i = W_1 + \ldots + W_i$ and if we assume that W_1 admits a density f_W, the jump intensity of the process X is $\lambda(t') = \frac{f_W(t')}{1 - \int_0^{t'} f_W(s)ds}$. Here t' denotes the *time since the last jump*. A common assumption we are going to adopt, is the independence between $\{Y_i\}_{i \in \mathbb{N}}$ and $\{W_i\}_{i \in \mathbb{N}}$.

We choose, in contrast to classical models, a non-constant premium rate to model the effect of a so-called dividend barrier $a > 0$ in a smooth way. A barrier at level $a > 0$ has the purpose that every excess of surplus of this level is distributed as a dividend to shareholders which allows to include economic considerations in insurance modeling. Mathematically, this means that the process X is reflected at level a. Now instead of directly reflecting the process we use the following construction. Fix $\varepsilon > 0$ and for some $\tilde{c} > 0$, define

$$c(x) = \begin{cases} \tilde{c}, & x \in [0, a - \varepsilon), \\ f(x), & x \in [a - \varepsilon, a], \\ 0, & x > a, \end{cases} \tag{12}$$

with some positive and twice continuously differentiable function f which fulfills $f(a - \varepsilon) = \tilde{c}, f(a) = 0, f'(a - \varepsilon) = f'(a) = f''(a - \varepsilon) = f''(a) = 0$. Altogether, we assume $c(\cdot) \in \mathscr{C}^2[0, a]$ with some Lipschitz constant $L > 0$ and $c'(a - \varepsilon) = c'(a) = 0, c''(a - \varepsilon) = c''(a) = 0, c' \leq 0$ and bounded derivatives c', c''. Then $\lim_{x \nearrow a} c(x) = 0$ and the process always stays below level a if started in $[0, a)$.

A concrete choice for f would be

$$\frac{c(a-x)^3 \left(15\varepsilon(x-a) + 6(a-x)^2 + 10\varepsilon^2\right)}{\varepsilon^5}. \tag{13}$$

In the following we do not specify f any further.

In this setting we add $X_0 = x$ into the definition of the time of ruin, i.e., $\tau_x = \inf\{t \geq 0 \,|\, X_t < 0,\, X_0 = x\}$.

Remark 2 In this model setting ruin can only take place at some jump time T_k and since the process is bounded a.s. we have that $P_x(\tau_x < \infty) = 1$. If an approximation to classical reflection of the process at level a is implemented, then the process virtually started above a is forced to jump down to $a - \varepsilon$ and continue from this starting value. Consequently, we put the focus on starting values $x \in [0, a)$.

In the remainder of this section we will study analytic properties of the discounted value function which in this framework takes the form

$$V(x) = \mathbb{E}_x \left(e^{-\delta \tau_x} w(|X_{\tau_x}|, X_{\tau_x -}) \right), \tag{14}$$

with $\delta > 0$ and a continuous penalty function $w : \mathbb{R}^+ \times [0, a) \to \mathbb{R}$.

To have a well defined function, typically the following integrability condition is used

$$\int_0^\infty \int_0^\infty |w(x, y)| f_Y(x + y) dy\, dx < \infty,$$

see [6]. Since our process is kept below level a and w is supposed to be continuous in both arguments we can naturally replace the above condition by

$$\sup_{z \in [0,a)} \int_0^\infty |w(|z - y|, z)| f_Y(y) dy =: M < \infty, \tag{15}$$

which we will assume in the following. The condition from Eq. (15) holds true for example, if $|w(x, y)| \leq (1 + |x| + |y|)^p$ and F_Y admits a finite p-th moment for some $p \geq 1$. The condition (15) is motivated by the observation that $X_{\tau_x -} \in [0, a)$ and $|X_{\tau_x}| = |X_{\tau_x -} - Y_{N_{\tau_x}}|$ where $Y_{N_{\tau_x}} \overset{d}{\sim} f_Y$. Consequently, we get

$$V(x) \leq \mathbb{E}_x \left(|w(|X_{\tau_x}|, X_{\tau_x -})| \right) \leq \sup_{z \in [0,a)} \int_0^\infty |w(|z - y|, z)| f_Y(y) dy.$$

Remark 3 From the construction of X we have that $\tilde{X} = (\tilde{X}_t)_{t \geq 0}$ with $\tilde{X}_t = (X_t, t'(t), t)$ is a piecewise-deterministic Markov process, see Davis [12]. Since the jump intensity depends on $t' = t - T_{N_t}$, one needs this additional component for the *Markovization* of X. But on the discrete time skeleton $\{T_i\}_{i \in \mathbb{N}}$ with $T_0 = 0$ the process $X = \{X_{T_k}\}_{k \in \mathbb{N}}$ has the Markov property.

Remark 4 In risk theory surplus models including a reflection at some level $a > 0$ with dynamics of the form

$$dX_t = c\, \mathbf{1}_{\{X_t < a\}} dt - dS_t, \quad X_0 = x \geq 0,$$

arise when studying dividend strategies which pay out every excess over the level a immediately to shareholders. This is motivated by the following observation: when studying ruin probabilities it is crucial having $\mathbb{E}_x(X_1 - x) > 0$, which results in $P(X_t < 0 \text{ for some } t > 0 \text{ or } \lim_{t\to\infty} X_t = \infty) = 1$. This says that on the favourable set $\{\omega \in \Omega \mid \tau_x(\omega) = \infty\}$ the surplus becomes arbitrarily large. As a reaction to this unrealistic behaviour, a shareholder participation via dividend payments is introduced. An overview on the dividend problem in risk theory and related results can for instance be found in [4]. In the present setting, we introduce a smoothed reflection to make relevant computations accessible to an application of QMC methods, a feature which does not show up in the corresponding literature. Results on a classical QMC treatment in the situation of a non-linear dividend barrier can be found in [3].

2.3 Analytic Properties and a Fixed Point Problem

We start with showing some elementary analytical properties of the function V defined in (14).

Theorem 1 *The function* $V : [0, a) \to \mathbb{R}$ *is bounded and continuous.*

Proof The boundedness of V follows directly from the assumption made in (15).

For proving continuity we split off the expectation defining V into two parts which we separately deal with. Let $x > y$ and observe

$$\begin{aligned}
|\,V(x) - V(y)\,| =& \left|\mathbb{E}\left[e^{-\delta\tau_x} w(|X^x_{\tau_x}|, X^x_{\tau_x -}) - e^{-\delta\tau_y} w(|X^y_{\tau_y}|, X^y_{\tau_y -})\right]\right| \\
\leq& \mathbb{E}\left[e^{-\delta\tau_x}\left|w(|X^x_{\tau_x}|, X^x_{\tau_x -}) - w(|X^y_{\tau_x}|, X^y_{\tau_x -})\right| \mathbf{1}_{\{\tau_x = \tau_y\}}\right] \\
&+ \mathbb{E}\left[\left|e^{-\delta\tau_x} w(|X^x_{\tau_x}|, X^x_{\tau_x -}) - e^{-\delta\tau_y} w(|X^y_{\tau_y}|, X^y_{\tau_y -})\right| \mathbf{1}_{\{\tau_x > \tau_y\}}\right] \\
=& A + B.
\end{aligned}$$

For A we fix some $T > 0$ and notice the following bound

$$\begin{aligned}
A \leq& \mathbb{E}\left[e^{-\delta\tau_x} |w(|X^x_{\tau_x}|, X^x_{\tau_x -}) - w(|X^y_{\tau_x}|, X^y_{\tau_x -})| \mathbf{1}_{\{\tau_x = \tau_y \leq T\}}\right] \\
&+ 2M\, P(\tau_x > T) \leq 2M.
\end{aligned} \tag{16}$$

Before going on we need some estimates on the difference of two paths, one starting in x and the other in y. For fixed $\omega \in \Omega$ we have that on $(0, T_1(\omega)$ the surplus fulfills $\frac{\partial X_t(\omega)}{\partial t} = c(X_t(\omega))$ with initial condition $X_0 = 0$, $T_1(\omega)$ is finite with probability one. Standard arguments on ordinary differential equations, see for instance Stoer and Bulirsch [38, Th. 7.1.1–7.1.8], yield that an appropriate solution exists and is continuously differentiable in t and continuous in the initial value x. We even get the bound $|X^x_t - X^y_t| \leq e^{Lt}\,|x - y|$ for fixed ω, where X^x_t denotes the path which starts in

x and $L > 0$ the Lipschitz constant of $c(\cdot)$. From these results we directly obtain for a given path

$$|X^x_{T_1-} - X^y_{T_1-}| \leq e^{LT_1}|x-y|,$$

which by iteration results in

$$|X^x_{T_n-} - X^y_{T_n-}| = |X^x_{T_n} - X^y_{T_n}| \leq e^{LT_n}|x-y|,$$

because $|X^x_{T_n} - X^y_{T_n}| = |X^x_{T_n-} - Y_n - (X^y_{T_n-} - Y_n)| = |X^x_{T_n-} - X^y_{T_n-}|$. Since ruin takes place at some claim occurrence time T_k we get that on $\{\omega \in \Omega \mid \tau_x = \tau_y \leq T\}$ the quantities $|X^x_{\tau_x}|$ and $X^x_{\tau_x-}$ converge to the corresponding quantities started in y, all possible differences are bounded by $e^{LT}|x-y|$. Therefore, sending y to x in (16) and then sending T to infinity, we get that A converges to zero because $P(\tau_x < \infty) = 1$ and bounded convergence. We can repeat the argument for $x \to y$ when using $P(\tau_y > T)$ in (16).

Now consider part B. We first observe that $B \leq 2MP(\tau_x > \tau_y)$. Consequently, we need to show that $P(\tau_x > \tau_y)$ tends to zero if $y \to x$ or $x \to y$. Again, fix $\omega \in \Omega$ for which $\tau_x(\omega) > \tau_y(\omega)$, this implies that there is a claim amount Y_n, occurring at some point in time T_n, for which

$$X^x_{T_n-}(\omega) \geq Y_n(\omega) > X^y_{T_n-}(\omega),$$

i.e., causing ruin for the path started in y, (X^y_t), but not causing ruin for the one started in x, (X^x_t). From the construction of the drift $c(\cdot)$, it is decreasing to zero, we have that, suppressing the ω dependence,

$$0 < Y_n - X^y_{T_n-} \leq X^x_{T_n-} - X^y_{T_n-} \leq x - y.$$

Since $X^y_{T_n-} \in [0, a)$ we have

$$P(\tau_x > \tau_y) \leq \sup_{q \in [0,a)} P(0 < Y - q \leq x - y) = \sup_{q \in [0,a)} \{F_Y(x - y + q) - F_Y(q)\},$$

which approaches zero whenever x and y tend to each other since F_Y is continuous.

□

Define for functions $f \in \mathscr{C}_b([0, a))$ the operator $\mathscr{A}$ by

$$\mathscr{A}f(x) := \mathbb{E}_x\left(e^{-\delta T_1} f(X_{T_1})\mathbf{1}_{\{T_1 < \tau_x\}} + e^{-\delta \tau_x} w(|X_{T_1}|, X_{T_1-})\mathbf{1}_{\{\tau_x = T_1\}}\right). \tag{17}$$

The Markov property of the sequence $\{X_{T_i}\}_{i \in \mathbb{N}}$ and the definition of V in (14) allow us to derive that $V = \mathscr{A}V$, or explicitly written

$$V(x) = \mathbb{E}_x\left[e^{-\delta T_1} V(X_{T_1})\mathbf{1}_{\{T_1 < \tau_x\}} + e^{-\delta T_1} w(|X_{T_1}|, X_{T_1-})\mathbf{1}_{\{\tau_x = T_1\}}\right].$$

We can state the following lemma.

Lemma 1 *If $\delta > 0$, the operator $\mathscr{A} : \mathscr{C}_b([0,a)) \to \mathscr{C}_b([0,a))$ defined in (17) is a contraction with respect to $||\cdot||_\infty$.*

Proof Let $f \in \mathscr{C}_b([0,a))$ be bounded by some constant M', then

$$\mathscr{A}f(x) = \mathbb{E}_x\left(e^{-\delta T_1}f(X_{T_1})\mathbf{1}_{\{T_1<\tau_x\}} + e^{-\delta\tau_x}w(|X_{T_1}|, X_{T_1}-)\mathbf{1}_{\{\tau_x=T_1\}}\right),$$

is bounded by $\max\{M, M'\}$. From the integral representation of $\mathscr{A}f(x)$ we get continuity in x,

$$\mathscr{A}f(x) = \int_0^\infty e^{-\delta t_1}f_W(t_1)\Bigg[\int_0^{X_{t_1-}} f(X_{t_1-}-y_1)dF_Y(y_1)+$$
$$\int_{X_{t_1-}}^\infty w(|X_{t_1-}-y_1|, X_{t_1-})dF_Y(y_1)\Bigg]dt_1,$$

where X_{t_1-} is the ODE's solution up to time t_1 with $X_0 = x$. From Stoer and Bulirsch [38, Th. 7.1.4] we have that X_{t_1-} is continuous in its initial value which shows that $\mathscr{A}f(x)$ is continuous in x.

Let $f,\ g \in \mathscr{C}_b([0,a))$, then we have for all $x \in [0,a)$ that

$$|\mathscr{A}(f-g)(x)| \le \int_0^\infty e^{-\delta t_1}f_W(t_1)\int_0^{X_{t_1}} |f(X_{t_1}-y_1)-g(X_{t_1}-y_1)|dF_Y(y_1)dt_1$$
$$\le ||f-g||_\infty \int_0^\infty e^{-\delta t_1}f_W(t_1)dt_1 = ||f-g||_\infty\mathbb{E}[e^{-\delta T_1}].$$

Since $\delta > 0$ and $T_1 > 0$ P-a.s., $\mathscr{A}$ is contractive with Lipschitz constant $\tilde{L} = \mathbb{E}[e^{-\delta T_1}] < 1$. □

For a possible application of quasi-Monte Carlo techniques we need to examine the structure of $\mathscr{A}$,

$$\mathscr{A}v(x) = \int_0^\infty e^{-\delta t_1}f_W(t_1)\int_0^{X_{t_1-}} v(X_{t_1-}-y_1)dF_Y(y_1)dt_1+$$
$$\int_0^\infty e^{-\delta t_1}f_W(t_1)\int_{X_{t_1-}}^\infty w(y_1-X_{t_1-}, X_{t_1-})dF_Y(y_1)dt_1$$
$$=: \mathscr{G}v(x) + \mathscr{H}(x).$$

For $n \in \mathbb{N}$ the probabilistic interpretation of iterated applications of $\mathscr{A}$ is $\mathscr{A}^n v(x) = \mathbb{E}_x\left(e^{-\delta T_n}v(X_{T_n})\mathbf{1}_{\{T_n<\tau_x\}} + e^{-\delta\tau_x}w(|X_{\tau_x}|, X_{\tau_x-})\mathbf{1}_{\{\tau_x\le T_n\}}\right)$. Using $\mathscr{G}$ and $\mathscr{H}$ we can write

$$\mathscr{A}^n v(x) = \mathscr{G}^n v(x) + \sum_{k=0}^{n-1} \mathscr{G}^k \mathscr{H}(x),$$

where $\mathscr{G}^n v(x) = \mathbb{E}_x(e^{-\delta T_n} v(X_{T_n}) \mathbf{1}_{\{T_n < \tau\}})$ and

$$\mathscr{G}^{k-1}\mathscr{H}(x) = \int_0^\infty \cdots \int_0^\infty \int_{X_{\bar{t}_k -}}^\infty \int_0^{X_{\bar{t}_{k-1}} -} \cdots \int_0^{X_{\bar{t}_1} -} \left(\prod_{i=1}^k e^{-\delta t_i} f_W(t_i)\right) w(y_k - X_{\bar{t}_k -}, X_{\bar{t}_k -}) dF_Y(y_k) \cdots dF_Y(y_1) dt_k \cdots dt_1.$$

Here, $\bar{t} := \sum_{i=1}^k t_i$ and represents the time of the k-th jump. We see that via $X_{\bar{t}_k} = X_{\bar{t}_{k-1}} - y_{k-1} + \int_{\bar{t}_{k-1}}^{\bar{t}_k} c(X_s) ds$ the path of the process depends on all integration variables $(t_1, \ldots, t_k, y_1, \ldots, y_k)$.

For dealing with the situation $\delta = 0$, i.e., when the contraction argument fails, we can use a probabilistic argument. Since $\lim_{n\to\infty} T_n = \infty$ and $P(\tau_x < \infty) = 1$ we have that $\lim_{n\to\infty} \mathscr{G}^n v(x) = \lim_{n\to\infty} \mathbb{E}_x \left(e^{-\delta T_n} v(X_{T_n}) \mathbf{1}_{\{T_n < \tau\}}\right) = 0$ for $v \in \mathscr{C}_b([0, a))$. Using $|\mathscr{A}^n v(x) - V(x)| = |\, \mathscr{G}^n v(x) - \mathscr{G}^n V(x)\,|$ we get $\lim_{n\to\infty} \mathscr{A}^n v(x) = V(x)$ pointwise, even in the case if $\delta = 0$.

In what follows we put the focus on the determination of $\mathscr{G}^k \mathscr{H}(x)$.

3 Approximation Procedure

For the application of QMC methods we need to transform in a first step the integration domain in

$$\mathscr{G}^{k-1}\mathscr{H}(x) = \int_0^\infty \cdots \int_0^\infty \int_{X_{\bar{t}_k -}}^\infty \int_0^{X_{\bar{t}_{k-1}} -} \cdots \int_0^{X_{\bar{t}_1} -} \left(\prod_{i=1}^k e^{-\delta t_i} f_W(t_i)\right) w(y_k - X_{\bar{t}_k -}, X_{\bar{t}_k -}) dF_Y(y_k) \cdots dF_Y(y_1) dt_k \cdots dt_1$$

to $[0, 1]^{2k}$. This is achieved by use of the following substitutions

$$\alpha_i := e^{-t_i} \Rightarrow t_i = -\log \alpha_i \qquad \text{for } i \in \{1, \ldots, k\}$$

$$\beta_i := \frac{y_i}{X_{\bar{t}_i -}} \Rightarrow y_i = X_{\bar{t}_i -} \beta_i \qquad \text{for } i \in \{1, \ldots, k-1\}$$

$$\beta_k := e^{X_{\bar{t}_k -}} e^{-y_k} \Rightarrow y_k = X_{\bar{t}_k -} - \log \beta_k.$$

Here it has to be taken into account that the values of the reserve process X have to be calculated recursively, i.e., $X_{\bar{t}_i-}$ depends on $t_1, \dots, t_i$ and $y_1, \dots, y_{i-1}$. Since the Jacobian matrix of this transformation has a lower triangular form, the determinant can easily be found as $\frac{1}{\alpha_1 \dots \alpha_k} X_{\bar{t}_1-} \cdots X_{\bar{t}_{k-1}-} \frac{1}{\beta_k}$. Altogether, we arrive at

$$\mathscr{G}^{k-1}\mathscr{H}(x) = \int_{[0,1]^{2k}} \prod_{i=1}^{k} \alpha_i^{\delta} f_W(t_i(\alpha_i)) \prod_{i=1}^{k} f_Y(y_i(\alpha_1, \dots, \alpha_i, \beta_1, \dots, \beta_i))$$
$$\frac{1}{\alpha_1 \dots \alpha_k} X_{\bar{t}_1-} \cdots X_{\bar{t}_{k-1}-} \frac{1}{\beta_k} w(-\log \beta_k, X_{\bar{t}_k-})\, d\alpha_1 \dots d\alpha_k d\beta_1 \dots d\beta_k.$$

Consequently, for recovering the Koksma-Hlawka type error bound we need to examine the variation of the integrand:

$$F(\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k) = \left(\prod_{i=1}^{k-1} \alpha_i^{\delta-1} f_W(-\log(\alpha_i)) \right) \left(\prod_{i=1}^{k-1} f_Y(\beta_i X_{\bar{t}_i-}) X_{\bar{t}_i-} \right) \cdot$$
$$\left(\alpha_k^{\delta-1} f_W(-\log(\alpha_k)) f_Y(X_{\bar{t}_k-} - \log(\beta_k)) \frac{1}{\beta_k} w(-\log \beta_k, X_{\bar{t}_k-}) \right). \qquad (18)$$

Here we denote by $\phi(t,s)$ the solution to $\frac{\partial}{\partial t} x(t) = c(x(t))$ with $x(0) = s$. Consequently, we can write

$$X_{\bar{t}_i-} = X_{\bar{t}_{i-1}-} - y_{i-1} + \phi(t_i, X_{\bar{t}_{i-1}-} - y_{i-1}).$$

Or in terms of α_i, putting $\hat{x}_{i-1} = X_{\bar{t}_{i-1}-} - y_{i-1} = X_{\bar{t}_{i-1}-}(1 - \beta_{i-1})$ and

$$X_{\bar{t}_i-} = \hat{x}_{i-1} + \phi(-\log(\alpha_i), \hat{x}_{i-1}). \qquad (19)$$

In the following proposition we show that with a particular choice of model parameters it is possible to apply results from [33] to show that the integrand in (18) is in some sense of finite variation. Its proof shows that probabilistic and deterministic model ingredients are considerably interconnected.

Theorem 2 *Let* $f_W(t) = \lambda e^{-\lambda t} \mathbf{1}_{\{t \geq 0\}}$ $(\lambda > 0)$, $f_Y(y) = \mu e^{-\mu y} \mathbf{1}_{\{y \geq 0\}}$ $(\mu > 0)$, $w \equiv 1$ *and* $c(\cdot)$ *be specified by (13). Then, under the assumption* $\lambda + \delta \geq 3$ *and* $\mu \geq 3$ *the variation* $\mathscr{V}_{\mathscr{K}}(F)$ *(see* (10) *with* $\mathscr{D} = \mathscr{K}$*) of* F*, defined in (18), is finite.*

Proof The main idea of the proof is the application of [33, Th. 3.12]. For this purpose we need to show that $M(F) = \sup\{\|\mathrm{Hess}(F,x)\| \mid x \in [0,1]^{2k}\}$, $\sup F$ and $\inf F$ are finite, with the implication

$$\mathscr{V}_{\mathscr{K}}(F) \leq \sup F - \inf F + M(F).$$

Since in this theorem the operator (matrix) norm $\|\mathrm{Hess}(F, x)\|$ is arbitrary we use the 2-norm and exploit the relation

$$\|\mathrm{Hess}(F, x)\|_2 \leq \left(\sum_{i=1}^{2k} \sum_{j=1}^{2k} [\mathrm{Hess}(F, x)]_{ij}^2 \right)^{\frac{1}{2}}.$$

We will show that $[\mathrm{Hess}(F, x)]_{ij}$ is finite for all $x \in [0, 1]^{2k}$. At first we observe that when taking derivatives with respect to α_i and β_j, the structure of (19) implies the appearance of the following terms:

$$\begin{aligned}
&\frac{\partial}{\partial t}\phi(t, s) = c(\phi(t, s)),\ \frac{\partial^2}{\partial t^2}\phi(t, s) = c'(\phi(t, s))c(\phi(t, s)),\\
&\frac{\partial}{\partial s}\phi(t, s) =: y(t, s) = e^{\int_0^t c'(\phi(u,s))du},\ \frac{\partial^2}{\partial t \partial s}\phi(t, s) = c'(\phi(t, s))y(t, s),\\
&\frac{\partial^2}{\partial s^2}\phi(t, s) =: z(t, s) = y(t, s) \int_0^t c''(\phi(u, s))y(u, s)du.
\end{aligned}$$

The functions y, z correspond to the first and second derivative of the ODE's solution with respect to the initial value. They can be derived from the associated first and second order variational equations (see [19]). From our assumptions on $c(\cdot)$ we have that y is bounded by one ($c' \leq 0$) and all other derivatives including z are bounded as well. The boundedness of z can be derived from the boundedness of $c''(\cdot)$ and an analysis of the growth behaviour of y.

For the structure of $[\mathrm{Hess}(F, x)]_{ij}$ we can derive the following

$$\left(\prod_{l=1}^{k} \alpha_l^{\delta+\lambda-a_{ij}} \beta_k^{\mu-b_{ij}} e^{-\mu(y_1+\cdots+y_{k-1}+X_{\bar{\tau}_k-})} \right) \cdot \\ Q_{ij}\left(\beta_1, \ldots, \beta_{k-1}, \phi, \frac{\partial}{\partial t}\phi, \frac{\partial^2}{\partial t^2}\phi, \frac{\partial}{\partial s}\phi, \frac{\partial^2}{\partial t \partial s}\phi, \frac{\partial^2}{\partial s^2}\phi \right),$$

for a_{ij}, $b_{ij} \in \{1, 2, 3\}$ and a function Q_{ij}. Q_{ij} is evaluated at the integration points and ϕ and its derivatives which themselves are evaluated in points of the form $(-\log(\alpha_l), \hat{x}_{l-1}) \in (0, \infty) \times [0, a)$ for $l \in \{1, \ldots, k\}$. If ϕ and its derivatives are considered to be variables, neglecting their dependence on the α_ls and β_ls, then Q_{ij} is a polynomial of degree k. The degree of the polynomial is produced by the recursive structure of the paths and its dependence on all previous jump times and sizes. From this inspection we get that under the conditions $\lambda + \delta \geq 3$ and $\mu \geq 3$ all entries of the Hessian matrix are bounded. Furthermore, the conditions on the parameters λ, δ, μ ensure that $\sup F$ is finite and $\inf F = 0$. □

Remark 5 We can combine the above result with the convergence rate from Banach's fixed point theorem and obtain for our specific situation

$$\begin{aligned}\left\|\sum_{k=0}^{n}\hat{\mathscr{G}}^k\mathscr{H} - V\right\|_\infty &\le \left\|\sum_{k=0}^{n}(\hat{\mathscr{G}}^k\mathscr{H} - \mathscr{G}^k\mathscr{H})\right\|_\infty + \|\mathscr{A}^n - V\|_\infty + \|\mathscr{G}^n v\|_\infty \\ &\le \sum_{k=0}^{n}\mathscr{V}_{\mathscr{K}}(F^k)\tilde{D}_{N_k} + \frac{\tilde{L}^n}{1-\tilde{L}}\|\mathscr{A}v - v\|_\infty + M'\left(\frac{\lambda}{\delta+\lambda}\right)^n.\end{aligned}$$

Here F^k denotes the integrand from (18) in dimension $2k$, $\tilde{D}_{N_k}$ the isotropic discrepancy of a pointset with N_k elements in $[0,1]^{2k}$ and $\hat{\mathscr{G}}^k\mathscr{H}$ is the QMC approximation for $\mathscr{G}^k\mathscr{H}$. For the last term we used that v is bounded by some $M' > 0$ and the fact the T_n follows a Gamma distribution $\Gamma(n,\lambda)$.

From the type of arguments we used for the proof of Theorem 2, we expect that the result holds true for Γ-distributed inter-claim times and jump heights and $w(y,z) = y^k z^l$ with similar conditions on the parameters. Hence the method is also applicable for this more general situation. A detailed study of this claim is part of future research.

4 Numerical Results

In this section, we evaluate the integrals from Sect. 3 by applying Monte Carlo and quasi-Monte Carlo methods for different choices of the penalty function w.

Note that in the general case, determining the Gerber-Shiu function analytically is a profoundly hard problem, since only certain parameter constellations allow for explicit results. For constant parameter settings, in particular constant drift c, inter-claim times and jumps following *phase-type* distributions and special choices of w, the problem can be handled by matrix-analytic methods. For an overview on these techniques see [6], or for exemplary results one may consult [5] and [28]. A main focus in the risk theoretic literature lies on asymptotical approximations as the initial value x becomes large, these results are referred to as Cramér-Lundberg type approximations, see [6] and [34]. In contrast to these probabilistic approximations, the literature on the numerical treatment of some examples of Gerber-Shiu functions is scarce. For a survey on the use of collocation methods we refer to [30].

4.1 The Discounted Time of Ruin

Letting $w(y,z) := 1$, we arrive at $V(x) = \mathbb{E}_x(e^{-\delta\tau_x}w(|X_{\tau_x}|, X_{\tau_x-})) = \mathbb{E}_x(e^{-\delta\tau_x})$ which is the discounted time of ruin. Lin et al. [28] found an analytic expression for this discounted time of ruin if both the inter-arrival times of the claims and the claim sizes are exponentially distributed. To have a reference value, we also adopt these assumptions and denote the parameters of the exponential distributions with λ for the parameter of the inter-arrival times and μ for the parameter of the claim sizes. The premium rate $c(\cdot)$ was chosen as in Sect. 2.2 with f from Eq. (13), with $\tilde{c} = 2$, $a = 3$ and ε was set to 0.001. Note that the results of Lin et al. [28] were proved for a reflected process in the classical sense, which means $c(x) = \tilde{c}$ for $x \leq a$ and $c(x) = 0$ for $x > a$. Since Theorem 2 requires a premium rate satisfying certain smoothness conditions, we cannot use a discontinuous c and thus have a methodic error in our simulations. However, we will see that this error is, at least for small ε, very small.

We list the parameters together with the approximation values for increasing numbers of (Q)MC points and $k = 20$ iterations of the algorithm in Fig. 1c, whereas Fig. 1d shows the approximation values for $k = 100$ iterations of the algorithm. Figure 1a, b shows the MC points (green) with 95% confidence intervals, together with QMC points from Sobol sequences (blue) and Halton sequences (orange).

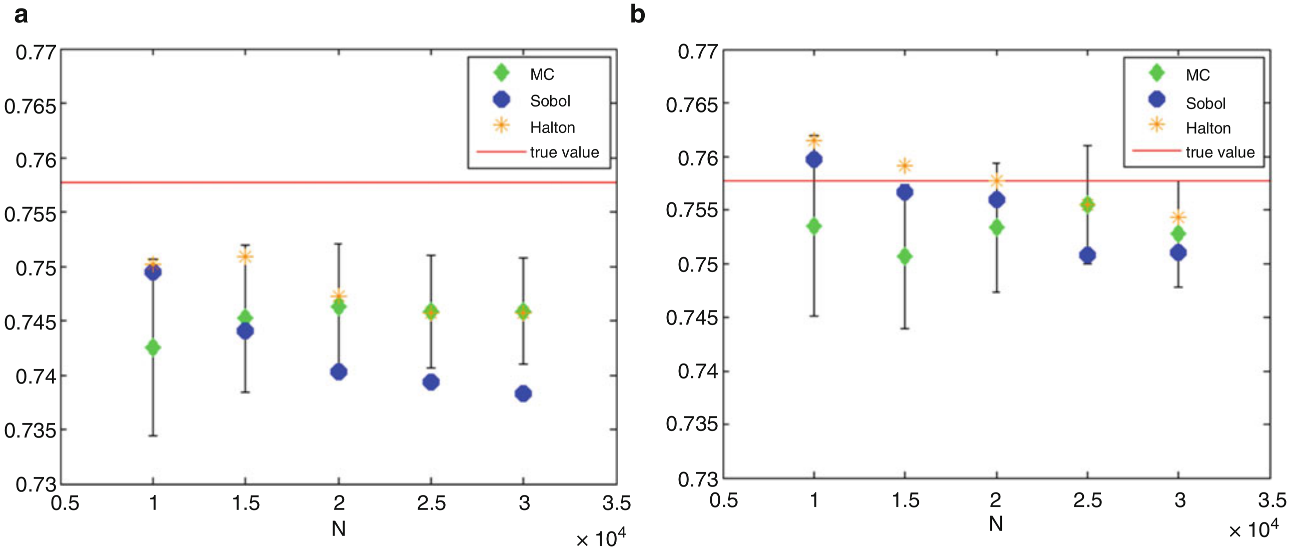
The red line at height 0.7577 marks the analytically found value for the reflected process. We use it as a reference value here but, again, remark that it is not the exact value for our smoothed process. As can be seen in Fig. 1a, the algorithm has not yet converged for $k = 20$, whereas Fig. 1b shows that $k = 100$ already yields a very good approximation. The computation time of the above example, for $k = 100$, was under 4 min for both choices of QMC sequences, whereas the MC method has a runtime of more than 30 min. The exact computation times are given in Fig. 2.

To illustrate the speed of convergence, we also plotted the absolute error, both for the MC approach as well as for QMC points (again taken from Sobol and Halton sequences) for varying numbers of points N. Figures 3 and 4 show the values obtained for $k = 40$ iterations of the algorithm. Obviously, $k = 40$ is also not yet enough to reach the actual value. But notice that the absolute error even for more iterations cannot converge to zero because of the smoothed reflection procedure. For both of the QMC methods, a scramble improved the results. In the Sobol case however, an "unlucky" choice in the scramble and the skip value (i.e. how many elements are dropped in the beginning) can lead to relatively high variation in the output, whereas the Halton set shows a more stable performance (compare Figs. 3 and 4).

For Fig. 5 we evaluated $k = 40$ iterations of the algorithm with $N = 30,000$ (Q)MC points for different starting values x, ranging from 0.7 to 2. As expected, the discounted time of ruin decreases for increasing x.

4.2 The Deficit at Ruin

If we set $w(y, z) := y$, and $\delta = 0$, we have $V(x) = \mathbb{E}_x(|X_{\tau_x}|)$, the expected deficit at ruin. We use the same premium rate $c(\cdot)$ as before and again choose exponential distributions for the inter-arrival times and claim sizes with parameters λ and μ



c

x	λ	μ	$w(y,z)$	k	δ
1.2	1	0.8	1	20	0.05
N:	10000	15000	20000	25000	30000
MC:	0.7425	0.7452	0.7463	0.7458	0.7459
Sobol:	0.7494	0.7440	0.7403	0.7394	0.7383
Halton:	0.7502	0.7509	0.7473	0.7488	0.7457

d

x	λ	μ	$w(y,z)$	k	δ
1.2	1	0.8	1	100	0.05
N:	10000	15000	20000	25000	30000
MC:	0.7535	0.7507	0.7534	0.7555	0.7527
Sobol:	0.7597	0.7566	0.7560	0.7508	0.7510
Halton:	0.7615	0.7591	0.7577	0.7555	0.7543

Fig. 1 (**a**) $k = 20$ iterations of the algorithm. (**b**) $k = 100$ iterations of the algorithm

k	MC points	Halton seq.	Sobol seq.
20	121	37	38
100	2436	198	200

Fig. 2 Times to obtain the plots in Fig. 1a and b in seconds

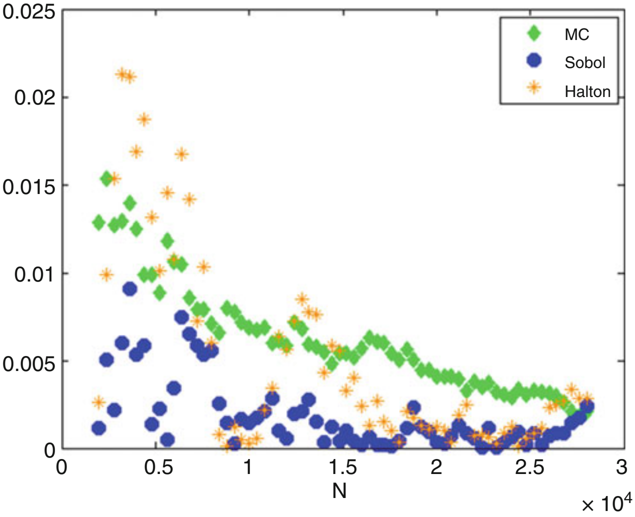


Fig. 3 "Lucky" choice of QMC points

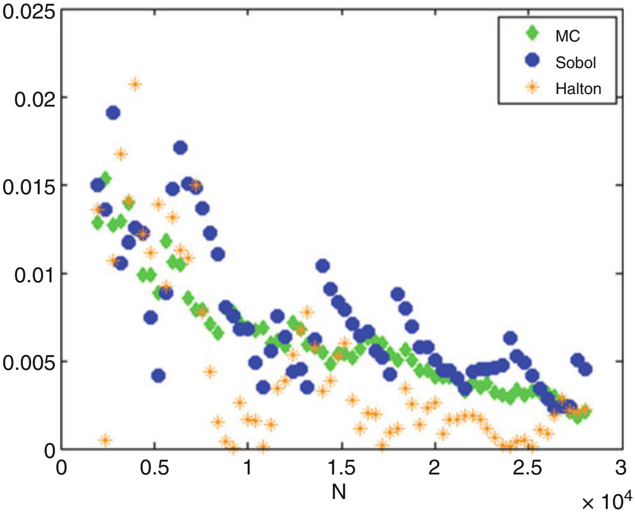


Fig. 4 "Unlucky" choice of QMC points

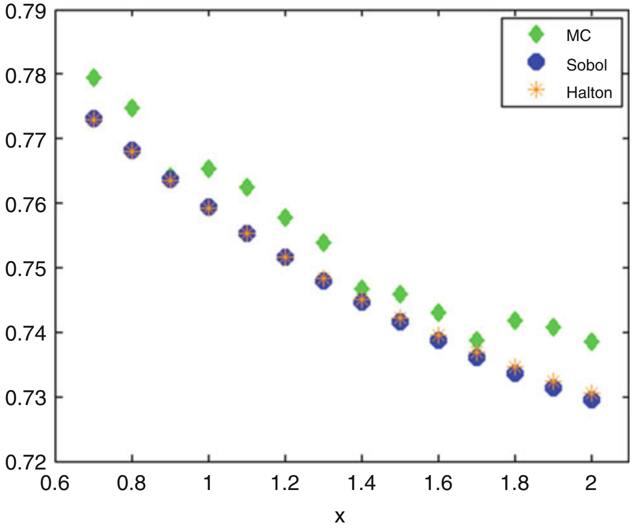


Fig. 5 Influence of the starting value

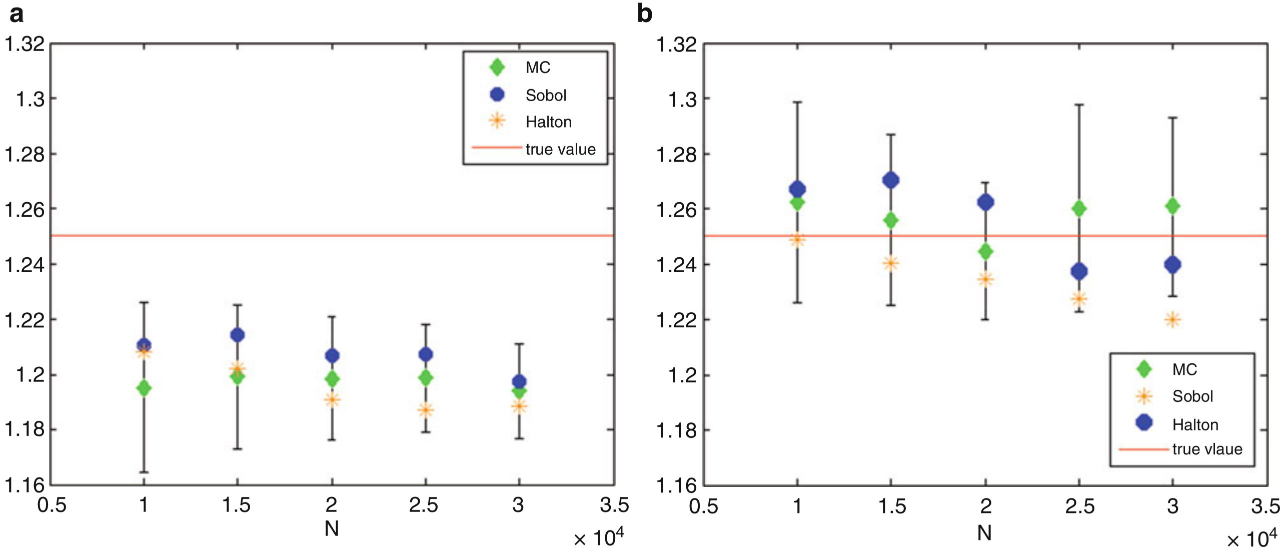
respectively, since also in this case the true value $\mathbb{E}_x(|X_{\tau_x}|) = \frac{1}{\mu}$ (for a classically reflected process) can be found in [28]. Figure 6a, b shows the results for $k = 20$ and $k = 100$ iterations respectively. The reference value is again shown as a red line, in our case at 1.25. The MC points are drawn in green, the Sobol points blue and the Halton points in orange. Figure 6c, d contains the precise values along with the corresponding parameters.

Note again the difference between Fig. 6a and b, resulting from a different number of iterations k. The computation times for these plots deviate very little from those given in Fig. 2.

Remark 6 We considered in our numerical examples two test cases for which explicit (approximate) reference values are available. Certainly our approach is not restricted to this particular choice of model ingredients—which are f_Y, f_W and the penalty function w.

Again, we plotted the absolute error for $k = 40$ iterations of the algorithm and a varying number of (Q)MC points N. Figure 7 shows the results using the same colorings as before.

While there are several approximation techniques for discounted penalty functions, it is precisely this flexibility that makes the (Q)MC approach favourable in many situations.



c

x	λ	μ	$w(y,z)$	k	δ
1.2	1	0.8	y	20	0
N:	10000	15000	20000	25000	30000
MC:	1.1952	1.1991	1.1986	1.1987	1.1939
Sobol:	1.2105	1.2142	1.2070	1.2074	1.1975
Halton:	1.2084	1.2019	1.1906	1.1872	1.1885

d

x	λ	μ	$w(y,z)$	k	δ
1.2	1	0.8	y	100	0
N:	10000	15000	20000	25000	30000
MC:	1.2624	1.2558	1.2446	1.2602	1.2607
Sobol:	1.2669	1.2704	1.2624	1.2373	1.2398
Halton:	1.2487	1.2404	1.2344	1.2273	1.2199

Fig. 6 (**a**) $k = 20$ iterations of the algorithm. (**b**) $k = 100$ iterations of the algorithm

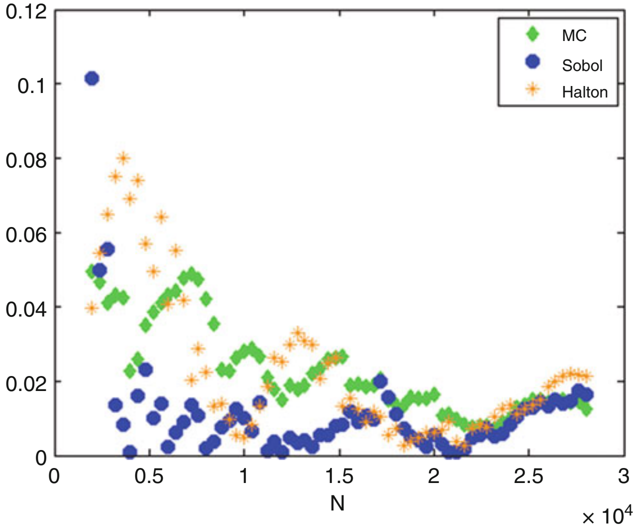


Fig. 7 The absolute error for the deficit at ruin

Acknowledgements The authors are supported by the Austrian Science Fund (FWF) Project F5510 (part of the Special Research Program (SFB) "Quasi-Monte Carlo Methods: Theory and Applications").

References

1. Aistleitner, C., Dick, J.: Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. Acta Arith. **167**(2), 143–171 (2015)
2. Aistleitner, C., Pausinger, F., Svane, A.M., Tichy, R.F.: On functions of bounded variation. Math. Proc. Camb. Philos. Soc. **162**(3), 405–418 (2017)
3. Albrecher, H., Kainhofer, R.: Risk theory with a nonlinear dividend barrier. Computing **68**(4), 289–311 (2002)
4. Albrecher, H., Thonhauser, S.: Optimality results for dividend problems in insurance. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. **103**(2), 295–320 (2009)
5. Albrecher, H., Constantinescu, C., Pirsic, G., Regensburger, G., Rosenkranz, M.: An algebraic operator approach to the analysis of Gerber–Shiu functions. Insur. Math. Econ. **46**(1), 42–51 (2010)
6. Asmussen, S., Albrecher, H.: Ruin Probabilities, 2nd edn. World Scientific, River Edge (2010)
7. Atkinson, K.E.: The numerical solution of Fredholm integral equations of the second kind. SIAM J. Numer. Anal. **4**(3), 337–348 (1967)
8. Brandolini, L., Colzani, L., Gigante, G., Travaglini, G.: A Koksma–Hlawka inequality for simplices. In: Trends in Harmonic Analysis, pp. 33–46. Springer, New York (2013)
9. Brandolini, L., Colzani, L., Gigante, G., Travaglini, G.: On the Koksma–Hlawka inequality. J. Complex. **29**(2), 158 – 172 (2013)
10. Brunner, H.: Iterated collocation methods and their discretizations for Volterra integral equations. SIAM J. Numer. Anal. **21**(6), 1132–1145 (1984)

11. Brunner, H.: Implicitly linear collocation methods for nonlinear Volterra equations. Appl. Numer. Math. **9**(3), 235–247 (1992)
12. Davis, M.H.A.: Markov Models and Optimization. Chapman and Hall, London (1993)
13. Dick, J., Kritzer, P., Kuo, F.Y., Sloan, I.H.: Lattice-Nyström method for Fredholm integral equations of the second kind with convolution type kernels. J. Complex. **23**(4), 752–772 (2007)
14. Drmota, M., Tichy, R.F.: Sequences, Discrepancies and Applications. Lecture Notes in Mathematics, vol. 1651. Springer, Berlin (1997)
15. Edelsbrunner, H., Pausinger, F.: Approximation and convergence of the intrinsic volume. Adv. Math. **287**, 674–703 (2016)
16. Gerber, H.U., Shiu, E.S.W.: On the time value of ruin. N. Am. Actuar. J. **2**(1), 48–78 (1998)
17. Gerber, H.U., Shiu, E.S.W.: The time value of ruin in a Sparre Andersen model. N. Am. Actuar. J. **9**(2), 49–84 (2005)
18. Götz, M.: Discrepancy and the error in integration. Monatsh. Math. **136**(2), 99–121 (2002)
19. Grigorian, A.: Ordinary Differential Equation. Lecture Notes (2009). Available at https://www.math.uni-bielefeld.de/~grigor/odelec2009.pdf
20. Harman, G.: Variations on the Koksma-Hlawka inequality. Unif. Distrib. Theory **5**(1), 65–78 (2010)
21. Hlawka, E.: Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. Ann. Mat. Pura Appl. **54**(1), 325–333 (1961)
22. Hua, L.K., Wang, Y.: Applications of Number theory to Numerical Analysis. Springer, Berlin; Kexue Chubanshe (Science Press), Beijing (1981). Translated from the Chinese
23. Ikebe, Y.: The Galerkin method for the numerical solution of Fredholm integral equations of the second kind. SIAM Rev. **14**(3), 465–491 (1972)
24. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Pure and Applied Mathematics. Wiley, New York (1974)
25. Kumar, S., Sloan, I.H.: A new collocation-type method for Hammerstein integral equations. Math. Comput. **48**(178), 585–593 (1987)
26. Kuo, F.Y.: Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. J. Complex. **19**(3), 301–320 (2003)
27. Leobacher, G., Pillichshammer, F.: Introduction to Quasi-Monte Carlo Integration and Applications. Compact Textbook in Mathematics. Birkhäuser/Springer, Cham (2014)
28. Lin, X.S., Willmot, G.E., Drekic, S.: The classical risk model with a constant dividend barrier: analysis of the Gerber–Shiu discounted penalty function. Insur. Math. Econ. **33**(3), 551–566 (2003)
29. Lundberg, F.: Approximerad framställning av sannolikhetsfunktionen. Aterförsäkring av kollektivrisker. Akad. Afhandling. Almqvist o. Wiksell, Uppsala (1903)
30. Makroglou, A.: Numerical solution of some second order integro-differential equations arising in ruin theory. In: Proceedings of the third Conference in Actuarial Science and Finance, pp. 2–5 (2004)
31. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
32. Owen, A.B.: Multidimensional variation for Quasi-Monte Carlo. In: Contemporary Multivariate Analysis and Design of Experiments. Series in Biostatistics, vol. 2, pp. 49–74. World Scientific, Hackensack (2005)
33. Pausinger, F., Svane, A.M.: A Koksma-Hlawka inequality for general discrepancy systems. J. Complex. **31**(6), 773–797 (2015)
34. Rolski, T., Schmidli, H., Schmidt, V., Teugels, J.: Stochastic processes for insurance and finance. Wiley, Chichester (1999)
35. Sloan, I.H.: A quadrature-based approach to improving the collocation method. Numer. Math. **54**(1), 41–56 (1988)
36. Sloan, I.H., Lyness, J.N.: The representation of lattice quadrature rules as multiple sums. Math. Comput. **52**(185), 81–94 (1989)

37. Sloan, I.H., Woźniakowski, H.: Tractability of multivariate integration for weighted Korobov classes. J. Complex. **17**(4), 697–721 (2001)
38. Stoer, J., Bulirsch, R.: Numerische Mathematik. 2, 4th edn. Springer-Lehrbuch. [Springer Textbook]. Springer, Berlin (2000)
39. Tichy, R.F.: Über eine zahlentheoretische Methode zur numerischen Integration und zur Behandlung von Integralgleichungen. Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II **193**(4–7), 329–358 (1984)
40. Twomey, S.: On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. J. Assoc. Comput. Mach. **10**(1), 97–101 (1963)

A Note on the Multidimensional Moment Problem

Liqun Qi

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In this note, we show that if a multidimensional sequence generates Hankel tensors and all the Hankel matrices, generated by this sequence, are positive semi-definite, then this sequence is a multidimensional moment sequence.

1 Introduction

The multidimensional moment problem is an important topic in mathematics [1, 2, 6–9, 12, 14]. In this note, we show that if a multidimensional sequence generates Hankel tensors and all the Hankel matrices, generated by this sequence, are positive semi-definite, then this sequence is a multidimensional moment sequence. We do this in Sect. 2. Some further questions are raised in Sect. 3.

We use small letters for scalars, bold letters for vectors, capital letters for matrices, and calligraphic letters for tensors.

2 The Multidimensional Moment Problem

Denote N for the set of all positive integers, and N_+ as the set of all nonnegative integers. According to [2, 12], a multidimensional sequence

$$S = \{b_{j_1 \cdots j_{n-1}} : j_1, \ldots, j_{n-1} \in N_+\} \tag{1}$$

L. Qi (✉)
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
e-mail: maqilq@polyu.edu.hk

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_48

is called a **multidimensional moment sequence** if there is a nonnegative measure μ on $\Re^{n-1}$ satisfying:

$$b_{j_1 \cdot\cdot j_{n-1}} = \int_{\Re^{n-1}} t_1^{j_1} \dots t_{n-1}^{j_{n-1}} d\mu, \text{ for } j_1, \dots, j_{n-1} \in N_+, \tag{2}$$

are all finite. For a given multidimensional sequence S defined by (1), is it a multidimensional moment sequence? i.e., Is there a nonnegative measure μ such that (2) holds? This problem is called the **multidimensional moment problem** [1, 2, 12].

For any $m \in N$, we may define a homogeneous polynomial of n variables and degree m:

$$f(\mathbf{x}) = \sum \{ b_{j_1 \cdot\cdot j_{n-1}} \frac{m!}{j_1! \dots j_{n-1}!(m - j_1 - \dots - j_{n-1})!} x_1^{j_1} \dots x_{n-1}^{j_{n-1}} x_0^{m-j_1-\dots-j_{n-1}}$$
$$: j_1, \dots, j_{n-1} \geq 0, j_1 + \dots + j_{n-1} \leq m \}. \tag{3}$$

According to [12], S is a multidimensional moment sequence if and only if for all $m, f(\mathbf{x})$ a **sum of** m**th power (SOM)** form.

A homogeneous polynomial $f(\mathbf{x})$ of n variables and degree m is corresponding to an mth order n-dimensional symmetric tensor $\mathscr{A} = (a_{i_1 \dots i_m})$, where

$$a_{i_1 \dots i_m} = b_{j_1 \cdot\cdot j_{n-1}}, \tag{4}$$

for $j_{n-1} \geq 0, j_1 + \dots + j_{n-1} \leq m$, if in $\{ i_1, \cdots, i_m \}$, the frequency of k is exactly j_k, $k = 1, \cdots, n-1$. Then $f(\mathbf{x})$ is an SOM form if and only if there are vectors $\mathbf{u}_k \in \Re^n$ for $k = 1 \dots, r$ such that

$$\mathscr{A} = \sum_{k=1}^{r} \mathbf{u}_k^m, \tag{5}$$

where for a vector $\mathbf{v} \in \Re^n$, $\mathbf{v}^m = (v_{i_1} \dots v_{i_m})$ denotes a symmetric rank-one tensor. Such a symmetric tensor is called a **completely decomposable tensor** in [11, 15].

Thus, a given multidimensional sequence S defined by (1), is a multidimensional moment sequence if and only if all the symmetric tensors $\mathscr{A}$ generated by it are completely decomposable tensors for all m. Note that when m is odd, a symmetric tensor is always completely decomposable [11].

Suppose now that for $j_1, \dots j_{n-1}, l_1, \dots, l_{n-1} \in N_+$, we have

$$b_{j_1 \cdot\cdot j_{n-1}} = b_{l_1 \dots l_{n-1}} \tag{6}$$

if

$$j_1 + 2j_2 + \dots + (n-1)j_{n-1} = l_1 + 2l_2 + \dots + (n-1)l_{n-1}. \tag{7}$$

By (4), for $i_1, \dots i_n, k_1, \dots, k_n \in N_+$, we have

$$a_{i_1 \dots i_m} = a_{k_1 \dots k_m} \tag{8}$$

as long as $i_1 + \dots + i_m = k_1 + \dots + k_m$. By [3–5, 10, 11, 15], such a tensor is called a **Hankel tensor**. Thus, we call a multidimensional sequence S satisfying (8) a **Hankel multidimensional sequence**.

For an mth order n-dimensional Hankel tensor $\mathscr{A} = (a_{i_1 \dots i_m})$, by [10], there is a **generating vector** $\mathbf{v} = (v_0, \dots, v_{mn})^\top$ such that $a_{i_1 \dots i_m} = v_{i_1 + \dots + i_m}$. If m is even, then $\mathbf{v}$ also generates a Hankel matrix A. If the associated Hankel matrix A is positive semi-definite, such a Hankel tensor $\mathscr{A}$ is called a **strong Hankel tensor**. By [15], a strong Hankel tensor is completely decomposable. An explicit decomposition expression of a strong Hankel tensor is given in [5].

Furthermore, by (6), we see that

$$v_{j_1 + 2j_2 + \dots + (n-1)j_{n-1}} = b_{j_1 \dots j_{n-1}}, \tag{9}$$

for $j_1, \dots, j_{n-1} \in N_+$, i.e., the components of $\mathbf{v}$ are independent from m. Thus, (9) defines an infinite sequence $V = \{v_k : k \in N_+\}$. This infinite sequence V generates a sequence of Hankel matrices $H_p = (h_{ij})$, with $i, j = 0, \dots, p-1, p \in N$, and $h_{ij} = v_{i+j}$ for $i, j \in N_+$.

By these, we have the following theorem.

Theorem 1 *Suppose that a given multidimensional sequence S defined by (1), satisfies (6), i.e., it is a Hankel multidimensional sequence. If all the Hankel tensors generated by V are strong Hankel tensors, i.e., all the Hankel matrices H_p generated by the sequence V are positive semi-definite, then S is a multidimensional moment sequence.*

This links the classical result for the Hamburger moment problem [6, 12], and gives an application of the results in [5, 10, 11, 15].

We may call such a sequence V a **strong Hankel sequence**. The well-known strong Hankel sequence is the Hilbert sequence $\{1, \frac{1}{2}, \frac{1}{3}, \dots\}$ [13].

3 Some Further Questions

The first question is: Are there any other Hankel multidimensional moment sequences, for which that infinite dimensional Hankel matrix H is not positive semi-definite? Another way to ask this question is as follows: Is there an infinite sequence $V = \{v_k : k \in N_+\}$, not all the Hankel matrices H_p generated by it are positive semi-definite, but all the n-dimensional Hankel tensors generated by it, for any order $m \in N$, are completely decomposable? Here $n \geq 3$ is a fixed positive integer. Thus, the simplest case is that $n = 3$. We thus may ask: Is there an infinite sequence $V = \{v_k : k \in N_+\}$, not all the Hankel matrices H_p generated by it are positive semi-definite, but all the 3-dimensional Hankel tensors, generated by it, for any order $m \in N$ are completely decomposable? In [11], a class of sixth order three dimensional truncated Hankel tensors are discussed. Such Hankel tensors are not strong Hankel tensors [11, 15], but still completely decomposable [15]. Can we build an infinite sequence V to answer this question, based upon such sixth order three dimensional truncated Hankel tensors?

If such a sequence V exists, then the next question is: what are the necessary and sufficient conditions such a sequence V should satisfy?

Note that until now, all the known positive semi-definite Hankel tensors are SOS [4, 11, 15], but there are positive semi-definite Hankel tensors which are not completely decomposable. This situation may make such a characterization somewhat subtle.

Another question is: Are there multidimensional moment sequences, which are not Hankel multidimensional sequences, i.e., condition (6) is not satisfied. If the answer to this question is "yes", then how to characterize such multidimensional sequences?

Acknowledgements The author is thankful to Weiyang Ding for his comments. The author was supported by the Hong Kong Research Grant Council (Grant No. PolyU 501212, 501913, 15302114 and 15300715).

References

1. Berg, C.: The multidimensional moment problem and semigroups, moments in Mathematics. Proc. Symp. Appl. Math. **37**, 110–124 (1987)
2. Berg, C., Christensen, J.P.R., Jensen C.U.: A remark on the multidimensional moment problem. Math. Ann. **243**, 163–169 (1979)
3. Chen, Y., Qi, L., Wang, Q.: Computing extreme eigenvalues of large scale Hankel tensors. J. Sci. Comput. **68**, 716–738 (2016)
4. Chen, Y., Qi, L., Wang, Q.: Positive semi-definiteness and sum-of-squares property of fourth order four dimensional Hankel tensors. J. Comput. Appl. Math. **302**, 356–368 (2016)
5. Ding, W., Qi, L., Wei, Y.: Inheritance properties and sum-of-squares decomposition of Hankel tensors: theory and algorithms. Bit Numer. Math. **57**, 169–190 (2017)
6. Hamberger, H.: Über eine Erweiterung des Stieltjesschen Momentproblems, Parts I, II, III. Math. Ann. **81**, 235–319 (1920), **82**, 20–164, 168–187 (1921)
7. Haviland, E.K.: On the momentum problem for distribution functions in more than one dimension. Am. J. Math. **57**, 562–568 (1935)
8. Haviland, E.K.: On the momentum problem for distribution functions in more than one dimension II. Am. J. Math. **58**, 164–168 (1936)
9. Kuhlmann, S., Marshall, M.: Positivity, sums of squares and the multi-dimensional moment problem. Trans. Am. Math. Soc. **354**, 4285–4301 (2002)
10. Qi, L.: Hankel tensors: associated Hankel matrices and Vandermonde decomposition. Commun. Math. Sci. **13**, 113–125 (2015)
11. Qi L., Luo, Z.: Tensor Analysis: Spectral Theory and Special Tensors. SIAM, Philadelphia (2017)
12. Reznick, B.: Sums of even powers of real linear forms. Mem. Am. Math. Soc. **96**(1992), Paper No. 463

13. Song, Y., Qi, L.: Infinite and finite dimensional Hilbert tensors. Linear Algebra Appl. **451**, 1–14 (2014)
14. Vasilescu, F.-H.: Hamburger and Stieltjes moment problems in several variables. Trans. Am. Math. Soc. **354**, 1265–1278 (2002)
15. Wang, Q., Li, G., Qi L., Xu, Y.: New classes of positive semi-definite Hankel tensors. Minimax Theory Appl. **2**, 231–248 (2017)

On a Novel Resonant Ermakov-NLS System: Painlevé Reduction

Colin Rogers and Wolfgang K. Schief

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract A novel resonant Ermakov-NLS system is introduced which admits symmetry reduction to a hybrid Ermakov-Painlevé II system. If the latter is Hamiltonian then combination with a characteristic Ermakov invariant provides an algorithmic integration procedure. The latter involves the isolation of positive solutions of a concomitant integrable Painlevé XXXIV equation. Explicit expressions for a multi-parameter class of wave packet representations for the original Ermakov-NLS system are obtained via the iterated application of a Bäcklund transformation admitted by the canonical Painlevé II equation.

1 Introduction

A variant of the nonlinear Schrödinger (NLS) equation of the type

$$i\Psi_t + \Delta\Psi + \nu|\Psi|^2 = s\frac{\Delta|\Psi|}{|\Psi|}\Psi \tag{1}$$

incorporating a de Broglie-Bohm quantum potential term $\Delta|\Psi|/|\Psi|$ as introduced in [8, 13] was derived 'ab initio' via Maxwell's equations in the context of the self-trapping of optical beams in [54]. In that nonlinear optics setting, the parameter $s < 1$ and reduction may be made to the standard NLS equation with the de Broglie-Bohm term removed. A 1+1-dimensional version of (1) was subsequently derived in [31] in connection with the Jackiw-Teitelbaum gravity model in general

C. Rogers · W. K. Schief (✉)
School of Mathematics and Statistics and Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems, The University of New South Wales, Sydney, NSW, Australia
e-mail: c.rogers@unsw.edu.au; w.schief@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_49

relativity. However, in that case, the parameter $s > 1$ and reduction to the canonical cubic NLS equation is not admitted and, accordingly, does not directly inherit such geometric properties as possession of an auto-Bäcklund transformation which is key in soliton theory (see, e.g., [43, 44]). The appearance of resonant solitonic behaviour involving novel fusion and fission phenomena in NLS equations (1) with de Broglie-Bohm potential and $s > 1$ has led to the terminology 'resonant' NLS equation for (1). NLS equations with underlying Hamiltonian structure have been set down in [32] which may be reduced to the 1+1-dimensional resonant NLS equation. The latter has subsequently been derived in a plasma physics context, where it describes the uni-axial propagation of long magneto-acoustic waves in a cold collisionless plasma subject to a transverse magnetic field [25]. In that setting, $s > 1$ and the resonant NLS model was shown to be equivalent to a canonical two-component system contained as a basic member of the AKNS hierarchy of integrable equations amenable to the inverse scattering transform method [2]. Therein, invariance under a novel Bäcklund-Darboux transformation was established along with a concomitant nonlinear superposition principle (permutability theorem). The latter was applied to analyse the solitonic nonlinear interaction of pairs of magneto-acoustic waves.

A 2+1-dimensional resonant Davey-Stewartson system was introduced in [52] in the context of a classical Korteweg-type capillarity model. A subsequent Painlevé analysis in [26] of this system is consistent with its integrability. An equivalent symmetric integrable 2+1-dimensional version of the Whitham-Broer-Kaup shallow water system was investigated in [41] and resonant solitonic interaction exhibited via a bilinear representation.

In [17], in a nonlinear optics setting, a class of symmetry reductions of the standard 1+1-dimensional NLS equation was introduced which resulted in the Painlevé II equation with zero parameter α. The existence of bounded dark solitons was revealed numerically for a restricted range of the ratio of nonlinearity to dispersion. Here, a natural extension of this class of similarity transformations is applied to a novel resonant Ermakov-Ray-Reid-NLS system and reduction made to an underlying hybrid Ermakov-Painlevé II system.

Nonlinear coupled systems of Ermakov-Ray-Reid type have their roots in the classical work of Ermakov [15] and were originally introduced by Ray and Reid in [33, 34]. The systems adopt the form

$$\ddot{u} + \omega(t)u = \frac{1}{u^2 v} S(v/u), \quad \ddot{v} + \omega(t)v = \frac{1}{v^2 u} T(u/v) \tag{2}$$

and admit a distinctive integral of motion, namely the invariant

$$\mathscr{E} = \frac{1}{2}(u\dot{v} - v\dot{u})^2 + \int^{v/u} S(z)\, dz + \int^{u/v} T(w)\, dw, \tag{3}$$

where, in the above, the dot indicates a derivative with respect to the independent variable t. Subsequently, 2+1-dimensional Ermakov-Ray-Reid systems were constructed in [48] and extensions to arbitrary order and dimension which admit characteristic Ermakov invariants were presented in [53]. Multi-component Ermakov-Ray-Reid systems were derived in the physical context of N-layer hydrodynamic systems in [42]. The importance of Ermakov-Ray-Reid systems in nonlinear optics is well-documented (see, e.g., [12, 18–20, 22, 50]) and literature cited therein). In that context, such systems have been derived, notably, to describe the evolution of the size and shape of the light spot and wave front in elliptic Gaussian beams. In addition, in recent years, integrable Ermakov-Ray-Reid systems have been shown to arise in a wide range of other physical contexts including warm-core oceanographic eddy theory [35], rotating shallow water hydrodynamics [39], anisentropic gasdynamics [45] and magnetogasdynamics [46]. They also arise in a spiralling elliptic soliton model in [14] and its extension in a Bose-Einstein setting in [1]. The Ermakov-Ray-Reid connections in the latter cases have been established in [51].

Here, the class of hybrid Ermakov-Painlevé II systems as obtained via a symmetry reduction of a coupled resonant NLS system is delimited under the constraint that it be Hamiltonian. An algorithmic solution procedure is presented which makes use of the Ermakov invariant associated with the Hamiltonian Ermakov-Painlevé II system. It involves, in particular, the Painlevé XXXIV equation in a squared amplitude $\Sigma > 0$. Thus, the isolation of positive solutions of Painlevé XXXIV is necessary to the method. In this regard, regions on which solutions of Painlevé XXXIV involving either Yablonskii-Vorob'ev polynomials or classical Airy functions are positive have been delimited in [7]. In that electrolytic context as described in [4–6, 9, 49], the solutions of Painlevé XXXIV determine the ion concentrations and so are necessarily positive.

2 A Resonant Ermakov-NLS System: Symmetry Reduction

Here, we introduce the two-component coupled Ermakov-NLS system

$$\begin{aligned}
i\Phi_z + \Delta\Phi - s\frac{\Delta|\Phi|}{|\Phi|}\Phi + \nu(|\Phi|^2 + |\Psi|^2)\Phi &= \frac{S(|\Psi|/|\Phi|)}{|\Phi|^3|\Psi|}\Phi \\
i\Psi_z + \Delta\Psi - s\frac{\Delta|\Psi|}{|\Psi|}\Psi + \nu(|\Phi|^2 + |\Psi|^2)\Psi &= \frac{T(|\Phi|/|\Psi|)}{|\Psi|^3|\Phi|}\Psi,
\end{aligned} \tag{4}$$

where $\Delta = \partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_n^2$, which incorporates de Broglie-Bohm potential terms $\Delta|\Phi|/|\Phi|$ and $\Delta|\Psi|/|\Psi|$. The quantities Φ and Ψ are scalar complex functions and S and T are real functions of their respective arguments, while s and ν denote real constants. A priori, all functions are defined locally with the independent variables $x_1, \ldots, x_n$ and z being real. For $S = T = 0$, system (4) constitutes a 'resonant' two-component Manakov system which may be transformed into the standard complex two-component Manakov system or its 'real' variant, depending on the value of the parameter s [38]. In particular, for $S = T = 0$ and $n = 1$, the above system is integrable. On the other hand, in [36], it has been shown that for $S = T = 0$, the resonant Manakov system (4) admits a symmetry reduction to a Ermakov-Painlevé system with particular non-vanishing associated Ermakov terms $S^*(|\Psi|/|\Phi|)$ and $T^*(|\Phi|/|\Psi|)$. Arbitrary such terms may be obtained by introducing the Ermakov terms $S(|\Psi|/|\Phi|)$ and $T(|\Phi|/|\Psi|)$ in (4). Thus, the above Ermakov-NLS system constitutes a natural extension of the resonant Manakov system. It is noted that, in 1974, Manakov [29], in the context of self-focussing electromagnetic wave propagation, introduced the celebrated integrable coupled NLS system which now bears his name. Later, Wai et al. [55] showed that the Manakov model accurately describes the propagation of stable pulses in optical devices consisting of birefringent fibres in the presence of random mode coupling. Empirical evidence for Manakov solitons in crystals was subsequently described by Kang et al. in [23]. The Manakov system has been also derived in a Kerr-type approximation of photoreactive crystals context by Kutuzov et al. in [24]. An N-component version of the Manakov system was shown to be integrable by Makhan'kov and Pashaev in [27]. The importance of such systems is current in connection with the construction of multimode optical fibre devices (see [30] and work cited therein).

A wave packet ansatz is now introduced into the $n+1$-dimensional nonlinear system (4) according to

$$\Phi = [\sigma(\xi) + i\tau(\xi)]e^{i\chi}, \quad \psi = [\phi(\xi) + i\psi(\xi)]e^{i\chi}, \tag{5}$$

where

$$\xi = \alpha z + \beta z^2 + \boldsymbol{\rho}\cdot\boldsymbol{x}, \quad \chi = \gamma z^3 + \delta z^2 + \zeta z + \epsilon z\,\boldsymbol{\rho}\cdot\boldsymbol{x} + \boldsymbol{\lambda}\cdot\boldsymbol{x} \tag{6}$$

and, as demonstrated below, the constant scalars $\alpha, \beta, \gamma, \delta, \zeta, \epsilon$ and vectors $\boldsymbol{\rho}, \boldsymbol{\lambda}$ have to be chosen appropriately. In the 1+1-dimensional case, in [17], symmetry representations of this type have been applied to the single component standard cubic NLS equation and reduction obtained to Painlevé II with zero parameter and interpreted in a nonlinear optics context. Here, we derive a nonlinear coupled dynamical system in the dependent variables σ, τ and ϕ, ψ and independent variable ξ. Thus, the linear structure of the variables ξ and χ in $\boldsymbol{x}$ implies that the Ermakov-NLS system subject to the constraints (5) and (6) may be formulated as

$$i\Phi_z + \Delta\Phi + A(\xi)\Phi = 0, \quad i\Psi_z + \Delta\Psi + B(\xi)\Psi = 0, \tag{7}$$

wherein the functions $A(\xi)$ and $B(\xi)$ may be read off the system (4). Insertion of the ansatz (5) into $(7)_1$ and subsequent separation of real and imaginary terms yield

$$\begin{aligned}
&\boldsymbol{\rho}^2\sigma_{\xi\xi} - [2\epsilon z\boldsymbol{\rho}^2 + 2\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda} + \alpha + 2\beta z]\tau_\xi \\
&\qquad - [3\gamma z^2 + 2\delta z + \zeta + \epsilon\,\boldsymbol{\rho}\cdot\boldsymbol{x} + (\epsilon z\boldsymbol{\rho} + \boldsymbol{\lambda})^2]\sigma + A(\xi)\sigma = 0 \\
&\boldsymbol{\rho}^2\tau_{\xi\xi} + [2\epsilon z\boldsymbol{\rho}^2 + 2\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda} + \alpha + 2\beta z]\sigma_\xi \\
&\qquad - [3\gamma z^2 + 2\delta z + \zeta + \epsilon\,\boldsymbol{\rho}\cdot\boldsymbol{x} + (\epsilon z\boldsymbol{\rho} + \boldsymbol{\lambda})^2]\tau + A(\xi)\tau = 0.
\end{aligned} \tag{8}$$

Collection of the terms independent of z leads to

$$\begin{aligned}\boldsymbol{\rho}^2\sigma_{\xi\xi} - k\tau_\xi - (\zeta + \epsilon\xi + \boldsymbol{\lambda}^2)\sigma + A(\xi)\sigma &= 0\\ \boldsymbol{\rho}^2\tau_{\xi\xi} + k\sigma_\xi - (\zeta + \epsilon\xi + \boldsymbol{\lambda}^2)\tau + A(\xi)\tau &= 0,\end{aligned} \tag{9}$$

where the constant k is defined by

$$k = \alpha + 2\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda}, \tag{10}$$

while the terms proportional to z give rise to

$$\begin{aligned}(2\epsilon\boldsymbol{\rho}^2 + 2\beta)\tau_\xi + (2\delta - \epsilon\alpha + 2\epsilon\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda})\sigma &= 0\\ (2\epsilon\boldsymbol{\rho}^2 + 2\beta)\sigma_\xi - (2\delta - \epsilon\alpha + 2\epsilon\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda})\tau &= 0.\end{aligned} \tag{11}$$

Here, terms of the type $\boldsymbol{\rho}\cdot\boldsymbol{x}$ in (8) have been eliminated in favour of ξ and z via $(6)_1$. The terms multiplying z^2 produce the constraint

$$3\gamma - \epsilon\beta + \epsilon^2\boldsymbol{\rho}^2 = 0 \tag{12}$$

and the conditions (11) are resolved by restricting the constants in (6) according to

$$\epsilon\boldsymbol{\rho}^2 + \beta = 0, \quad 2\delta - \epsilon\alpha + 2\epsilon\,\boldsymbol{\rho}\cdot\boldsymbol{\lambda} = 0. \tag{13}$$

Finally, for reasons of symmetry, the second Eq. $(7)_2$ does not impose any additional constraints on the available constants and may be formulated as the pair

$$\begin{aligned}\boldsymbol{\rho}^2\phi_{\xi\xi} - k\psi_\xi - (\zeta + \epsilon\xi + \boldsymbol{\lambda}^2)\phi + B(\xi)\phi &= 0\\ \boldsymbol{\rho}^2\psi_{\xi\xi} + k\phi_\xi - (\zeta + \epsilon\xi + \boldsymbol{\lambda}^2)\psi + B(\xi)\psi &= 0.\end{aligned} \tag{14}$$

Hence, we are left with the analysis of the pairs of ordinary differential equations (9) and (14), wherein

$$\begin{aligned}A(\xi) &= \nu(|\Phi|^2 + |\Psi|^2) - s\boldsymbol{\rho}^2\frac{|\Phi|_{\xi\xi}}{|\Phi|} - \frac{S(|\Psi|/|\Phi|)}{|\Phi|^3|\Psi|}\\ B(\xi) &= \nu(|\Phi|^2 + |\Psi|^2) - s\boldsymbol{\rho}^2\frac{|\Psi|_{\xi\xi}}{|\Psi|} - \frac{T(|\Phi|/|\Psi|)}{|\Psi|^3|\Phi|}\end{aligned} \tag{15}$$

with $|\Phi|^2 = \sigma^2 + \tau^2$ and $|\Psi|^2 = \phi^2 + \psi^2$.

We first observe that the pairs (9) and (14), in turn, yield

$$\begin{aligned}\boldsymbol{\rho}^2(\tau\sigma_{\xi\xi} - \sigma\tau_{\xi\xi}) - k(\sigma\sigma_\xi + \tau\tau_\xi) &= 0\\ \boldsymbol{\rho}^2(\psi\phi_{\xi\xi} - \phi\psi_{\xi\xi}) - k(\phi\phi_\xi + \psi\psi_\xi) &= 0\end{aligned} \tag{16}$$

so that the coupled system (9), (14) admits the first integrals

$$\begin{aligned}\mathscr{I} &= \boldsymbol{\rho}^2(\tau\sigma_\xi - \sigma\tau_\xi) - \frac{k}{2}(\sigma^2+\tau^2)\\ \mathscr{J} &= \boldsymbol{\rho}^2(\psi\phi_\xi - \phi\psi_\xi) - \frac{k}{2}(\phi^2+\psi^2).\end{aligned} \tag{17}$$

Furthermore, it is readily verified that

$$\begin{aligned}|\Phi||\Phi|_{\xi\xi} &= \sigma\sigma_{\xi\xi} + \tau\tau_{\xi\xi} + \frac{(\tau\sigma_\xi - \sigma\tau_\xi)^2}{\sigma^2+\tau^2}\\ |\Psi||\Psi|_{\xi\xi} &= \phi\phi_{\xi\xi} + \psi\psi_{\xi\xi} + \frac{(\psi\phi_\xi - \phi\psi_\xi)^2}{\phi^2+\psi^2}\end{aligned} \tag{18}$$

which, in conjunction with the first integrals (17), shows that the linear combinations $\sigma\cdot(9)_1 + \tau\cdot(9)_2$ and $\phi\cdot(14)_1 + \psi\cdot(14)_2$ may be expressed entirely in terms of the amplitudes $|\Phi|$ and $|\Psi|$, namely

$$\begin{aligned}|\Phi|_{\xi\xi} + [c_1 + c_2\xi + c_3(|\Phi|^2+|\Psi|^2)]|\Phi| &= \frac{1}{(1-s)}\left(\frac{\mathscr{I}^2}{\boldsymbol{\rho}^4|\Phi|^3} + \frac{S(|\Psi|/|\Phi|)}{\boldsymbol{\rho}^2|\Phi|^2|\Psi|}\right)\\ |\Psi|_{\xi\xi} + [c_1 + c_2\xi + c_3(|\Phi|^2+|\Psi|^2)]|\Psi| &= \frac{1}{(1-s)}\left(\frac{\mathscr{J}^2}{\boldsymbol{\rho}^4|\Psi|^3} + \frac{T(|\Phi|/|\Psi|)}{\boldsymbol{\rho}^2|\Psi|^2|\Phi|}\right),\end{aligned} \tag{19}$$

where, in the above,

$$c_1 = \frac{1}{(1-s)\boldsymbol{\rho}^2}\left(\frac{k^2}{4\boldsymbol{\rho}^2} - \zeta - \boldsymbol{\lambda}^2\right), \quad c_2 = \frac{\epsilon}{(s-1)\boldsymbol{\rho}^2}, \quad c_3 = \frac{\nu}{(1-s)\boldsymbol{\rho}^2} \tag{20}$$

and it is assumed that $s \neq 1$.

In the sequel, we set $\xi^* = c_1 + c_2\xi$, whence, with $c_2^2 = 2$ without loss of generality, (19) reduces to the hybrid Ermakov-Painlevé II system

$$\begin{aligned}|\Phi|_{\xi^*\xi^*} + \frac{\xi^*}{2}|\Phi| + \epsilon^*(|\Phi|^2+|\Psi|^2)|\Phi| &= \frac{S^*(|\Psi|/|\Phi|)}{|\Phi|^2|\Psi|}\\ |\Psi|_{\xi^*\xi^*} + \frac{\xi^*}{2}|\Psi| + \epsilon^*(|\Phi|^2+|\Psi|^2)|\Psi| &= \frac{T^*(|\Phi|/|\Psi|)}{|\Psi|^2|\Phi|},\end{aligned} \tag{21}$$

where $\epsilon^* = c_3/2$ and

$$\begin{aligned}S^*\left(\frac{|\Psi|}{|\Phi|}\right) &= \frac{1}{2(1-s)}\left[\frac{\mathscr{I}^2}{\boldsymbol{\rho}^4}\frac{|\Psi|}{|\Phi|} + \frac{1}{\boldsymbol{\rho}^2}S\left(\frac{|\Psi|}{|\Phi|}\right)\right]\\ T^*\left(\frac{|\Phi|}{|\Psi|}\right) &= \frac{1}{2(1-s)}\left[\frac{\mathscr{J}^2}{\boldsymbol{\rho}^4}\frac{|\Phi|}{|\Psi|} + \frac{1}{\boldsymbol{\rho}^2}T\left(\frac{|\Phi|}{|\Psi|}\right)\right].\end{aligned} \tag{22}$$

It is observed that, even in the case of the resonant Manakov system corresponding to $S = T = 0$, system (21) contains non-vanishing Ermakov terms S^* and T^*. Therefore, as indicated at the beginning of this section, by virtue of the algebraic form of S^* and T^*, the inclusion of the Ermakov terms S and T turns out to be natural.

3 An Integrable Ermakov-Painleve II System

Standard Ermakov-Ray-Reid systems (1) which admit a Hamiltonian have been delimited in [39, 50]. The Ermakov invariant together with the Hamiltonian, in that case, allow the algorithmic integration of such Ermakov-Ray-Reid systems. Here, the non-autonomous Ermakov-Painlevé II system (21) is considered subject to the Hamiltonian-type conditions

$$\frac{1}{|\Phi|^2|\Psi|}S^*\left(\frac{|\Psi|}{|\Phi|}\right) = -\frac{\partial V}{\partial|\Phi|}, \quad \frac{1}{|\Psi|^2|\Phi|}T^*\left(\frac{|\Phi|}{|\Psi|}\right) = -\frac{\partial V}{\partial|\Psi|}, \tag{23}$$

whence, it may be shown to adopt the form (cf. [39, 50])

$$\begin{aligned}
&|\Phi|_{\xi^*\xi^*} + \left[\frac{\xi^*}{2} + \epsilon^*(|\Phi|^2 + |\Psi|^2)\right]|\Phi| = \frac{2}{|\Phi|^3}J\left(\frac{|\Psi|}{|\Phi|}\right) + \frac{|\Psi|}{|\Phi|^4}J'\left(\frac{|\Psi|}{|\Phi|}\right) \\
&|\Psi|_{\xi^*\xi^*} + \left[\frac{\xi^*}{2} + \epsilon^*(|\Phi|^2 + |\Psi|^2)\right]|\Psi| = -\frac{1}{|\Phi|^3}J'\left(\frac{|\Psi|}{|\Phi|}\right),
\end{aligned} \tag{24}$$

where the prime denotes a derivative with respect to the argument $|\Psi|/|\Phi|$. Since the system (21) is of Ermakov form (1) if one formally sets $\omega(\xi^*) = \xi^*/2 + \epsilon^*(|\Phi|^2 + |\Psi|^2)$, the associated Ermakov invariant (2) is still applicable and, in the current Hamiltonian case, we obtain the invariant

$$\mathscr{E} = \frac{1}{2}(|\Phi||\Psi|_{\xi^*} - |\Psi||\Phi|_{\xi^*})^2 + \frac{|\Phi|^2 + |\Psi|^2}{|\Phi|^2}J\left(\frac{|\Psi|}{|\Phi|}\right). \tag{25}$$

On use of the identity,

$$\begin{aligned}
&(|\Phi|^2 + |\Psi|^2)(|\Phi|^2_{\xi^*} + |\Psi|^2_{\xi^*}) \\
&\qquad -(|\Phi||\Psi|_{\xi^*} - |\Psi||\Phi|_{\xi^*})^2 = (|\Phi||\Phi|_{\xi^*} + |\Psi||\Psi|_{\xi^*})^2,
\end{aligned} \tag{26}$$

the above Ermakov invariant adopts the form

$$\mathscr{E} = \left[\frac{1}{2}(|\Phi|^2_{\xi^*} + |\Psi|^2_{\xi^*}) + \frac{J(|\Psi|/|\Phi|)}{|\Phi|^2}\right]\Sigma - \frac{1}{8}\Sigma^2_{\xi^*}, \qquad \Sigma = |\Phi|^2 + |\Psi|^2, \tag{27}$$

while an appropriate linear combination of the pair (24) delivers the relation

$$\frac{1}{2}(|\Phi|^2_{\xi^*} + |\Psi|^2_{\xi^*})_{\xi^*} + \frac{1}{2}\left[\frac{\xi^*}{2} + \epsilon^*\Sigma\right]\Sigma_{\xi^*} = -\left[\frac{J(|\Psi|/|\Phi|)}{|\Phi|^2}\right]_{\xi^*}. \tag{28}$$

Now, elimination of J between the latter two relations, importantly, produces the canonical integrable Painlevé XXXIV equation in standard form [10], namely

$$\Sigma_{\xi^*\xi^*} - \frac{1}{2\Sigma}\Sigma^2_{\xi^*} + \xi^*\Sigma + 2\epsilon^*\Sigma^2 - \frac{4\mathscr{E}}{\Sigma} = 0, \tag{29}$$

if we set $\epsilon^{*2} = 1$ without loss of generality and make the identification

$$\mathscr{E} = -\frac{(\pm\alpha^* - \epsilon^*/2)^2}{8} < 0. \tag{30}$$

It is noted that, with $\Omega = \Sigma^{1/2}$, (29) may be formulated as the single component hybrid Ermakov-Painlevé II equation

$$\Omega_{\xi^*\xi^*} + \left[\frac{\xi^*}{2} + \epsilon^*\Omega^2\right]\Omega = \frac{2\mathscr{E}}{\Omega^3}. \tag{31}$$

In the present context, it is seen that attention must be restricted to regions in which the solutions Σ of Painlevé XXXIV are positive. The physical importance of positive solutions of Painlevé XXXIV arises naturally in the setting of two-ion electrodiffusion. Thus, in the electrolytic context of [7], the ion concentrations, which are necessarily positive, were shown to be determined by Painlevé XXXIV. This positivity constraint and attendant conditions were treated therein.

3.1 Determination of the Amplitudes $|\Phi|$ and $|\Psi|$

To find $|\Psi|/|\Phi|$ so that, together with a known solution $\Sigma > 0$ of the Painlevé XXXIV equation, the variables $|\Phi|$ and $|\Psi|$ in the Ermakov-Painlevé II system (24) are determined, return is made to the Ermakov invariant relation (25). Thus, on introduction of Λ according to

$$\Lambda = \frac{2|\Phi||\Psi|}{|\Phi|^2 + |\Psi|^2}, \tag{32}$$

it is readily shown that

$$\Lambda_{\xi^*} = 2\frac{|\Phi|^2 - |\Psi|^2}{(|\Phi|^2 + |\Psi|^2)^2}(|\Phi||\Psi|_{\xi^*} - |\Psi||\Phi|_{\xi^*}), \tag{33}$$

whence the Ermakov invariant (25) may be formulated as

$$\mathscr{E} = \frac{1}{8}\frac{(|\Phi|^2+|\Psi|^2)^4}{(|\Phi|^2-|\Psi|^2)^2}\Lambda^2_{\xi^*} + \frac{|\Phi|^2+|\Psi|^2}{|\Phi|^2} J\left(\frac{|\Psi|}{|\Phi|}\right). \tag{34}$$

Now, in terms of the new independent variable $\bar{\xi}$ defined by

$$d\bar{\xi} = \Sigma^{-1} d\xi^*, \tag{35}$$

the Ermakov invariant (34) shows that

$$\Lambda^2_{\bar{\xi}} = 8\left(\frac{1-(|\Psi|/|\Phi|)^2}{1+(|\Psi|/|\Phi|)^2}\right)^2 \left[\mathscr{E} - \left(1+\left(\frac{|\Psi|}{|\Phi|}\right)^2\right) J\left(\frac{|\Psi|}{|\Phi|}\right)\right]. \tag{36}$$

The relation (32) gives $|\Psi|/|\Phi|$ in terms of Λ according to

$$\frac{|\Psi|}{|\Phi|} = \frac{1 \pm \sqrt{1-\Lambda^2}}{\Lambda} \tag{37}$$

so that Λ is obtained as an implicit function of $\bar{\xi}$ via

$$\pm\frac{1}{2^{3/2}}\int L(\Lambda)\, d\Lambda = \bar{\xi}, \quad L(\Lambda) = \sqrt{\frac{\Lambda}{(1-\Lambda^2)[\mathscr{E}\Lambda - 2\mathscr{L}(\Lambda)]}}, \tag{38}$$

where

$$\mathscr{L}(\Lambda) = \frac{|\Psi|}{|\Phi|} J\left(\frac{|\Psi|}{|\Phi|}\right). \tag{39}$$

Hence, corresponding to positive solutions Σ of the Painlevé XXXIV equation (29) and Λ determined by (38) for appropriate $J(|\Psi|/|\Phi|)$ via (39), the squared amplitudes of Φ and Ψ in the Hamiltonian Ermakov-Painleve II system (24) are given by the relations

$$|\Phi|^2 = \frac{1}{2}(1 \pm \sqrt{1-\Lambda^2})\Sigma, \quad |\Psi|^2 = \frac{1}{2}(1 \mp \sqrt{1-\Lambda^2})\Sigma. \tag{40}$$

3.2 Determination of the Phases of Φ and Ψ

To complete the solution procedure for σ, τ and ϕ, ψ in the original wave packet representations (5), we return to the first integrals (17) involving $\mathscr{I}$ and $\mathscr{J}$. These yield, in turn,

$$c_2 \boldsymbol{\rho}^2 \frac{d}{d\xi^*}\left(\tan^{-1}\frac{\sigma}{\tau}\right) - \frac{k}{2} = \frac{\mathscr{I}}{\sigma^2+\tau^2} \tag{41}$$

and

$$c_2\boldsymbol{\rho}^2 \frac{d}{d\xi^*}\left(\tan^{-1}\frac{\phi}{\psi}\right) - \frac{k}{2} = \frac{\mathscr{J}}{\phi^2+\psi^2}, \tag{42}$$

whence, on integration,

$$c_2\boldsymbol{\rho}^2 \tan^{-1}\frac{\sigma}{\tau} = \frac{k}{2}\xi^* + \mathscr{I}\int\frac{d\xi^*}{|\Phi|^2}, \quad c_2\boldsymbol{\rho}^2 \tan^{-1}\frac{\phi}{\psi} = \frac{k}{2}\xi^* + \mathscr{J}\int\frac{d\xi^*}{|\Psi|^2}. \tag{43}$$

Thus, the quantities σ, τ and ϕ, ψ are given by the relations

$$\begin{aligned}
\sigma &= \pm|\Phi|\sin\left[\frac{1}{c_2\boldsymbol{\rho}^2}\left(\frac{k}{2}\xi^* + \mathscr{I}\int\frac{d\xi^*}{|\Phi|^2}\right)\right]\\
\tau &= \pm|\Phi|\cos\left[\frac{1}{c_2\boldsymbol{\rho}^2}\left(\frac{k}{2}\xi^* + \mathscr{I}\int\frac{d\xi^*}{|\Phi|^2}\right)\right]\\
\phi &= \pm|\Psi|\sin\left[\frac{1}{c_2\boldsymbol{\rho}^2}\left(\frac{k}{2}\xi^* + \mathscr{J}\int\frac{d\xi^*}{|\Psi|^2}\right)\right]\\
\psi &= \pm|\Psi|\cos\left[\frac{1}{c_2\boldsymbol{\rho}^2}\left(\frac{k}{2}\xi^* + \mathscr{J}\int\frac{d\xi^*}{|\Psi|^2}\right)\right],
\end{aligned} \tag{44}$$

where $|\Phi|^2$ and $|\Psi|^2$ are determined by (40). In the above, on use of the relations (35) and (38), the integrals in the above phases may be reformulated as

$$\begin{aligned}
\int\frac{d\xi^*}{|\Phi|^2} &= \int\frac{2\,d\bar{\xi}}{1\pm\sqrt{1-\Lambda^2}} = \pm\frac{1}{2^{1/2}}\int\frac{L(\Lambda)}{1\pm\sqrt{1-\Lambda^2}}d\Lambda\\
\int\frac{d\xi^*}{|\Psi|^2} &= \int\frac{2\,d\bar{\xi}}{1\mp\sqrt{1-\Lambda^2}} = \pm\frac{1}{2^{1/2}}\int\frac{L(\Lambda)}{1\mp\sqrt{1-\Lambda^2}}d\Lambda.
\end{aligned} \tag{45}$$

4 Solution Generation Techniques

In the preceding, we have established that wave packet solutions of the Ermakov-NLS system (4) may, in principle, be obtained, by isolating positive solutions Σ of the Painlevé XXXIV equation (29) and, subsequently, performing associated quadratures (35), (38) and (45). Here, we briefly demonstrate how this can be done in concrete terms. It is noted that the challenges posed by the numerical treatment of the classical Painleve I-VI equations have been described in [16]. The Cauchy problem for Painlevé I-VI has been investigated in [3]. The numerical treatment of the hybrid Ermakov-Painleve II system of the present paper remains to be investigated.

For convenience, we focus on the case $\epsilon^* = -1$ and the Ermakov invariant $\mathscr{E} = -(\alpha^* + \frac{1}{2})^2/8$ (cf. (30)) so that the Painlevé equation XXXIV (29) adopts the form

$$\Sigma_{\xi^*\xi^*} = \frac{\Sigma_{\xi^*}^2}{2\Sigma} - \xi^*\Sigma + 2\Sigma^2 - \frac{(\alpha^* + \frac{1}{2})^2}{2\Sigma}. \tag{46}$$

All other cases may be treated in a similar manner. The latter may be formulated as the pair of first-order differential equations

$$Y_{\xi^*} = -Y^2 - \frac{\xi^*}{2} + \Sigma, \quad \Sigma_{\xi^*} = 2Y\Sigma + \alpha^* + \frac{1}{2}. \tag{47}$$

Indeed, elimination of the auxiliary variable Y leads to (46). On the other hand, if we regard $(47)_1$ as a definition of Σ then the Painlevé II equation

$$Y_{\xi^*\xi^*} = 2Y^3 + \xi^* Y + \alpha^* \tag{48}$$

results. Hence, we have retrieved the classical link between the Painlevé II and XXXIV equations (see [10, 11] and references therein). This connection may be exploited to generate algebraically an infinite sequence of solutions of the Painlevé XXXIV equation by means of the Bäcklund transformation for the Painlevé II equation. Specifically, for any solution Y_{α^*} of the Painlevé II equation corresponding to the parameter α^* and an associated solution Σ_{α^*} of the Painlevé XXXIV equation, another solution Y_{α^*+1} of the Painlevé II equation corresponding to the parameter $\alpha^* + 1$ is given by [11]

$$Y_{\alpha^*+1} = -Y_{\alpha^*} - \frac{\alpha^* + \frac{1}{2}}{\Sigma_{\alpha^*}}. \tag{49}$$

The action of this Bäcklund transformation on the solutions of Painlevé XXXIV may be formulated as

$$\Sigma_{\alpha^*+1} = -\Sigma_{\alpha^*} + 2Y_{\alpha^*+1}^2 + \xi^*. \tag{50}$$

Thus, firstly, if we start with the solution $Y_0 = 0$ of the Painlevé II equation, corresponding to the parameter $\alpha^* = 0$ and the solution $\Sigma_0 = \xi^*/2$ of the Painlevé XXXIV equation then their Bäcklund transforms are given by $Y_1 = -1/\xi^*$ and $\Sigma_1 = 2/\xi^{*2} + \xi^*/2$ respectively. Iterative application of the Bäcklund transformation leads to a sequence of rational solutions of the Painlevé II and XXXIV equations which may be expressed in terms of Yablonskii-Vorob'ev polynomials [10] and, as required in the current Ermakov-NLS context, the solutions Σ_{α^*}, $\alpha^* \in \mathbb{N}$ of the Painlev'e XXXIV equation may be shown to be positive in suitable regions [7]. Secondly, it is well known that

$$Y_{\frac{1}{2}} = -\frac{\mathsf{A}_{\xi^*}}{\mathsf{A}}, \tag{51}$$

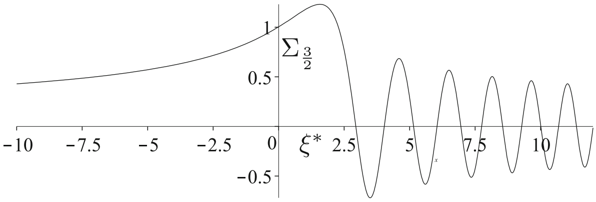


Fig. 1 The solution $\Sigma_{\frac{3}{2}}$ of the Painlevé XXXIV equation

where A is a solution of the Airy equation

$$\mathsf{A}_{\xi^*\xi^*} + \frac{\xi^*}{2}\mathsf{A} = 0, \tag{52}$$

constitutes a solution of the Painlevé II equation associated with the parameter $\alpha^* = 1/2$. The associated solution of the Painlevé XXXIV equation is given by

$$\Sigma_{\frac{1}{2}} = 2Y_{\frac{1}{2}}^2 + \xi^*, \tag{53}$$

as may be concluded from $(47)_1$. Hence, iterative application of the above Bäcklund transformation to this type of solution leads to an infinite sequence of solutions Σ_{α^*}, $\alpha^* = \mathbb{N} + \frac{1}{2}$ of the Painlevé XXXIV equation which are parametrised in terms of Airy functions [10] and, once again, may be shown to satisfy the positivity requirement in appropriate regions [7] if one makes the choice $\mathsf{A} = \mathsf{Ai}(-2^{-1/3}\xi^*)$, where Ai denotes the Airy function of the first kind. Specifically, evaluation of (49) and (50) for $\alpha^* = 1/2$ produces

$$\Sigma_{\frac{3}{2}} = 2\frac{1 - 2\xi^* R - 4R^3}{(2R^2 + \xi^*)^2}, \quad R = \frac{\mathsf{A}_{\xi^*}}{\mathsf{A}}. \tag{54}$$

As depicted in Fig. 1, $\Sigma_{\frac{3}{2}}$ has an infinite number of zeros $\xi_0^* < \xi_1^* < \xi_2^* < \cdots$ but it is positive for $\xi^* < \xi_0^*$ with $\Sigma_{\frac{3}{2}} \to 0$ as $\xi^* \to -\infty$.

The next step in the procedure is to express Λ encoded in (38) as an implicit function of $\bar{\xi}$ as a function of ξ^*. To this end, the relation (35) has to be integrated so that $\bar{\xi}$ becomes a known function of ξ^*. Remarkably, it turns out that the action of the above Bäcklund transformation may be extended to the variable $\bar{\xi}$. Thus, if $\bar{\xi}_{\alpha^*}$ is associated with a solution Σ_{α^*} of the Painlevé equation XXXIV via

$$d\bar{\xi}_{\alpha^*} = \Sigma_{\alpha^*}^{-1} d\xi^* \tag{55}$$

then it may be shown [47] that, up to an additive constant,

$$\bar{\xi}_{\alpha^*+1} = \frac{(\alpha^* + \frac{1}{2})\bar{\xi}_{\alpha^*} + \ln(\Sigma_{\alpha^*+1}\Sigma_{\alpha^*})}{\alpha^* + \frac{3}{2}} \tag{56}$$

is the solution of

$$d\bar{\xi}_{\alpha^*+1} = \varSigma^{-1}_{\alpha^*+1}\, d\xi^*, \tag{57}$$

where $\varSigma_{\alpha^*+1}$ is the Bäcklund transform of $\varSigma_{\alpha^*}$. Accordingly, for the above-mentioned rational and Airy-type solutions $\varSigma$ of the Painlevé XXXIV equation (29), the change of variables (35) may be achieved explicitly. For instance, for the Airy-type seed solution $\varSigma_{\frac{1}{2}}$, (55) may be integrated directly to obtain $\bar{\xi}_{\frac{1}{2}} = \ln(2\mathsf{A}^2_{\xi^*} + \xi^*\mathsf{A}^2)$ so that (56) results in

$$\bar{\xi}_{\frac{3}{2}} = \frac{1}{2}\ln[2(1 - 2\xi^* R - 4R^3)\mathsf{A}^2]. \tag{58}$$

Finally, in order to complete the solution procedure, the quadratures (38) and (45) need to be addressed. Here, we focus on the canonical case $\mathscr{L}(\varLambda) = \text{const}$. By definition (39), this case corresponds to

$$J\left(\frac{|\varPsi|}{|\varPhi|}\right) = \frac{|\varPhi|}{|\varPsi|}\mathscr{L} \tag{59}$$

so that comparison of the Ermakov-Painlevé II systems (21) and (24) reveals that $S^* = T^* = \mathscr{L}$. It is noted that conventional Ermakov-Ray-Reid systems (2) with $S = T =$ const arise in moving shoreline analysis in 2+1-dimensional shallow water theory [39], variational approximation in nonlinear optics [51] and in the study of integrable structure in modulated NLS models [37, 40]. The latter arise notably in a nonlinear optics context, in particular, in connection with soliton management [28] in the context of the propagation of Bloch waves in optical lattices (see, e.g., [56]), wherein modulation was related to the classical Ermakov equation. Now, reality of the integral (38) dictates that

$$\mathscr{L} = \frac{1}{2}\mathscr{A}\mathscr{E}, \quad \mathscr{A} > 1 \tag{60}$$

so that we obtain the elliptic integral

$$\int L(\varLambda)\, d\varLambda = \frac{1}{\sqrt{-\mathscr{E}}}\int \frac{\varLambda}{\sqrt{(1-\varLambda^2)(\mathscr{A}-\varLambda)\varLambda}} d\varLambda. \tag{61}$$

In terms of the incomplete elliptic integrals of the first and third kinds [21]

$$\mathsf{F}(z,\kappa) = \varPi(z,0,\kappa), \quad \varPi(z,a,\kappa) = \int_0^z \frac{1}{(1-at^2)\sqrt{1-t^2}\sqrt{1-\kappa^2 t^2}} dt, \tag{62}$$

we find that, up to an additive constant of integration,

$$\int L(\Lambda)\,d\Lambda = \frac{1}{\sqrt{-\mathscr{L}}}[\Pi(z,a,\kappa) - \mathsf{F}(z,\kappa)], \tag{63}$$

where

$$z = \sqrt{\frac{2\Lambda}{1+\Lambda}}, \quad a = \frac{1}{2}, \quad \kappa = \sqrt{\frac{1+\mathscr{A}}{2\mathscr{A}}}. \tag{64}$$

It is emphasised that the relation (38), that is,

$$\Pi(z,a,\kappa) - \mathsf{F}(z,\kappa) = \pm\sqrt{-8\mathscr{L}}\,\bar{\xi}, \tag{65}$$

is invertible due to the positivity of the integrand $L(\Lambda)$. Hence, the function

$$\Lambda : [0, \bar{\xi}_{max}] \to [0,1], \quad \bar{\xi}_{max} = \frac{\Pi(a,\kappa) - \mathsf{F}(\kappa)}{\sqrt{-8\mathscr{L}}} \tag{66}$$

is well defined. Here, we have chosen the plus sign in (65) and $\mathsf{F}(\kappa)$ and $\Pi(a,\kappa)$ constitute the complete elliptic integrals of the first and third kinds respectively. Moreover, $\Lambda(\bar{\xi})$ may be defined for all $\bar{\xi}$ by considering the even periodic extension of (66) obtained by including the additive constant of integration and exploiting the arbitrary sign in (65). Even though this extended function is not differentiable everywhere since $d\Lambda/d\bar{\xi} \sim 1/\sqrt{\Lambda}$ for small Λ, the quantity Λ^2, which is the key ingredient in the squares (40) of the amplitudes $|\Phi|$ and $|\Psi|$, is differentiable everywhere (cf. Fig. 2). It is noted that we may also incorporate a global constant of integration $\bar{\xi}_t$ in (65) corresponding to $\bar{\xi} \to \bar{\xi} + \bar{\xi}_t$.

The remaining integrals (45) may also be expressed in terms of elliptic integrals. Indeed, one may directly verify that

$$\sqrt{-\mathscr{L}}\int \frac{L(\Lambda)}{1 \pm \sqrt{1-\Lambda^2}}d\Lambda = \mathsf{F}(z,\kappa) - 2\mathsf{E}(z,\kappa) - \frac{\sqrt{2}}{\sqrt{\mathscr{A}}}\frac{\sqrt{\mathscr{A}-\Lambda}}{\sqrt{\Lambda}}\left(\frac{\sqrt{1-\Lambda}}{\sqrt{1+\Lambda}} \mp 1\right), \tag{67}$$

where

$$\mathsf{E}(z,\kappa) = \int_0^z \frac{\sqrt{1-\kappa^2 t^2}}{\sqrt{1-t^2}}dt \tag{68}$$

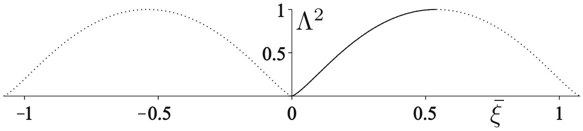


Fig. 2 The even periodic extension (dotted) of Λ^2 (solid) for $\mathscr{A} = 2$ and $\mathscr{E} = -1/2$

denotes the incomplete elliptic integral of the second kind. Once again, an additive constant of integration may be included. We conclude by remarking that the quadratures (38) and (45) may also be performed explicitly in the case $\mathscr{L}(\varLambda) \sim \varLambda^{-1}$ which, in particular, captures the resonant Manakov system ($S = T = 0$) for equal first integrals $\mathscr{I} = \mathscr{J}$.

As an illustration, we now briefly examine the nature of the wave packet solutions of the Ermakov-NLS system associated with the solution $\varSigma_{\frac{3}{2}}$ of the Painlevé XXXIV equation. To this end, we observe that the signs in the expressions (40) for the (squares of the) amplitudes $|\varPhi|$ and $|\varPsi|$ do not have to be fixed globally but may vary from region to region as long as, for any fixed ξ^*, $|\varPhi|$ and $|\varPsi|$ are associated with opposite signs so that $|\varPhi|^2 + |\varPsi|^2 = \varSigma$. Accordingly, we consider the quantities

$$Q_{\pm} = \frac{1}{2}(1 \pm \sqrt{1 - \varLambda^2})\varSigma_{\frac{3}{2}}. \tag{69}$$

For $\mathscr{A} = 2$ (and $\mathscr{E} = -1/2$ due to $\alpha^* = 3/2$), these are depicted in Fig. 3 for the relevant region $\xi^* \leq \xi_0^*$ on which $\varSigma_{\frac{3}{2}} \geq 0$. It is seen that if we define the squares $|\varPhi|^2$ and $|\varPsi|^2$ by 'alternating' between Q_+ and Q_- then $|\varPhi|^2$ and $|\varPsi|^2$ may be regarded as encoding two wave trains which are enveloped by $\varSigma$. Indeed, since there exists an infinite sequence of points $\cdots < \xi_{-2}^* < \xi_{-1}^* < \xi_0^*$ at which $Q_+ = Q_- = \varSigma_{\frac{3}{2}}/2$, we may set

$$\left.\begin{aligned} |\varPhi|^2 &= \frac{1}{2}(1 + (-1)^m\sqrt{1 - \varLambda^2})\varSigma_{\frac{3}{2}}, \\ |\varPsi|^2 &= \frac{1}{2}(1 - (-1)^m\sqrt{1 - \varLambda^2})\varSigma_{\frac{3}{2}}, \end{aligned}\right. \quad \xi^* \in [\xi_{m-1}^*, \xi_m^*], \tag{70}$$

which are differentiable everywhere. Thus, the solution $\varSigma_{\frac{3}{2}}$ of the Painlevé XXXIV equation gives rise to wave train solutions of the Ermakov-NLS system (4) which travel at constant speed if $\beta = 0$ or accelerate uniformly if $\beta \neq 0$.

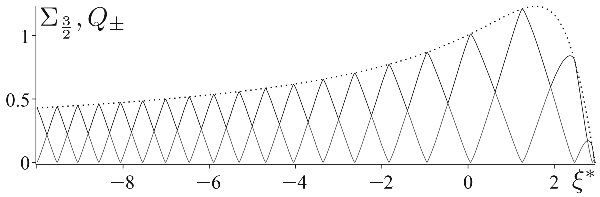


Fig. 3 The positive part of the solution $\varSigma_{\frac{3}{2}}$ (dotted) of the Painlevé XXXIV equation and its constituents Q_+ (black) and Q_- (grey) for $\mathscr{A} = 2$ and $\bar{\xi}_t = 0.3$, encoding the squared amplitudes $|\varPhi|^2$ and $|\varPsi|^2$ via suitable 'glueing'

References

1. Abdullaev, Y., Desyatnikov, A.S., Ostravoskaya, E.A.: Suppression of collapse for matter waves with orbital angular momentum. J. Opt. **13**, 064023 (2011)
2. Ablowitz, M.J., Kaup, D.J., Newell, A.C., Segur, H.: Nonlinear evolution equations of physical significance. Phys. Rev. Lett. **31**, 125–127 (1973)
3. Abramov, A.A., Yukhno, L.F.: A method for the numerical solution of the Painlevé equations. Comput. Math. Math. Phys. **53**, 540–563 (2013)
4. Amster, P., Rogers, C.: On a Ermakov-Painlevé II reduction in three-ion electrodiffusion. A Dirichlet boundary value problem. Discrete Contin. Dyn. Syst. **35**, 3277–3292 (2015)
5. Bass, L.K.: Electrical structures of interfaces in steady electrolysis. Trans. Faraday Soc. **60**, 1656–1669 (1964)
6. Bass, L.K.: Irreversible interactions between metals and electrolytes. Proc. R. Soc. Lond. A **277**, 125–136 (1964)
7. Bass, L., Nimmo, J.J.C., Rogers, C., Schief, W.K.: Electrical structures of interfaces: a Painlevé II model. Proc. R. Soc. Lond. A **466**, 2117–2136 (2010)
8. Bohm, D.: A suggested interpretation of the quantum theory in terms of "hidden" variables. I and II. Phys. Rev. **85**, 166–193 (1952)
9. Bracken, A.J., Bass, L., Rogers, C.: Bäcklund flux-quantization in a model of electrodiffusion based on Painlevé II. J. Phys. A Math. Theor. **45**, 105204 (2012)
10. Clarkson, P.A.: Painlevé equations. Nonlinear special functions. J. Comput. Appl. Math. **153**, 127–140 (2003)
11. Conte, R. (ed.): The Painlevé Property: One Century Later. Springer, New York (1999)
12. Cornolti, F., Lucchesi, M., Zambon, B.: Elliptic Gaussian beam self-focussing in nonlinear media. Opt. Commun. **75**, 129–135 (1990)
13. de Broglie, L.: La mécanique ondulatoire et la structure atomique de la matiére et du rayonnement. J. Phys. Radium **8**, 225–241 (1927)
14. Desyatnikov, A.S., Buccoliero, D., Dennis, M.R., Kivshar, Y.S.: Suppression of collapse for spiralling elliptic solitons. Phys. Rev. Lett. **104**, 053902-1–053902-4 (2010)
15. Ermakov, V.P.: Second-order differential equations: conditions of complete integrability. Univ. Izy. Kiev **20**, 1–25 (1880)
16. Fornberg, B., Weideman, J.A.C.: A numerical methodology for the Painlevé equations. Oxford Centre for Collaborative Applied Mathematics Report **11/06** (2011)
17. Giannini, J.A., Joseph, R.I.: The role of the second Painlevé transcendent in nonlinear optics. Phys. Lett. A **141**, 417–419 (1989)
18. Goncharenko, A.M., Logvin, Y.A., Samson, A.M., Shapovalov, P.S., Turovets, S.I.: Ermakov Hamiltonian systems in nonlinear optics of elliptic Gaussian beams. Phys. Lett. A **160**, 138–142 (1991)
19. Goncharenko, A.M., Logvin, Y.A., Samson, A.M., Shapovalov, P.S.: Rotating ellipsoidal Gaussian beams in nonlinear media. Opt. Commun. **81**, 225–230 (1991)
20. Goncharenko, A.M., Kukushkin, V.G., Logvin, Y.A., Samson, A.M.: Self-focussing of two orthogonally polarised light beams in a nonlinear medium. Opt. Quant. Electron. **25**, 97–104 (1999)
21. Gradshteyn, I.S., Ryzhik, I.M.: In: Jeffrey, A., Zwillinger, D. (eds.) Table of Integrals, Series, and Products, 6th edn. Academic Press, San Diego (2000)
22. Guilano, C.R., Marburger, J.H., Yariv, A.: Enhancement of self-focussing threshold in sapphire with elliptical beams. Appl. Phys. Lett. **21**, 58–60 (1972)
23. Kang, J.U., Stegeman, G.I., Aitchison, J.S., Akhmediev, N.: Nonlinear pulse propagation in birefringent optical fibres. Phys. Rev. Lett. **76**, 3699–3702 (1996)
24. Kutuzov, V., Petnikova, V.M., Shuvalov, V.V., Vysloukh, V.A.: Cross-modulation coupling of incoherent soliton models in photorefractive crystals. Phys. Rev. E **57**, 6056–6065 (1998)

25. Lee, J.H., Pashaev, O.K., Rogers, C., Schief, W.K.: The resonant nonlinear Schrödinger equation in cold plasma physics: application of Bäcklund-Darboux transformations and superposition principles. J. Plasma Phys. **73**, 257–272 (2007)
26. Liang, Z.F., Tang, X.Y.: Painlevé analysis and exact solutions of the resonant Davey-Stewartson system. Phys. Lett. A **274**, 110–115 (2009)
27. Makhan'kov, V.G., Pashaev, O.K.: Nonlinear Schrödinger equation with noncompact isogroup. Theor. Math. Phys. **53**, 979–987 (1982)
28. Malomed, B.A.: Soliton Management in Periodic Systems. Springer, New York (2006)
29. Manakov, S.V.: On the theory of two-dimensional stationary self-focussing of electromagnetic waves. Sov. Phys. JETP **38**, 248–553 (1974)
30. Mecozzi, A., Antonelli, C., Shtaif, M.: Nonlinear propagation in multi-mode fibers in the strong coupling regime (2012). arXiv: 1203.6275.v2 [physics optics]
31. Pashaev, O.K., Lee, J.H.: Resonance solitons as black holes in Madelung fluid. Mod. Phys. Lett. A **17**, 1601–1619 (2002)
32. Pashaev, O.K., Lee, J.H., Rogers, C.: Soliton resonances in a generalised nonlinear Schrödinger equation. J. Phys. A Math. Theor. **41**, 452001 (9pp) (2008)
33. Ray, J.R.: Nonlinear superposition law for generalised Ermakov systems. Phys. Lett. A **78**, 4–6 (1980)
34. Reid, J.L., Ray, J.R.: Ermakov systems, nonlinear superposition and solution of nonlinear equations of motion. J. Math. Phys. **21**, 1583–1587 (1980)
35. Rogers, C.: Elliptic warm-core theory. Phys. Lett. A **138**, 267–273 (1989)
36. Rogers, C.: A novel Ermakov-Painleve II system: N+1-dimensional coupled NLS and elastodynamic reductions. Stud. Appl. Math. **133**, 214–231 (2014)
37. Rogers, C.: Gausson-type representations in nonlinear physics: Ermakov modulation. Phys. Scr. **89**, 105208 (8pp) (2014)
38. Rogers, C.: Integrable substructure in a Korteweg capillarity model. A Kármán-Tsien type constitutive relation. J. Nonlinear Math. Phys. **21**, 74–88 (2014)
39. Rogers, C., An, H.: Ermakov-Ray-Reid systems in 2+1-dimensional rotating shallow water theory. Stud. Appl. Math. **125**, 275–299 (2010)
40. Rogers, C., An, H.: On a 2+1-dimensional Madelung system with logarithmic and with Bohm quantum potentials: Ermakov reduction. Phys. Scr. **84**, 045004 (7pp) (2011)
41. Rogers, C., Pashaev, O.K.: On a 2+1-dimensional Whitham-Broer-Kaup system: a resonant NLS connection. Stud. Appl. Math. **127**, 114–152 (2011)
42. Rogers, C., Schief, W.K.: Multi-component Ermakov systems: structure and linearisation. J. Math. Anal. Appl. **198**, 194–220 (1996)
43. Rogers, C., Schief, W.K.: Intrinsic geometry of the NLS equation and its auto-Bäcklund transformation. Stud. Appl. Math. **101**, 267–287 (1998)
44. Rogers, C., Schief, W.K.: Bäcklund and Darboux Transformations. Geometry and Modern Applications in Soliton Theory. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (2002)
45. Rogers, C., Schief, W.K.: On the integrability of a Hamiltonian reduction of a 2+1-dimensional non-isothermal rotating gas cloud system. Nonlinearity **24**, 3165–3178 (2011)
46. Rogers, C., Schief, W.K.: The pulsrodon in 2+1-dimensional magneto-gasdynamics. Hamiltonian structure and integrability. J. Math. Phys. **52**, 083701 (2011)
47. Rogers, C., Schief, W.K.: On Ermakov-Painlevé II systems. Integrable reduction. Meccanica **51**, 2967–2974 (2016)
48. Rogers, C., Hoenselaers, C., Ray, J.R.: On 2+1-dimensional Ermakov systems. J. Phys. A Math. Gen. **26**, 2625–2633 (1993)
49. Rogers, C., Bassom, A.P., Schief, W.K.: On a Painlevé II model in steady electrolysis: application of a Bäcklund transformation. J. Math. Anal. Appl. **240**, 367–381 (1999)
50. Rogers, C., Malomed, B., Chow, K., An, H.: Ermakov-Ray-Reid systems in nonlinear optics. J. Phys. A Math. Theor. **43**, 455214 (2010)
51. Rogers, C., Malomed, B., An, H.: Ermakov-Ray-Reid reductions of variational approximations in nonlinear optics. Stud. Appl. Math. **129**, 389–413 (2012)

52. Rogers, C., Yip, L.P., Chow, K.W.: A resonant Davey-Stewartson capillarity model system. Soliton generation. Int. J. Nonlinear Sci. Numer. Simul. **10**, 397–405 (2009)
53. Schief, W.K., Rogers, C., Bassom, A.: Ermakov systems of arbitrary order and dimension. Structure and linearisation. J. Phys. A Math. Gen. **29**, 903–911 (1996)
54. Wagner, W.G., Haus, H.A., Marburger, J.H.: Large scale self-trapping of optical beams in the paraxial ray approximation. Phys. Rev. **175**, 256–266 (1968)
55. Wai, P.K.A., Menyuk, C.R., Chen, H.H.: Stability of solitons in randomly varying birefringent fibers. Opt. Lett. **16**, 1231–1233 (1991)
56. Zhang, J.F., Li, Y.S., Meng, J., Wo, L., Malomed, B.A.: Matter-wave solitons and finite amplitude Bloch waves in optical lattices with a spatially modulated linearity. Phys. Rev. A **82**, 033614 (2010)

An Upper Bound of the Minimal Dispersion via Delta Covers

Daniel Rudolf

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract For a point set of n elements in the d-dimensional unit cube and a class of test sets we are interested in the largest volume of a test set which does not contain any point. For all natural numbers n, d and under the assumption of the existence of a δ-cover with cardinality $|\Gamma_\delta|$ we prove that there is a point set, such that the largest volume of such a test set without any point is bounded above by $\frac{\log|\Gamma_\delta|}{n} + \delta$. For axis-parallel boxes on the unit cube this leads to a volume of at most $\frac{4d}{n}\log(\frac{9n}{d})$ and on the torus to $\frac{4d}{n}\log(2n)$.

1 Introduction and Main Results

For a point set P of n elements in the unit cube $[0, 1]^d$ and for a set $\mathscr{B}$ of measurable subsets of $[0, 1]^d$ the quantity of interest is the *dispersion*, given by

$$\operatorname{disp}(P, \mathscr{B}) := \sup_{P\cap B=\emptyset,\, B\in\mathscr{B}} \lambda_d(B). \tag{1}$$

Here λ_d denotes the d-dimensional Lebesgue measure and $\mathscr{B}$ is called set of test sets. The dispersion measures the size of the largest hole which does not contain any point of P. The shape of the hole is specified by the set of test sets. We are interested in point sets with best possible upper bounds of the dispersion, which thus allow only small holes without any point. Of course, any estimate of $\operatorname{disp}(P, \mathscr{B})$ depends on n, d and $\mathscr{B}$.

Classically, the dispersion of a point set P was introduced by Hlawka [12] as the radius of the largest ball, with respect to some metric, which does not contain any

D. Rudolf (✉)
Institut für Mathematische Stochastik, University of Goettingen, Göttingen, Germany
e-mail: daniel.rudolf@uni-goettingen.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_50

point of P. This quantity appears in the setting of quasi-Monte Carlo methods for optimization, see [14] and [15, Chapter 6]. The notion of the dispersion from (1) was introduced by Rote and Tichy in [21] to allow more general test sets. There the focus is on the dependence of n (the cardinality of the point set) of $\operatorname{disp}(P, \mathscr{B})$. In contrast to that, we are also interested in the behavior with respect to the dimension.

There is a well known relation to the star-discrepancy, namely, the dispersion is a lower bound of this quantity. For further literature, open problems, recent developments and applications related to this topic we refer to [5, 6, 15, 16, 19].

For the test sets we focus on axis-parallel boxes. Point sets with small dispersion with respect to such axis-parallel boxes are useful for the approximation of rank-one tensors, see [2, 17]. In computational geometry, given a point configuration the problem of finding the largest empty axis-parallel box is well studied. Starting with [13] for $d = 2$, there is a considerable amount of work for $d > 2$, see [7, 8] and the references therein. Given a large dataset of points, the search for empty axis-parallel boxes is motivated by the fact that such boxes may reveal natural constraints in the data and thus unknown correlations, see [9].

The *minimal dispersion*, given by

$$\operatorname{disp}_{\mathscr{B}}(n, d) := \inf_{P \subset [0,1]^d, |P| = n} \operatorname{disp}(P, \mathscr{B}),$$

quantifies the best possible behavior of the dispersion with respect to n, d and $\mathscr{B}$. Another significant quantity is the *inverse of the minimal dispersion*, that is, the minimal number of points $N_{\mathscr{B}}(d, \varepsilon)$ with minimal dispersion at most $\varepsilon \in (0, 1)$, i.e.,

$$N_{\mathscr{B}}(d, \varepsilon) = \min\{n \in \mathbb{N} \mid \operatorname{disp}_{\mathscr{B}}(n, d) \leq \varepsilon\}.$$

By virtue of a result of Blumer et al. [4, Lemma A2.1, Lemma A2.2 and Lemma A2.4] one obtains

$$\operatorname{disp}_{\mathscr{B}}(n, d) \leq \frac{2d_{\mathscr{B}}}{n} \log_2\left(\frac{6n}{d_{\mathscr{B}}}\right) \quad \text{for} \quad n \geq d_{\mathscr{B}}, \tag{2}$$

or stated differently

$$N_{\mathscr{B}}(d, \varepsilon) \leq 8 d_{\mathscr{B}} \varepsilon^{-1} \log_2(13\varepsilon^{-1}), \tag{3}$$

where $\log_2$ is the dyadic logarithm and $d_{\mathscr{B}}$ denotes the VC-dimension[1] of $\mathscr{B}$. The dependence on d is hidden in the VC-dimension $d_{\mathscr{B}}$. For example, for the set of test sets of axis-parallel boxes

$$\mathscr{B}_{\text{ex}} = \{\Pi_{k=1}^{d}[x_k, y_k) \subseteq [0, 1]^d \mid x_k < y_k,\ k = 1, \ldots, d\},$$

[1] The VC-dimension is the cardinality of the largest subset T of $[0, 1]^d$ such that the set system $\{T \cap B \mid B \in \mathscr{B}\}$ contains all subsets of T.

it is well known that $d_{\mathscr{B}_{\rm ex}} = 2d$. However, the concept of VC-dimension is not as easy to grasp as it might seem on the first glance and it is also not trivial to prove upper bounds on $d_{\mathscr{B}}$ depending on $\mathscr{B}$. For instance, for *periodic axis-parallel boxes*, which coincide with the interpretation of considering the torus instead of the unit cube, given by

$$\mathscr{B}_{\rm per} = \{\Pi_{k=1}^{d} I_k(\boldsymbol{x},\boldsymbol{y}) \mid \boldsymbol{x} = (x_1,\dots,x_d), \boldsymbol{y} = (y_1,\dots,y_d) \in [0,1]^d\}$$

with

$$I_k(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} (x_k, y_k) & x_k < y_k \\ [0,1] \setminus [y_k, x_k] & y_k \leq x_k, \end{cases}$$

the dependence on d in $d_{\mathscr{B}_{\rm per}}$ is not obvious. The conjecture here is that $d_{\mathscr{B}_{\rm per}}$ behaves similar as $d_{\mathscr{B}_{\rm ex}}$, i.e., linear in d, but we do not have a proof for this fact.

The aim of this paper is to prove an estimate similar to (2) based on the concept of a δ-cover of $\mathscr{B}$. For a discussion about δ-covers, bracketing numbers and VC-dimension we refer to [10]. Let $\mathscr{B}$ be a set of measurable subsets of $[0,1]^d$. A *δ-cover for $\mathscr{B}$* with $\delta > 0$ is a finite set $\Gamma_\delta \subseteq \mathscr{B}$ which satisfies

$$\forall B \in \mathscr{B} \quad \exists L_B, U_B \in \Gamma_\delta \quad \text{with} \quad L_B \subseteq B \subseteq U_B$$

such that $\lambda_d(U_B \setminus L_B) \leq \delta$. The main abstract theorem is as follows.

Theorem 1 *For a set of test sets $\mathscr{B}$ assume that for $\delta > 0$ the set Γ_δ is a δ-cover of $\mathscr{B}$. Then*

$$\mathrm{disp}_{\mathscr{B}}(n,d) \leq \frac{\log|\Gamma_\delta|}{n} + \delta. \tag{4}$$

The cardinality of the δ-cover plays a crucial role in the upper bound of the minimal dispersion. Thus, to apply the theorem to concrete sets of test sets one has to construct suitable, not too large, δ-covers.

For $\mathscr{B}_{\rm ex}$ the best results on δ-covers we know are due to Gnewuch, see [10]. As a consequence of the theorem and a combination of [10, Formula (1), Theorem 1.15, Lemma 1.18] one obtains

Corollary 1 *For $\mathscr{B}_{\rm ex}$ and $n > 2d$ we have*

$$\mathrm{disp}_{\mathscr{B}_{\rm ex}}(n,d) \leq \frac{4d}{n} \log\left(\frac{9n}{d}\right). \tag{5}$$

(For $n \leq 2d$ the trivial estimate $\mathrm{disp}_{\mathscr{B}_{\rm ex}}(n,d) \leq 1$ *applies.) In particular,*

$$N_{\mathscr{B}_{\rm ex}}(\varepsilon,d) \leq 8d\varepsilon^{-1}\log(33\varepsilon^{-1}). \tag{6}$$

Obviously, this is essentially the same as the estimates (2) and (3) in the setting of $\mathscr{B}_{\text{ex}}$. Let us discuss how those estimates fit into the literature. From [1, Theorem 1 and (4)] we know that

$$\frac{\log_2 d}{4(n+\log_2 d)} \leq \operatorname{disp}_{\mathscr{B}_{\text{ex}}}(n,d) \leq \frac{1}{n} \min\left\{2^{7d+1}, 2^{d-1}\Pi_{i=1}^{d-1} p_i\right\}, \tag{7}$$

where p_i denotes the ith prime. The upper bound $2^{7d+1}/n$ is due to Larcher based on suitable (t,m,d)-nets and for $d \geq 54$ improves the super-exponential estimate $2^{d-1}\Pi_{i=1}^{d-1} p_i/n$ of Rote and Tichy [21, Proposition 3.1] based on the Halton sequence. The order of convergence with respect to n is optimal, but the dependence on d in the upper bound is exponential. In the estimate of Corollary 1 the optimal order in n is not achieved, but the dependence on d is much better. Already for $d = 5$ it is required that n must be larger than $5 \cdot 10^{72}$ to obtain a smaller upper bound from (7) than from (5). By rewriting the result of Larcher in terms of $N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d)$ the dependence on d can be very well illustrated, one obtains

$$N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d) \leq 2^{7d+1}\varepsilon^{-1}.$$

Here, for fixed ε there is an exponential dependence on d, whereas in the estimate of (6) there is a linear dependence on d. Summarizing, according to $N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d)$ the result of Corollary 1 reduces the gap with respect to d, we obtain[2]

$$(1/4-\varepsilon)\varepsilon^{-1}\log_2 d \leq N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d) \leq 8d\varepsilon^{-1}\log(33\varepsilon^{-1}).$$

As already mentioned for $\mathscr{B}_{\text{per}}$ the estimates (2) and (3) are not applicable, since we do not know the VC-dimension. We construct a δ-cover in Lemma 2 below and obtain the following estimate as a consequence of the theorem. Note that, since $\mathscr{B}_{\text{ex}} \subset \mathscr{B}_{\text{per}}$, we cannot expect something better than in Corollary 1.

Corollary 2 *For $\mathscr{B}_{\text{per}}$ and $n \geq 2$ we have*

$$\operatorname{disp}_{\mathscr{B}_{\text{per}}}(n,d) \leq \frac{4d}{n}\log(2n). \tag{8}$$

[2] After acceptance of the current paper a new upper bound of $N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d)$ was proven in [22]. From [22] one obtains for $\varepsilon \in (0,1/4)$ that

$$N_{\mathscr{B}_{\text{ex}}}(\varepsilon,d) \leq c_\varepsilon \log_2 d$$

with $c_\varepsilon = \varepsilon^{-(\varepsilon^{-2}+2)}(4\log\varepsilon^{-1}+1)$ for $\varepsilon^{-1} \in \mathbb{N}$. In particular, it shows that the lower bound cannot be improved with respect to the dimension. Note that the dependence on ε^{-1} is not as good as in (6).

In particular,

$$N_{\mathscr{B}_{\text{per}}}(\varepsilon, d) \le 8d\varepsilon^{-1}[\log(8d) + \log \varepsilon^{-1}]. \tag{9}$$

Indeed, the estimates of Corollary 2 are not as good as the estimates of Corollary 1. By adding the result of Ullrich [23, Theorem 1] one obtains

$$\min\{1, d/n\} \le \operatorname{disp}_{\mathscr{B}_{\text{per}}}(n, d) \le \frac{4d}{n} \log(2n),$$

or stated differently,

$$d\varepsilon^{-1} \le N_{\mathscr{B}_{\text{per}}}(\varepsilon, d) \le 8d\varepsilon^{-1}[\log(8d) + \log \varepsilon^{-1}]. \tag{10}$$

In particular, (10) illustrates the dependence on the dimension, namely, for fixed $\varepsilon \in (0, 1)$ Corollary 2 gives, except of a $\log d$ term, the right dependence on d.

In the rest of the paper we prove the stated results and provide a conclusion.

2 Auxiliary Results, Proofs and Remarks

For the proof of Theorem 1 we need the following lemma.

Lemma 1 *For $\delta > 0$ let Γ_δ be a δ-cover of $\mathscr{B}$. Then, for any point set $P \subset [0, 1]^d$ with n elements we have*

$$\operatorname{disp}(P, \mathscr{B}) \le \delta + \max_{A \cap P = \emptyset,\, A \in \Gamma_\delta} \lambda_d(A).$$

Proof Let $B \in \mathscr{B}$ with $B \cap P = \emptyset$. Then, there are $L_B, U_B \in \Gamma_\delta$ with $L_B \subseteq B \subseteq U_B$ such that

$$\lambda_d(B \setminus L_B) \le \lambda_d(U_B \setminus L_B) \le \delta.$$

In particular, $L_B \cap P = \emptyset$ and

$$\operatorname{disp}(P, \mathscr{B}) \le \sup_{P \cap B = \emptyset,\, B \in \mathscr{B}} (\lambda_d(U_B \setminus L_B) + \lambda_d(L_B)) \le \delta + \max_{A \cap P = \emptyset,\, A \in \Gamma_\delta} \lambda_d(A).$$

□

Remark 1 In the proof we actually only used that there is a set $L_B \subseteq B$ with $\lambda_d(B \setminus L_B) \le \delta$. Thus, instead of considering δ-covers it would be enough to work with set systems which approximate B from below up to δ.

By probabilistic arguments similar to those of [3, Section 8.1] we prove the main theorem. As in [11, Theorem 1 and Theorem 3] for the star-discrepancy, it also turns out that such arguments are useful for studying the dependence on the dimension of the dispersion.

Proof of Theorem 1 By Lemma 1 it is enough to show that there is a point set P which satisfies

$$\max_{A\cap P=\emptyset,\, A\in\Gamma_\delta} \lambda_d(A) \leq \frac{\log|\Gamma_\delta|}{n}. \tag{11}$$

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(X_i)_{1\leq i\leq n}$ be an iid sequence of uniformly distributed random variables mapping from $(\Omega, \mathscr{F}, \mathbb{P})$ into $[0,1]^d$. We consider the sequence of random variables as "point set" and prove that with high probability the desired property (11) is satisfied. For $(c_n)_{n\in\mathbb{N}} \subset (0,1)$ we have

$$\begin{aligned}
\mathbb{P}\Big(\max_{A\in\Gamma_\delta,\, A\cap\{X_1,\ldots,X_n\}=\emptyset} \lambda_d(A) \leq c_n\Big) &= \mathbb{P}\Big(\bigcap_{A\in\Gamma_\delta} \{\mathbf{1}_{A\cap\{X_1,\ldots,X_n\}=\emptyset}\cdot\lambda_d(A)\leq c_n\}\Big)\\
&= 1-\mathbb{P}\Big(\bigcup_{A\in\Gamma_\delta} \{\mathbf{1}_{A\cap\{X_1,\ldots,X_n\}=\emptyset}\cdot\lambda_d(A) > c_n\}\Big)\\
&\geq 1-\sum_{A\in\Gamma_\delta}\mathbb{P}\Big(\mathbf{1}_{A\cap\{X_1,\ldots,X_n\}=\emptyset}\cdot\lambda_d(A) > c_n\Big)\\
&> 1-|\Gamma_\delta|(1-c_n)^n.
\end{aligned}$$

By the fact that $1-|\Gamma_\delta|^{-1/n}\leq \frac{\log|\Gamma_\delta|}{n}$ and by choosing $c_n = \frac{\log|\Gamma_\delta|}{n}$ we obtain

$$\mathbb{P}\Big(\max_{A\in\Gamma_\delta,\, A\cap\{X_1,\ldots,X_n\}=\emptyset} \lambda_d(A) \leq \frac{\log|\Gamma_\delta|}{n}\Big) > 0.$$

Thus, there exists a realization of $(X_i)_{1\leq i\leq n}$, say $(\boldsymbol{x}_i)_{1\leq i\leq n} \subset [0,1]^d$, so that for $P = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ the inequality (11) is satisfied. □

Remark 2 By Lemma 1 and the same arguments as in the proof of the theorem one can see that a point set of iid uniformly distributed random variables $X_1,\ldots,X_n$ satisfies a "good dispersion bound" with high probability. In detail,

$$\begin{aligned}
\mathbb{P}\left(\operatorname{disp}(\{X_1,\ldots,X_n\},\mathscr{B})\leq 2\delta\right) &\geq \mathbb{P}\Big(\max_{A\in\Gamma_\delta,\, A\cap\{X_1,\ldots,X_n\}=\emptyset}\lambda_d(A)\leq\delta\Big)\\
&> 1-|\Gamma_\delta|(1-\delta)^n.
\end{aligned}$$

In particular, for confidence level $\alpha\in(0,1]$ and

$$n := \frac{\log(|\Gamma_\delta|\alpha^{-1})}{\delta} \geq \frac{\log(|\Gamma_\delta|\alpha^{-1})}{\log(1-\delta)^{-1}}$$

the probability that the random point set has dispersion smaller than 2δ is strictly larger than $1-\alpha$. This implies

$$N_{\mathscr{B}}(d,\varepsilon) \leq 2\varepsilon^{-1}\log|\Gamma_{\varepsilon/2}|, \tag{12}$$

where the dependence on d is hidden in $|\Gamma_{\varepsilon/2}|$.

In the spirit of [18–20] we are interested in *polynomial tractability* of the minimal dispersion, that is, $N_{\mathscr{B}}(d,\varepsilon)$ may not grow faster than polynomial in ε^{-1} and d. The following corollary is a consequence of the theorem and provides a condition on the δ-cover for such polynomial tractability.

Corollary 3 *For $\delta \in (0,1)$ and the set of test sets $\mathscr{B}$ let Γ_δ be a δ-cover satisfying*

$$\exists c_1 \geq 1 \;\&\; c_2, c_3 \geq 0 \quad s.t. \quad |\Gamma_\delta| \leq (c_1 d^{c_2}\delta^{-1})^{c_3 d}.$$

Then, for $n > c_3 d$ one has

$$\mathrm{disp}_{\mathscr{B}}(n,d) \leq \frac{c_3 d}{n}\Big[\log\Big(\frac{c_1 d^{c_2-1} n}{c_3}\Big)+1\Big].$$

Proof Set $\delta = c_3 d/n$ in (4) and the assertion follows. □

This implies the result of Corollary 1.

Proof of Corollary 1 By [10, Formula (1), Theorem 1.15, Lemma 1.18] one has

$$|\Gamma_\delta| \leq \frac{1}{2}(2\delta^{-1}+1)^{2d}\cdot\frac{(2d)^{2d}}{(d!)^2} \leq (6\mathrm{e}\delta^{-1})^{2d}.$$

Here the last inequality follows mainly by $d! > \sqrt{2\pi d}(d/\mathrm{e})^d$ and the assertion is proven by Corollary 3 with $c_1 = 6\mathrm{e}$, $c_2 = 0$, $c_3 = 2$. □

For $\mathscr{B}_{\mathrm{per}}$ we need to construct a δ-cover.

Lemma 2 *For $\mathscr{B}_{\mathrm{per}}$ with $\delta > 0$ and $m = \lceil 2d/\delta\rceil$ the set*

$$\Gamma_\delta = \big\{\Pi_{k=1}^d I_k(\boldsymbol{a},\boldsymbol{b}) \mid \boldsymbol{a},\boldsymbol{b} \in G_m\big\}$$

with

$$G_m = \{(a_1,\ldots,a_d) \in [0,1]^d \mid a_k = i/m,\ i = 0,\ldots,m;\ k = 1,\ldots,d\}$$

is a δ-cover and satisfies $|\Gamma_\delta| = (m+1)^{2d}$.

Proof For arbitrary $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$ with $\boldsymbol{x} = (x_1, \dots, x_d)$ and $\boldsymbol{y} = (y_1, \dots, y_d)$ there are

$$\boldsymbol{a} = (a_1, \dots, a_d) \in G_m, \quad \bar{\boldsymbol{a}} = (\bar{a}_1, \dots, \bar{a}_d) \in G_m$$
$$\boldsymbol{b} = (b_1, \dots, b_d) \in G_m, \quad \bar{\boldsymbol{b}} = (\bar{b}_1, \dots, \bar{b}_d) \in G_m,$$

such that

$$a_k \le x_k \le \bar{a}_k \le a_k + 1/m, \qquad b_k \le y_k \le \bar{b}_k \le b_k + 1/m.$$

Define $B(\boldsymbol{x}, \boldsymbol{y}) = \Pi_{k=1}^d I_k(\boldsymbol{x}, \boldsymbol{y})$ and note that it is enough to find $L_B, U_B \in \Gamma_\delta$ with $L_B \subseteq B(\boldsymbol{x}, \boldsymbol{y}) \subseteq U_B$ and $\lambda_d(U_B \setminus L_B) \le \delta$. For any coordinate $k \in \{1, \dots, d\}$ we distinguish four cases illustrated in Fig. 1:

1. Case: $|x_k - y_k| \le 1/m$ and $x_k < y_k$:
   Define $I_k^L = \emptyset$ and $I_k^U = (a_k, \bar{b}_k)$. (Here $I_k^L = [0,1] \setminus [0,1] = \emptyset$.)
2. Case: $|x_k - y_k| \le 1/m$ and $x_k \ge y_k$:
   Define $I_k^L = [0,1] \setminus [b_k, \bar{a}_k]$ and $I_k^U = [0,1] \setminus [a_k, a_k]$. (Here $I_k^U = [0,1] \setminus \{a_k\}$.)
3. Case: $|x_k - y_k| > 1/m$ and $x_k < y_k$:
   Define $I_k^L = (\bar{a}_k, b_k)$ and $I_k^U = (a_k, \bar{b}_k)$.
4. Case: $|x_k - y_k| > 1/m$ and $x_k \ge y_k$:
   Define $I_k^L = [0,1] \setminus [b_k, \bar{a}_k]$ and $I_k^U = [0,1] \setminus [\bar{b}_k, a_k]$.

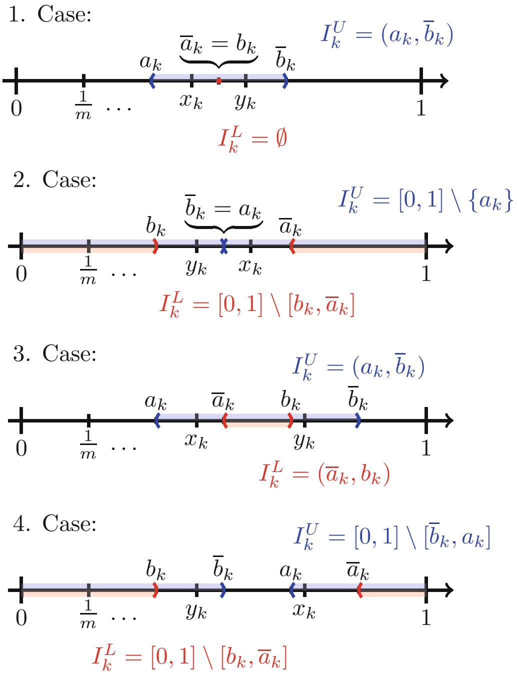


Fig. 1 The four cases from the proof of Lemma 2 to show the existence of I_k^L, I_k^U such that $I_k^L \subseteq I_k(\boldsymbol{x}, \boldsymbol{y}) \subseteq I_k^U$ and $\lambda_1(I_k^U \setminus I_k^L) \le 2/m$ are illustrated

In all cases we have $I_k^L \subseteq I_k(\boldsymbol{x}, \boldsymbol{y}) \subseteq I_k^U$ as well as $\lambda_1(I_k^U \setminus I_k^L) \leq 2/m$. For $L_B = \Pi_{i=1}^d I_i^L \in \Gamma_\delta$ and $U_B = \Pi_{i=1}^d I_i^U \in \Gamma_\delta$ the inclusion property with respect to $B(x, y)$ does hold and

$$\begin{aligned}\lambda_d(U_B \setminus L_B) &= \Pi_{i=1}^d \lambda_1(I_i^U) - \Pi_{i=1}^d \lambda_1(I_i^L) \\ &= \sum_{k=1}^d \left[\Pi_{i=1}^{k-1} \lambda_1(I_i^L)(\lambda_1(I_k^U) - \lambda_1(I_k^L))\Pi_{i=k+1}^d \lambda_1(I_i^U)\right] \leq \frac{2d}{m}.\end{aligned}$$

By the choice of m the right-hand side $2d/m$ is bounded by δ and the assertion is proven. □

Now we easily can prove an upper bound of the minimal dispersion according to $\mathscr{B}_{\text{per}}$ as formulated in Corollary 2.

Proof of Corollary 2 By the previous lemma we know that there is a δ-cover with cardinality bounded by $(4d\delta^{-1})^{2d}$. Then by Corollary 3 with $c_1 = 4$, $c_2 = 1$ and $c_3 = 2$ the proof is finished. □

3 Conclusion

Based on δ-covers we provide in the main theorem an estimate of the minimal dispersion similar to the one of (2). In the case where the VC-dimension of the set of test sets is not known, but a suitable δ-cover can be constructed our Theorem 1 leads to new results, as illustrated for $\mathscr{B}_{\text{per}}$. One might argue, that we only show existence of "good" point sets. However, Remark 2 tells us that a uniformly distributed random point set has small dispersion with high probability. As far as we know, an explicit construction of such point sets is not known.

Acknowledgements The author thanks Aicke Hinrichs, David Krieg, Erich Novak and Mario Ullrich for fruitful discussions to this topic.

References

1. Aistleitner, C., Hinrichs, A., Rudolf, D.: On the size of the largest empty box amidst a point set. Discrete Appl. Math. **230**, 146–150 (2017)
2. Bachmayr, M., Dahmen, W., DeVore, R., Grasedyck, L.: Approximation of high-dimensional rank one tensors. Constr. Approx. **39**(2), 385–395 (2014)
3. Beck, J., Chen, W.: Irregularities of Distribution. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge (1987)
4. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.: Learnability and the Vapnik-Chervonenkis dimension. J. Assoc. Comput. Mach. **36**(4), 929–965 (1989)

5. Dick, J., Pillichshammer, F.: Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge (2010)
6. Dick, J., Rudolf, D., Zhu, H.: Discrepancy bounds for uniformly ergodic Markov chain quasi-Monte Carlo. Ann. Appl. Probab. **26**, 3178–3205 (2016)
7. Dumitrescu, A., Jiang, M.: On the largest empty axis-parallel box amidst *n* points. Algorithmica **66**(2), 225–248 (2013)
8. Dumitrescu, A., Jiang, M.: Perfect vector sets, properly overlapping partitions, and largest empty box (2016, Preprint). Available at https://arxiv.org/abs/1608.06874
9. Edmonds, J., Gryz, J., Liang, D., Miller, R.: Mining for empty spaces in large data sets. Theor. Comput. Sci. **296**(3), 435–452 (2003)
10. Gnewuch, M.: Bracketing numbers for axis-parallel boxes and applications to geometric discrepancy. J. Complex. **24**, 154–172 (2008)
11. Heinrich, S., Novak, E., Wasilkowski, G., Woźniakowski, H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arith. **96**, 279–302 (2001)
12. Hlawka, E.: Abschätzung von trigonometrischen Summen mittels diophantischer Approximationen. Österreich. Akad. Wiss. Math.-Naturwiss. Kl. S.-B. II **185**, 43–50 (1976)
13. Naamad, A., Lee, D., Hsu, W.: On the maximum empty rectangle problem. Discrete Appl. Math. **8**(3), 267–277 (1984)
14. Niederreiter, H.: A quasi-Monte Carlo method for the approximate computation of the extreme values of a function. Studies in Pure Mathematics, pp. 523–529. Birkhäuser, Basel (1983)
15. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. Society for Industrial and Applied Mathematics, Philadelphia (1992)
16. Novak, E.: Some results on the complexity of numerical integration. In: Cools, R., Nuyens, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2014, pp. 161–183. Springer, Berlin (2016)
17. Novak, E., Rudolf, D.: Tractability of the approximation of high-dimensional rank one tensors. Constr. Approx. **43**(1), 1–13 (2016)
18. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Vol. 1: Linear information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society (EMS), Zürich (2008)
19. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Vol. 2: Standard information for functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
20. Novak, E., Woźniakowski, H.: Tractability of multivariate problems. Vol. 3: Standard information for operators. EMS Tracts in Mathematics, vol. 18. European Mathematical Society (EMS), Zürich (2012)
21. Rote, G., Tichy, R.: Quasi-monte carlo methods and the dispersion of point sequences. Math. Comput. **23**(8–9), 9–23 (1996)
22. Sosnovec, J.: A note on minimal dispersion of point sets in the unit cube (2017, Preprint). Available at https://arxiv.org/abs/1707.08794
23. Ullrich, M.: A lower bound for the dispersion on the torus. Math. Comput. Simul. **143**, 186–190 (2018)

A Local Inverse Formula and a Factorization

Gilbert Strang and Shev MacNamara

With congratulations to Ian Sloan!

Abstract When a matrix has a banded inverse there is a remarkable formula that quickly computes that inverse, using only local information in the original matrix. This local inverse formula holds more generally, for matrices with sparsity patterns that are examples of chordal graphs or perfect eliminators. The formula has a long history going back at least as far as the completion problem for covariance matrices with missing data. Maximum entropy estimates, log-determinants, rank conditions, the Nullity Theorem and wavelets are all closely related, and the formula has found wide applications in machine learning and graphical models. We describe that local inverse and explain how it can be understood as a matrix factorization.

1 Introduction

Here is the key point in two sentences. If a square matrix M has a tridiagonal inverse, then M^{-1} can be determined from the tridiagonal part M_0 of the original M. The formula for M^{-1} is "local" and fast to compute—it uses only 1×1 and 2×2 submatrices (assumed invertible) along the main diagonal of M_0.

Outside of M_0, the entries of M could be initially unknown ("missing data"). They are determined by the requirement that M^{-1} is tridiagonal. That requirement maximizes the determinant of the completed matrix M: the entropy.

G. Strang (✉)
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: gs@math.mit.edu

S. MacNamara
University of Technology Sydney, Ultimo, NSW, Australia
e-mail: Shev.MacNamara@uts.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_51

This theory (developed by others) extends to all chordal matrices: the nonzero positions (i, j) in M_0 correspond to edges of a chordal graph. In applications these come primarily from one-dimensional differential and integral equations. We believe that this special possibility of a local inverse should be more widely appreciated. It suggests a fast preconditioner for more general problems.

In our first examples of this known (and surprising) formula, M^{-1} will be block tridiagonal:

$$M^{-1} = \begin{pmatrix} B_{11} & B_{12} & \mathbf{0} \\ B_{21} & B_{22} & B_{23} \\ \mathbf{0} & B_{32} & B_{33} \end{pmatrix}. \tag{1}$$

This imposes a strong condition on M itself, which we identify now. M will be written in the same block form with n square blocks along the main diagonal: $n = 3$ above. When all blocks are 1×1, the entries of M^{-1} are known to be cofactors of M divided by the determinant of M. Those zero cofactors (away from the three central diagonals) mean that M is "*semiseparable*:" all submatrices that don't cross the main diagonal have rank 1.

In other words, all the 2×2 submatrices of M (that do not cross the main diagonal) will be singular. There is a well-developed theory for these important matrices [20] that allows wider bands for M_0 and M^{-1}.

Here are equivalent conditions on M that make M^{-1} *block tridiagonal*. The key point is that the entries in M_0 determine all other entries in M. Those entries are shown explicitly in condition 2.

1. The completion from M_0 to M maximizes the determinant of M (the *entropy*).
2. The completion for $n = 3$ is given by

$$M = \begin{pmatrix} M_{11} & M_{12} & \boldsymbol{M_{12}M_{22}^{-1}M_{23}} \\ M_{21} & M_{22} & M_{23} \\ \boldsymbol{M_{32}M_{22}^{-1}M_{21}} & M_{32} & M_{33} \end{pmatrix}. \tag{2}$$

   Applying this rule recursively outward from the main diagonal, M is obtained from M_0 for any matrix size n.
3. The completed entries M_{13} and M_{31} minimize the ranks of

$$\begin{pmatrix} M_{12} & M_{13} \\ M_{22} & M_{23} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} M_{21} & M_{22} \\ M_{31} & M_{32} \end{pmatrix}.$$

This extends the "zero cofactor" condition to the block case. For ordinary tridiagonal matrices M, all 2×2 blocks (except those crossing the main diagonal) have rank 1. The Nullity Theorem [19] says that the dimensions of the null spaces of these 2×2 block matrices match the number of columns in the zero blocks in M^{-1}.

Early motivation for these matrix problems came in statistics where a covariance matrix might have missing entries because complete data was not available. For instance in finance, perhaps two assets are not traded frequently enough or not on sufficiently comparable time-scales to provide data that would lead to a sensible estimate of a covariance. An example of an incomplete covariance matrix could be (although more complicated patterns of missing entries are possible)

$$\begin{pmatrix} M_{11} & M_{12} & \textbf{?} \\ M_{21} & M_{22} & M_{23} \\ \textbf{?} & M_{32} & M_{33} \end{pmatrix}.$$

A naïve first remedy is to replace the missing entries by zeros, but that is usually not a good idea; amongst other issues that choice is not guaranteed to always result in a positive definite completion. Dempster [4] suggested completing to a covariance matrix by instead inserting zeros in the inverse matrix in positions that correspond to missing values in the original incomplete covariance matrix. For this example, that leads to a sparsity pattern for M^{-1} displayed in (1), and eventually to the completion to M displayed in (2). In general, the entries of the inverse covariance matrix (the *concentration matrix* or the *precision matrix*) can be interpreted as the *information*, so setting these entries to zero reflects the situation that in the absence of data we have no information. More than that, in the multivariate Gaussian case where a vector $x \in \mathbb{R}^d$ has probability density

$$p(x) = \frac{1}{\sqrt{(2\pi)^d}\sqrt{\det M}} \exp(-x^\top M^{-1} x/2),$$

the entropy $\int p(x) \log p(x) \mathrm{d}x$ (an integral in d-dimensions) of the distribution is maximized by maximizing the determinant. The zeros in the inverse matrix are a consequence of maximizing the determinant [8] subject to the constraint of being consistent with the initial data. That seems intuitively satisfying because maximizing an entropy corresponds in some sense to assuming as little as possible while remaining consistent with the partial data. This also leads to the maximum likelihood estimator.

Zeros in the concentration matrix M^{-1} correspond to *conditional independence* and the non-zero pattern of the concentration matrix corresponds to edges in the graph of the associated Gaussian Markov Random Field [15]. Extending these ideas to estimate covariance matrices in *high dimensions* is an important and active line of work, connecting to methods that impose sparsity on the concentration matrix via l_1-regularization while still optimizing log-determinant objective functions [7, 13].

Other references include: Gohberg et al. [5, 6], Johnson and Lundquist [8, 9], Lauritzen [11, page 145], Speed and Kiiveri [15] and Strang and Nguyen [16, 19].

2 The Local Inverse Formula

At first sight it is hard to believe that the inverse of an $n \times n$ matrix (or block matrix) can be found from "local inverses." But if M^{-1} is a tridiagonal (or block tridiagonal) matrix, that statement is true. The only inverses you need are 1×1 and 2×2 along the main diagonal of M. The 1×1 inverses, $M_{2,2}^{-1}, \ldots, M_{n-1,n-1}^{-1}$, come from the diagonal blocks. The 2×2 inverses Z_i^{-1} come from adjacent blocks:

$$Z_i^{-1} = \begin{pmatrix} M_{i,i} & M_{i,i+1} \\ M_{i+1,i} & M_{i+1,i+1} \end{pmatrix}^{-1} .$$

In other words, we only need the tridiagonal part of M to find the tridiagonal matrix M^{-1}.

From $n-2$ inverses $M_{i,i}^{-1}$ and $n-1$ inverses Z_i^{-1}, here is the **local inverse formula** (when $n = 3$):

$$M^{-1} = \begin{pmatrix} \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}^{-1} & \\ & \end{pmatrix} + \begin{pmatrix} & \\ & \begin{pmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{pmatrix}^{-1} \end{pmatrix} - \begin{pmatrix} & & \\ & M_{22}^{-1} & \\ & & \end{pmatrix} \tag{3}$$

$$= \begin{pmatrix} \begin{pmatrix} & \\ & Z_1^{-1} & \\ & & \end{pmatrix} & \\ & \end{pmatrix} + \begin{pmatrix} & \\ & \begin{pmatrix} & \\ & Z_2^{-1} \end{pmatrix} \end{pmatrix} - \begin{pmatrix} & & \\ & M_{22}^{-1} & \\ & & \end{pmatrix} .$$

We emphasize that M itself need not be tridiagonal. It rarely is. The construction of M does *start* with a tridiagonal matrix, M_0. That matrix is completed to M in such a way that M^{-1} is tridiagonal. It becomes reasonable to expect that M^{-1} depends only on the starting tridiagonal matrix M_0. But still the simplicity of the local inverse formula is unexpected and attractive.

This local formula for M^{-1} can be established in several ways. Direct matrix multiplication will certainly succeed. Johnson and Lundquist [9] show how this 3×3 block case extends by iteration to larger matrices (with wider bands or general chordal structures, described next). The present paper looks at the triangular *LDU* factorization—which produces banded or chordal factors. And we view the matrix algebra in the $A^{\top}CA$ framework that is fundamental to applied mathematics.

The generalisation to matrices M^{-1} with five non-zero block diagonals is straightforward. Thus M_0 is pentadiagonal, and its extension to M is determined so that M^{-1} is also pentadiagonal. Then the local inverse formula goes directly from M_0 to M^{-1}, bypassing the completed matrix M. The formula involves the 2×2 inverses Z_i^{-1} together with the 3×3 inverses Y_i^{-1}. The submatrices Y_i come from three adjacent rows and columns $(i, i+1, i+2)$ of M_0 and M. The local inverse formula assembles the inverse as before (displayed for $n = 4$):

$$M^{-1} = \begin{pmatrix} \begin{pmatrix} Y_1^{-1} \end{pmatrix} & \\ & \end{pmatrix} + \begin{pmatrix} & \\ & \begin{pmatrix} Y_2^{-1} \end{pmatrix} \end{pmatrix} - \begin{pmatrix} & & \\ & (Z_2^{-1}) & \\ & & \end{pmatrix}$$

$$= \text{5-diagonal matrix.}$$

The formula extends to wider bands in M_0 in a natural way. Beyond that come 'staircase matrices' that are unions of overlapping square submatrices Y_i centered on the main diagonal. The sizes of the Y_i can vary and the overlaps (intersections) are the Z_i. The inverse formula remains correct.

The ultimate extension is to *chordal matrices* M_0 and M^{-1} [9]. Their non-zero entries produce a *chordal graph* [1, 2, 10]. Beyond that we cannot go. Two equivalent definitions of the class of chordal matrices are:

- Suppose M_0 has non-zero entries in positions $(i_0, i_1), (i_1, i_2), \ldots, (i_m, i_0)$. If $m \geq 4$ then that closed path has a "shortcut" chord from an i_J to an $i_L \neq i_{J+1}$ for which $M_0(i_J, i_L) \neq 0$.
- There are permutations P and $Q^{\top}$ of the rows and columns of M_0 so that the matrix $A = PM_0Q^{\top}$ allows "*perfect elimination with no fill-in*:"

$$A = LDU = \text{(lower triangular) (diagonal) (upper triangular)}$$

$$\text{with } L_{ij} = 0 \text{ and } U_{ij} = 0 \text{ whenever } A_{ij} = 0.$$

We may assume [14] that M_0 comes in this perfect elimination order. Then it is completed to M in such a way that M has the same elimination order as M_0.

3 Completion of *M* and Triangular Factorizations

When does a 3×3 block matrix M have a tridiagonal inverse? If the tridiagonal part of M itself is prescribed, the entries in the upper right and lower left corners are determined by the requirement that the corresponding entries in M^{-1} are zero:

$$M = \begin{pmatrix} M_{11} & M_{12} & \boldsymbol{M_{12}M_{22}^{-1}M_{23}} \\ M_{21} & M_{22} & M_{23} \\ \boldsymbol{M_{32}M_{22}^{-1}M_{21}} & M_{32} & M_{33} \end{pmatrix}. \tag{4}$$

It is this completed matrix M (also in (2)) that multiplies the matrix in (3) to give the identity matrix and verify the local inverse formula. Suppose M is block upper triangular: call it U, with unit diagonal blocks. Then the matrices and the local inverse formula become particularly simple. Here are the incomplete U_0, the completed U and the inverse U^{-1}:

$$U_0 = \begin{pmatrix} I & U_{12} & \mathbf{?} \\ 0 & I & U_{23} \\ 0 & 0 & I \end{pmatrix}$$

$$U = \begin{pmatrix} I & U_{12} & \boldsymbol{U_{12}U_{23}} \\ 0 & I & U_{23} \\ 0 & 0 & I \end{pmatrix} \tag{5}$$

$$U^{-1} = \begin{pmatrix} I & -U_{12} & \mathbf{0} \\ 0 & I & -U_{23} \\ 0 & 0 & I \end{pmatrix}.$$

The local inverse formula separates U^{-1} in three parts:

$$U^{-1} = \begin{pmatrix} I & -U_{12} & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & -U_{23} \\ 0 & 0 & I \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6}$$

We vainly hoped that this simple idea could apply to each factor of $M = LDU$ and produce factors of M^{-1}. That idea was destined to fail—the correct factors mix upper with lower (just as elimination does). Still it would be attractive to understand the general chordal case through its triangular factors. The key property of "no fill-in" distinguishes chordal matrices in such a beautiful way.

Example

Consider the 3 × 3 matrix

$$M_0 \equiv \frac{1}{4}\begin{pmatrix} 3 & 2 & \mathbf{?} \\ 2 & 4 & 2 \\ \mathbf{?} & 2 & 3 \end{pmatrix}$$

that is completed to

$$M = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix},$$

with inverse

$$M^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

In this symmetric example, $U = L^\top$ and $M = LDU = LDL^\top$ where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{3} & \frac{1}{2} & 1 \end{pmatrix} \qquad \text{and} \qquad D = \frac{1}{12}\begin{pmatrix} 9 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

Notice these examples of L and of U satisfy the formats displayed in (5) and in (6). Our example illustrates these formats in the scalar case but those formats remain true in the block matrix case.

4 The $A^\top CA$ Framework: A Matrix Factorization Restatement

Applied mathematics is a broad subject far too diverse to be summarized by merely one equation. Nevertheless, $A^\top CA$ offers a matrix framework to understand a great many of the classical topics, including: least squares and projections, positive definite matrices and the Singular Value Decomposition, Laplace's equation with $A^\top = -\text{div}$ and the Laplacian as $-\nabla^2 = -\text{div(grad)} = A^\top A$, networks and graph Laplacians [17]. The assembly process used in the finite element method also fits this framework [21]. It is therefore satisfying to place the local inverse formula in this framework.

Consider a square $n \times n$ invertible matrix M such that the inverse matrix satisfies the 'local inverse formula.' We will express the local inverse formula as the following factorization of the inverse matrix

$$M^{-1} = A^\top C^{-1} A \tag{7}$$

and factorization of the original matrix

$$M = G^\top CG. \tag{8}$$

Such factorizations are often represented by commutative diagrams. Here we represent $x = M^{-1}b$ as

$$\begin{tikzcd}
\mathbb{R}^m \arrow[d, "A^\top"'] & \mathbb{R}^m \arrow[l, "C^{-1}"'] \\
\mathbb{R}^n & \mathbb{R}^n \arrow[l, "M^{-1}"] \arrow[u, "A"']
\end{tikzcd}$$

and we reverse the directions of all four arrows to represent $Mx = b$ as

$$\begin{tikzcd}
\mathbb{R}^m \arrow[r, "C"] & \mathbb{R}^m \arrow[d, "G^\top"] \\
\mathbb{R}^n \arrow[u, "G"] \arrow[r, "M"'] & \mathbb{R}^n
\end{tikzcd}$$

Notice that in this approach we *start with the inverse matrix* M^{-1}, and then we *invert the inverse* to arrive at the original matrix: $(M^{-1})^{-1} = M$. To describe the factorizations we must identify the matrices C, A and G, but first we introduce notation.

Our setting is that the non-zero sparsity pattern of M^{-1} is a chordal graph on n nodes, with a clique tree (sometimes called a *junction tree*) on c_b nodes that represent the c_b *maximal cliques* (square submatrices of M_0 with no missing entries). There are c_b ‘blocks’ and c_o ‘overlaps’ in the corresponding local inverse formula. Let $c = c_b + c_o$ be the sum of these counts. Denote these block matrices by C_k for $k = 1, \ldots, c$. Order these matrices so that all the c_b blocks that correspond to maximal cliques come first, and all the $c - c_b$ blocks that correspond to overlaps come last. Let d_k denote the size of clique k so that C_k is a $d_k \times d_k$ matrix, and let

$$m = \sum_{k=1}^{c} d_k = \sum_{k=1}^{c_b} d_k + \sum_{k=c_b+1}^{c} d_k.$$

Note that $m > n$.

Define the $m \times m$ block diagonal matrix

$$C \equiv \begin{pmatrix} C_1 & & & & & & \\ & C_2 & & & & & \\ & & \ddots & & & & \\ & & & C_{c_b} & & & \\ & & & & -C_{c_b+1} & & \\ & & & & & \ddots & \\ & & & & & & -C_c \end{pmatrix}.$$

The minus signs in front of the blocks in the bottom right corner correspond to overlaps.

Because each block C_k corresponds to a subset of nodes in the original graph, each row $(i = 1, \ldots, m)$ of C corresponds to a node $(j = 1, \ldots, n)$ in the original graph. Define the $m \times n$ matrix A of $0s$ and $1s$ to encode this correspondence:

$$A_{i,j} \equiv \begin{cases} 1 & \text{if node j corresponds to row i of } C \\ 0 & \text{otherwise.} \end{cases}$$

Note that each row of A contains precisely one non-zero entry and that entry is 1. The total number of non-zero entries in A is m. Each column of A contains one or more $1s$.

It is a necessary condition for the local inverse formula to apply that all of the blocks C_k be separately invertible. Then C is an invertible matrix, and C^{-1} is the block diagonal matrix with blocks $C_1^{-1}, \ldots, C_c^{-1}$. With these definitions, the factorization $M^{-1} = A^\top C^{-1} A$ in (7) is simply matrix notation for the local inverse formula: M^{-1} is "the sum of the inverses of the blocks, minus the inverses of the overlaps."

It remains to describe the factorization $M = G^\top C G$ in (8). Intuitively, this is arrived at by reversing the directions of the arrows in the commutative diagram for M^{-1}. It is easy to see that replacing C^{-1} by C will reverse the direction of the arrow at the top of the diagram in a way that correctly inverts the action of C^{-1}.

It is not so easy to see that we can find matrices $G^\top$ and G such that the directions of the arrows corresponding to A and to $A^\top$ are reversed with the desired effect. Indeed, at first glance that seems to be tantamount to finding the 'inverse' of the A matrix, but that is impossible because $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is not a square matrix. However, there is redundancy in the action of A. Although A maps from a smaller n-dimensional space to a larger m-dimensional space, the matrix only has n columns, so the column space reached by A is only an n-dimensional subspace of $\mathbb{R}^m$. (Columns of A are independent because each row contains precisely one 1.) This makes it possible to choose $G^\top$ so that we only 'invert' on the subspace that we need to. A possible choice is the pseudoinverse of A, i.e.

$$F \equiv (A^\top A)^{-1} A^\top.$$

Note that $FA = I_n$ is the $n \times n$ identity matrix (but $AF \neq I_m$). So F is a left inverse for A. Instead of F, we could choose another left inverse of A, namely $G^\top$, where G is the matrix described next.

The last step is to now find the matrix G that will 'undo' the effect of $A^\top$. Note that in our factorization, the matrix $A^\top$ acts only on the range of $C^{-1}A$ (and not on all of $\mathbb{R}^m$). In other words, in our factorization, it is the column space of $C^{-1}A$ that is the 'input space' to $A^\top$. So we only need to invert on that subspace, by

$$G \equiv C^{-1} A M.$$

This choice makes it clear that G has two desirable properties:

- the columns of G are linear combinations of the columns of $C^{-1}A$, so the range of G is in the n-dimensional subspace of $\mathbb{R}^m$ that is reached by $C^{-1}A$, and
- $G^\top C G = G^\top C(C^{-1}A)M = (G^\top A)M = (I)M = M.$

The second property, that $L_A C G = M$, is not unique to our choice of G—it holds for any matrix L_A that is a left inverse of A. (Then we have $L_A C G = L_A C(C^{-1}A)M = (L_A A)M = (I)M = M$.) We have already seen that F is a left inverse of A so F is a possible choice for a factorization to recover the original matrix, i.e. $FCG = M$. To see that $G^\top$ is also a left inverse of A, recall the definition $G \equiv C^{-1}AM$. By the rule for a transpose of a product, $G^\top = M^\top A^\top (C^{-1})^\top$. So

$$G^\top A = M^\top A^\top (C^{-1})^\top A = M^\top (A^\top C^{-1} A)^\top = M^\top (M^{-1})^\top = I,$$

as required.

These choices also have three more notable properties: $A^\top G = (A^\top C^{-1} A)M = M^{-1}M = I$, $FG = (A^\top A)^{-1}$, and AF is a projection matrix.

Example

We now exhibit the $A^\top CA$ factorization for the same 3×3 matrix example that we used earlier in (7) to demonstrate the LDU factorization

$$M \equiv \frac{1}{4} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix} \quad \text{with inverse} \quad M^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

The graph of non-zeros is a line of three nodes. The maximal cliques are

$$C_1 \equiv \frac{1}{4}\begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix} \quad \text{and} \quad C_2 \equiv \frac{1}{4}\begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}.$$

The overlap is

$$C_3 \equiv \frac{1}{4}\begin{pmatrix} 4 \end{pmatrix} = (1),$$

so in this example $m = 2 + 2 + 1 = 5$. The $m \times m$ block diagonal matrix C is

$$C \equiv \frac{1}{4}\begin{pmatrix} 3 & 2 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 2 & 0 \\ 0 & 0 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & -4 \end{pmatrix}.$$

The matrix that sends ‘node space’ (the three columns could correspond to the three nodes) to ‘clique space’ (rows $1, 2, 3, 4, 5$ correspond to nodes $1, 2, 2, 3, 2$) is

$$A \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Direct matrix multiplication confirms that $A^\top C^{-1} A$ does indeed give M^{-1}, as expected from the local inverse formula. In this example

$$F \equiv (A^\top A)^{-1} A^\top = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

and

$$G \equiv C^{-1} A M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \\ -\frac{1}{2} & -1 & -\frac{1}{2} \end{pmatrix}$$

and $FCG = G^{\top}CG = M$. Typically and in this example, the local inverse formula only applies in going from M to M^{-1}, and

$$M \neq A^{\top} CA = \frac{1}{4}\begin{pmatrix} 3 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 3 \end{pmatrix}.$$

5 Applications

We now showcase by example some of the especially elegant applications of the local inverse formula.

Example: A Toeplitz Matrix

Complete the missing entries in M_0 to arrive at a first example via

$$M_0 = \begin{pmatrix} 2 & -1 & ? & ? \\ -1 & 2 & -1 & ? \\ ? & -1 & 2 & -1 \\ ? & ? & -1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & -1 & \frac{1}{2} & -\frac{1}{4} \\ -1 & 2 & -1 & \frac{1}{2} \\ \frac{1}{2} & -1 & 2 & -1 \\ -\frac{1}{4} & \frac{1}{2} & -1 & 2 \end{pmatrix} = M$$

so that the completed matrix has an inverse with zeros in the locations where entries were missing in the original matrix:

$$M^{-1} = \frac{1}{6}\begin{pmatrix} 4 & 2 & 0 & 0 \\ 2 & 5 & 2 & 0 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix} \tag{9}$$

$$= \frac{1}{6}\begin{pmatrix} 4 & 2 & & \\ 2 & 4 & & \\ & & & \\ & & & \end{pmatrix} + \frac{1}{6}\begin{pmatrix} & & & \\ & 4 & 2 & \\ & 2 & 4 & \\ & & & \end{pmatrix} + \frac{1}{6}\begin{pmatrix} & & & \\ & & & \\ & & 4 & 2 \\ & & 2 & 4 \end{pmatrix}$$

$$- \begin{pmatrix} & & & \\ & \frac{1}{2} & & \\ & & & \\ & & & \end{pmatrix} - \begin{pmatrix} & & & \\ & & & \\ & & \frac{1}{2} & \\ & & & \end{pmatrix}. \tag{10}$$

The local inverse formula assembles M^{-1} in (10) from the inverses of the three repeating blocks in M

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}^{-1} = \frac{1}{6}\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

and subtracting the inverses, $(2)^{-1} = 1/2$, of the two overlaps.

To appreciate the significance of those zeros in M^{-1} in (9), it helps to recall that the derivative of the determinant with respect to the entries of the matrix is given by a cofactor up to scaling by the determinant (this result comes quickly from the cofactor expansion of the determinant along one row of the matrix, for example). This leads to an especially simple form of derivative of the *log-determinant*, which in the symmetric case is simply the corresponding entry of the inverse matrix:

$$\frac{\partial}{\partial a_{ij}} \log \det M = (M^{-1})_{ij}.$$

The zeros in the inverse matrix, such as appear in (9), correspond to setting derivatives to zero, which corresponds to a local optimum. The log-determinant is convex on the cone of symmetric positive definite matrices so a local optima is also a global maximum in this case.

This first example suggests a second example, by generalizing to a doubly infinite Toeplitz matrix [5]. A *Toeplitz matrix* is constant along diagonals: the (i,j) entry is a function of $(i-j)$, so specifying one row of the matrix completely specifies all entries of the matrix. In the doubly infinite Toeplitz case, the entries of a row are the Fourier series of an associated function s known as the *symbol* of the matrix. The matrix completion problem becomes a problem of Fourier series for functions. We must complete the missing Fourier coefficients for a function s so that the Fourier series of the reciprocal function $1/s$ has zero coefficients corresponding to missing entries in the Fourier series of s. For example,

$$\begin{pmatrix} & \ddots & \\ \cdots ?\ ?\ {-1}\ 2\ {-1}\ ?\ ? \cdots \\ & & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} & \ddots & \\ \cdots 0\ 0\ \ ?\ \ ?\ \ ?\ \ 0\ 0 \cdots \\ & & \ddots \end{pmatrix}$$

is completed to

$$\begin{pmatrix} & \ddots & \\ \cdots -\tfrac{1}{4}\ \tfrac{1}{2}\ {-1}\ 2\ {-1}\ \tfrac{1}{2}\ -\tfrac{1}{4} \cdots \\ & & \ddots \end{pmatrix}^{-1} = \frac{1}{6}\begin{pmatrix} & \ddots & \\ \cdots 0\ 0\ \ 2\ \ 5\ \ 2\ \ 0\ 0 \cdots \\ & & \ddots \end{pmatrix}.$$

The general principle is to complete the symbol

$$s(x) = \sum_{-\infty}^{\infty} a_k e^{ikx} \qquad \text{with inverse} \qquad \frac{1}{s(x)} = \sum_{-\infty}^{\infty} b_k e^{ikx}$$

so that $\mathbf{b_k = 0}$ **when** $\mathbf{a_k}$ **was not specified.** This maximizes the log-determinant

$$\int_0^{2\pi} \log \sum_{-\infty}^{\infty} a_k e^{ikx} \mathrm{d}x$$

amongst symmetric positive definite Toeplitz matrices.

Banded Matrices with Banded Inverse

In very exceptional cases [16] a banded matrix can have a banded inverse. Then the local inverse formula applies in 'both directions' (leading to a class of 'chordal matrices with chordal inverse'). This will give a (new?) algorithm for the analysis and synthesis steps in a discrete wavelet transform (known as a filter bank) [3, 12, 18]. Here is an example of one of the celebrated Daubechies wavelets in this framework.

Example: A Daubechies Wavelet

Set

$$s = \sqrt{3}, \quad B_1 = \begin{pmatrix} 1+s & 3+s \\ -1+s & 3-s \end{pmatrix}, \text{ and } \quad B_2 = \begin{pmatrix} 3-s & 1-s \\ -3-s & 1+s \end{pmatrix}.$$

Notice B_1 and B_2 are singular. Set

$$t' = \begin{pmatrix} -(3+s) & 1+s & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad t = \sqrt{32}\frac{t'}{||t'||_2}$$

and

$$b' = \big(0\ 0\ 0\ 0\ (1+s)\ 3+s\big) \quad \text{and} \quad b = \sqrt{32}\frac{b'}{||b'||_2}.$$

With these definitions a matrix corresponding to a Daubechies *D4* wavelet is

$$M = \frac{1}{\sqrt{32}}\begin{pmatrix} & t & \\ B_1 & B_2 & \mathbf{0} \\ \mathbf{0} & B_1 & B_2 \\ & b & \end{pmatrix}$$

$$= \begin{pmatrix} -0.8660 & 0.5000 & 0 & 0 & 0 & 0 \\ 0.4830 & 0.8365 & 0.2241 & -0.1294 & 0 & 0 \\ 0.1294 & 0.2241 & -0.8365 & 0.4830 & 0 & 0 \\ 0 & 0 & 0.4830 & 0.8365 & 0.2241 & -0.1294 \\ 0 & 0 & 0.1294 & 0.2241 & -0.8365 & 0.4830 \\ 0 & 0 & 0 & 0 & 0.5000 & 0.8660 \end{pmatrix}.$$

Then, as desired for a wavelet basis, M is orthogonal so $M^{-1} = M^{\top}$ is also banded. (There are also important non-orthogonal wavelets with banded M and M^{-1}.)

Early motivation for the local inverse formula came from problems with covariance matrices, which are symmetric positive definite. But the local inverse formula can also apply to matrices that are not symmetric positive definite, as in this Daubechies wavelet matrix example.

More interestingly in the context of our present article, in this example, *the local inverse formula applies in both directions*. We have

$$A^{\top}C^{-1}A = M^{-1} \tag{11}$$

(this is the local inverse formula that we have come to expect when M^{-1} is chordal) *and*

$$M = A^{\top}CA$$

(this is not a local inverse formula, and it happens only in the special case that the nonzero pattern of M is subordinate to the same chordal graph associated with M^{-1}).

The matrix A is

$$A = \begin{pmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0 \end{pmatrix}.$$

The matrix C is block diagonal with blocks, in this order,

$$C_1 = \begin{pmatrix} -0.8660 & 0.5000 & 0 \\ 0.4830 & 0.8365 & 0.2241 \\ 0.1294 & 0.2241 & -0.8365 \end{pmatrix},\ C_2 = \begin{pmatrix} 0.8365 & 0.2241 & -0.1294 \\ 0.2241 & -0.8365 & 0.4830 \\ 0 & 0.4830 & 0.8365 \end{pmatrix},$$

$$C_3 = \begin{pmatrix} -0.8365 & 0.4830 & 0 \\ 0.4830 & 0.8365 & 0.2241 \\ 0.1294 & 0.2241 & -0.8365 \end{pmatrix},\ C_4 = \begin{pmatrix} 0.8365 & 0.2241 & -0.1294 \\ 0.2241 & -0.8365 & 0.4830 \\ 0 & 0.5000 & 0.8660 \end{pmatrix},$$

$$-C_5 = \begin{pmatrix} -0.8365 & -0.2241 \\ -0.2241 & 0.8365 \end{pmatrix},\ -C_6 = \begin{pmatrix} 0.8365 & -0.4830 \\ -0.4830 & -0.8365 \end{pmatrix},$$

$$-C_7 = \begin{pmatrix} -0.8365 & -0.2241 \\ -0.2241 & 0.8365 \end{pmatrix}.$$

For the special class of matrices for which the local inverse formula applies in both directions, and analogous to the way a block diagonal C is defined from that part of M corresponding to the chordal graph of M_0, we could also define a block diagonal matrix D from that part of M^{-1} corresponding to the same chordal graph. Then

$$A^{\top}D^{-1}A = M \tag{12}$$

(compared to (11), here (12) is the local inverse formula in the *opposite* direction, by assembling M from inverses of blocks and overlaps in M^{-1}) *and*

$$M^{-1} = A^{\top}DA.$$

In this example D is the same as $C^{\top}$, but there are other examples for which the local inverse formula applies in both directions where $D \neq C^{\top}$.

Acknowledgements The authors gratefully acknowledge a grant from The Mathworks that made this work possible.

References

1. Bartlett, P.: Undirected graphical models: chordal graphs, decomposable graphs, junction trees, and factorizations (2009). https://people.eecs.berkeley.edu/~bartlett/courses/2009fall-cs281a/
2. Blair, J.R.S., Peyton, B.: An Introduction to Chordal Graphs and Clique Trees. In: Graph Theory and Sparse Matrix Computation. The IMA Volumes in Mathematics and Its Applications, vol. 56, pp. 1–29. Springer, New York (1993)
3. Daubechies, I.: Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia (1992)
4. Dempster, A.P.: Covariance selection. Biometrics **28**, 157–175 (1972)
5. Dym, H., Gohberg, I.: Extensions of band matrices with band inverses. Linear Algebra Appl. **36**, 1–24 (1981)
6. Eidelman, Y., Gohberg, I., Haimovici, I.: Separable Type Representations of Matrices and Fast Algorithms, vol. 1. Springer, Basel (2013)
7. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
8. Johnson, C.R.: Matrix Completion Problems: A Survey. In: Johnson, C.R. (ed.) Matrix Theory and Applications, pp. 69–87. American Mathematical Society, Providence (1989)
9. Johnson, C.R., Lundquist, M.: Local inversion of matrices with sparse inverses. Linear Algebra Appl. **277**, 33–39 (1998)
10. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
11. Lauritzen, S.: Graphical Models. Oxford University Press, Oxford (1996)
12. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, Boston (1998)
13. Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., et al.: High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. Electron. J. Stat. **5**, 935–980 (2011)
14. Rose, D.: Triangulated graphs and the elimination process. J. Math. Anal. Appl. **32**, 597–609 (1970)
15. Speed, T.P., Kiiveri, H.T.: Gaussian Markov distributions over finite graphs. Ann. Stat. **14**(1), 138–150 (1986)
16. Strang, G.: Fast transforms: banded matrices with banded inverses. Proc. Natl. Acad. Sci. U. S. A. **107**(28), 12413–12416 (2010)
17. Strang, G.: Introduction to Linear Algebra. Cambridge Press, Wellesley (2016)
18. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Cambridge Press, Wellesley (1996)
19. Strang, G., Nguyen, T.: The interplay of ranks of submatrices. SIAM Rev. **46**(4), 637–646 (2004)

20. Vandebril, R., van Barel, M., Mastronardi, N.: Matrix Computations and Semiseparable Matrices, vol. 1. Johns Hopkins, Baltimore (2007)
21. Wathen, A.J.: An analysis of some Element-by-Element techniques. Comput. Methods Appl. Mech. Eng. **74**, 271–287 (1989)

Ian Sloan's Legacy in Integral Equation Methods

Thanh Tran

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In almost four decades, from the early 1970s until the first decade of this century, Ian Sloan has contributed immensely in the area of integral equation methods for elliptic boundary value problems. A search on MathSciNet with entries "Author=Sloan" and "Title=integral equation" reveals 44 papers. This review article sheds some lights on this historic path.

1 Introduction

Trained as a theoretical physicist, Ian H. Sloan published his first Physics paper in 1964 [16] and his potentially last Physics paper in 1977 [21]. Within a span of slightly more than a decade, he published nearly 50 articles in this area of research. This is no doubt a target that many researchers wish to achieve.

However, starting from 1968 Sloan has gradually moved away from Theoretical Physics and ventured into Numerical Analysis. His first stop was error analysis for numerical methods of integral equations. For four decades, he has played a leading role in the area of integral equation methods for elliptic boundary value problems which can be reformulated as elliptic or strongly elliptic pseudo-differential equations.

The aim of this article is to look into this historical path to highlight Sloan's major contributions in the development of numerical methods for integral equations. This is not a rigorous mathematical review article, namely that we shall not discuss in depth the methods and the mathematics behind them. We would rather emphasise motivations which instigate these developments and their significance.

T. Tran (✉)
School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW, Australia
e-mail: thanh.tran@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_52

2 The Early Days: Motivation from Physics

Ian Sloan's first paper in integral equations is probably a joint paper with Elvin Moore in Journal of Physics B in 1968 [26] in which the authors studied the problem of electron-hydrogen collisions. Later, together with his postdoc Sadhan Adhikari,[1] Sloan solved singular integral equations of the Lippmann-Schwinger type arising from the study of the two-body t-matrix in the three-body problem [1, 25]. The degenerate-kernel method suggested in [25] is in fact an extension of Sloan's almost concurrent joint paper (with Brian Burn and N. Datyner) for Fredholm integral equations of the second kind [33].

The authors of [33] consider the equation

$$y(t) = f(t) + \lambda \int_a^b K(t,s)y(s)\,ds, \quad a \le t \le b, \tag{1}$$

or, in the operator form,

$$y = f + \lambda \mathscr{K} y, \tag{2}$$

where y and f are real- or complex-valued functions in $L^2(a,b)$, the kernel K is square-integrable in $(a,b) \times (a,b)$, and $\mathscr{K}$ is an integral operator defined for any $v \in L^2(a,b)$ by

$$\mathscr{K} v(t) := \int_a^b K(t,s)v(s)\,ds, \quad a \le t \le b. \tag{3}$$

The integral equation arising from the three-body problem that Sloan and Adhikari considered in [1, 25], namely

$$T(\boldsymbol{p}, \boldsymbol{p}'; s) = V(\boldsymbol{p}, \boldsymbol{p}') + \int_0^\infty \frac{V(\boldsymbol{p}, \boldsymbol{p}'')T(\boldsymbol{p}'', \boldsymbol{p}'; s)}{s - \boldsymbol{p}''^2/(2\mu)}\,d\boldsymbol{p}'',$$

is of Lippmann-Schwinger type and is a special case of (1).

The article [33], though published in a physics journal, is written in a language that is familiar and easily understandable to pure mathematicians. Soon, Sloan started to publish in mathematical journals [17–20]. The next section reviews some of these early contributions.

[1] Adhikari later became a full Professor at the Institute of Theoretical Physics, UNESP São Paulo State University, Brazil.

3 A Turn of Career Path: Numerical Analysis

Being a very successful researcher in Physics, in 1970s Sloan gradually changed his research field. Within a year or so he published three papers [18–20] in top journals in numerical analysis, journals that numerical analysts at all times aim to publish in: Numerische Mathematik, Mathematics of Computation, and SIAM Journal on Numerical Analysis. These papers analyse errors in the degenerate-kernel approximation and iterated Galerkin method.

3.1 Degenerate-Kernel Methods

In order to understand the analysis in [18, 20], a recall of the degenerate-kernel method is required. In a degenerate-kernel approach, one seeks to approximate the solution y of (2) by y_N which is the solution to

$$y_N = f + \lambda \mathscr{K}_N y_N, \tag{4}$$

where $\mathscr{K}_N$ is an integral operator defined in the same manner as $\mathscr{K}$, see (3), with the kernel $K(t,s)$ replaced by a finite-rank operator

$$K_N(t,s) = \sum_{n=1}^{N} \alpha_n(t)\beta_n(s). \tag{5}$$

Here the functions α_n are assumed to be linearly independent.

The advantage of the degenerate form (5) is to reduce Eq. (4) to a system of algebraic linear equations. This can be easily seen by seeking the solution y_N in the form

$$y_N(x) = f(x) + \sum_{n=1}^{N} a_n \alpha_n(x).$$

Substituting this into (4) one finds that the coefficients a_n satisfy the linear system

$$a_n - \lambda \sum_{m=1}^{N} \gamma_{nm} a_m = \lambda f_n, \quad n = 1, \ldots, N, \tag{6}$$

where

$$f_n = (f, \beta_n) \quad \text{and} \quad \gamma_{nm} = (\alpha_m, \beta_n).$$

Here, $(\cdot,\cdot)$ is the L^2-inner product defined by

$$(u,v) = \int_a^b u(s)\overline{v(s)}\,ds, \quad u,v \in L^2(a,b).$$

This method is not new. Indeed, earlier works by Russian mathematicians [8, 13] have studied this topic. The authors of these monographs suggest to define $K_N(t,s)$ so as to construct $\mathscr{K}_N$ to be a good approximation to $\mathscr{K}$, as an operator. For example, one can define $K_N(t,s)$ as the Taylor polynomial of degree N of $K(t,s)$; see [8, 13].

Contrarily, Sloan and his co-authors in [33] try to approximate $\mathscr{K}$ in the context this operator actually occurs in the integral equation. In other words, they aim at finding a good approximation of $\mathscr{K}y$ by $\mathscr{K}_N y_N$. Thus these authors construct $\mathscr{K}_N$ in such a way that

$$\mathscr{K}_N u = \mathscr{K} u \quad \forall u \in \mathfrak{V}_N, \tag{7}$$

where $\mathfrak{V}_N$ is a finite-dimensional space (to be specified later) that contains a good approximation to y. Needless to say, if $y \in \mathfrak{V}_N$ then the approximation is exact.

For (7) to hold, Sloan et al define K_N by

$$K_N(t,s) = \sum_{m,n=1}^{N} \mathscr{K}u_n(t) D_{nm} \overline{v_m(s)}, \tag{8}$$

where $\{u_1,\ldots,u_N\}$ is a linearly independent set of functions in $L^2(a,b)$ to be chosen for their suitability as a basis set for approximating y, the set $\{v_1,\ldots,v_N\}$ is a second linearly independent set in $L^2(a,b)$ to be specified later, and where the coefficients D_{nm} are entries of a matrix whose inverse has entries defined by

$$(D^{-1})_{mn} = (u_n, v_m), \quad m,n = 1,\ldots,N. \tag{9}$$

The invertibility of D imposes a condition on the choice of the set $\{v_1,\ldots,v_N\}$. It is true that, regardless of the choice of this set, the definition (8) of $K_N(t,s)$ implies (7) with $\mathfrak{V}_N = \operatorname{span}\{u_1,\ldots,u_N\}$. Indeed, for any $j = 1,\ldots,N$ we have

$$\mathscr{K}_N u_j(t) = \sum_{n=1}^{N}\sum_{m=1}^{N} D_{nm}(u_j, v_m)\mathscr{K}u_n(t) = \sum_{n=1}^{N} \delta_{nj}\mathscr{K}u_n(t) = \mathscr{K}u_j(t),$$

where δ_{nj} is the Kronecker delta. A simple choice of v_j is $v_j = u_j, j = 1,\ldots,N$. This choice ensures the invertibility of D because the matrix with entries defined by (9) is a Gram matrix. For convergence analysis, it is assumed in [20] that $\{u_n\}$ and $\{v_n\}$ are complete sets in $L^2(a,b)$.

There is another benefit of defining $\mathscr{K}_N$ by (8), which can be seen in the error analysis of the approximation. Let P_N be the orthogonal projection into $\mathfrak{V}_N$. Then it follows from (7) that

$$\mathscr{K}_N P_N = \mathscr{K} P_N,$$

and hence that

$$(\mathscr{K} - \mathscr{K}_N)y = (\mathscr{K} - \mathscr{K}_N)(y - P_N y).$$

This implies

$$\|(\mathscr{K} - \mathscr{K}_N)y\| \le \|\mathscr{K} - \mathscr{K}_N\| \|y - P_N y\|,$$

where the norms are the usual norm in $L^2(a,b)$ and the corresponding operator norm. A conventional analysis for degenerate-kernel methods makes use of the estimate

$$\|(\mathscr{K} - \mathscr{K}_N)y\| \le \|\mathscr{K} - \mathscr{K}_N\| \|y\|.$$

The difference can be seen in the following theorem.

Theorem 1 ([20, Theorem 1]) *Let $\Pi_N : L^2(a,b) \to \mathfrak{V}_N$ be a linear operator defined by*

$$\Pi_N u = \sum_{i=1}^{N} \Big(\sum_{j=1}^{N} D_{ij}(u, v_j) \Big) u_i. \tag{10}$$

If there exists $M > 0$ such that

$$\|\Pi_N\| \le M, \tag{11}$$

then

$$\|y - y_N\| \le \beta_N \|y - P_N y\|, \tag{12}$$

where

$$\beta_N = |\lambda| \|(I - \lambda \mathscr{K}_N)^{-1}\| \|\mathscr{K} - \mathscr{K}_N\| \to 0 \quad as \quad N \to \infty.$$

It can be seen that

$$\Pi_N u = u \quad \forall u \in \mathfrak{V}_N.$$

Hence $\Pi_N^2 = \Pi_N$, i.e., Π_N is a projection onto $\mathfrak{V}_N$. However, in general Π_N is different from the orthogonal projection P_N.

More discussion for the simple choice $v_i = u_i$, $i = 1, \dots, N$, is worthwhile. The definitions (10) and (9) yield, for $n = 1, \dots, N$,

$$\begin{aligned}(\varPi_N u, u_n) &= \sum_{i=1}^{N} \Big(\sum_{j=1}^{N} D_{ij}(u, u_j) \Big)(u_i, u_n) = \sum_{j=1}^{N} \Big(\sum_{i=1}^{N} (D^{-1})_{ni} D_{ij} \Big)(u, u_j) \\ &= \sum_{j=1}^{N} \delta_{nj}(u, u_j) = (u, u_n).\end{aligned}$$

This means $\varPi_N = P_N$, so that (11) holds due to $\|\varPi_N\| = 1$. Moreover, the representation $y_N = \sum_{n=1}^{N} a_n u_n$, (4), and (7) imply

$$y_N = f + \lambda \mathscr{K} y_N = f + \lambda \sum_{n=1}^{N} a_n \mathscr{K} u_n, \qquad (13)$$

so that the coefficients a_n satisfy the system of N linear equations

$$\sum_{n=1}^{N} \big[(u_n, u_m) - \lambda (\mathscr{K} u_n, u_m) \big] a_n = (f, u_m), \quad m = 1, \dots, N. \qquad (14)$$

(In fact the same system holds true for any choice of v_i, not necessarily $v_i = u_i$.)

The method described above is very similar to the method of moment described in [8, 13], for which the solution $y_N^{\text{mom}} = \sum_{n=1}^{N} b_n u_n \in \mathfrak{V}_N$ satisfies

$$y_N^{\text{mom}} = P_N f + \lambda P_N \mathscr{K} y_N^{\text{mom}},$$

or

$$(y_N^{\text{mom}} - \lambda \mathscr{K} y_N^{\text{mom}}, u_m) = (f, u_m), \quad m = 1, \dots, N.$$

Substituting the representation of y_N^{mom} into the above system results in the same linear system (14), so that $a_n = b_n$. This and (13) imply

$$y_N = f + \lambda \mathscr{K} y_N^{\text{mom}}. \qquad (15)$$

In other words, y_N and y_N^{mom} are related by a single iteration of the integral equation or more precisely of the operator $(fI + \lambda \mathscr{K})$ with I being the identity operator.

More can be said about the errors $\| y - y_N\|$ and $\| y - y_N^{\text{mom}}\|$. Recalling that P_N is the orthogonal projection from $L^2(a, b)$ onto $\mathfrak{V}_N$ and that $y - y_N^{\text{mom}} = P_N (y - y_N^{\text{mom}}) + (I - P_N)(y - y_N^{\text{mom}})$ we have (due to orthogonality and $y_N^{\text{mom}} = P_N y_N^{\text{mom}}$)

$$\begin{aligned} \| y - y_N^{\text{mom}}\|^2 &= \| P_N(y - y_N^{\text{mom}})\|^2 + \|(I - P_N)(y - y_N^{\text{mom}})\|^2 \\ &= \| P_N(y - y_N^{\text{mom}})\|^2 + \| y - P_N y\|^2 \\ &\geq \| y - P_N y\|^2. \end{aligned}$$

This together with (12) gives

$$\| y - y_N\| \leq \beta_N \| y - y_N^{\text{mom}}\|, \tag{16}$$

which shows that for sufficiently large N, the error in the method suggested by Sloan et al. in [18, 33] is smaller than that of the method of moment suggested in [8, 13].

Numerical experiments carried out in [18, 33] underline the theoretical result proved in Theorem 1 and support the above observation.

Motivated by Sloan [18], in the same year Sloan wrote two other papers, one published in Mathematics of Computation [19] which proposed an iteration method to improve the approximation of solution of integral equations, and another in SIAM Journal on Numerical Analysis [20] which discussed the same issue for eigenvalue problems. That is the topic of the next subsection.

3.2 Iterated Galerkin

The paper [19] is concerned with the approximate solution of the equation

$$y = f + \mathscr{K} y, \tag{17}$$

where f and y belong to a separable Hilbert space H, and $\mathscr{K}$ is a compact linear operator in H. Note that the integral operator $\mathscr{K}$ in (2) is a compact linear operator in $H = L^2(a, b)$ if the kernel K is square-integrable in $(a, b) \times (a, b)$.

To generalise his observation (15) and (16), Sloan shows that from an approximation of y, either by the best approximation, or the Galerkin or collocation solution of (17), one can obtain a better approximation by use of an iteration of the form (15).

More precisely, let $\{u_i\}$ be a complete sequence in H (which exists due to the assumption that H is separable), and let $\mathfrak{V}_N = \text{span}\{u_1, \ldots, u_N\}$. Denoting by $P_{\mathfrak{V}_N}$ the orthogonal projection from H onto $\mathfrak{V}_N$, it is known that $P_{\mathfrak{V}_N} y = \text{argmin}_{z \in \mathfrak{V}_N} \| y - z\|$. It is proved in [19, Theorem 1] that if $y_N^{(1)} = P_{\mathfrak{V}_N} y$ and

$$y_N^{(2)} = f + \mathscr{K} y_N^{(1)} \tag{18}$$

then

$$\| y - y_N^{(2)} \| \leq \alpha_N \| y - y_N^{(1)} \|,$$

where $\alpha_N = \| \mathscr{K} - \mathscr{K} P_N \| \to 0$ as $N \to \infty$. This means an iteration of the form (18) yields an approximation that converges faster.

Similarly, [19, Theorem 3] shows that if $y_N^{(1)} \in \mathfrak{V}_N$ is the (Bubnov- or Petrov-) Galerkin solution to (17), and $y_N^{(2)}$ is defined by (18), then

$$\| y - y_N^{(2)} \| \leq \beta_N \| y - y_N^{(1)} \|,$$

where $\beta_N \to 0$ as $N \to \infty$. An explicit form for β_N is given in [19]. It is noted that the computation of $y_N^{(2)}$ for Galerkin solutions can be carried out without extra work, since the Galerkin methods already require the calculation of the quantities $\mathscr{K} u_i$, $i = 1, \ldots, N$. For instance, in the case of the Bubnov-Galerkin method one finds $y_N^{(1)} = \sum_{i=1}^N a_{N,i} u_i \in \mathfrak{V}_N$ by solving

$$\sum_{i=1}^{N} a_{N,i} [(u_i, u_j) - (\mathscr{K} u_i, u_j)] = (f, u_j), \quad j = 1, \ldots, N.$$

A repeated iteration of the form (18) gives an even faster convergence. Superconvergence of the Galerkin iterates for this type of equations is analysed in a joint paper with Thomée [28].

The paper [20] performs the same study for the approximation of the solution of the eigenvalue problem

$$y = \lambda \mathscr{K} y,$$

where K is a compact linear operator in a complex Banach space E. We omit the details.

All the above-mentioned equations are Fredholm integral equations of the second kind. It is in the late 1980s that Sloan switched his interest to Fredholm integral equations of the first kind. For almost a decade, he spent efforts to develop quadrature-based approaches to improve the collocation method. It can be said that with these achievements he put his stamp on the integral equation method.

4 The Pinnacle: The Qualocation Method

Sloan coined the term the *qualocation method* in his 1988 paper [22] and the nomenclature technically means a *quadrature-based generalisation of the collocation method*. Some American author [35] amusingly refers to this method as the *koalacation* method, because according to them the method was introduced and mostly developed in Australia.

In fact, in [22] the author's address was *Mathematisches Institut A, Universität Stuttgart, Federal Republic of Germany* (with his permanent address being the School of Mathematics, University of New South Wales, Australia). This indicates that the work was carried out during Sloan's sabbatical leave sometime in 1987 at Stuttgart, where Wolfgang Wendland, one of the founders of theories of boundary integral equation methods, was based. One can guess that Wendland's works on first-kind Fredholm integral equations stimulated Sloan to change his interests from Fredholm integral equations of the second kind (inherited from his Physics years) to equations of the first kind.

Roughly speaking, the qualocation method is a compromise between the Galerkin and collocation methods, which aims to achieve the benefits of the Galerkin method at the cost comparable to the collocation method.

Before starting to discuss qualocation methods, it is worth reviewing the two well-known collocation and Galerkin methods.

4.1 The Collocation and Galerkin Methods

Some notations are required. We denote the inner product for 1-periodic functions by $(u, v) := \int_0^1 u\overline{v}$. With I denoting the interval $[0, 1) = \mathbb{R}/\mathbb{Z}$ and $\phi_n : I \to \mathbb{C}$ the exponential function $\phi_n(x) = e^{\mathrm{i}2\pi nx}$ we define the Fourier coefficients of a 1-periodic function $u : I \to \mathbb{C}$ by $\hat{u}(n) = (u, \phi_n)$. For any $s \in \mathbb{R}$, the Sobolev space H^s is defined by

$$H^s := \{v : [0,1) \to \mathbb{C} \ : \ \|v\|_s^2 := |\hat{v}(0)|^2 + \sum_{n\in\mathbb{Z}} |n|^{2s} |\hat{v}(n)|^2 < \infty\}.$$

In [4], Chandler and Sloan considered the equation

$$\mathscr{L}u = f, \tag{19}$$

in which $\mathscr{L}$ is a pseudo-differential operator (ψdo) of order $\beta \in \mathbb{R}$. More precisely,

$$\mathscr{L} = \mathscr{L}_0 + \mathscr{K}, \tag{20}$$

where the principal part $\mathscr{L}_0$ is defined by

$$\mathscr{L}_0 v := \sum_{n\in\mathbb{Z}} [n]_\beta \hat{v}(n) \phi_n \quad \forall v \in H^s, \tag{21}$$

with $[n]_\beta$ defined by

$$[n]_\beta := \begin{cases} 1, & n = 0, \\ |n|^\beta, & n \neq 0, \end{cases} \tag{22}$$

or

$$[n]_\beta := \begin{cases} 1, & n = 0, \\ (\operatorname{sign} n)|n|^\beta, & n \neq 0. \end{cases} \tag{23}$$

Hence, $\mathscr{L}_0$ is an isometry from H^s to $H^{s-\beta}$ for all $s \in \mathbb{R}$. The operator $\mathscr{K}$ is required to be a mapping from H^s to H^t for all $s, t \in \mathbb{R}$. In practice, it is an integral operator with a smooth kernel, which is a compact perturbation of $\mathscr{L}_0$ and plays only a minor role in the analysis.

Examples of the ψdo $\mathscr{L}$ can be found in [15]. We present one example arising from the solution of the Dirichlet problem for the Laplacian, i.e,

$$\begin{aligned} \Delta\Phi &= 0 \qquad \text{in } \Omega, \\ \Phi &= \Phi_D \qquad \text{on } \Gamma. \end{aligned} \tag{24}$$

Here Ω is a simply connected bounded domain in $\mathbb{R}^2$ with smooth boundary Γ. It is well known that for $\boldsymbol{x} \in \Omega$, the potential $\Phi(\boldsymbol{x})$ can be represented as

$$\Phi(\boldsymbol{x}) = \frac{1}{2\pi}\int_\Gamma \frac{\partial\Phi}{\partial\boldsymbol{n}}(\boldsymbol{y}) \log\frac{\alpha}{|\boldsymbol{x}-\boldsymbol{y}|}\,d\sigma_{\boldsymbol{y}} - \frac{1}{2\pi}\int_\Gamma \Phi_D(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{n}_{\boldsymbol{y}}}\log\frac{\alpha}{|\boldsymbol{x}-\boldsymbol{y}|}\,d\sigma_{\boldsymbol{y}}, \tag{25}$$

for any positive parameter α. By passing to the limit when $\boldsymbol{x}$ approaches a point on the boundary Γ and using the jump relations (see e.g. [9]) one can prove that $U := \partial\Phi/\partial\boldsymbol{n}$ satisfies

$$\frac{1}{2\pi}\int_\Gamma U(\boldsymbol{y})\log\frac{\alpha}{|\boldsymbol{x}-\boldsymbol{y}|}\,d\sigma_{\boldsymbol{y}} = F(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma, \tag{26}$$

where

$$F(\boldsymbol{x}) = \frac{1}{2}\Phi_D(\boldsymbol{x}) + \frac{1}{2\pi}\int_\Gamma \Phi_D(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{n}_{\boldsymbol{y}}}\log\frac{\alpha}{|\boldsymbol{x}-\boldsymbol{y}|}\,d\sigma_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Gamma.$$

Hence, finding Φ reduces to finding U by solving the boundary integral Eq. (26), a Fredholm integral equation of the first kind. Introducing a parametrisation of the smooth curve Γ by a 1-periodic smooth function $\gamma : \mathbb{R} \to \Gamma$ which satisfies $|\gamma'(x)| \neq 0$, we can rewrite (26) as

$$\int_0^1 \log\frac{\alpha}{|\gamma(x)-\gamma(y)|}u(y)\,dy = f(x), \quad x \in [0,1], \tag{27}$$

where

$$u(x) = \frac{1}{2\pi} U[\gamma(x)]|\gamma'(x)| \quad \text{and} \quad f(x) = F[\gamma(x)], \quad x \in \mathbb{R}. \tag{28}$$

Since

$$\begin{aligned}\log \frac{\alpha}{|\gamma(x) - \gamma(y)|} &= \log \frac{\alpha}{|2\sin(\pi(x-y))|} + \log \left|\frac{\gamma(x) - \gamma(y)}{2\sin(\pi(x-y))}\right| \\ &= \log \alpha + \sum_{m=1}^{\infty} \frac{1}{m} \cos(2\pi m(x-y)) + \log \left|\frac{\gamma(x) - \gamma(y)}{2\sin(\pi(x-y))}\right|,\end{aligned}$$

Eq. (27) can be written in the form (19) with $\mathscr{L}_0$ defined by

$$\mathscr{L}_0 v(x) = (\log \alpha)\ \hat{v}(0) + \sum_{n\in\mathbb{Z}} \frac{\hat{v}(n)}{2|n|} e^{\mathrm{i}2\pi n x} \tag{29}$$

and $\mathscr{K}$ by

$$\mathscr{K} u(x) = \int_0^1 u(y) \log \left|\frac{\gamma(x) - \gamma(y)}{2\sin(\pi(x-y))}\right| dy. \tag{30}$$

Thus $\mathscr{L}$ is a ψdo of order $\beta = -1$. The parameter α is chosen to be greater than the logarithmic capacity of Γ so that $\mathscr{L}$ satisfies

$$(\mathscr{L} v, v) \geq c\|v\|^2_{H^{-1/2}} \quad \forall v \in H^{-1/2}.$$

For an explanation of the logarithmic capacity, see [6].

It is noted that if Γ is the unit circle then $\mathscr{K} = 0$ and $\mathscr{L} = \mathscr{L}_0$. It is also noted that the Fourier mode $v(x) = e^{\mathrm{i}2\pi k x}$ for $k \neq 0$ is an eigenfunction of $\mathscr{L}_0$ with eigenvalue $1/(2|k|)$.

We now return to the general Eq. (19). Two classical methods to solve this equation are the collocation method and the Galerkin method. Consider for simplicity a uniform partition of the (periodic) interval $[0, 1)$ by $x_k = kh$ with $h = 1/N$ being the step-size, and use the periodic labelling convention, $x_{k+N} = x_k$ for all k. We denote by $\mathfrak{S}_h^r$ the space of smoothest splines of order $r \geq 1$ on the partition $\{x_k\}$, namely, $\mathfrak{S}_h^r$ contains functions $v_h \in C^{r-2}$ which are polynomials of degree not greater than $r-1$ on each sub-interval $[x_k, x_{k+1}]$. The space $\mathfrak{S}_h^1$ contains piecewise-constant functions, whereas $\mathfrak{S}_h^2$ contains continuous piecewise-linear functions.

The standard collocation method approximates u by $u_h^{\mathrm{c}} \in \mathfrak{S}_h^r$ satisfying

$$\mathscr{L} u_h^{\mathrm{c}}(x_k^{\mathrm{c}}) = f(x_k^{\mathrm{c}}), \quad k = 0, 1, \ldots, N-1, \tag{31}$$

where

$$x_k^{\mathrm{c}} = \begin{cases} x_k & \text{if } r \text{ is even,} \\ (x_k + x_{k+1})/2 & \text{if } r \text{ is odd.} \end{cases}$$

The Galerkin method approximates u by $u_h^{\mathrm{G}} \in \mathfrak{S}_h^r$ satisfying

$$(\mathscr{L}u_h^{\mathrm{G}}, v) = (f, v) \quad \forall v \in \mathfrak{S}_h^r. \tag{32}$$

Under some conditions on r so that both the collocation and Galerkin methods are well defined, then both methods achieve the optimal error estimate

$$\|u - u_h^{\mathrm{c}}\|_s \le Ch^{t-s}\|u\|_t \quad \text{and} \quad \|u - u_h^{\mathrm{G}}\|_s \le Ch^{t-s}\|u\|_t.$$

However, for the collocation method it is required that

$$\beta \le s \le t \le r, \quad s < r - 1/2, \quad \text{and} \quad \beta + 1/2 < t,$$

whereas for the Galerkin method s and t satisfy

$$\beta - r \le s \le t \le r \quad \text{and} \quad s < r - 1/2.$$

Therefore, the highest orders of convergence for the two methods are

$$\|u - u_h^{\mathrm{c}}\|_\beta \le Ch^{r-\beta}\|u\|_r \quad \text{and} \quad \|u - u_h^{\mathrm{G}}\|_{\beta-r} \le Ch^{2r-\beta}\|u\|_r. \tag{33}$$

Note that for negative values of β, both norms on the left-hand sides of the estimates in (33) are negative norms. The importance of a higher order of convergence in a negative norm can be seen by considering again the example discussed above with the Dirichlet problem for the Laplacian in Ω, which entails the solution to the logarithmic-kernel integral Eq. (27). Recall that for this problem $\beta = -1$; see (21). With $r = 2$ so that u_h^{c} and u_h^{G} are continuous piecewise-linear functions, (33) becomes

$$\|u - u_h^{\mathrm{c}}\|_{-1} \le Ch^3\|u\|_2 \quad \text{and} \quad \|u - u_h^{\mathrm{G}}\|_{-3} \le Ch^5\|u\|_2. \tag{34}$$

On the other hand, it follows from (25)–(28) that

$$\Phi(\boldsymbol{x}) = \int_0^1 u(y) \log \frac{\alpha}{|\boldsymbol{x} - \gamma(y)|}\, dy - \frac{1}{2\pi} \int_\Gamma \Phi_D(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{n}_{\boldsymbol{y}}} \log \frac{\alpha}{|\boldsymbol{x} - \boldsymbol{y}|}\, d\sigma_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega.$$

Hence, the potential Φ being the solution of (24) can be approximated by Φ_h^{c} and Φ_h^{G} computed from u_h^{c} and u_h^{G}, respectively, as follows:

$$\Phi_h^{\mathrm{c}}(\boldsymbol{x}) = \int_0^1 u_h^{\mathrm{c}}(y) \log \frac{\alpha}{|\boldsymbol{x} - \gamma(y)|}\, dy - \frac{1}{2\pi} \int_\Gamma \Phi_D(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{n}_{\boldsymbol{y}}} \log \frac{\alpha}{|\boldsymbol{x} - \boldsymbol{y}|}\, d\sigma_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega,$$

$$\Phi_h^{\mathrm{G}}(\boldsymbol{x}) = \int_0^1 u_h^{\mathrm{G}}(y) \log \frac{\alpha}{|\boldsymbol{x} - \gamma(y)|}\, dy - \frac{1}{2\pi} \int_\Gamma \Phi_D(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{n}_{\boldsymbol{y}}} \log \frac{\alpha}{|\boldsymbol{x} - \boldsymbol{y}|}\, d\sigma_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega.$$

Hölder's inequality and (34) give, for $\boldsymbol{x} \in \Omega$,

$$|\Phi(\boldsymbol{x}) - \Phi_h^{\mathrm{c}}(\boldsymbol{x})| \le \left\| \log \frac{\alpha}{|\boldsymbol{x} - \gamma(\cdot)|} \right\|_1 \|u - u_h^{\mathrm{c}}\|_{-1} \le Ch^3 \|u\|_2,$$

$$|\Phi(\boldsymbol{x}) - \Phi_h^{\mathrm{G}}(\boldsymbol{x})| \le \left\| \log \frac{\alpha}{|\boldsymbol{x} - \gamma(\cdot)|} \right\|_3 \|u - u_h^{\mathrm{G}}\|_{-3} \le Ch^5 \|u\|_2.$$

Clearly, the Galerkin method yields a better approximation for $\Phi(\boldsymbol{x})$. However, this method is harder to implement than the collocation method, as the left-hand side of (32) involves the evaluation of two integrals, compared to one integral evaluation in (31).

The aim of the qualocation method is to achieve at least the same order of convergence obtained by the Galerkin method, at the cost comparable to the collocation method. Between 1987 and 2000, one witnesses different stages of development of qualocation. In his review paper [24] Sloan names them *first-generation* and *second-generation* qualocation methods.

4.2 First-Generation Qualocation

The qualocation method is a discrete version of the Galerkin method in which the outer integral in (32) is approximated by a composite quadrature rule determined by points ξ_j and weights w_j satisfying

$$0 \le \xi_1 < \cdots < \xi_J < 1, \quad w_j > 0 \quad \text{and} \quad \sum_{j=1}^{J} w_j = 1. \tag{35}$$

The qualocation rule is defined by

$$Q_N(f) := h \sum_{k=0}^{N-1} \sum_{j=1}^{J} w_j f(x_k + h\xi_j) \approx \int_0^1 f(x)\, dx, \tag{36}$$

which in turn allows us to define the discrete inner product

$$(f, g)_h := Q_N(f\overline{g}). \tag{37}$$

Hence, instead of solving (32) we solve

$$(\mathscr{L}u_h^{\mathrm{q}}, v_h)_h = (f, v_h)_h \quad \forall v_h \in \mathfrak{S}_h^r \tag{38}$$

for $u_h^{\mathrm{q}} \in \mathfrak{S}_h^r$. The initial results for this method were first obtained in [22, 30] and later generalised in [4].

There is great freedom in the choice of the quadrature rule Q_N in (36). Instead of the usual choices such as composite Simpson's rule or composite Gaussian rule, Sloan aims at finding rules most profitable to the problem at heart, namely $\mathscr{L}$ and $\mathfrak{S}_h^r$.[2] It should be emphasised that in the implementation of the Galerkin method, i.e., in solving (32), in general one has to apply a quadrature rule at least to the outer integral in this equation. The development of the qualocation method aims to choose a rule most beneficial to the solution.

Defining

$$\Lambda_N := \left\{ \mu \in \mathbb{Z} : -\frac{N}{2} < \mu \le \frac{N}{2} \right\},$$

we note that [2] for any $v \in \mathfrak{S}_h^r$ there holds

$$\hat{v}(m) = \left(\frac{\mu}{m}\right)^r \hat{v}(\mu) \quad \text{if } m \equiv \mu, m \ne 0.$$

The idea of qualocation is to choose the quadrature so that for $\mu \in \Lambda_N$, the Fourier coefficients $\hat{u}_h^{\mathrm{q}}(\mu) - \hat{u}(\mu)$ behave like $O\left((\mu/N)^\nu\right))$ for a large value of ν. More details on choices of Q_N depending on $\mathscr{L}$ and $\mathfrak{S}_h^r$ can be found in [4].

To illustrate the method we go back to our example of the logarithmic-kernel integral Eq. (27), with $\mathfrak{S}_h^r = \mathfrak{S}_h^2$. For this problem, a 2-point rule (i.e., $J = 2$) is used, namely,

$$\xi_1 = 0, \quad \xi_2 = 1/2, \quad w_1 = 3/7, \quad \text{and} \quad w_2 = 4/7. \tag{39}$$

Analogously to (34) we now have

$$\|u - u_h^{\mathrm{q}}\|_{-4} \le Ch^5 \|u\|_4, \tag{40}$$

[2] We have seen this unconventional approach in Sloan's development of degenerate-kernel methods, where he aims at finding a good approximation of $\mathscr{K}y$ by $\mathscr{K}_N y_N$, differently to the traditional approach of finding a good approximation $K_N(t, s)$ to the kernel $K(t, s)$; see Sect. 3.1 and (7).

which means the qualocation method in this case achieves the same highest order of convergence as the Galerkin method. Interestingly, the same estimate is achieved for the same rule (39) with $\mathfrak{S}_h^r = \mathfrak{S}_h^1$. In this case, the qualocation method performs better than the Galerkin method, as the latter achieves an order of convergence of 3 only; cf. (33).

It is not surprising that the advantage of the qualocation over the Galerkin method is obtained at the cost of an extra requirement on the regularity of the exact solution u. In general, with suitable choice of ξ_j and w_j, $j = 1, \dots, J$, so that the qualocation method is well defined and stable, analogously to (33), the method achieves the following highest order of convergence:

$$\|u - u_h^{\mathrm{q}}\|_{\beta-b} \le Ch^{r+b-\beta}\|u\|_{r+b} \tag{41}$$

for some additional order of convergence $b \ge 0$. Details of the derivation of different rules can be found in [4]. It is noted that if $b > r$ then the qualocation method achieves a higher order of convergence than the Galerkin method; cf. (33). For the example discussed above which yields the estimate (40), $b = 2$ when $r = 2$ and $b = 3$ when $r = 1$.

Before moving to the next part on the second-generation qualocation methods, we note that both the Galerkin and qualocation methods can be defined with a test space different from the trial space. We seek $u_h^{\mathrm{G}}, u_h^{\mathrm{q}} \in \mathfrak{S}_h^r$ satisfying

$$(\mathscr{L}u_h^{\mathrm{G}}, v_h) = (f, v_h) \quad \text{and} \quad (\mathscr{L}u_h^{\mathrm{q}}, v_h)_h = (f, v_h)_h \quad \forall v_h \in \mathfrak{S}_h^{r'}. \tag{42}$$

This version of the Galerkin method is called the Petrov-Galerkin method.

4.3 Second-Generation Qualocation

The second-generation qualocation rules are defined for ψdo $\mathscr{L}$ defined with principal symbols being combinations of (22) and (23). The operator has the form

$$\mathscr{L}v := b_+\mathscr{L}_+^\beta v + b_-\mathscr{L}_-^\beta v + \mathscr{K}v, \tag{43}$$

where $b_\pm$ are 1-periodic complex-valued C^∞ functions, and $\mathscr{L}_+^\beta$ is defined by (21), (22), whereas $\mathscr{L}_-^\beta$ is defined by (21), (23). The operator $\mathscr{K}$ can include any combination of ψdo's of lower order:

$$\mathscr{K} := \sum_{i=0}^{\infty} a_{i,+}\mathscr{L}_+^{\beta-i} + a_{i,-}\mathscr{L}_-^{\beta-i} + \mathscr{K}',$$

where $a_{i,\pm}$ belongs to C^∞ with only a finite number of the $a_{i,\pm} \in C^\infty$ allowed to be nonzero, and $\mathscr{K}'$ is an integral operator with a kernel which is a C^∞ function of both variables.

An example is the singular integral equation

$$A(\boldsymbol{x})U(\boldsymbol{a})+\frac{B(\boldsymbol{x})}{\mathrm{i}\pi}\int_{\Gamma}\frac{U(\boldsymbol{y})}{\boldsymbol{y}-\boldsymbol{x}}\,d\boldsymbol{y}+C(\boldsymbol{x})\int_{\Gamma}K(\boldsymbol{x},\boldsymbol{y})U(\boldsymbol{y})\,d\boldsymbol{y} = F(\boldsymbol{x}), \quad \boldsymbol{x}\in\Gamma, \tag{44}$$

where Γ is a smooth curve in the complex plane, A, B, C are smooth complex-valued functions, and K is a given weakly singular kernel of the form

$$K(\boldsymbol{x},\boldsymbol{y}) := \log|\boldsymbol{x}-\boldsymbol{y}| + K'(\boldsymbol{x},\boldsymbol{y})$$

with K' a C^∞ function of both variables. With Γ parametrised by the C^∞ function γ and with $a(x) := A(\gamma(e^{2\pi\mathrm{i}x})$ and $b(x) := B(\gamma(e^{2\pi\mathrm{i}x})$.

A convergence of the type (41) is proved for the general Eq. (43); see [31, 32].

4.4 Tolerant Qualocation

It is observed from (41) that the extra order of convergence is obtained at the expense of extra smoothness requirement on the exact solution. The tolerant qualocation methods developed in [34] for first generation methods, and in [29] for second generation methods remove this requirement. The remedy is to replace the quadrature on the right-hand side of (38) by an exact integral. More precisely, we approximate the solution u of (19) by $u_h^{\mathrm{tq}} \in \mathfrak{S}_h^r$ satisfying

$$(\mathscr{L}u_h^{\mathrm{tq}}, v_h)_h = (f, v_h) \quad \forall v_h \in \mathfrak{S}_h^{r'}. \tag{45}$$

This seemingly small change has profound effects. In the first place, it turns out that it eliminates the extra smoothness requirement: the smoothness requirement is now exactly the same as in the corresponding Petrov-Galerkin method. However, this small change necessitates a redesign of the qualocation method, and fresh convergence analysis, even though the techniques are traditional for the analysis of the collocation and qualocation methods.

The following error estimate is obtained for tolerant qualocation methods

$$\|u_h^{\mathrm{tq}} - u\|_s \le ch^{t-s}\|u\|_t$$

where

$$\beta - b \le s \le t \le r, \quad s < r - 1/2, \quad \text{and} \quad \beta + 1/2 < t.$$

Here the additional order of convergence b satisfies $0 < b \le r'$ and is obtained by an appropriate choice of the quadrature rule (36).

In the implementation, the exact integral on the right-hand side of (45) can be calculated by using an appropriate Gaussian quadrature, as in the case of the Galerkin method.

5 Other Contributions

Almost contemporaneously with the development of the qualocation method, Sloan and his co-authors made other contributions in the analyses of the equations

$$-\int_{\Gamma} \log|\boldsymbol{t}-\boldsymbol{s}|\, y(\boldsymbol{s})\, d\ell_s = f(\boldsymbol{t}), \quad \boldsymbol{t} \in \Gamma$$

and

$$-\int_{\Gamma} \log|\boldsymbol{t}-\boldsymbol{s}| z(\boldsymbol{s})\, d\ell_s + \omega = f(\boldsymbol{t}), \quad \boldsymbol{t} \in \Gamma, \quad \int_{\Gamma} z(\boldsymbol{s})\, d\ell_s = b,$$

where Γ is a rectifiable open or closed curve in the plane, y is the unknown in the first equation, whereas z and ω are unknowns in the second, while f is a given function and b is a given real number.

In [27] Sloan and Spence develop a robust yet conceptually simple analysis of the Galerkin method for the above equations. The method is robust in the sense that it copes easily with Γ being an open arc, a smooth and closed curve, or the boundary of a region with corners and cusps. Their approach abandons the coercivity property which is employed in previous works by e.g. Le Roux [11], Hsiao and Wendland [7], Richter [14], Wendland [36], and Chandler [3]. As a consequence, they are no longer concerned with the special function spaces that one usually has to resort to in the presence of corners; see e.g. Costabel and Stephan [5] and McLean [12].

Another approach for open curves, closed curves, and polygons, using Fourier analysis is developed by Yan and Sloan in [37]. They also consider mesh grading in the case of domains with corners [38]. An analysis for the first-kind integral equation arising from the Helmholtz equation is carried out together with Kress in [10].

Sloan's survey paper [23] presents a nice introduction to boundary integral equation methods and summarises the above-mentioned results.

6 Conclusion

Over a period of almost four decades, starting as a theoretical physicist and by using unconventional approaches, Ian Sloan has played a leading role and contributed significantly in the area of numerical methods for boundary integral equations. His tireless research activities do not stop there. After qualocation he moved to numerical integration, lattice rules, quasi-Monte-Carlo methods, and the many other topics.

Acknowledgements I am deeply indebted to Ian Sloan for his invaluable guides in the early years of my career. **Happy 80th birthday, Ian.**

References

1. Adhikari, S.K., Sloan, I.H.: Separable operator expansions for the t-matrix. Nucl. Phys. A **241**, 429–442 (1975)
2. Arnold, D.N.: A spline-trigonometric Galerkin method and an exponentially convergent boundary integral method. Math. Comput. **41**, 383–397 (1983)
3. Chandler, G.A.: Numerical analysis of the boundary integral method. In: Mathematical Programming and Numerical Analysis Workshop (Canberra, 1983), Proceedings of the Centre for Mathematics and Its Applications, vol. 6, pp. 211–230. Australian National University, Canberra (1984)
4. Chandler, G., Sloan, I.H.: Spline qualocation methods for boundary integral equations. Numer. Math. **58**, 537–567 (1990)
5. Costabel, M., Stephan, E.P.: Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation. In: Mathematical Models and Methods in Mechanics. Banach Center Publications, vol. 15, pp. 175–251. Polish Scientific Publishers, Warszawa (1985)
6. Hille, E.: Analytic Function Theory, vol. II. Ginn, Boston (1962)
7. Hsiao, G.C., Wendland, W.L.: A finite element method for some integral equations of the first kind. J. Math. Anal. Appl. **58**, 449–481 (1977)
8. Kantorovich, L.V., Krylov, V.I.: Approximate Methods of Higher Analysis (Translated from the 3rd Russian edition by C. D. Benster). Interscience Publishers, New York/P. Noordhoff, Groningen (1958)
9. Kress, R.: Linear Integral Equations. Springer, New York (1999)
10. Kress, R., Sloan, I.H.: On the numerical solution of a logarithmic integral equation of the first kind for the Helmholtz equation. Numer. Math. **66**(2), 199–214 (1993)
11. Le Roux, M.N.: Equations intégrales pour le problème du potentiel électrique dans le plan. C. R. Acad. Sci. Paris Sér. A **278**, 541–544 (1974)
12. McLean, W.: A spectral Galerkin method for a boundary integral equation. Math. Comput. **47**(176), 597–607 (1986)
13. Mikhlin, S.G., Smolitskiy, K.L.: Approximate Methods for Solution of Differential and Integral Equations (Translated from the Russian by Scripta Technica, Inc. Translation editor, R.E. Kalaba). Modern Analytic and Computational Methods in Science and Maathematics, vol. 5. American Elsevier Publishing, New York (1967)
14. Richter, G.R.: Numerical solution of integral equations of the first kind with nonsmooth kernels. SIAM J. Numer. Anal. **15**(3), 511–522 (1978)
15. Saranen, J., Vainikko, G.: Periodic Integral and Pseudodifferential Equations with Numerical Approximation. Springer Monograph in Mathematics. Springer, Berlin/New York (2002)
16. Sloan, I.H.: The method of polarized orbitals for the elastic scattering of slow electrons by ionized helium and atomic hydrogen. Proc. R. Soc. A (London) **281**, 151–163 (1964)
17. Sloan, I.H.: Convergence of degenerate-kernel methods. J. Aust. Math. Soc. Ser. B **19**(4), 422–431 (1975/76)
18. Sloan, I.H.: Error analysis for a class of degenerate-kernel methods. Numer. Math. **25**(3), 231–238 (1975/76)
19. Sloan, I.H.: Improvement by iteration for compact operator equations. Math. Comput. **30**(136), 758–764 (1976)
20. Sloan, I.H.: Iterated Galerkin method for eigenvalue problems. SIAM J. Numer. Anal. **13**(5), 753–760 (1976)

21. Sloan, I.H.: Three-body collisions involving breakup. In: Devins, D. (ed.) Momentum Wave Functions, AIP Conference Proceedings, pp. 187–194 (1977)
22. Sloan, I.H.: A quadrature-based approach to improving the collocation method. Numer. Math. **54**, 41–56 (1988)
23. Sloan, I.H.: Error analysis of boundary integral methods. Acta Numer. **1**, 287–339 (1992)
24. Sloan, I.H.: Qualocation. J. Comput. Appl. Math. **125**(1–2), 461–478 (2000). Numerical analysis 2000, vol. VI, Ordinary differential equations and integral equations
25. Sloan, I.H., Adhikari, S.K.: Method for Lippmann-Schwinger equations. Nucl. Phys. A **235**, 352–360 (1974)
26. Sloan, I.H., Moore, E.J.: Integral equation approach to electron-hydrogen collisions. J. Phys. B (Proc. Phys. Soc.) **1**(3), 414–422 (1968)
27. Sloan, I.H., Spence, A.: The Galerkin method for integral equations of the first kind with logarithmic kernel: theory. IMA J. Numer. Anal. **8**(1), 105–122 (1988)
28. Sloan, I.H., Thomée, V.: Superconvergence of the Galerkin iterates for integral equations of the second kind. J. Integr. Equ. **9**(1), 1–23 (1985)
29. Sloan, I.H., Tran, T.: The tolerant qualocation method for variable-coefficient elliptic equations on curves. J. Integr. Eqn. Appl. **13**, 73–98 (2001)
30. Sloan, I.H., Wendland, W.L.: A quadrature-based approach to improving the collocation method for splines of even degree. Z. Anal. Anwend. **8**(4), 361–376 (1989)
31. Sloan, I.H., Wendland, W.L.: Qualocation methods for elliptic boundary integral equations. Numer. Math. **79**, 451–483 (1998)
32. Sloan, I.H., Wendland, W.L.: Spline qualocation methods for variable-coefficient elliptic equations on curves. Numer. Math. **83**, 497–533 (1999)
33. Sloan, I.H., Burn, B.J., Datyner, N.: A new approach to the numerical solution of integral equations. J. Comput. Phys. **18**, 92–105 (1975)
34. Tran, T., Sloan, I.H.: Tolerant qualocation – a qualocation method for boundary integral equations with reduced regularity requirement. J. Integr. Eqn. Appl. **10**, 85–115 (1998)
35. Wahlbin, L.B.: Superconvergence in Galerkin Finite Element Methods. Lecture Notes in Mathematics, vol. 1605. Springer, Berlin/Heidelberg (1995)
36. Wendland, W.L.: Boundary element methods and their asymptotic convergence. In: Theoretical Acoustics and Numerical Techniques. CISM Courses and Lectures, vol. 277, pp. 135–216. Springer, Vienna (1983)
37. Yan, Y., Sloan, I.H.: On integral equations of the first kind with logarithmic kernels. J. Integr. Equ. Appl. **1**(4), 549–579 (1988)
38. Yan, Y., Sloan, I.H.: Mesh grading for integral equations of the first kind with logarithmic kernel. SIAM J. Numer. Anal. **26**(3), 574–587 (1989)

A Qualocation Method for Parabolic Partial Integro-Differential Equations in One Space Variable

Lok Pati Tripathi, Amiya K. Pani, and Graeme Fairweather

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract In this article, a qualocation method is formulated and analyzed for parabolic partial integro-differential equations in one space variable. Using a new Ritz–Volterra type projection, optimal rates of convergence are derived. Based on the second-order backward differentiation formula, a fully discrete scheme is formulated and a convergence analysis is derived. Results of numerical experiments are presented which support the theoretical results.

1 Introduction

During the last several decades, much attention has been devoted to the formulation, analysis and implementation of collocation methods involving smoothest splines for the approximate solution of second-order two-point boundary value problems (TPBVPs) and for the spatial discretization in time dependent partial differential equations. Invariably, these methods use C^2 cubic splines and are of suboptimal accuracy. Oft ignored is the work of de Boor [14] who proved that C^2 cubic nodal spline approximations for TPBVPs cannot be more than second order accurate when one would expect fourth order accuracy. Also overlooked are the fundamental works of Archer [4, 5] and Daniel and Swartz [18] who devised optimal nodal spline collocation methods based on a perturbed differential equation. Optimal nodal

L. P. Tripathi
Department of Mathematics, IIT Goa, Ponda, Goa, India

A. K. Pani
Department of Mathematics, Industrial Mathematics Group, IIT Bombay, Mumbai, India
e-mail: akp@math.iitb.ac.in

G. Fairweather (✉)
Mathematical Reviews, American Mathematical Society, Ann Arbor, MI, USA

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_53

cubic spline collocation methods for TPBVPs and elliptic problems have also been developed in [1, 2, 6, 7, 10]. Similar work was carried out by Houstis et al. [25] for C^1 quadratic spline collocation for TPBVPS, by Christara [16] for elliptic problems in two space variables and by Christara et al. [17] for parabolic problems in one space variable; see also [12, 22].

The drawback of these optimal methods is that, in general, they require the spatial mesh to be uniform. In [29], a cubic spline method for linear second-order TPBVPs, called a *qualocation method* is formulated and analyzed. This method can be viewed as a Petrov–Galerkin method using a cubic spline trial space, a piecewise linear test space, and a simple quadrature rule for the integration (compound Simpson's rule or compound two-point Gauss quadrature), and may also be considered a discrete version of the H^1-Galerkin method. It is proved that the error in the W_p^i norm for $i = 0, 1, 2, 1 \leq p \leq \infty$ is of order $4-i$. A key feature of this method is that it allows an arbitrary mesh. The results in [29] were generalized to higher order TPBVPS using higher order smoothest splines in [23, 24]. Jones and Pani [26] considered qualocation for a semilinear second-order TPBVP and derived optimal estimates in the W_p^i norm for $i = 0, 1, 2, 1 \leq p \leq \infty$. Pani [28] considered qualocation for the spatial discretization in the numerical solution of a semilinear parabolic problem in one space variable and both the linearized backward Euler method and the extrapolated Crank–Nicolson scheme for the time-stepping, and established optimal estimates. This method is further generalized to a one dimensional Stefan problem in [27] and optimal error estimates derived. For the solution of elliptic, parabolic and hyperbolic equations in two space variables using a qualocation-like approach for the spatial discretization, see [8, 9, 11, 13].

In this paper, we formulate a qualocation method for the parabolic partial integro-differential equations in one space variable of the form

$$u_t(x,t) - \mathscr{A}u(x,t) = \mathscr{B}(t)u(x,t) + f(x,t), (x,t) \in I \times J, \tag{1a}$$

subject to the boundary conditions

$$u(0,t) = u(1,t) = 0, \quad t \in J, \tag{1b}$$

and the initial condition

$$u(x,0) = u_0(x), \quad x \in I, \tag{1c}$$

where $I := (0,1)$, $J := (0,T]$ with $T < \infty$, and the operators $\mathscr{A}$ and $\mathscr{B}$ are of the form of

$$\mathscr{A}u(x,t) := u_{xx}(x,t) - b(x)u_x(x,t) - c(x)u(x,t), \quad \mathscr{B}(t)u(x,t) := \int_0^t B(t,s)u(x,s)ds,$$

with

$$B(t,s)u(x,s) := u_{xx}(x,s) + b_1(x;t,s)u_x(x,s) + c_1(x;t,s)u(x,s).$$

We assume that u_0, f, b, c, b_1, and c_1 are sufficiently smooth.

The major contributions of this article are the following.

- Using the Ritz–Volterra projection introduced in Sect. 3, we derive optimal error estimates for the semidiscrete case in Sect. 4.
- Replacing the time derivative by the second order backward differentiation formula (BDF2) and the integral in time by compound trapezoidal rule, we obtain a fully discrete scheme for which optimal estimates are established in Sect. 5.
- Finally in Sect. 6, we present the results of numerical experiments which confirm our theoretical findings.

2 Preliminaries

For $I = (0, 1)$, $m \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty]$, the spaces $C^m(I)$, $C(I) = C^0(I)$, $W^{m,p}(I)$, $W_0^{m,p}(I)$, $L^p(I) = W^{0,p}(I)$, $H^m(I) = W^{m,2}(I)$ and $H_0^m(I) = W_0^{m,2}(I)$ are the standard function spaces introduced in [3]. Further, for a Banach space $\mathscr{X}$, the spaces $W^{m,p}(0, T; \mathscr{X})$, $T > 0$, $L^p(0, T; \mathscr{X}) = W^{0,p}(0, T; \mathscr{X})$ and $H^m(0, T; \mathscr{X}) = W^{m,2}(0, T; \mathscr{X})$ denote the standard Banach valued ($\mathscr{X}$-valued) function spaces introduced in [19]. The norm corresponding to a Banach space $\mathscr{X}$ will be denoted by $\|\cdot\|_{\mathscr{X}}$. We shall frequently use the following notations also

$$\|\cdot\| = \|\cdot\|_{L^2}, \ \|\cdot\|_m = \|\cdot\|_{H^m} \text{ and } \|\cdot\|_{m,p} = \|\cdot\|_{W^{m,p}}.$$

The weak formulation of the problem (1a)–(1c), which is appropriate for the H^1-Galerkin formulation, is defined to be a function $u : [0, T] \to H^2 \cap H_0^1$ such that

$$-(u_t, v_{xx}) + (\mathscr{A}u, v_{xx}) = -(\mathscr{B}(t)u, v_{xx}) - (f, v_{xx}), \ v \in H^2 \cap H_0^1, \ t \in J, \tag{2a}$$

$$u(0) = u_0. \tag{2b}$$

Given $M \geq 1$, let $0 = x_0 < x_1 < \cdots < x_M = 1$ be an arbitrary partition of $[0, 1]$ with the property that $h \to 0$ as $n \to \infty$, where

$$I_i = [x_{i-1}, x_i], \quad h_i = x_i - x_{i-1}, \quad i = 1, \cdots, M,$$

and $h = \max_{1\leq i\leq M} h_i$.

Let $S_h := \{\chi \in C^2(\bar{I}) : \chi|_{I_i} \in P_3, \ i = 1, \cdots, M\}$,

$S_h^0 := \{\chi \in S_h : \chi(0) = \chi(1) = 0\}$,

and

$$T_h := \{\chi_{xx} : \ \chi \in S_h^0\} = \{v \in C(\bar{I}) : v|_{I_i} \in P_1, \ i = 1, \cdots, M\},$$

where P_m is the space of polynomials of degree $\leq m$.

In the standard H^1-Galerkin procedure for the solution of (1a)–(1c), we seek a function $\bar{u}_h : [0, T] \to S_h^0$ satisfying

$$-(\bar{u}_{h,t}, \chi_{xx}) + (\mathscr{A}\bar{u}_h, \chi_{xx}) + (\mathscr{B}(t)\bar{u}_h, \chi_{xx}) = -(f, \chi_{xx}),\ \chi \in S_h^0,\ t \in J, \tag{3}$$

with $u_h(0)$ given.

In practice, the integrals in (3) are rarely evaluated exactly. Therefore, we replace the exact inner product $(\cdot, \cdot)$ by the discrete approximation $\langle \cdot, \cdot \rangle$, where

$$\langle v, w \rangle = Q_h(vw),$$

and Q_h is the fourth-order composite 2-point Gauss quadrature rule given by

$$Q_h g = \frac{1}{2} \sum_{i=1}^{M} h_i [g(x_{i,1}) + g(x_{i,2})], \tag{4}$$

with

$$x_{i,\ell} = \frac{1}{2}(x_i + x_{i-1}) + (-1)^{\ell} \frac{h_i}{2\sqrt{3}}, \quad \ell = 1, 2.$$

The resulting scheme is a quadrature based modification of the collocation method often called a qualocation method. For the quadrature error

$$\epsilon_h(g) := \int_0^1 g(x)dx - Q_h(g),$$

a use of Peano Kernel Theorem [20] yields the following error bound:

$$|\epsilon_h(g)| \leq C \sum_{i=1}^{M} h_i^4 \|g^{(4)}\|_{L^1(I_i)}. \tag{5}$$

The qualocation approximation of (2a)–(2b) is then defined to be a function u_h : $[0, T] \to S_h^0$ satisfying

$$-\langle u_{h,t}, \chi_{xx} \rangle + \langle \mathscr{A}u_h, \chi_{xx} \rangle + \langle \mathscr{B}(t)u_h, \chi_{xx} \rangle = -\langle f, \chi_{xx} \rangle, \quad \chi \in S_h^0, \quad t \in J, \tag{6a}$$

$$u_h(0) = u_{0h}, \tag{6b}$$

which is equivalent to find a function $u_h : [0, T] \to S_h^0$ satisfying

$$-\langle u_{h,t}, v \rangle + \langle \mathscr{A}u_h, v \rangle + \langle \mathscr{B}(t)u_h, v \rangle = -\langle f, v \rangle, \quad v \in T_h, \quad t \in J, \tag{7a}$$

$$u_h(0) = u_{0h}, \tag{7b}$$

where u_{0h} is a suitable approximation of u_0 in S_h^0 to be defined later. In the convergence analysis, we use (6a)–(6b) whereas we employ (7a)–(7b) in computations.

The semidiscrete problem (6a)–(6b) leads to a system of linear integro-differential equations. An application of Picard's theorem yields the existence of a unique solution of (6a)–(6b) for $t \in J$.

The following result which is used to derive some basic inequalities is stated without proof. For a proof, see [21].

Lemma 1 *For all f and g in S_h,*

$$-\langle f, g_{xx}\rangle = (f_x, g_x) - fg_x|_0^1 + \frac{1}{1080}\sum_{i=1}^{M}(f_{xxx,i})(g_{xxx,i})h_i^5,$$

where $f_{xxx,k}$ is the (constant) value of the third derivative of f and g in I_k.

From Lemma 1,

$$-\langle v, v_{xx}\rangle \geq \|v_x\|^2, \quad v \in S_h^0.$$

From (5), it follows that

$$\langle p, 1\rangle = \int_0^1 p dx, \quad p \in P_3.$$

Hence,

$$\langle v_{xx}, v_{xx}\rangle := [[v_{xx}]]^2 = \|v_{xx}\|^2, \quad v \in S_h. \tag{8}$$

Moreover, from [21],

$$[[v_x]]^2 \leq \|v_x\|^2, \quad v \in S_h. \tag{9}$$

Throughout this paper, C denotes a generic positive constant whose dependence can be traced from the proof.

3 The Ritz–Volterra Type Projection and Related Estimates

Let $\tilde{u} : [0, T] \to S_h^0$ be the projection of u defined by

$$\langle \mathscr{A}(u - \tilde{u}), \chi_{xx}\rangle + \langle \mathscr{B}(t)(u - \tilde{u}), \chi_{xx}\rangle = 0, \ \forall \chi \in S_h^0. \tag{10}$$

In order to prove the existence of a unique solution $\tilde{u}(t) \in S_h^0$ of (10) for a given $u(t),\ t \in [0, T]$, we choose $\{\phi_i\}_{i=1}^{M+1}$ as a basis for S_h^0, and write $\tilde{u}(t) = \sum_{j=1}^{M+1} \alpha_j(t)\phi_j$. On substituting $\tilde{u}(t)$ in (10) and choosing $\chi = \phi_i,\ i = 1, 2, \ldots, M + 1$, we obtain the system

$$E\alpha(t) + \int_0^t F(t, s)\alpha(s)ds = G(t), \tag{11}$$

where

$$\alpha(t) := [\alpha_1(t), \alpha_2(t), \ldots, \alpha_{M+1}(t)]^T, \quad E = (E_{i,j})_{i,j=1}^{M+1}, \quad E_{i,j} = \left\langle \mathscr{A}\phi_j, \phi_{ixx} \right\rangle,$$
$$F(t, s) = (F_{i,j}(t, s))_{i,j=1}^{M+1}, \quad F_{i,j}(t, s) = \left\langle B(t, s)\phi_j, \phi_{ixx} \right\rangle,$$
$$G(t) = (G_i(t))_{i=1}^{M+1}, \quad G_i(t) = \langle \mathscr{A}u(t) + \mathscr{B}(t)u(t), \phi_{ixx} \rangle .$$

For sufficiently small h, E is invertible (see [29]) and hence (11) can be written as a system of linear Volterra equations of the form

$$\alpha(t) + \int_0^t \widetilde{F}(t, s)\alpha(s)ds = \widetilde{G}(t),$$

where $\widetilde{F} = E^{-1}F$ and $\widetilde{G} = E^{-1}G$. Using Picard's theorem there exists a unique solution $\alpha(t)$ of this system. Thus, for sufficiently small h, the problem (10) has a unique solution.

In the following lemma, we derive estimates of $\eta := u - \tilde{u}$.

Lemma 2 *Let $u \in W^{2,\infty}(0, T; W^{6,p}(I)),\ p \in [1, \infty]$, and η satisfy (10). Then, for sufficiently small h and $i = 0, 1, 2$,*

$$\left\| \frac{\partial^j}{\partial t^j}\eta(t) \right\|_{i,p} \leq Ch^{4-i} \left(\sum_{l=0}^{j} \left\| \frac{\partial^l}{\partial t^l}u(t) \right\|_{6,p} + \int_0^t \|u(s)\|_{6,p} ds \right), \quad j = 0, 1, 2, \ \ p \in [1, \infty].$$

Proof With

$$\mathscr{A}^*\psi := \psi_{xx} + (b\psi)_x - c\psi,$$

let ψ be the unique solution of

$$\mathscr{A}^*\psi = \phi \ \text{ in } I, \tag{12a}$$
$$\psi(0) = \psi(1) = 0, \tag{12b}$$

satisfying

$$\|\psi\|_{2,q} \leq C\|\mathscr{A}^*\psi\|_{L^q} = C\|\phi\|_{L^q}, \tag{13}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

For $\chi \in S_h^0$ and $\phi \in L^q$, (12a) and (10) yield

$$\begin{aligned}(\eta, \phi) &= (\mathscr{A}\eta, \psi)\\ &= (\mathscr{A}\eta, \psi - \chi_{xx}) + (\mathscr{A}\eta, \chi_{xx}) - \langle \mathscr{A}\eta, \chi_{xx}\rangle + \langle \mathscr{A}\eta, \chi_{xx}\rangle\\ &= (\mathscr{A}\eta, \psi - \chi_{xx}) + \epsilon_h((\mathscr{A}\eta)\chi_{xx}) + \epsilon_h((\mathscr{B}(t)\eta)\chi_{xx})\\ &\quad - (\mathscr{B}(t)\eta, \chi_{xx} - \psi) - (\mathscr{B}(t)\eta, \psi).\end{aligned}$$

With $\chi_{xx} = \psi_h = I_h\psi$, where $I_h\psi$ is the piecewise linear interpolant of ψ, and using estimate (a) in [29, Lemma 4.2], we obtain

$$\begin{aligned}|(\eta, \phi)| &\leq \|\mathscr{A}\eta\|_{L^p}\|\psi - I_h\psi\|_{L^q} + |\epsilon_h((\mathscr{A}\eta)(I_h\psi))| + |(\epsilon_h((\mathscr{B}(t)\eta)(I_h\psi))|\\ &\quad + \|\mathscr{B}(t)\eta\|_{L^p}\|\psi - I_h\psi\|_{L^q} + \left|\int_0^t (\eta, B^*(t,s)\psi)ds\right|\\ &\leq C\Bigg\{h^2\|\eta\|_{2,p}\|\psi\|_{2,q} + (h\|\eta\|_{L^p} + h^4\|u\|_{6,p})\|I_h\psi\|_{1,q}\\ &\quad + \left(h\int_0^t \|\eta(s)\|_{L^p}ds + h^4\int_0^t \|u(s)\|_{6,p}ds\right)\|I_h\psi\|_{1,q}\\ &\quad + h^2\left(\int_0^t \|\eta\|_{2,p}ds\right)\|\psi\|_{2,q} + \left(\int_0^t \|\eta\|_{L^p}ds\right)\|\psi\|_{2,q}\Bigg\}.\end{aligned}$$

Note that

$$\|I_h\psi\|_{1,q} \leq \|I_h\psi - \psi\|_{1,q} + \|\psi\|_{1,q} \leq Ch\|\psi\|_{2,q} + \|\psi\|_{1,q} \leq C\|\psi\|_{2,q}. \tag{14}$$

Using (13), we obtain

$$\begin{aligned}\|\eta\|_{L^p} &\leq C\Bigg[h^2\left(\|\eta\|_{2,p} + \int_0^t \|\eta\|_{2,p}ds\right) + h\left(\|\eta\|_{L^p} + \int_0^t \|\eta\|_{L^p}ds\right)\\ &\quad + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\Bigg] + C\int_0^t \|\eta\|_{L^p}ds.\end{aligned}$$

If h is chosen so that $(1 - Ch) > 0$, then an application of Gronwall's Lemma yields

$$\|\eta\|_{L^p} \leq C\left[h^2\left(\|\eta\|_{2,p} + \int_0^t \|\eta\|_{2,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]. \tag{15}$$

Now we need to estimate $\|\eta_{xx}\|_{L^p}$. To this end, let P_h be the L^2 projection operator onto T_h defined by

$$(w - P_h w, \chi_{xx}) = 0, \qquad \chi \in S_h^0. \tag{16}$$

Note that, from [15], the L^2-projection is stable in L^p, i.e.,

$$\| P_h w\|_{L^p} \le C\|w\|_{L^p} \quad \forall\ w \in L^p. \tag{17}$$

Now,

$$\|\eta_{xx}\|_{L^p} = \|u_{xx} - \tilde{u}_{xx}\|_{L^p} \le \|u_{xx} - P_h u_{xx}\|_{L^p} + \| P_h u_{xx} - \tilde{u}_{xx}\|_{L^p}. \tag{18}$$

From [15],

$$\|u_{xx} - P_h u_{xx}\|_{L^p} \le Ch^2\|u_{xx}\|_{2,p} \le Ch^2\|u\|_{4,p}. \tag{19}$$

For the second term on the right of (18),

$$\begin{aligned}
(P_h u_{xx} - \tilde{u}_{xx}, \chi_{xx}) &= (u_{xx} - \tilde{u}_{xx}, \chi_{xx}) = (\eta_{xx}, \chi_{xx}) \\
&= (\mathscr{A}\eta + b\eta_x + c\eta, \chi_{xx}) \\
&= (\mathscr{A}\eta, \chi_{xx}) - \langle \mathscr{A}\eta, \chi_{xx}\rangle - \langle \mathscr{B}(t)\eta, \chi_{xx}\rangle + (\mathscr{B}(t)\eta, \chi_{xx}) \\
&\quad - (\mathscr{B}(t)\eta, \chi_{xx}) + (b\eta_x + c\eta, \chi_{xx}),
\end{aligned}$$

on using (10). Using estimate (b) in [29, Lemma 4.2], we obtain

$$\begin{aligned}
(P_h u_{xx} - \tilde{u}_{xx}, \chi_{xx}) &= \epsilon_h((\mathscr{A}\eta)(\chi_{xx})) + \epsilon_h((\mathscr{B}(t)\eta)(\chi_{xx})) \\
&\quad - (\mathscr{B}(t)\eta, \chi_{xx}) + ((b\eta_x + c\eta), \chi_{xx}) \\
&\le C\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p} ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p} ds\right)\right]\right. \\
&\quad \left. + \|\eta\|_{1,p} + \int_0^t \|\eta\|_{2,p} ds\right\} \|\chi_{xx}\|_{L^q}.
\end{aligned}$$

Let $\tilde{\phi}$ be an arbitrary element of L^q. Then, using the definition of $P_h\tilde{\phi}$ in (16), it follows that

$$\begin{aligned}
|(P_h u_{xx} - \tilde{u}_{xx}, \tilde{\phi})| &= |(P_h u_{xx} - \tilde{u}_{xx}, P_h\tilde{\phi})| \\
&\leq C\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right. \\
&\qquad \left. + \|\eta\|_{1,p} + \int_0^t \|\eta\|_{2,p}ds\right\} \| P_h\tilde{\phi}\|_{L^q} \\
&\leq C\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right. \\
&\qquad \left. + \|\eta\|_{1,p} + \int_0^t \|\eta\|_{2,p}ds\right\} \|\tilde{\phi}\|_{L^q},
\end{aligned}$$

where in the last step we have used (17) with p replaced by q. Hence,

$$\begin{aligned}
\| P_h u_{xx} - \tilde{u}_{xx}\|_{L^p} &\leq C\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right. \\
&\qquad \left. + \|\eta\|_{1,p} + \int_0^t \|\eta\|_{2,p}ds\right\}. \qquad (20)
\end{aligned}$$

On combining (18)–(20), we find that

$$\begin{aligned}
\|\eta_{xx}\|_{L^p} \leq C&\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right. \\
&\left. + h^2\|u\|_{4,p} + \|\eta\|_{1,p} + \int_0^t \|\eta\|_{2,p}ds\right\}.
\end{aligned}$$

Note that

$$\|\eta\|_{2,p} \leq \|\eta\|_{1,p} + \|\eta_{xx}\|_{L^p},$$

and from the interpolation inequality in [29, (4.4)]), we obtain

$$\|\eta\|_{1,p} \leq C\left(h^{-1}\|\eta\|_{L^p} + h\|\eta\|_{2,p}\right). \qquad (21)$$

Thus, for sufficiently small h,

$$\begin{aligned}
\|\eta\|_{2,p} \leq C&\left\{\left[h^3\left(\|u\|_{5,p} + \int_0^t \|u\|_{5,p}ds\right) + h^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right. \\
&\left. + h^2\|u\|_{4,p} + h^{-1}\|\eta\|_{L^p} + \int_0^t \|\eta\|_{2,p}ds\right\},
\end{aligned}$$

On using (15), we obtain

$$\|\eta\|_{2,p} \leq C\left\{\left[h^2\|u\|_{4,p} + h^3\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right]\right.$$
$$\left. +h\|\eta\|_{2,p} + \int_0^t \|\eta\|_{2,p}ds\right\}.$$

With h chosen so that $1 - Ch > 0$, an application of Gronwall's Lemma yields

$$\begin{aligned}\|\eta\|_{2,p} &\leq C\left[h^2\left(\|u\|_{4,p} + \int_0^t \|u\|_{4,p}ds\right) + h^3\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right)\right] \\ &\leq Ch^2\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right). \end{aligned} \tag{22}$$

Therefore, on using this estimate in (15), we obtain

$$\|\eta\|_{L^p} \leq Ch^4\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right). \tag{23}$$

Moreover, from (21)–(23),

$$\|\eta\|_{i,p} \leq Ch^{4-i}\left(\|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right), \ i = 0, 1, 2. \tag{24}$$

Now, for $\phi \in L^q$ and $\chi \in S_h^0$,

$$\begin{aligned}(\eta_t, \phi) &= (\eta_t, \mathscr{A}^*\psi) = (\mathscr{A}\eta_t, \psi) \\ &= (\mathscr{A}\eta_t, \psi - \chi_{xx}) + (\mathscr{A}\eta_t, \chi_{xx}) - \langle \mathscr{A}\eta_t, \chi_{xx}\rangle \\ &\quad - \langle B(t,t)\eta, \chi_{xx}\rangle - \left\langle \int_0^t B_t(t,s)\eta(s)ds, \chi_{xx}\right\rangle \\ &\quad + (B(t,t)\eta, \chi_{xx}) + \left(\int_0^t B_t(t,s)\eta(s)ds, \chi_{xx}\right) \\ &\quad - (B(t,t)\eta, \chi_{xx} - \psi) - \left(\int_0^t B_t(t,s)\eta(s)ds, \chi_{xx} - \psi\right) \\ &\quad - (B(t,t)\eta, \psi) - \left(\int_0^t B_t(t,s)\eta(s)ds, \psi\right).\end{aligned}$$

Let $\chi_{xx} = I_h\psi \in T_h$, then again by using estimate (a) in [29, Lemma 4.2], we obtain

$$\begin{aligned}
(\eta_t, \phi) &= (\mathscr{A}\eta_t, \psi) \\
&= (\mathscr{A}\eta_t, \psi - I_h\psi) + \epsilon_h((\mathscr{A}\eta_t)(I_h\psi)) + \epsilon_h((B(t,t)\eta)(I_h\psi)) \\
&\quad + \epsilon_h\left(\left(\int_0^t B_t(t,s)\eta(s)ds\right)(I_h\psi)\right) \\
&\quad - (B(t,t)\eta, I_h\psi - \psi) - \left(\int_0^t B_t(t,s)\eta(s)ds, I_h\psi - \psi\right) \\
&\quad - (\eta, B^*(t,t)\psi) - \int_0^t (\eta(s), B_t^*(t,s)\psi)ds \\
&\leq C\Big\{h^2\|\eta_t\|_{2,p}\|\psi\|_{2,q} + \Big(h\|\eta_t\|_{L^p} + h^4\|u_t\|_{6,p} + h\|\eta\|_{L^p} + h^4\|u\|_{6,p} \\
&\quad + h\int_0^t \|\eta\|_{L^p}ds + h^4\int_0^t \|u\|_{6,p}ds\Big)\|I_h\psi\|_{1,q} \\
&\quad + \Big(h^2\|\eta\|_{2,p} + h^2\int_0^t \|\eta\|_{2,p}ds + \|\eta\|_{L^p} + \int_0^t \|\eta\|_{L^p}ds\Big)\|\psi\|_{2,q}\Big\}.
\end{aligned}$$

Using (13), (14) and (24), we obtain

$$|(\eta_t, \phi)| \leq C\Big[h^2\|\eta_t\|_{2,p} + h\|\eta_t\|_{L^p} + h^4\Big(\|u_t\|_{6,p} + \|u\|_{6,p} + \int_0^t \|u\|_{6,p}\Big)\Big]\|\phi\|_{L^q}.$$

Again, for sufficiently small h, we find that

$$\|\eta_t\|_{L^p} \leq C\Big[h^2\|\eta_t\|_{2,p} + h^4\Big(\|u_t\|_{6,p} + \|u\|_{6,p} + \int_0^t \|u\|_{6,p}\Big)\Big].$$

By using similar steps as in the estimation of $\|\eta\|_{2,p}$, we arrive at

$$\|\eta_t\|_{2,p} \leq Ch^2\Big(\|u_t\|_{6,p} + \|u\|_{6,p} + \int_0^t \|u\|_{6,p}\Big).$$

Hence,

$$\|\eta_t\|_{i,p} \leq Ch^{4-i}\left(\|u_t\|_{6,p} + \|u\|_{6,p} + \int_0^t \|u\|_{6,p}ds\right),\ i = 0, 1, 2.$$

Now, a similar procedure yields the estimates of higher order time derivatives. This completes the rest of the proof. □

4 Error Estimates for the Semi-Discrete Scheme

We write e as

$$e = u - u_h = (u - \tilde{u}) - (u_h - \tilde{u}) = \eta - \theta, \tag{25}$$

and note that it is sufficient to estimate θ as estimates of η are known from Lemma 2.

From (2a) and (6a) and using the projection (10), we obtain

$$-\langle \theta_t, \chi_{xx}\rangle + \langle \mathscr{A}\theta, \chi_{xx}\rangle = -\langle \mathscr{B}(t)\theta(t), \chi_{xx}\rangle - \langle \eta_t, \chi_{xx}\rangle, \qquad \chi \in S_h. \tag{26}$$

Lemma 3 *Let u_h be the solution of (6a)–(6b) with $u_h(0) = u_{0h} = \tilde{u}(x, 0)$. Then*

$$\|\theta\|_{L^\infty(0,\, t;\, H^1)} \le Ch^4 \left(\|u\|_{L^2(0,\, t;\, W^{6,\infty})} + \|u_t\|_{L^2(0,\, t;\, W^{6,\infty})}\right).$$

Proof Choose $\chi = \theta$ in (26) to obtain

$$\begin{aligned} -\langle \theta_t, \theta_{xx}\rangle + \langle \theta_{xx}, \theta_{xx}\rangle &= -\langle \mathscr{B}(t)\theta(t), \theta_{xx}\rangle - \langle \eta_t, \theta_{xx}\rangle + \langle b\theta_x + c\theta, \theta_{xx}\rangle \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{27}$$

From Lemma 1, the first term on the left hand side of (27) becomes

$$\begin{aligned} -\langle \theta_t, \theta_{xx}\rangle &= (\theta_{tx}, \theta_x) + \frac{1}{1080}\sum_{i=1}^{M} h_i^5 (\theta_{txxx,i})(\theta_{xxx,i}) \\ &= \frac{1}{2}\frac{d}{dt}\left(\|\theta_x\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5 (\theta_{xxx,i})^2\right). \end{aligned}$$

Upon integration with respect to time from 0 to t, we obtain

$$-\int_0^t \langle \theta_\tau(\tau), \theta_{xx}(\tau)\rangle\, d\tau \ge \frac{1}{2}\|\theta_x(t)\|^2.$$

From (8), it follows that

$$\langle \theta_{xx}(t), \theta_{xx}(t)\rangle = [[\theta_{xx}(t)]]^2 = \|\theta_{xx}(t)\|^2.$$

For I_1, using (8) and (9) together with the Poincaré inequality for $\theta \in S_h^0$, we arrive at

$$|I_1| \le C(\epsilon)\int_0^t (\|\theta_{xx}(\tau)\|^2 + \|\theta_x(\tau)\|^2) d\tau + \epsilon\|\theta_{xx}(t)\|^2.$$

To estimate the term I_2, we apply Young's inequality,

$$ab \le a^2/2\epsilon + \epsilon b^2/2, \quad a, b \in R, \quad \epsilon > 0 \tag{28}$$

together with (8) to obtain

$$|I_2| \le C(\epsilon)\|\eta_t(t)\|^2_{L^\infty} + \epsilon\|\theta_{xx}(t)\|^2.$$

Finally, for the estimation of I_3, using (8) and (9) and the Poincaré inequality for $\theta \in S_h^0$, we obtain

$$|I_3| \le C(\epsilon)\|\theta_x(t)\|^2 + \epsilon\|\theta_{xx}(t)\|^2.$$

On combining these estimates and integrating with respect to time from 0 to t, it follows that

$$\|\theta_x(t)\|^2 + (2-6\epsilon)\int_0^t \|\theta_{xx}(\tau)\|^2 d\tau \le C(\epsilon)\left[\int_0^t\int_0^\tau \|\theta_{xx}(\tau')\|^2 d\tau' d\tau \right.$$
$$\left. + \int_0^t \|\theta_x(\tau)\|^2 d\tau + \int_0^t \|\eta_\tau(\tau)\|^2_{L^\infty} d\tau\right].$$

Choosing $\epsilon = 1/6$ and using Lemma 2, we obtain

$$\|\theta_x(t)\|^2 + \int_0^t \|\theta_{xx}(\tau)\|^2 d\tau \le Ch^8 \int_0^t \{\|u(\tau)\|^2_{W^{6,\infty}} + \|u_\tau(\tau)\|^2_{W^{6,\infty}}\} d\tau$$
$$+ C\left(\int_0^t \|\theta_x(\tau)\|^2 d\tau + \int_0^t\int_0^\tau \|\theta_{xx}(\tau')\|^2 d\tau' d\tau\right).$$

An application of Gronwall's Lemma completes the proof. □

Lemma 4 *Let u_h be the solution of (6a)–(6b) with $u_h(0) = u_{0h} = \tilde{u}(x,0)$. Then*

$$\|\theta_{xx}\|_{L^\infty(0,\,t;\,L^2)} \le Ch^4\left[\sum_{l=0}^{1}\left\|\frac{\partial^l u}{\partial t^l}\right\|_{L^\infty(0,\,t;\,W^{6,\infty})} + \sum_{l=0}^{2}\left\|\frac{\partial^l u}{\partial t^l}\right\|_{L^2(0,\,t;\,W^{6,\infty})}\right].$$

Proof Choose $\chi = \theta_t$ in (26). Then integrating from 0 to t gives

$$-\int_0^t \langle\theta_\tau(\tau), \theta_{\tau xx}(\tau)\rangle\, d\tau + \int_0^t \langle\theta_{xx}(\tau), \theta_{\tau xx}(\tau)\rangle\, d\tau$$
$$= -\int_0^t \langle\mathscr{B}(\tau)\theta(\tau) + \eta_\tau(\tau) - b\theta_x(\tau) - c\theta(\tau), \theta_{\tau xx}(\tau)\rangle\, d\tau.$$

Using integration by parts on right hand side yields

$$-\int_0^t \langle \theta_\tau(\tau), \theta_{\tau xx}(\tau)\rangle \, d\tau + \int_0^t \langle \theta_{xx}(\tau), \theta_{\tau xx}(\tau)\rangle \, d\tau = \int_0^t \Big\langle \int_0^\tau B_\tau(\tau,\tau^{'})\theta(\tau^{'})d\tau^{'}$$

$$+ B(\tau,\tau)\theta(\tau) + \eta_{\tau\tau}(\tau) - b\theta_{\tau x}(\tau) - c\theta_\tau(\tau), \theta_{xx}(\tau)\Big\rangle d\tau$$

$$- \langle \mathscr{B}(t)\theta(t) + \eta_t(t) - b\theta_x(t) - c\theta(t), \theta_{xx}(t)\rangle = I_1 + I_2.$$

By Lemma 1,

$$-\langle \theta_t, \theta_{txx}\rangle = (\theta_{tx}, \theta_{tx}) + \frac{1}{1080}\sum_{i=1}^{M} h_i^5(\theta_{txxx,i})^2 \ge \|\theta_{tx}\|^2,$$

and

$$\langle \theta_{xx}, \theta_{txx}\rangle = \frac{1}{2}\frac{d}{dt}\langle \theta_{xx}, \theta_{xx}\rangle = \frac{1}{2}\frac{d}{dt}[[\theta_{xx}]] = \frac{1}{2}\frac{d}{dt}\|\theta_{xx}\|^2.$$

To estimate I_1 and I_2, we use Young's inequality (28) together with (8) and (9) to obtain

$$|I_1| \le C(\epsilon)\int_0^t \left(\|\theta_x(\tau)\|^2 + \|\theta_{xx}(\tau)\|^2 + \|\eta_{\tau\tau}(\tau)\|^2_{L^\infty}\right) d\tau + \epsilon\int_0^t \|\theta_{\tau x}(\tau)\|^2 d\tau,$$

and

$$|I_2| \le C(\epsilon)\left(\int_0^t (\|\theta_x(\tau)\|^2 + \|\theta_{xx}(\tau)\|^2)d\tau + \|\eta_t(t)\|^2_{L^\infty} + \|\theta_x(t)\|^2\right) + \epsilon\|\theta_{xx}(t)\|^2.$$

Combining these estimates, we arrive at

$$(1-\epsilon)\int_0^t \|\theta_{\tau x}(\tau)\|^2 d\tau + \left(\frac{1}{2} - \epsilon\right)\|\theta_{xx}(t)\|^2 \le C(\epsilon)\Big[\|\eta_t(t)\|^2_{L^\infty}$$

$$+ \int_0^t \|\eta_{\tau\tau}(\tau)\|^2_{L^\infty} d\tau + \|\theta_x(t)\|^2 + \int_0^t \|\theta_x(\tau)\|^2 d\tau\Big] + C(\epsilon)\int_0^t \|\theta_{xx}(\tau)\|^2 d\tau$$

Choose ϵ appropriately so that $(1-2\epsilon) = \frac{1}{4}$. Then, the use of Lemmas 2 and 3 together with Gronwall's Lemma completes the rest of the proof. □

Using Lemmas 3, 4 and 2, we have the following error estimate for the semi-discrete problem.

Theorem 1 *Let $u \in H^2(0,T;W^{6,\infty}(I))$ and u_h be the solution of (6a)–(6b) with $u_h(0) = u_{0h} = \tilde{u}(x,0)$. Then*

$$\|e\|_{L^\infty(J;\,H^j(I))} \leq Ch^{4-j}, \quad j = 0,1,2,$$

and

$$\|e\|_{L^\infty(J;\,W^{j,\infty}(I))} \leq Ch^{4-j}, \quad j = 0,1.$$

5 Second Order Backward Difference (BDF2) Scheme

Let

$$0 = t_0 < t_1 < \cdots < t_N = T;\ t_n - t_{n-1} = k,\ 1 \leq n \leq N,$$

be a uniform partition of $[0,T]$. Then the fully discrete scheme based on the second-order backward differentiation formula in time and qualocation in space takes the form: find $U_h^n \in S_h^0,\ 1 \leq n \leq N$, such that $U_h^0 = u_{0h}$, and

$$\left\langle D_t U_h^n, v\right\rangle - \left\langle \mathscr{A} U_h^n, v\right\rangle = \left\langle \mathscr{B}_k(t_n) U_h^n, v\right\rangle + \langle f^n, v\rangle \ \ \forall\ v \in T_h, \tag{29}$$

where

$$f^n(x) := f(x,t_n),$$

$$\mathscr{B}_k(t_n)U_h^n := k\sum_{j=0}^{n} w_j B(t_n,t_j)U_h^j, \qquad w_j := \begin{cases} 0.5, & \text{if } j = 0, n, \\ 1, & \text{if } 1 \leq j \leq n-1, \end{cases}$$

$$\text{and}\ \ D_t U_h^n := \begin{cases} \bar{\partial}_t U_h^n = \dfrac{U_h^n - U_h^{n-1}}{k} & \text{if } n = 1, \\ \dfrac{3}{2}\bar{\partial}_t U_h^n - \dfrac{1}{2}\bar{\partial}_t U_h^{n-1} = \dfrac{3U_h^n - 4U_h^{n-1} + U_h^{n-2}}{2k} & \text{if } n \geq 2. \end{cases}$$

At each time step, the discrete problem (29) gives rise to a system of linear algebraic equations which is easily shown to be nonsingular. Thus the solution of (29) is unique.

5.1 Error Estimates for BDF2 Scheme

Since $T_h = \{\chi_{xx} :\ \chi \in S_h^0\}$, we can rewrite (29) as

$$\left\langle D_t U_h^n, \chi_{xx}\right\rangle - \left\langle \mathscr{A} U_h^n, \chi_{xx}\right\rangle = \left\langle \mathscr{B}_k(t_n) U_h^n, \chi_{xx}\right\rangle + \langle f^n, \chi_{xx}\rangle\,,\ \forall\ \chi \in S_h^0. \tag{30}$$

If $u^n(x) := u(x, t_n),\ 0 \le n \le N$, then, for $2 \le n \le N$, (1a)–(1c) yields

$$\begin{aligned}\langle D_t u^n, \chi_{xx}\rangle - \langle \mathscr{A} u^n, \chi_{xx}\rangle = \langle \mathscr{B}_k(t_n)u^n, \chi_{xx}\rangle + \langle f^n, \chi_{xx}\rangle \\ - \langle \tau^n(u), \chi_{xx}\rangle + \langle \varepsilon^n(B(t_n, \cdot)u), \chi_{xx}\rangle\,, \\ \forall\ \chi \in S_h^0,\end{aligned} \tag{31}$$

where

$$\varepsilon^n(\phi) := \int_0^{t_n} \phi(t)dt - k\sum_{j=0}^{n} w_j \phi(t_j),$$

and

$$\tau^n(\phi) := \phi_t(t_n) - D_t\phi(t_n).$$

Furthermore, using Taylor series, we obtain, for $\phi \in W^{2,1}(J)$,

$$\varepsilon^n(\phi) = -\frac{1}{2}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} (t - t_{j-1})(t_j - t)\phi_{tt}(t)dt, \tag{32}$$

and for $\phi \in W^{3,1}(J)$,

$$\tau^n(\phi) = \begin{cases} \dfrac{1}{k}\displaystyle\int_{t_{n-1}}^{t_n} (t - t_{n-1})\phi_{tt}(t)dt, & \text{if } n = 1, \\ \dfrac{1}{k}\displaystyle\int_{t_{n-1}}^{t_n} (t - t_{n-1})^2\phi_{ttt}(t)dt - \dfrac{1}{4k}\displaystyle\int_{t_{n-2}}^{t_n} (t - t_{n-2})^2\phi_{ttt}(t)dt, & \text{if } n \ge 2. \end{cases} \tag{33}$$

We now write

$$e_h^n = u(t_n) - U_h^n = \big(u(t_n) - \tilde{u}(t_n)\big) - \big(\tilde{u}(t_n) - U_h^n\big) = \eta^n - \Theta^n.$$

On subtracting (30) from (31) and using (10) at $t = t_n$, we have

$$\begin{aligned}- \langle D_t\Theta^n, \chi_{xx}\rangle + \langle \mathscr{A}\Theta^n, \chi_{xx}\rangle = - \langle \mathscr{B}_k(t_n)\Theta^n, \chi_{xx}\rangle + \langle \sigma^n, \chi_{xx}\rangle\,, \\ \forall\ \chi \in S_h^0,\end{aligned} \tag{34}$$

where

$$\sigma^n := -\tau^n(u) + \varepsilon^n(B(t_n, \cdot)u) - \varepsilon^n(B(t_n, \cdot)\eta) - D_t\eta^n.$$

Lemma 5 *Let $U_h^0 = \tilde{u}(x,0)$. Then there exists a positive constant k_0 such that for $0 < k \le k_0$,*

$$\|\Theta_x^n\|^2 + k\sum_{j=1}^{n} \|\Theta_{xx}^j\|^2 \le C\Big[k^4 \left(\|u\|^2_{W^{2,\infty}(0,\ k;\ W^{1,\infty})} + \|u\|^2_{H^3(0,\ t_n;\ L^\infty)} + \|u\|^2_{H^2(0,\ t_n;\ W^{2,\infty})}\right) + k^4h^4\|u\|^2_{H^2(0,\ t_n;\ W^{6,\infty})} + h^8\|u\|^2_{H^1(0,\ t_n;\ W^{6,\infty})}\Big],\ 1 \le n \le N.$$

Proof Let $n \ge 2$ and set $\chi = \Theta^n$ in (34). Then

$$\begin{aligned} -\left\langle D_t\Theta^n, \Theta_{xx}^n\right\rangle + \left\langle \Theta_{xx}^n, \Theta_{xx}^n\right\rangle &= -\left\langle \mathscr{B}_k(t_n)\Theta^n, \Theta_{xx}^n\right\rangle + \langle b\Theta_x^n + c\Theta^n, \Theta_{xx}^n\rangle + \left\langle \sigma^n, \Theta_{xx}^n\right\rangle \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{35}$$

From Lemma 1 and the relation

$$\begin{aligned} 2(3a-4b+c, a) &= a^2 - b^2 + (2a-b)^2 - (2b-c)^2 + (a-2b+c)^2 \\ &\ge a^2 - b^2 + (2a-b)^2 - (2b-c)^2, \quad a, b, c \in \mathbb{R}, \end{aligned}$$

the first term on the left hand side of (35) can be estimated as

$$\begin{aligned} -\left\langle D_t\Theta^n, \Theta_{xx}^n\right\rangle &= (D_t\Theta_x^n, \Theta_x^n) + \frac{1}{1080}\sum_{i=1}^{M} h_i^5 (D_t\Theta_{xxx,i}^n)(\Theta_{xxx,i}^n) \\ &\ge \frac{1}{4}\bar{\partial}_t \left(\|\Theta_x^n\|^2 + \|2\Theta_x^n - \Theta_x^{n-1}\|^2\right) \\ &\quad + \frac{1}{1080}\sum_{i=1}^{M} \frac{h_i^5}{4}\bar{\partial}_t \left(|\Theta_{xxx,i}^n|^2 + |2\Theta_{xxx,i}^n - \Theta_{xxx,i}^{n-1}|^2\right). \end{aligned}$$

On multiplying by k on both sides and then summing from $n = 2$ to $m \le N$, we obtain

$$\begin{aligned} -k\sum_{n=2}^{m} \left\langle D_t\Theta^n, \Theta_{xx}^n\right\rangle &\ge \frac{1}{4}\left(\|\Theta_x^m\|^2 + \|2\Theta_x^m - \Theta_x^{m-1}\|^2\right) \\ &\quad + \frac{1}{1080}\sum_{i=1}^{M} \frac{h_i^5}{4}\left(|\Theta_{xxx,i}^m|^2 + |2\Theta_{xxx,i}^m - \Theta_{xxx,i}^{m-1}|^2\right) \\ &\quad - \frac{5}{4}\Big(\|\Theta_x^1\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta_{xxx,i}^1|^2\Big) \\ &\ge \frac{1}{4}\|\Theta_x^m\|^2 - \frac{5}{4}\Big(\|\Theta_x^1\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta_{xxx,i}^1|^2\Big). \end{aligned}$$

For the second term on right hand side of (35), we have

$$\langle \Theta^n_{xx}, \Theta^n_{xx} \rangle = [[\Theta^n_{xx}]]^2 = \|\Theta^n_{xx}\|^2.$$

To estimate the terms I_1 and I_2, the use of (8) and (9) together with the Poincaré inequality for $\Theta^n \in S_h^0$ and Young's inequality yields

$$\begin{aligned}
|I_1| &\le |\langle k\sum_{j=1}^{n} w_j B(t_n, t_j)\Theta^j, \Theta^n_{xx}\rangle| \\
&\le C(\epsilon)\left(k(\|\Theta^1_{xx}\|^2 + \|\Theta^1_x\|^2) + k\sum_{j=2}^{n}(\|\Theta^j_{xx}\|^2 + \|\Theta^j_x\|^2)\right) + \epsilon\|\Theta^n_{xx}\|^2, \\
|I_2| &\le C(\epsilon)\|\Theta^n_x\|^2 + \epsilon\|\Theta^n_{xx}\|^2.
\end{aligned}$$

For I_3, using Young's inequality, we have

$$|I_3| \le C(\epsilon)\|\sigma^n\|^2_{L^\infty} + \epsilon\|\Theta^n_{xx}\|^2.$$

On combining these estimates in (35) and summing from $n = 2$ to $m \le N$, it follows that

$$\begin{aligned}
\|\Theta^m_x\|^2 + (4 - 12\epsilon)k\sum_{n=2}^{m}\|\Theta^n_{xx}\|^2 \le C(\epsilon)\Bigg(&\|\Theta^1_x\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta^1_{xxx,i}|^2 + k\|\Theta^1_{xx}\|^2 \\
&+ k\sum_{n=2}^{m}\|\sigma^n\|^2_{L^\infty} + k\sum_{n=2}^{m}\Big(\|\Theta^n_x\|^2 + k\sum_{j=2}^{n}\|\Theta^j_{xx}\|^2\Big)\Bigg).
\end{aligned}$$

Choose $\epsilon = \frac{1}{4}$ and $k_1 > 0$ so that $1 - kC(\epsilon) > 0$, for $0 < k \le k_1$. Then, for $0 < k \le k_1$, it follows that, for $2 \le m \le N$,

$$\begin{aligned}
\|\Theta^m_x\|^2 + k\sum_{n=2}^{m}\|\Theta^n_{xx}\|^2 \le C\Big[&\|\Theta^1_x\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta^1_{xxx,i}|^2 + k\|\Theta^1_{xx}\|^2 + k\sum_{n=2}^{m}\|\sigma^n\|^2_{L^\infty} \\
&+ k\sum_{n=2}^{m-1}\big(\|\Theta^n_x\|^2 + k\sum_{j=2}^{n}\|\Theta^j_{xx}\|^2\big)\Big],
\end{aligned}$$

where we have used the summation convention $\sum_{n=M}^{N} = 0$ for $N < M$. An application of Gronwall's Lemma then yields, for $2 \le m \le N$,

$$\begin{aligned}&\|\Theta_x^m\|^2 + k\sum_{n=2}^{m}\|\Theta_{xx}^n\|^2\\ &\quad\le C\left(\|\Theta_x^1\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta_{xxx,i}^1|^2 + k\|\Theta_{xx}^1\|^2 + k\sum_{n=2}^{m}\|\sigma^n\|_{L^\infty}^2\right).\end{aligned} \tag{36}$$

To estimate the first two terms on right hand side, let $n = 1$ and $\chi = \Theta^1$ in (34). Then

$$\begin{aligned}-\left\langle D_t\Theta^1, \Theta_{xx}^1\right\rangle + \|\Theta_{xx}^1\|^2 &= -\left\langle \mathscr{B}_k(t_1)\Theta^1, \Theta_{xx}^1\right\rangle + \langle b\Theta_x^1 + c\Theta^1, \Theta_{xx}^1\rangle\\ &\quad + \left\langle \sigma^1, \Theta_{xx}^1\right\rangle.\end{aligned} \tag{37}$$

Note that

$$\begin{aligned}-\left\langle D_t\Theta^1, \Theta_{xx}^1\right\rangle &= -\left\langle \bar{\partial}_t\Theta^1, \Theta_{xx}^1\right\rangle = -\frac{1}{k}\left\langle \Theta^1, \Theta_{xx}^1\right\rangle\\ &= \frac{1}{k}\left(\|\Theta_x^1\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta_{xxx,i}^1|^2\right),\end{aligned}$$

$$-\left\langle \mathscr{B}_k(t_1)\Theta^1, \Theta_{xx}^1\right\rangle = -\left\langle \frac{k}{2}B(t_1, t_1)\Theta^1, \Theta_{xx}^1\right\rangle \le Ck(\|\Theta_{xx}^1\|^2 + \|\Theta_x^1\|^2),$$

and

$$\begin{aligned}\left\langle \sigma^1, \Theta_{xx}^1\right\rangle &= \left\langle -\tau^1(u), \Theta_{xx}^1\right\rangle + \left\langle \varepsilon^1(B(t_1,\cdot)u) - \varepsilon^1(B(t_1,\cdot)\eta) - D_t\eta^1, \Theta_{xx}^1\right\rangle\\ &\le C\|\left(\tau^1(u)\right)_x\|_{L^\infty}\|\Theta_x^1\|\\ &\quad + C\|\varepsilon^1(B(t_1,\cdot)u) - \varepsilon^1(B(t_1,\cdot)\eta) - D_t\eta^1\|_{L^\infty}\|\Theta_{xx}^1\|.\end{aligned} \tag{38}$$

To estimate the first term on right hand side of (38), we use

$$\begin{aligned}|\langle\tau^1(u), \Theta_{xx}^1\rangle| &= \left|\langle\tau^1(u) - I_h(\tau^1(u)), \Theta_{xx}^1\rangle + \langle I_h(\tau^1(u)), \Theta_{xx}^1\rangle\right|\\ &= \left|\sum_{i=1}^{M}\langle\tau^1(u) - I_h(\tau^1(u)), \Theta_{xx}^1\rangle_i - \sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}\left(\frac{\partial}{\partial x}I_h(\tau^1(u))\right)\Theta_x^1 dx\right|\\ &\le C\sum_{i=1}^{M}\left(h_i\|(\tau^1(u))_x\|_{L^\infty(I_i)}\|\Theta_{xx}^1\|_{L^2(I_i)} + \left\|\frac{\partial}{\partial x}I_h(\tau^1(u))\right\|_{L^\infty(I_i)}\|\Theta_x^1\|_{L^2(I_i)}\right),\end{aligned}$$

where $I_h : C[0,1] \to T_h$ is the piecewise linear interpolation operator. Using the inverse inequality yields

$$|\langle \tau^1(u), \Theta^1_{xx}\rangle| \le C\|(\tau^1(u))_x\|_{L^\infty}\|\Theta^1_x\|.$$

By using Young's inequality in (38), we have

$$\begin{aligned}\left\langle \sigma^1, \Theta^1_{xx}\right\rangle &\le \frac{\epsilon}{k}\|\Theta^1_x\|^2 + \epsilon\|\Theta^1_{xx}\|^2\\ &\quad + C(\epsilon)\left(k\|\left(\tau^1(u)\right)_x\|^2_{L^\infty} + \left\|\varepsilon^1(B(t_1,\cdot)u) - \varepsilon^1(B(t_1,\cdot)\eta) - \frac{1}{k}\int_0^k \eta_s(s)ds\right\|^2_{L^\infty}\right).\end{aligned}$$

On combining the above estimates in (37) with $\epsilon = \frac{1}{4}$, there exists $k_2 > 0$ such that, for $0 < k \le k_2$,

$$\begin{aligned}\|\Theta^1_x\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta^1_{xxx,i}|^2 + k\|\Theta^1_{xx}\|^2 &\le Ck\Bigg(k\|\left(\tau^1(u)\right)_x\|^2_{L^\infty}\\ &\quad + \left\|\varepsilon^1(B(t_1,\cdot)u) - \varepsilon^1(B(t_1,\cdot)\eta) - \frac{1}{k}\int_0^k \eta_s(s)ds\right\|^2_{L^\infty}\Bigg)\end{aligned}$$

Use of (32), (33) and Lemma 2 yields

$$\begin{aligned}&\|\Theta^1_x\|^2 + \frac{1}{1080}\sum_{i=1}^{M} h_i^5|\Theta^1_{xxx,i}|^2 + k\|\Theta^1_{xx}\|^2\\ &\quad\le Ck\Bigg(k^3\|u_{tt}\|^2_{L^\infty(0,\,k;\,W^{1,\infty})} + k^5\|u_{tt}\|^2_{L^2(0,\,k;\,W^{2,\infty})} + k^5h^4\sum_{i=0}^{2}\left\|\frac{\partial^i u}{\partial t^i}\right\|^2_{L^2(0,\,k;\,W^{6,\infty})}\\ &\qquad + h^8\left(\|u_t\|^2_{L^\infty(0,\,k;\,W^{6,\infty})} + \|u\|^2_{L^2(0,\,k;\,W^{6,\infty})}\right)\Bigg) \qquad (39)\end{aligned}$$

Using (32), (33) and Lemma 2, the last term in (36) can be estimated as follows:

$$\begin{aligned}k\sum_{n=2}^{m}\|\sigma^n\|^2_{L^\infty} &\le C\Bigg(k^4\|u_{ttt}\|^2_{L^2(0,\,t_m;\,L^\infty)} + k^4\|u_{tt}\|^2_{L^2(0,\,t_m;\,W^{2,\infty})}\\ &\quad + k^4h^4\sum_{i=0}^{2}\left\|\frac{\partial^i u}{\partial t^i}\right\|^2_{L^2(0,\,t_m;\,W^{6,\infty})} + h^8\sum_{i=0}^{1}\left\|\frac{\partial^i u}{\partial t^i}\right\|^2_{L^2(0,\,t_m;\,W^{6,\infty})}\Bigg) \qquad (40)\end{aligned}$$

Combining (36), (39) and (40) completes the proof with $k_0 = \min(k_1, k_2)$. □

Lemma 6 *With $U_h^0 = \tilde{u}(x,0)$, there exists a positive constant k_0 such that for $0 < k \leq k_0$*

$$k\sum_{j=1}^{n} \|\bar{\partial}_t \Theta_{hx}^j\|^2 + \|\Theta_{hxx}^n\|^2 \leq C\Bigg[k\|\bar{\partial}_t \Theta_x^1\|^2 + k\sum_{i=1}^{M} h_i^5 |\bar{\partial}_t \Theta_{xxx,i}^1|^2 + \|\Theta_{xx}^1\|^2$$
$$+ k^4\Big(\|u\|^2_{W^{2,\infty}(0,\, k;\, W^{1,\infty})} + \|u\|^2_{W^{3,\infty}(0,\, t_n;\, L^\infty)}$$
$$+ \|u\|^2_{H^4(0,\, t_n;\, L^\infty)} + \|u\|^2_{H^3(0,\, t_n;\, W^{2,\infty})}\Big) + k^4 h^4 \|u\|^2_{H^2(0,\, t_n;\, W^{6,\infty})}$$
$$+ h^8\Big(\|u\|^2_{W^{1,\infty}(0,\, t_n;\, W^{6,\infty})} + \|u\|^2_{H^2(0,\, t_n;\, W^{6,\infty})}\Big)\Bigg], \quad 1 \leq n \leq N.$$

Proof With $\chi = k\bar{\partial}_t \Theta^n = \Theta^n - \Theta^{n-1}$ in (34) and summing from $n = 2$ to $m \leq N$, we obtain

$$-k\sum_{n=2}^{m} \langle D_t \Theta^n, \bar{\partial}_t \Theta_{xx}^n \rangle + k\sum_{n=2}^{m} \langle \Theta_{xx}^n, \bar{\partial}_t \Theta_{xx}^n \rangle = -k\sum_{n=2}^{m} \langle \mathscr{B}_k(t_n)\Theta^n, \bar{\partial}_t \Theta_{xx}^n \rangle$$
$$-k\sum_{n=2}^{m} \langle b\Theta_x^n + c\Theta^n, \bar{\partial}_t \Theta_{xx}^n \rangle + k\sum_{n=3}^{m} \langle \sigma^n, \bar{\partial}_t \Theta_{xx}^n \rangle + \langle \sigma^2, \Theta_{xx}^2 - \Theta_{xx}^1 \rangle.$$

Using the summation by parts formula,

$$k\sum_{n=M}^{N} A_n(\bar{\partial}_t B_n) = -k\sum_{n=M}^{N} (\bar{\partial}_t A_n) B_{n-1} + A_N B_M - A_{M-1} B_{M-1} \;:\; A_n, \quad B_n \in \mathbb{R},$$

on the right hand side yields

$$\begin{aligned}
&-k\sum_{n=2}^{m} \langle D_t \Theta^n, \bar{\partial}_t \Theta_{xx}^n \rangle + k\sum_{n=2}^{m} \langle \Theta_{xx}^n, \bar{\partial}_t \Theta_{xx}^n \rangle \\
&\quad = \Big(k\sum_{n=2}^{m} \langle \bar{\partial}_t(\mathscr{B}_k(t_n)\Theta^n), \Theta_{xx}^{n-1} \rangle - \langle \mathscr{B}_k(t_m)\Theta^m, \Theta_{xx}^m \rangle + \langle \mathscr{B}_k(t_1)\Theta^1, \Theta_{xx}^1 \rangle \Big) \\
&\quad + \Big(k\sum_{n=2}^{m} \langle \bar{\partial}_t(b\Theta_x^n + c\Theta^n), \Theta_{xx}^{n-1} \rangle - \langle b\Theta_x^m + c\Theta^m, \Theta_{xx}^m \rangle + \langle b\Theta_x^1 + c\Theta^1, \Theta_{xx}^1 \rangle \Big) \\
&\quad - \Big(k\sum_{n=3}^{m} \langle \bar{\partial}_t \sigma^n, \Theta_{xx}^{n-1} \rangle + \langle \sigma^m, \Theta_{xx}^m \rangle - \langle \sigma^2, \Theta_{xx}^1 \rangle \Big) \\
&\quad = I_1 + I_2 + I_3.
\end{aligned} \tag{41}$$

Using Lemma 1 and the relation

$$2(a-b,a)=a^2-b^2+(a-b)^2\geq a^2-b^2,\ a,b\in\mathbb{R}, \tag{42}$$

we find that

$$\begin{aligned}-\left\langle D_t\Theta^n,\bar{\partial}_t\Theta^n_{xx}\right\rangle &= (D_t\Theta^n_x,\bar{\partial}_t\Theta^n_x)+\frac{1}{1080}\sum_{i=1}^{M}h_i^5(D_t\Theta^n_{xxx,i})(\bar{\partial}_t\Theta^n_{xxx,i})\\ &= \|\bar{\partial}_t\Theta^n_x\|^2+\frac{1}{2}(\bar{\partial}_t\Theta^n_x-\bar{\partial}_t\Theta^{n-1}_x,\bar{\partial}_t\Theta^n_x)\\ &\quad+\frac{1}{1080}\sum_{i=1}^{M}h_i^5\left(|\bar{\partial}_t\Theta^n_{xxx,i}|^2+\frac{1}{2}(\bar{\partial}_t\Theta^n_{xxx,i}-\bar{\partial}_t\Theta^{n-1}_{xxx,i},\bar{\partial}_t\Theta^n_{xxx,i})\right)\\ &\geq \frac{5}{4}\|\bar{\partial}_t\Theta^n_x\|^2-\frac{1}{4}\|\bar{\partial}_t\Theta^{n-1}_x\|^2\\ &\quad+\frac{1}{1080}\sum_{i=1}^{M}h_i^5\left(\frac{5}{4}|\bar{\partial}_t\Theta^n_{xxx,i}|^2-\frac{1}{4}|\bar{\partial}_t\Theta^{n-1}_{xxx,i}|^2\right).\end{aligned}$$

Hence

$$-k\sum_{n=2}^{m}\left\langle D_t\Theta^n,\bar{\partial}_t\Theta^n_{xx}\right\rangle\geq k\sum_{n=2}^{m}\|\bar{\partial}_t\Theta^n_x\|^2-\frac{1}{4}k\Big(\|\bar{\partial}_t\Theta^1_x\|^2+\frac{1}{1080}\sum_{i=1}^{M}h_i^5|\bar{\partial}_t\Theta^1_{xxx,i}|^2\Big).$$

A use of (8) with (42) shows

$$\begin{aligned}k\sum_{n=2}^{m}\langle\Theta^n_{xx},\bar{\partial}_t\Theta^n_{xx}\rangle &\geq \frac{1}{2}\sum_{n=2}^{m}\left([[\Theta^n_{xx}]]^2-[[\Theta^{n-1}_{xx}]]^2\right)=\frac{1}{2}\left([[\Theta^m_{xx}]]^2-[[\Theta^1_{xx}]]^2\right)\\ &=\frac{1}{2}\left(\|\Theta^m_{xx}\|^2-\|\Theta^1_{xx}\|^2\right).\end{aligned}$$

To estimate $I_1,\ I_2$ and I_3, we use Young's inequality (28) together with (8) and (9) to obtain

$$\begin{aligned}|I_1| &\leq Ck\sum_{n=2}^{m}(\|\bar{\partial}_t(\mathscr{B}_k(t_n)\Theta^n)\|^2+\|\Theta^{n-1}_{xx}\|^2)+C(\epsilon)\|\mathscr{B}_k(t_m)\Theta^m\|^2+\epsilon\|\Theta^m_{xx}\|^2\\ &\quad+Ck(\|\Theta^1_{xx}\|^2+\|\Theta^1_x\|^2)\\ &\leq Ck\sum_{n=2}^{m}\Big(k\sum_{j=1}^{n-1}(\|\Theta^j_{xx}\|^2+\|\Theta^j_x\|^2)+\|\Theta^{n-1}_{xx}\|^2+\|\Theta^{n-1}_x\|^2+\|\Theta^n_{xx}\|^2+\|\Theta^n_x\|^2\Big)\end{aligned}$$

$$
\begin{aligned}
&+ C(\epsilon)k\sum_{n=1}^{m}(\|\Theta^n_{xx}\|^2 + \|\Theta^n_x\|^2) + \epsilon\|\Theta^m_{xx}\|^2 + Ck(\|\Theta^1_x\|^2 + \|\Theta^1_{xx}\|^2)\\
&\le C(\epsilon)\Big(k\|\Theta^1_{xx}\|^2 + k\|\Theta^1_x\|^2 + k\sum_{n=2}^{m}\|\Theta^n_x\|^2\Big) + (\epsilon + C(\epsilon)k)\|\Theta^m_{xx}\|^2,
\end{aligned}
$$

$$
\begin{aligned}
|I_2| &\le C(\epsilon)\Big(k\|\Theta^1_{xx}\|^2 + k\sum_{n=2}^{m-1}\|\Theta^n_{xx}\|^2\Big) + \epsilon k\sum_{n=2}^{m}\|\bar{\partial}_t\Theta^n_x\|^2\\
&\quad + C(\epsilon)\|\Theta^m_x\|^2 + \epsilon\|\Theta^m_{xx}\|^2 + Ck(\|\bar{\partial}_t\Theta^1_x\|^2 + \|\Theta^1_{xx}\|^2)\\
&\le C(\epsilon)\Big(k\|\bar{\partial}_t\Theta^1_x\|^2 + k\|\Theta^1_{xx}\|^2 + k^2\|\bar{\partial}_t\Theta^m_x\|^2 + k\sum_{n=2}^{m-1}\|\Theta^n_{xx}\|^2\Big)\\
&\quad + \epsilon\|\Theta^m_{xx}\|^2 + \epsilon k\sum_{n=2}^{m}\|\bar{\partial}_t\Theta^n_x\|^2.
\end{aligned}
$$

and

$$
|I_3| \le C(\epsilon)\Big(\|\Theta^1_{xx}\|^2 + \|\sigma^2\|^2_{L^\infty} + \|\sigma^m\|^2_{L^\infty} + k\sum_{n=3}^{m}\|\bar{\partial}_t\sigma^n\|^2_{L^\infty} + k\sum_{n=2}^{m-1}\|\Theta^n_{xx}\|^2\Big) + \epsilon\|\Theta^m_{xx}\|^2.
$$

On combining these estimates in (41) with $\epsilon = \frac{1}{12}$, it follows that

$$
\begin{aligned}
k\sum_{n=2}^{m}\|\bar{\partial}_t\Theta^n_x\|^2 + (1 - Ck)\|\Theta^m_{xx}\|^2 &\le C\Big(k\|\bar{\partial}_t\Theta^1_x\|^2 + k\sum_{i=1}^{M}h_i^5|\bar{\partial}_t\Theta^1_{xxx,i}|^2 + \|\Theta^1_{xx}\|^2\\
&\quad + \|\Theta^m_x\|^2 + k\sum_{n=2}^{m}\|\Theta^n_x\|^2 + k\sum_{n=2}^{m-1}\|\Theta^n_{xx}\|^2\\
&\quad + \|\sigma^2\|^2_{L^\infty} + \|\sigma^m\|^2_{L^\infty} + k\sum_{n=3}^{m}\|\bar{\partial}_t\sigma^n\|^2_{L^\infty}\Big)\\
&\le C\Big(k\|\bar{\partial}_t\Theta^1_x\|^2 + k\sum_{i=1}^{M}h_i^5|\bar{\partial}_t\Theta^1_{xxx,i}|^2 + \|\Theta^1_{xx}\|^2 + \max_{2\le n\le m}\|\Theta^n_x\|^2\\
&\quad + \max_{2\le n\le m}\|\sigma^n\|^2_{L^\infty} + k\sum_{n=3}^{m}\|\bar{\partial}_t\sigma^n\|^2_{L^\infty} + k\sum_{n=2}^{m-1}\|\Theta^n_{xx}\|^2\Big).
\end{aligned}
$$

There exists a number $k_0 > 0$ such that $\forall k \le k_0$, $1 - Ck > 0$ and hence,

$$k\sum_{n=2}^{m} \|\bar{\partial}_t \Theta_x^n\|^2 + \|\Theta_{xx}^m\|^2 \le C\Big(k\|\bar{\partial}_t \Theta_x^1\|^2 + k\sum_{i=1}^{M} h_i^5 |\bar{\partial}_t \Theta_{xxx,i}^1|^2 + \|\Theta_{xx}^1\|^2 + \max_{2\le n\le m} \|\Theta_x^n\|^2$$
$$+ \max_{2\le n\le m} \|\sigma^n\|_{L^\infty}^2 + k\sum_{n=3}^{m} \|\bar{\partial}_t \sigma^n\|_{L^\infty}^2 + k\sum_{n=2}^{m-1} \|\Theta_{xx}^n\|^2\Big).$$

Apply Gronwall's Lemma to obtain

$$k\sum_{n=2}^{m} \|\bar{\partial}_t \Theta_x^n\|^2 + \|\Theta_{xx}^n\|^2 \le C\Big(k\|\bar{\partial}_t \Theta_x^1\|^2 + k\sum_{i=1}^{M} h_i^5 |\bar{\partial}_t \Theta_{xxx,i}^1|^2 + \|\Theta_{xx}^1\|^2 + \max_{2\le n\le m} \|\Theta_x^n\|^2$$
$$+ \max_{2\le n\le m} \|\sigma^n\|_{L^\infty}^2 + k\sum_{n=3}^{m} \|\bar{\partial}_t \sigma^n\|_{L^\infty}^2\Big). \tag{43}$$

Now, using (32), (33) and Lemma 2, one can easily derive the following estimates:

$$\|\sigma^n\|_{L^\infty}^2 \le C\Bigg(k^4 \|u_{ttt}\|_{L^\infty(t_{n-2},\, t_n;\, L^\infty)}^2 + k^4 \|u_{tt}\|_{L^2(0,\, t_n;\, W^{2,\infty})}^2$$
$$+ k^4 h^4 \sum_{i=0}^{2} \left\| \frac{\partial^i u}{\partial t^i} \right\|_{L^2(0,\, t_n;\, W^{6,\infty})}^2$$
$$+ h^8 \left(\sum_{i=0}^{1} \left\| \frac{\partial^i u}{\partial t^i} \right\|_{L^\infty(t_{n-2},\, t_n;\, W^{6,\infty})}^2 + \|u\|_{L^2(0,\, t_n;\, W^{6,\infty})}^2 \right)\Bigg) \tag{44}$$

and

$$k\sum_{n=3}^{m} \|\bar{\partial}_t \sigma^n\|_{L^\infty}^2 \le C\Bigg(k^4 \|u_{tttt}\|_{L^2(0,\, t_m;\, L^\infty)}^2 + k^4 \|u_{ttt}\|_{L^2(0,\, t_m;\, W^{2,\infty})}^2$$
$$+ k^4 h^4 \sum_{i=0}^{2} \left\| \frac{\partial^i u}{\partial t^i} \right\|_{L^2(0,\, t_m;\, W^{6,\infty})}^2 + h^8 \sum_{i=0}^{2} \left\| \frac{\partial^i u}{\partial t^i} \right\|_{L^2(0,\, t_m;\, W^{6,\infty})}^2 \Bigg). \tag{45}$$

Thus, the use of Lemma 5, (44) and (45) in (43) yields the required result. □

Finally, Lemmas 2, 5 and 6 with (39) give the following convergence estimate for the fully discrete scheme.

Theorem 2 *Let* $u \in H^2(0, T; W^{6,\infty}(I)) \cap H^3(0, T; W^{2,\infty}(I)) \cap H^4(0, T; L^\infty(I))$ *and* U_h^n, $1 \le n \le N$, *be the solution of (29) with* $U_h^0 = u_{0h} = \tilde{u}(x, 0)$. *Then there exists a positive constant* k_0 *such that for* $0 < k \le k_0$,

$$\max_{1\le n\le N} \|u(\cdot, t_n) - U_h^n\|_{H^m(I)} \le C(k^2 + h^{4-m}),\ m = 0, 1,$$

and

$$\max_{1\le n\le N} \|u(\cdot,t_n) - U_h^n\|_{H^2(I)} \le C(k^{1+\ell} + h^2).$$

Moreover,

$$\max_{1\le n\le N} \|u(\cdot,t_n) - U_h^n\|_{L^\infty(I)} \le C(k^2 + h^4),$$

and

$$\max_{1\le n\le N} \|u(\cdot,t_n) - U_h^n\|_{W^{1,\infty}(I)} \le C(k^{1+\ell} + h^3),$$

where $\ell = \frac{1}{2}$, however, if the solution at first time level, i.e. U_h^1, is obtained by using Crank-Nicolson method instead of backward Euler then the above estimates hold for $\ell = 1$.

6 Numerical Experiments

In this section, we present numerical results obtained using the fully discrete scheme (29). In our test problem,

$$b(x) = x, \quad c(x) = e^{-x}, \quad b_1(x;t,s) = -xe^{2s-t}, \quad c_1(x;t,s) = e^{-x+2s-t},$$

and the forcing function $f(x,t)$ is chosen so that $u(x,t) = x(x-1)e^{x-t}$ is the exact solution of (1a)–(1c). In addition, consider uniform partitions

$$0 = x_0 < x_1 < \cdots < x_M = 1;\ x_i - x_{i-1} = h,\ 1 \le i \le M,$$

and

$$0 = t_0 < t_1 < \cdots < t_N = T;\ t_n - t_{n-1} = k,\ 1 \le n \le N,$$

of $I = (0,1)$ and $J = (0,T],\ T = 1$, respectively, and select $k = h^2$ (i.e., $N = M^2$) to verify the rate of convergence with respect to the norms

$$\max_{1\le n\le N} \|u(\cdot,t_n) - U_h^n\|_m \approx \max_{1\le n\le N} \sqrt{\left(\sum_{s=0}^{m}\sum_{i=1}^{M}\frac{h_i}{2}\sum_{l=1}^{4} w_l \left(\frac{\partial^s}{\partial x^s}\left(u(\xi_{i,l},t_n) - U_h^n(\xi_{i,l})\right)\right)^2\right)}$$
$$:= \mathscr{E}_h^m;\quad m = 0,1,2,$$

$$\max_{1\le n\le N}\|u(\cdot,t_n)-U_h^n\|_{m,\infty}\approx\max_{1\le n\le N}\sum_{s=0}^{m}\left(\max_{\substack{0\le i\le M\\0\le j\le 100}}\left|\frac{\partial^s}{\partial x^s}\big(u(x_{i,j},t_n)-U_h^n(x_{i,j})\big)\right|\right)$$
$$:=\mathscr{E}_h^{m,\infty};\ \ m=0,1,$$

and

$$\max_{\substack{1\le n\le N\\0\le i\le M}}\left|\frac{\partial^m}{\partial x^m}\big(u(x_i,t_n)-U_h^n(x_i)\big)\right|:=\mathscr{E}_h^{m,\text{nodal}};\ \ m=0,1.$$

where

$$x_{i,0}=x_i,\ 0\le i\le M,\quad x_{i,j}-x_{i,j-1}=\frac{h_{i+1}}{100},\ 0\le i\le M-1,\ 1\le j\le 100,$$
$$\xi_{i,l}:=\frac{x_{i-1}+x_i}{2}+\frac{h_i}{2}\zeta_l,$$

and

$l=$	1	2	3	4
$\zeta_l=$	$\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}$	$-\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}$	$\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}$	$-\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}$
$w_l=$	$\frac{18+\sqrt{30}}{36}$	$\frac{18+\sqrt{30}}{36}$	$\frac{18-\sqrt{30}}{36}$	$\frac{18-\sqrt{30}}{36}$

The corresponding convergence rate is determined from the formula

$$\mathscr{R}^m:=\log_2\left(\frac{\mathscr{E}_h^m}{\mathscr{E}_{h/2}^m}\right),\ m=0,1,2.$$

The rates of convergence corresponding to $\mathscr{E}_h^{m,\text{nodal}}$ and $\mathscr{E}_h^{m,\infty}$ are denoted by $\mathscr{R}^{m,\text{nodal}}$ and $\mathscr{R}^{m,\infty}$, respectively.

To approximate the integrals involved in the error term $\max_{1\le n\le N}\|u(\cdot,t_n)-U_h^n\|_m$ without degrading the actual rate of convergence, we use a 4-point Gauss quadrature formula the nodes and weights of which are given in preceding table.

In Table 1 we present the errors in various norms together with corresponding convergence rates. These results support the estimates given in Theorem 2. The third block of Table 2 shows the superconvergence in first spatial derivative at nodal points.

Table 1 Errors and corresponding rates of convergence

$M \downarrow$	$\mathscr{E}_h^0$	$\mathscr{R}^0$	$\mathscr{E}_h^1$	$\mathscr{R}^1$	$\mathscr{E}_h^2$	$\mathscr{R}^2$	$\mathscr{E}_h^{0,\infty}$	$\mathscr{R}^{0,\infty}$	$\mathscr{E}_h^{1,\infty}$	$\mathscr{R}^{1,\infty}$
8	4.50e−05	0.0	3.03e−04	0.0	1.38e−02	0.0	6.68e−05	0.0	9.31e−04	0.0
16	3.49e−06	3.7	3.52e−05	3.1	3.46e−03	2.0	5.22e−06	3.7	1.08e−04	3.1
32	2.38e−07	3.9	4.24e−06	3.1	8.63e−04	2.0	3.56e−07	3.9	1.24e−05	3.1
64	1.54e−08	4.0	5.23e−07	3.0	2.16e−04	2.0	2.29e−08	4.0	1.45e−06	3.1

Table 2 Nodal errors and corresponding rates of convergence

$M \downarrow$	$\mathscr{E}_h^{0,\text{nodal}}$	$\mathscr{R}^{0,\text{nodal}}$	$\mathscr{E}_h^{1,\text{nodal}}$	$\mathscr{R}^{1,\text{nodal}}$
8	5.31e−05	0.0	2.47e−04	0.0
16	4.30e−06	3.6	2.12e−05	3.5
32	2.96e−07	3.9	1.60e−06	3.7
64	1.91e−08	4.0	1.10e−07	3.9

Acknowledgements The authors gratefully acknowledge the research support of the Department of Science and Technology, Government of India, through the National Programme on Differential Equations: Theory, Computation and Applications, DST Project No.SERB/F/1279/2011-2012. LPT acknowledges the support through Institute Post-Doctoral Fellowship of IIT-Bombay. Support was also received by GF from IIT Bombay while a Distinguished Guest Professor at that institution. The authors wish to thank the referees for their insightful comments and constructive suggestions, which significantly improved the manuscript.

References

1. Abushama, A.A., Bialecki, B.: Modified nodal cubic spline collocation for biharmonic equations. Numer. Algorithms. **43**, 331–353 (2006)
2. Abushama, A.A., Bialecki, B.: Modified nodal cubic spline collocation for Poisson's equation. SIAM J. Numer. Anal. **46**, 331–353 (2008)
3. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, vol. 140, 2nd ed. Elsevier/Academic, Amsterdam (2003)
4. Archer, D.: Some collocation methods for differential equations. Ph.D. thesis, Rice University, Houston, Texas (1973)
5. Archer, D.: An $O(h^4)$ cubic spline collocation method for quasilinear parabolic equations. SIAM J. Numer. Anal. **14**, 620–637 (1977)
6. Bialecki, B., Wang, Z.: Modified nodal spline collocation for elliptic equations. Numer. Methods Partial Differ. Equ. **28**, 1817–1839 (2012)
7. Bialecki, B., Fairweather, G., Karageorghis, A.: Matrix decomposition algorithms for modified spline collocation for Helmholtz problems. SIAM J. Sci. Comput. **24**, 1733–1753 (2003)
8. Bialecki, B., Ganesh, M., Mustapha, K.: A Petrov-Galerkin method with quadrature for elliptic boundary value problems. IMA J. Numer. Anal. **24**, 157–177 (2004)
9. Bialecki, B., Ganesh, M., Mustapha, K.: A Crank-Nicolson Petrov-Galerkin method with quadrature for semi-linear parabolic problems. Numer. Methods Partial Differ. Equ. **21**, 918–937 (2005)
10. Bialecki, B., Fairweather, G., Karageorghis, A.: Optimal superconvergent one step nodal cubic spline collocation methods. SIAM J. Sci. Comput. **27**, 575–598 (2005)
11. Bialecki, B., Ganesh, M., Mustapha, K.: A Petrov-Galerkin method with quadrature for semi-linear hyperbolic problems. Numer. Methods Partial Differ. Equ. **22**, 1052–1069 (2006)

12. Bialecki, B., Fairweather, G., Karageorghis, A., Nguyen, Q.N.: Optimal superconvergent one step quadratic spline collocation methods. BIT Numer. Math. **48**, 449–472 (2008)
13. Bialecki, B., Ganesh, M., Mustapha, K.: An ADI Petrov-Galerkin method with quadrature for parabolic problems. Numer. Methods Partial Differ. Equ. **25**, 1129–1148 (2009)
14. de Boor, C.: The method of projections as applied to the numerical solution of two point boundary value problems using cubic splines. Ph.D. thesis, University of Michigan, Ann Arbor, Michigan (1966)
15. de Boor, C.: A bound on the L_∞-norm of the L_2-approximation by splines in terms of a global mesh ratio. Math. Comput. **30**, 765–771 (1976)
16. Christara, C.C.: Quadratic spline collocation methods for elliptic partial differential equations. BIT Numer. Math. **34**, 33–61 (1994)
17. Christara, C.C., Chen, T., Dang, D.M.: Quadratic spline collocation for one-dimensional parabolic partial differential equations. Numer. Algorithms. **53**, 511–553 (2010)
18. Daniel, J.W., Swartz, B.K.: Extrapolated collocation for two-point boundary value problems using cubic splines. J. Inst. Math. Appl. **16**, 161–174 (1975)
19. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems I, vol. 5, 2nd ed. Springer, Berlin (2000)
20. Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration. Academic, New York (1975)
21. Douglas, J. Jr., Dupont, T.: Collocation Methods for Parabolic Equations in a Single Space Variable. Lecture Notes in Mathematics, vol. 385. Springer, New York/Berlin (1974)
22. Fairweather, G., Karageorghis, A., Maack, J.: Compact optimal quadratic spline collocation methods for Poisson and Helmholtz problems: formulation and numerical verification. Technical Report TR/03/2010, Department of Mathematics and Statistics, University of Cyprus (2010)
23. Grigorieff, R.D., Sloan, I.H.: High-order spline Petrov-Galerkin methods with quadrature. ICIAM/GAMM 95 (Hamburg, 1995). Z. Angew. Math. Mech. **76**(1), 15–18 (1996)
24. Grigorieff, R.D., Sloan, I.H.: Spline Petrov-Galerkin methods with quadrature. Numer. Funct. Anal. Optimiz. **17**, 755–784 (1996)
25. Houstis, E.N., Christara, C.C., Rice, J.R.: Quadratic-spline collocation methods for two-point boundary value problems. Int. J. Numer. Methods Eng. **26**, 935–952 (1988)
26. Jones, D.L., Pani, A.K.: A qualocation method for a semilinear second-order two-point boundary value problem. In: Brokate, M., Siddiiqi, A.H. (eds.) Current Applications in Science, Technology and Industry. Pitman Notes in Mathematics, vol. 377, pp. 128–144. Addison Wesley Longman, Reading, MA (1998)
27. Jones, D.L., Pani, A.K.: A qualocation method for a unidimensional single phase semilinear Stefan problem. IMA J. Numer. Anal. **25**, 139–159 (2005)
28. Pani, A.K.: A qualocation method for parabolic partial differential equations. IMA J. Numer. Anal. **19**, 473–495 (1999)
29. Sloan, I.H., Tran, D., Fairweather, G.: A fourth-order cubic spline method for linear second-order two-point boundary value problems. IMA J. Numer. Anal. **13**, 591–607 (1993)

Analysis of Framelet Transforms on a Simplex

Yu Guang Wang and Houying Zhu

Dedicated to Ian H. Sloan on the occasion of his 80th birthday with our gratitude for his constant supervision, support and encouragement.

Abstract In this paper, we construct framelets associated with a sequence of quadrature rules on the simplex T^2 in $\mathbb{R}^2$. We give the framelet transforms—decomposition and reconstruction of the coefficients for framelets of a function on T^2. We prove that the reconstruction is exact when the framelets are tight. We give an example of construction of framelets and show that the framelet transforms can be computed as fast as FFT.

1 Introduction

Multiresolution analysis on a simplex T^2 in $\mathbb{R}^2$ has many applications such as in numerical solution of PDEs and computer graphics [6, 10, 11]. In this paper, we construct framelets (or a framelet system) on T^2, following the framework of [18], and give the transforms of coefficients for framelets.

Framelets are localised functions associated with quadrature rules of T^2. Each framelet is scaled at a level $j, j = 0, 1, \ldots$ and translated at a node of a quadrature rule of level j. The *framelet coefficients* for a square-integrable function f on the simplex are the inner products of the framelets with f on T^2. We give the *framelet transforms* which include the *decomposition* and *reconstruction* of the coefficients for framelets. Since the framelets are well-localised, see e.g. [15], the decomposition

Y. G. Wang (✉)
Department of Mathematics and Statistics, La Trobe University, Melbourne, VIC, Australia
e-mail: y.wang@latrobe.edu.au

H. Zhu
School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, Australia
e-mail: houying.zhu@unimelb.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_54

gives all *approximate* and *detailed* information of the function f. This plays an important role in signal processing on the simplex.

For levels j and $j+1$, the decomposition estimates the framelet coefficients of level $j+1$ by the coefficients of level j. The reconstruction is the inverse, which estimates the coefficients of level j by the level $j+1$. Such framelet transforms are significant as by decompositions or reconstructions, we are able to estimate high-level framelet coefficients from the bottom level 0, or the inverse.

We show that when the quadrature rules and masks have good properties, the reconstruction is exact and invertible with the decomposition, see Sect. 4. We also show that the framelet transforms can be computed as fast as the FFTs, see Sect. 6.

We construct framelets using the tensor-product form of Jacobi polynomials and triangular Kronecker lattices [2] with equal weights, see Sect. 5.

2 Framelets on Simplex

In the paper, we consider the *simplex* (or the *triangle*)

$$T^2 := \{\boldsymbol{x} := (x_1, x_2) | x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}.$$

Let $L_2(T^2)$ be the space of complex-valued square integrable functions on T^2 with respect to the normalized Lebesgue area measure μ on $\mathbb{R}^2$ (i.e. $\int_{T^2} \mathrm{d}\mu(\boldsymbol{x}) = 1$), provided with inner product $\langle f, g \rangle := \langle f, g \rangle_{L_2(T^2)} := \int_{T^2} f(\boldsymbol{x}) \mathrm{d}\mu(\boldsymbol{x})$, where $\overline{g}$ is the complex conjugate of g, and endowed with the induced L_2-norm $\|f\|_{L_2(T^2)} := \sqrt{\langle f, f \rangle}$ for $f \in L_2(T^2)$.

For $\ell \geq 0$, let $\mathscr{V}_\ell := \mathscr{V}_\ell(T^2)$ be the space of orthogonal polynomials of degree ℓ with respect to the inner product $\langle \cdot, \cdot \rangle_{L_2(T^2)}$. The dimension of $\mathscr{V}_\ell$ is $\ell + 1$, see [9]. The elements of $\mathscr{V}_\ell$ are said to be the *polynomials* of degree ℓ on T^2. The union of all polynomial spaces $\cup_{\ell=0}^{\infty} \mathscr{V}_\ell$ is dense in $L_2(T^2)$.

As a compact Riemannian manifold, the simplex T^2 has the Laplace-Beltrami operator

$$\Delta := \sum_{i=1}^{2} x_i(1 - x_i) \frac{\partial^2}{\partial x_i^2} - 2 \sum_{1 \leq i < j \leq 2} x_i x_j \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{2} (1 - 3x_i) \frac{\partial}{\partial x_i},$$

with polynomials P_ℓ in $\mathscr{V}_\ell$ as the eigenfunctions and with (square-rooted) eigenvalues $\lambda_\ell := \sqrt{\ell(\ell + 2)}$:

$$-\Delta P_\ell = \lambda_\ell^2 P_\ell, \quad \ell \in \mathbb{N}_0,$$

where $\mathbb{N}_0 := \{0, 1, 2, \dots\}$, see [1].

Let $L_1(\mathbb{R})$ be the space of absolutely integrable functions on $\mathbb{R}$ with respect to the Lebesgue measure and let $l_1(\mathbb{Z})$ be the set of l_1 summable sequences on $\mathbb{Z}$. For

$r \geq 1$, let $\Psi := \{\alpha; \beta^1, \ldots, \beta^r\}$ be a set of $(r+1)$ functions in $L_1(\mathbb{R})$, which are associated with a *filter bank* $\boldsymbol{\eta} := \{a; b_1, \ldots, b_r\} \subset l_1(\mathbb{Z})$ satisfying

$$\widehat{\alpha}(2\xi) = \widehat{a}(\xi)\widehat{\alpha}(\xi), \quad \widehat{\beta^n}(2\xi) = \widehat{b_n}(\xi)\widehat{\alpha}(\xi), \quad n = 1, \ldots, r,\ \xi \in \mathbb{R}, \tag{1}$$

where $\widehat{g}(\xi) := \int_{\mathbb{R}} g(x)e^{-2\pi i x\xi}\,\mathrm{d}x$, $\xi \in \mathbb{R}$ is the Fourier transform for $g \in L_1(\mathbb{R})$ and $\widehat{h}(\xi) := \sum_{k\in\mathbb{Z}} h_k e^{-2\pi \mathrm{i}\xi}$ is the Fourier series of a sequence $h := (h_k)_{k\in\mathbb{Z}}$ in $l_1(\mathbb{Z})$. Here, the sequences a and b_n are said to be *low-pass (mask)* and *high-pass (mask)* respectively.

We introduce the continuous and semi-discrete framelets on the simplex following the construction and notation of [8, 18]. The *continuous framelets* on the simplex T^2 are, for $j \in \mathbb{N}_0$,

$$\begin{aligned} \boldsymbol{\varphi}_{j,\boldsymbol{y}}(\boldsymbol{x}) &:= \sum_{\ell=0}^{\infty} \widehat{\alpha}\left(\frac{\lambda_\ell}{2^j}\right)\overline{P_\ell(\boldsymbol{y})}P_\ell(\boldsymbol{x}), \\ \boldsymbol{\psi}_{j,\boldsymbol{y}}^n(\boldsymbol{x}) &:= \sum_{\ell=0}^{\infty} \widehat{\beta^n}\left(\frac{\lambda_\ell}{2^j}\right)\overline{P_\ell(\boldsymbol{y})}P_\ell(\boldsymbol{x}), \quad n = 1, \ldots, r. \end{aligned} \tag{2}$$

The continuous framelets in (2) are analogues of continuous wavelets in $\mathbb{R}$. The level "j" indicates the "dilation" scale and "$\boldsymbol{y}$" is the point at which the framelet is "translated".

Let $Q_N := \{(w_j, \boldsymbol{x}_j)\}_{j=1}^N$, which is a set of N pairs of weights w_j in $\mathbb{R}\backslash\{0\}$ and points $\boldsymbol{x}_j$ on T^2. We define the quadrature rule

$$Q_N[f] := \sum_{k=1}^{N} w_j f(\boldsymbol{x}_j)$$

for continuous functions f on T^2. Let $Q_{N_j} := \{(\omega_{j,k}, \boldsymbol{x}_{j,k})\}_{k=1}^{N_j}$, $j \in \mathbb{N}_0$, be a sequence of such quadrature rules. For $j = 0, 1, \ldots$, the *semi-discrete framelets* $\boldsymbol{\varphi}_{j,k}$ and $\boldsymbol{\psi}_{j,k'}^n$ associated with quadrature rules Q_{N_j} are defined as the continuous framelets $\boldsymbol{\varphi}_{j,\boldsymbol{y}}$ and $\boldsymbol{\psi}_{j,\boldsymbol{y}}^n$ translated at $\boldsymbol{x}_{j,k}$ and $\boldsymbol{x}_{j+1,k'}$ respectively. That is, for $k = 1, \ldots, N_j$,

$$\boldsymbol{\varphi}_{j,k}(\boldsymbol{x}) := \sqrt{\omega_{j,k}}\,\boldsymbol{\varphi}_{j,\boldsymbol{x}_{j,k}}(\boldsymbol{x}) = \sqrt{\omega_{j,k}}\sum_{\ell=0}^{\infty}\widehat{\alpha}\left(\frac{\lambda_\ell}{2^j}\right)\overline{P_\ell(\boldsymbol{x}_{j,k})}P_\ell(\boldsymbol{x}), \tag{3}$$

and for $k' = 1, \ldots, N_{j+1}$ and $n = 1, \ldots, r$,

$$\begin{aligned} \boldsymbol{\psi}_{j,k'}^n(\boldsymbol{x}) &:= \sqrt{\omega_{j+1,k'}}\,\boldsymbol{\psi}_{j,\boldsymbol{x}_{j+1,k'}}^n \\ &= \sqrt{\omega_{j+1,k'}}\sum_{\ell=0}^{\infty}\widehat{\beta^n}\left(\frac{\lambda_\ell}{2^j}\right)\overline{P_\ell(\boldsymbol{x}_{j+1,k'})}P_\ell(\boldsymbol{x}). \end{aligned} \tag{4}$$

We say $\boldsymbol{\varphi}_{j,k}$ and $\boldsymbol{\psi}_{j,k'}^n$ are *low-pass framelet* and *high-pass framelet* respectively.

Note that here we use the $Q_{N_{j+1}}$ for high-passes due to the scale of $\boldsymbol{\psi}^n_{j,k'}$ is at $j+1$. This will be clear in Sect. 5.

We also use the notation $\boldsymbol{\psi}^n_{j,k}$ for $\boldsymbol{\psi}^n_{j,k'}$ if no confusion arises.

The framelets $\boldsymbol{\varphi}_{j,k}$ and $\boldsymbol{\psi}^n_{j,k}$ corresponding to the low-pass a and high-pass b_n carry the information of approximations and details in framelet transforms, as we will show below.

3 Decomposition for Framelets

In practice, we need to estimate the framelet coefficients of high levels from low-level coefficients. This can be achieved by the *decomposition* of framelets.

The decomposition for framelets can be realized by the operations of convolution and downsampling as we introduce now.

Let $h \in l_1(\mathbb{Z})$ be a mask satisfying that the support of the Fourier series $\widehat{h}$ of h is a subset of $[0, 1/2]$. Let $l(N)$ be the set of complex-valued sequences with supports in $[0, N]$. Let $Q_{N_j} := \{(\omega_{j,k}, \boldsymbol{x}_{j,k})\}_{k=1}^{N_j}$, $j \in \mathbb{N}_0$, be the quadrature rules for framelets. Let $l(Q_{N_j})$ be the set of sequences v in $l(N_j)$ satisfying that there exists a sequence u in $l_1(\mathbb{Z})$ such that

$$(\mathrm{v})_k = \sqrt{\omega_{j,k}} \sum_{\ell=0}^{\infty} \mathrm{u}_\ell \, P_\ell(\boldsymbol{x}_{j,k}), \quad k = 1, \ldots, N_j.$$

We let $\widehat{\mathrm{v}}_\ell := \mathrm{u}_\ell$ (with abuse of notation) be the (generalized) *Fourier coefficients* of v for the orthonormal basis P_ℓ and the quadrature rule Q_{N_j} on T^2.

The *(discrete) convolution* $\mathrm{v} *_j h$ of a sequence v with the mask h is a sequence in $l(Q_{N_j})$ given by

$$(\mathrm{v} *_j h)_k := \sum_{\ell=0}^{\infty} \widehat{\mathrm{v}}_\ell \, \widehat{h}\left(\frac{\lambda_\ell}{2^j}\right) \sqrt{\omega_{j,k}} \, P_\ell(\boldsymbol{x}_{j,k}), \quad k = 1, \ldots, N_j. \tag{5}$$

Then, the Fourier coefficients of $\mathrm{v} *_j h$ are $(\widehat{\mathrm{v} *_j h})_\ell = \widehat{\mathrm{v}}_\ell \, \widehat{h}\left(\frac{\lambda_\ell}{2^j}\right)$, $\ell \in \mathbb{N}_0$.

The *downsampling* $\mathrm{v}{\downarrow_j}$, $j \geq 1$, of a sequence $\mathrm{v} \in l(Q_{N_j})$ is a sequence in $l(Q_{N_{j-1}})$ given by

$$(\mathrm{v}{\downarrow_j})_k := \sum_{\lambda_\ell \leq 2^{j-1}} \widehat{\mathrm{v}}_\ell \, \sqrt{\omega_{j,k}} \, P_\ell(\boldsymbol{x}_{j,k}), \quad k = 1, \ldots, N_{j-1}. \tag{6}$$

For semi-discrete framelets in (3) and (4), the inner products $\langle f, \boldsymbol{\varphi}_{j,k} \rangle$ and $\langle f, \boldsymbol{\psi}^n_{j,k'} \rangle$, $n = 1, \ldots, r$, $j \in \mathbb{N}_0$, $k = 1, \ldots, N_j$ and $k' = 1, \ldots, N_{j+1}$, are said to be *framelet coefficients* for f. For convenience, we let v_j and w^n_j denote the vectors of the framelet coefficients for f:

$$(\mathrm{v}_j)_k := \langle f, \boldsymbol{\varphi}_{j,k}\rangle, \quad (\mathrm{w}_j^n)_{k'} := \langle f, \boldsymbol{\psi}_{j,k'}^n\rangle. \tag{7}$$

The Fourier coefficients of a function $f \in L_2(T^2)$ are $\widehat{f_\ell} := \langle f, P_\ell\rangle$, $\ell \in \mathbb{N}_0$. Let h be a mask in $l_1(\mathbb{Z})$ and $h^\star$ be the mask whose Fourier series is conjugate to the Fourier series of h.

The following proposition shows the decomposition for framelet coefficients between adjacent levels.

Proposition 1 *Let the framelet coefficients for semi-discrete framelets in* (3) *and* (4) *be given by* (7)*, where the supports of* $\widehat{\alpha}$ *and* $\widehat{\beta^n}$ *are subsets of* $[0, 1/2]$. *For* $j = 1, 2, \ldots$*, the decomposition from level* j *into level* $j-1$ *is*

$$\mathrm{v}_{j-1} = (\mathrm{v}_j *_j a^\star)\downarrow_j, \quad \mathrm{w}_{j-1}^n = \mathrm{v}_j *_j b_n^\star, \quad n = 1, \ldots, r. \tag{8}$$

Proof For $f \in L_2(T^2)$, by the orthonormality of P_ℓ and (7),

$$\begin{aligned}
(\mathrm{v}_{j-1})_k &= \sqrt{\omega_{j-1,k}} \sum_{\ell=0}^{\infty} \overline{\widehat{\alpha}\left(\frac{\lambda_\ell}{2^{j-1}}\right)} \widehat{f}_\ell \, P_\ell(\boldsymbol{x}_{j-1,k}), \quad k = 1, \ldots, N_{j-1},\\
(\mathrm{w}_{j-1}^n)_{k'} &= \sqrt{\omega_{j,k'}} \sum_{\ell=0}^{\infty} \overline{\widehat{\beta^n}\left(\frac{\lambda_\ell}{2^{j-1}}\right)} \widehat{f}_\ell \, P_\ell(\boldsymbol{x}_{j,k'}), \quad k' = 1, \ldots, N_j, \ n = 1, \ldots, r.
\end{aligned}$$

For low-pass, by (1), (5) and (6), for $k = 1, \ldots, N_{j-1}$,

$$\begin{aligned}
(\mathrm{v}_{j-1})_k &= \sqrt{\omega_{j-1,k}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{f}_\ell \, \overline{\widehat{\alpha}\left(\frac{\lambda_\ell}{2^{j-1}}\right)} P_\ell(\boldsymbol{x}_{j-1,k})\\
&= \sqrt{\omega_{j-1,k}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{f}_\ell \, \overline{\widehat{\alpha}\left(\frac{\lambda_\ell}{2^{j}}\right)} \, \overline{\widehat{a}\left(\frac{\lambda_\ell}{2^{j}}\right)} P_\ell(\boldsymbol{x}_{j-1,k})\\
&= \sqrt{\omega_{j-1,k}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell \, \overline{\widehat{a}\left(\frac{\lambda_\ell}{2^{j}}\right)} P_\ell(\boldsymbol{x}_{j-1,k})\\
&= \big((\mathrm{v}_j *_j a^\star)\downarrow_j\big)_k.
\end{aligned}$$

For high-passes, for $k' = 1, \ldots, N_j$ and $n = 1, \ldots, r$,

$$\begin{aligned}
(\mathrm{w}_{j-1}^n)_{k'} &= \sqrt{\omega_{j,k'}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{f}_\ell \, \overline{\widehat{\beta^n}\left(\frac{\lambda_\ell}{2^{j-1}}\right)} P_\ell(\boldsymbol{x}_{j,k'})\\
&= \sqrt{\omega_{j,k'}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{f}_\ell \, \overline{\widehat{\beta^n}\left(\frac{\lambda_\ell}{2^{j}}\right)} \, \overline{\widehat{b_n}\left(\frac{\lambda_\ell}{2^{j}}\right)} P_\ell(\boldsymbol{x}_{j,k'})
\end{aligned}$$

$$= \sqrt{\omega_{j,k'}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell \, \overline{\widehat{b_n}\left(\frac{\lambda_\ell}{2^j}\right)} P_\ell(\boldsymbol{x}_{j,k'})$$

$$= (\mathrm{v}_j *_j b_n^\star)_{k'}.$$

These give (8). □

4 Reconstruction for Tight Framelets

We say the set of framelets $\{\boldsymbol{\varphi}_{j,k}, \boldsymbol{\psi}^n_{j,k'} | n = 1, \ldots, r, \;\; k = 1, \ldots, N_j, \;\; k' = 1, \ldots, N_{j+1}, \;\; j \in \mathbb{N}_0\}$ a *tight frame* for $L_2(T^2)$ if the framelets are all in $L_2(T^2)$, and in the L_2 sense,

$$f = \sum_{k=1}^{N_0} \langle f, \boldsymbol{\varphi}_{j,k} \rangle \boldsymbol{\varphi}_{j,k} + \sum_{j=0}^{\infty} \sum_{k'=1}^{N_{j+1}} \sum_{n=1}^{r} \langle f, \boldsymbol{\psi}^n_{j,k'} \rangle \boldsymbol{\psi}^n_{j,k'} \quad \text{for all}\, f \in L_2(T^2),$$

or equivalently,

$$\|f\|^2_{L_2(T^2)} = \sum_{k=1}^{N_0} \left|\langle f, \boldsymbol{\varphi}_{j,k} \rangle\right|^2 + \sum_{j=0}^{\infty} \sum_{k'=1}^{N_{j+1}} \sum_{n=1}^{r} \left|\langle f, \boldsymbol{\psi}^n_{j,k'} \rangle\right|^2 \quad \text{for all}\, f \in L_2(T^2).$$

The framelets are then said to be *(semi-discrete) tight framelets*.

If the framelets are tight on the simplex, a function in $L_2(T^2)$ can be represented using the framelet coefficients. The following property as a consequence of [18, Theorem 2.4] shows that the tightness of framelets is equivalent to a multiscale representation of framelets of a level by lower levels.

Proposition 2 ([18]) *The semi-discrete framelets in* (3) *and* (4) *are tight if and only if for all $f \in L_2(T^2)$, the following identities hold:*

$$\lim_{j \to \infty} \sum_{k=1}^{N_j} \left|\langle f, \boldsymbol{\varphi}_{j,k} \rangle\right|^2 = \|f\|^2_{L_2(T^2)},$$

$$\sum_{k=1}^{N_{j+1}} \left|\langle f, \boldsymbol{\varphi}_{j+1,k} \rangle\right|^2 = \sum_{k=1}^{N_j} \left|\langle f, \boldsymbol{\varphi}_{j,k} \rangle\right|^2 + \sum_{k=1}^{N_{j+1}} \sum_{n=1}^{r} \left|\langle f, \boldsymbol{\psi}^n_{j,k} \rangle\right|^2, \quad j \in \mathbb{N}_0. \tag{9}$$

The condition in (9) implies that high-level framelet coefficients can be estimated by low levels. This then gives the *reconstruction* for framelets.

The reconstruction depends on the property of the quadrature rules Q_{N_j} for framelets. A quadrature rule $Q_N := \{(w_j, \boldsymbol{x}_j)\}_{j=1}^N$ is said to be exact for polynomials up to degree ℓ if for $\ell' = 0, \dots, \ell$,

$$\int_{T^2} p_{\ell'}(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}) = \sum_{j=1}^{N} w_j p_{\ell'}(\boldsymbol{x}_j) \quad \text{for all } p_{\ell'} \in \mathscr{V}_{\ell'}.$$

When the quadrature rule Q_{N_j}, $j \in \mathbb{N}_0$, for framelets $\boldsymbol{\varphi}_{j,k}$ and $\boldsymbol{\psi}^n_{j,k}$ is exact for polynomials up to degree 2^j, the tightness of the framelets is equivalent to the following condition on masks:

$$\lim_{j\to\infty} \widehat{a}\left(\frac{\lambda_\ell}{2^j}\right) = 1, \quad \left|\widehat{a}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2 + \sum_{n=1}^{r} \left|\widehat{b_n}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2 = 1 \quad \text{for } j, \ell \in \mathbb{N}_0, \tag{10}$$

see [18, Theorem 2.1 and Corollary 2.6].

The *upsampling* $\mathrm{v}\!\uparrow_j$, $j \geq 1$, of a sequence $\mathrm{v} \in l(Q_{N_{j-1}})$ is a sequence in $l(Q_{N_j})$ given by

$$(\mathrm{v}\!\uparrow_j)_k := \sum_{\lambda_\ell \leq 2^{j-2}} \widehat{\mathrm{v}}_\ell \sqrt{\omega_{j,k}}\, P_\ell(\boldsymbol{x}_{j,k}), \quad k = 1, \dots, N_j,$$

where $\widehat{\mathrm{v}}_\ell$ are the Fourier coefficients of v for basis P_ℓ and quadrature rule $Q_{N_{j-1}}$ on T^2.

The reconstruction involving the operations of convolution and upsampling is given by the following proposition.

Proposition 3 *Let the framelet coefficients for semi-discrete framelets in* (3) *and* (4) *be given by* (7)*, where the supports of* $\widehat{\alpha}$ *and* $\widehat{\beta^n}$ *are subsets of* $[0, 1/2]$*, and* (10) *holds. Then, for* $j \geq 1$*, the reconstruction from level* $j-1$ *to level* j *is*

$$\mathrm{v}_j = (\mathrm{v}_{j-1}\!\uparrow_j) *_j a + \sum_{n=1}^{r} \mathrm{w}^n_{j-1} *_j b_n. \tag{11}$$

Proof By Proposition 1, for $k = 1, \dots, N_j$,

$$((\mathrm{v}_{j-1}\!\uparrow_j) *_j a)_k = \sqrt{\omega_{j,k}} \sum_{\lambda_\ell \leq 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell \left|\widehat{a}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2 P_\ell(\boldsymbol{x}_{j,k})$$

and

$$(\mathrm{w}^n_{j-1} *_j b_n)_k = \sqrt{\omega_{j,k}} \sum_{\lambda_\ell \leq 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell \left|\widehat{b_n}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2 P_\ell(\boldsymbol{x}_{j,k}), \quad n = 1, \dots, r.$$

These give

$$
\begin{aligned}
&\left((\mathrm{v}_{j-1}\uparrow_j) *_j a + \sum_{n=1}^{r} \mathrm{w}^n_{j-1} *_j b_n\right)_k \\
&\quad= \sqrt{\omega_{j,k}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell \left(\left|\widehat{a}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2 + \sum_{n=1}^{r}\left|\widehat{b_n}\left(\frac{\lambda_\ell}{2^j}\right)\right|^2\right) P_\ell(\boldsymbol{x}_{j,k}) \\
&\quad= \sqrt{\omega_{j,k}} \sum_{\lambda_\ell \le 2^{j-1}} \widehat{(\mathrm{v}_j)}_\ell\, P_\ell(\boldsymbol{x}_{j,k}) \\
&= (\mathrm{v}_j)_k,
\end{aligned}
$$

thus proving (11). □

Remark 1 [18, Theorem 3.1] proves (11) for general Riemannian manifolds when the quadrature rule is exact for polynomials up to degree 2^j and under condition (10). Here we do not require that the quadrature rules of the framelets satisfy the polynomial exactness.

Repeatedly using the decomposition and reconstruction in Propositions 1 and 3 gives multi-level framelet transforms. Figure 1 illustrates the decomposition and reconstruction for levels $0, \ldots, j$. In the decomposition, each level is decomposed into a low-pass (framelet) coefficient and r high-pass coefficients of the next (lower)

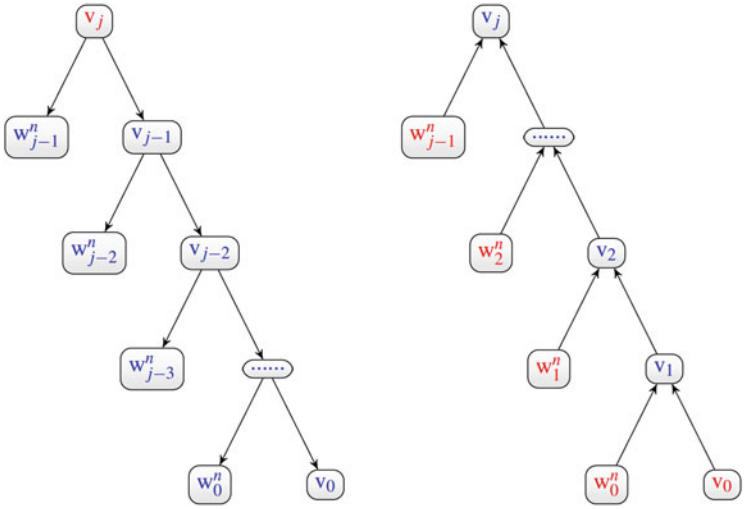


Fig. 1 The left diagram illustrates the decomposition of the framelets coefficients which computes all coefficients in lower levels by v_j. The right shows the reconstruction of framelet coefficients v_j from the coefficients v_0 and $\mathrm{w}^n_0, \ldots, \mathrm{w}^n_{j-1}$ of lower levels

level. In the reconstruction, the low-pass coefficient at level j is estimated by the low-pass coefficient and high-pass coefficients at the level $j-1$.

5 Constructive Examples

From the above analysis, the construction of semi-discrete framelets needs an orthonormal basis for $L_2(T^2)$ and appropriate masks and quadrature rules.

Orthonormal Bases One orthonormal basis can be constructed by the tensor product of Jacobi polynomials, see [9, Proposition 2.4.1]. For $\tau, \gamma > -1$ and $\ell \geq 0$, let $P_\ell^{(\tau,\gamma)}(t)$ be the Jacobi polynomial of degree ℓ with respect to the weight $(1-t)^\tau(1+t)^\gamma$ on [-1,1]. For $\boldsymbol{x} := (x_1, x_2) \in T^2$ and $\ell \in \mathbb{N}_0$ and $m = 0, \ldots, \ell$, let

$$\begin{aligned} P_{\ell,m}(\boldsymbol{x}) := \sqrt{(\ell+1)(2m+1)}\, & P_{\ell-m}^{(2m+1,0)}(2x_1-1)(1-x_1)^m \\ & \times P_m^{(0,0)}\left(\frac{2x_2}{1-x_1}-1\right). \end{aligned} \tag{12}$$

Then $\{P_{\ell,m} | m = 0, \ldots, \ell\}$ is an orthonormal basis of $\mathcal{V}_\ell$ and $\{P_{\ell,m} | m = 0, \ldots, \ell, \ell \geq 0\}$ forms an orthonormal basis of $L_2(T^2)$.

Sun [17] constructs another orthonormal basis for $L_2(T^2)$, which is useful in discrete Fourier analysis on T^2, see [13, 14].

Masks We give an example of masks with two high-passes. Let

$$\nu(t) := t^4(35 - 84t + 70t^2 - 20t^3), \quad t \in \mathbb{R}.$$

By [5, Chapter 4], the masks a, b_1 and b_2 can be defined by their Fourier series as

$$\widehat{a}(\xi) := \begin{cases} 1, & |\xi| < \frac{1}{8}, \\ \cos\left(\frac{\pi}{2}\nu(8|\xi|-1)\right), & \frac{1}{8} \leq |\xi| \leq \frac{1}{4}, \\ 0, & \frac{1}{4} < |\xi| \leq \frac{1}{2}, \end{cases} \tag{13}$$

$$\widehat{b_1}(\xi) := \begin{cases} 0, & |\xi| < \frac{1}{8}, \\ \sin\left(\frac{\pi}{2}\nu(8|\xi|-1)\right), & \frac{1}{8} \leq |\xi| \leq \frac{1}{4}, \\ \cos\left(\frac{\pi}{2}\nu(4|\xi|-1)\right), & \frac{1}{4} < |\xi| \leq \frac{1}{2}. \end{cases} \tag{14}$$

$$\widehat{b_2}(\xi) := \begin{cases} 0, & |\xi| < \frac{1}{4}, \\ \sin\left(\frac{\pi}{2}\nu(4|\xi|-1)\right), & \frac{1}{4} \leq |\xi| \leq \frac{1}{2}, \end{cases} \tag{15}$$

which satisfy (10).

The corresponding scaling functions are

$$\widehat{\alpha}(\xi) = \begin{cases} 1, & |\xi| < \frac{1}{4}, \\ \cos\left(\frac{\pi}{2}\nu(4|\xi|-1)\right), & \frac{1}{4} \le |\xi| \le \frac{1}{2}, \\ 0, & \text{else}, \end{cases} \tag{16}$$

$$\widehat{\beta^1}(\xi) = \begin{cases} \sin\left(\frac{\pi}{2}\nu(4|\xi|-1)\right), & \frac{1}{4} \le |\xi| < \frac{1}{2}, \\ \cos^2\left(\frac{\pi}{2}\nu(2|\xi|-1)\right), & \frac{1}{2} \le |\xi| \le 1, \\ 0, & \text{else}, \end{cases} \tag{17}$$

$$\widehat{\beta^2}(\xi) = \begin{cases} 0, & |\xi| < \frac{1}{2}, \\ \cos\left(\frac{\pi}{2}\nu(2|\xi|-1)\right)\sin\left(\frac{\pi}{2}\nu(2|\xi|-1)\right), & \frac{1}{2} \le |\xi| \le 1, \\ 0, & \text{else}. \end{cases} \tag{18}$$

Here, supp $\widehat{\alpha} \subseteq [0, 1/2]$ and supp $\widehat{\beta^n} \subseteq [1/4, 1]$, $n = 1, 2$. This means that the scaling of the framelet $\boldsymbol{\varphi}_{j,k}$ in (3) with the low-pass scaling function in (16) is half of the scaling of the framelets $\boldsymbol{\psi}^1_{j,k}$ and $\boldsymbol{\psi}^2_{j,k}$ in (4) with high-pass scaling functions in (17) and (18). The high-pass framelets thus need to use a quadrature rule at the level $j+1$, one level higher than $\boldsymbol{\varphi}_{j,k}$.

Figure 2 shows the Fourier series of masks a, b_1 and b_2 in (13), (14) and (15).

Quadrature Rules We use triangular Kronecker lattices of Basu and Owen [2] with equal weights as the quadrature rules for framelets, which are shifted lattice points intersecting with the simplex. For the quadrature rule Q_{N_j} of framelets, we use the triangular Kronecker lattice with at least 2^{2j} nodes, which are the translation points of the low-pass framelets $\boldsymbol{\varphi}_{j,k}$ at level j and those of high-pass framelets

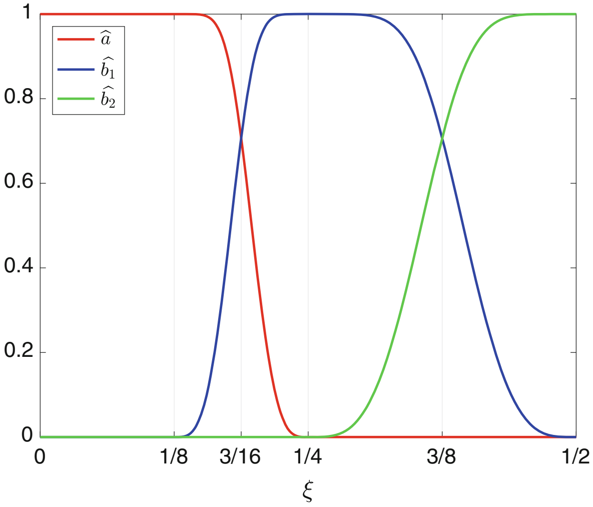


Fig. 2 The red curve shows the Fourier series of the low-pass mask $\widehat{a}$ in (13) which has support in $[0, 1/4]$. The blue and green curves show the Fourier series of high-pass masks $\widehat{b_1}$ and $\widehat{b_2}$ in (14) and (15) whose supports are subsets of $[0, 1/2]$

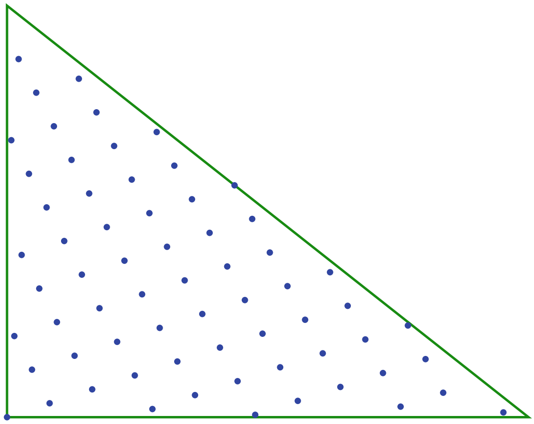


Fig. 3 Triangular Kronecker lattice with 65 nodes for framelets $\boldsymbol{\varphi}_{3,k}$ and $\boldsymbol{\psi}^{n}_{2,k}$

$\boldsymbol{\psi}^{n}_{j-1,k'}$ at level $j-1$. Figure 3 shows the triangular Kronecker lattice with $N = 65$ nodes on T^2 used for framelets at levels 2 and 3.

Framelets Using the orthonormal basis in (12), scaling functions in (16)–(18) and triangular Kronecker lattices with equal weights, the framelets are, for $j \in \mathbb{N}_0$,

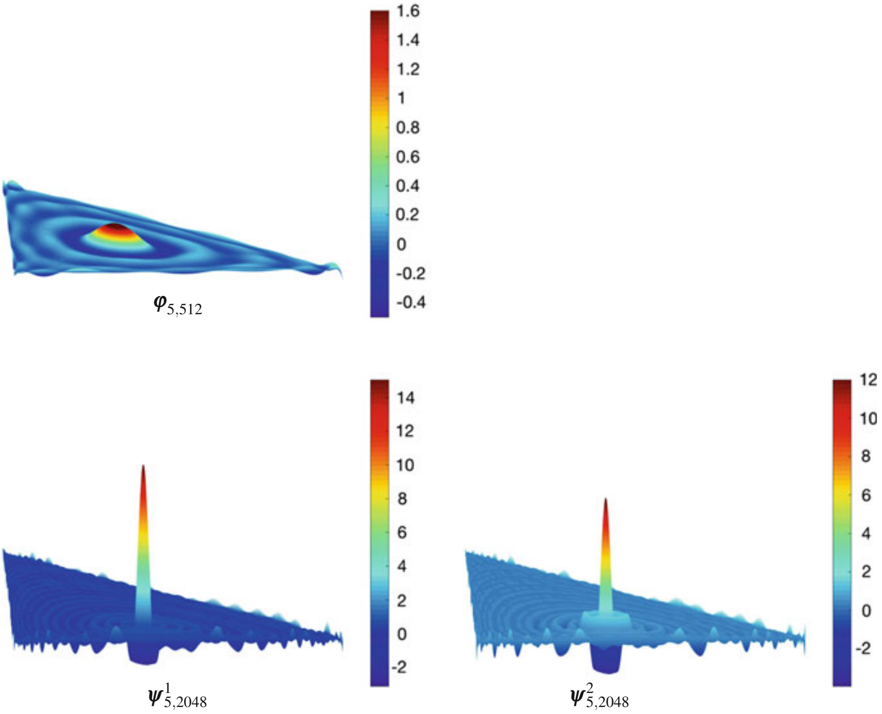
$$\boldsymbol{\varphi}_{j,k}(\boldsymbol{x}) = \frac{1}{\sqrt{N_j}} \sum_{\ell=0}^{\infty} \sum_{m=0}^{\ell} \widehat{\alpha}\left(\frac{\sqrt{\ell(\ell+2)}}{2^j}\right) \overline{P_{\ell,m}(\boldsymbol{x}_{j,k})} P_{\ell,m}(\boldsymbol{x}), \quad k = 1, \ldots, N_j, \tag{19}$$

and for $n = 1, 2$,

$$\boldsymbol{\psi}^{n}_{j,k'}(\boldsymbol{x}) = \frac{1}{\sqrt{N_{j+1}}} \sum_{\ell=0}^{\infty} \sum_{m=0}^{\ell} \widehat{\beta^n}\left(\frac{\sqrt{\ell(\ell+2)}}{2^j}\right) \overline{P_{\ell,m}(\boldsymbol{x}_{j+1,k'})} P_{\ell,m}(\boldsymbol{x}), \quad k' = 1, \ldots, N_{j+1}. \tag{20}$$

Figure 4 shows the framelets $\boldsymbol{\varphi}_{j,k}$, $\boldsymbol{\psi}^{1}_{j,k'}$ and $\boldsymbol{\psi}^{2}_{j,k'}$ at level $j = 5$ with $k = 512$ and $k' = 2048$, using orthonormal basis (12) and scaling functions (16)–(18), translated at the triangular Kronecker lattice points $\boldsymbol{x}_{5,512}, \boldsymbol{x}_{6,2048}$ and $\boldsymbol{x}_{6,2048}$. The total number of low-pass framelets $\boldsymbol{\varphi}_{j,k}$ at level $j = 5$ is $N_5 = 1025$ and the total number of high-pass framelets $\boldsymbol{\psi}^{n}_{j,k'}$, $n = 1$ or 2, at level $j = 5$ is $N_6 = 4097$. The pictures show that the framelets $\boldsymbol{\varphi}_{5,512}$, $\boldsymbol{\psi}^{1}_{5,2048}$ and $\boldsymbol{\psi}^{2}_{5,2048}$ are radial functions on T^2 with centers at the translation points $\boldsymbol{x}_{5,512}, \boldsymbol{x}_{6,2048}$ and $\boldsymbol{x}_{6,2048}$ respectively.

We observe that the high-pass framelets $\boldsymbol{\psi}^{1}_{5,2048}$ and $\boldsymbol{\psi}^{2}_{5,2048}$ are highly concentrated at the translation point $\boldsymbol{x}_{6,2048}$, and are more localised than the low-pass framelet at the same level. This illustrates that the high-pass framelets can be used to depict details of a function on T^2 in multiresolution analysis.



$\boldsymbol{\varphi}_{5,512}$

$\boldsymbol{\psi}^1_{5,2048}$

$\boldsymbol{\psi}^2_{5,2048}$

Fig. 4 The three pictures show framelets $\boldsymbol{\varphi}_{5,512}$, $\boldsymbol{\psi}^1_{5,2048}$ and $\boldsymbol{\psi}^2_{5,2048}$ given by (19) and (20) at level $j = 5$

6 Fast Computing

The framelet transforms on T^2 can be represented by discrete Fourier transforms on the simplex. This implies a fast computational strategy of the decomposition and reconstruction for framelets.

We use the notation of Sects. 3 and 4. Let $j \in \mathbb{N}_0$ and let Λ_j be the largest integer ℓ such that $\lambda_\ell \le 2^{j-1}$. The *discrete Fourier transform* (DFT) for a sequence $\mathrm{u} \in l(\Lambda_j)$ is the sequence $\mathbf{F}_j\mathrm{u}$ in $l(N_j)$ such that

$$(\mathbf{F}_j\mathrm{u})_k := \sum_{\ell=0}^{\Lambda_j} \mathrm{u}_\ell \sqrt{\omega_{j,k}}\, P_\ell(\boldsymbol{x}_{j,k}), \quad k = 0, \ldots, N_j. \tag{21}$$

The *adjoint discrete Fourier transform* (adjoint DFT) $\mathbf{F}_j^*$ of a sequence $\mathrm{v} \in l(N_j)$ is the sequence $\mathbf{F}_j^*\mathrm{v}$ in $l(\Lambda_j)$ such that

$$(\mathbf{F}_j^*\mathrm{v})_\ell := \sum_{k=0}^{N_j} \mathrm{v}_k \sqrt{\omega_{j,k}}\, \overline{P_\ell(\boldsymbol{x}_{j,k})}, \quad \ell = 0, \ldots, \Lambda_j. \tag{22}$$

The DFTs on the simplex in (21) and (22) using the orthonormal basis P_ℓ are the analogues of DFTs for square-integrable periodic functions on $\mathbb{R}$ which use the orthogonal basis $e^{2\pi i \ell' x}$, $\ell' \in \mathbb{Z}$.

By (21) and (22), we can rewrite the decomposition in (8) and reconstruction in (11) as

$$\mathrm{v}_{j-1} = \mathbf{F}_{j-1}(\widehat{\mathrm{v}_j *_j a^\star}), \quad \mathrm{w}_{j-1}^n = \mathbf{F}_j(\widehat{\mathrm{v}_j *_j (b_n)^\star}), \quad n = 1, \ldots, r$$

and

$$\mathrm{v}_j = \big(\mathbf{F}_j^*(\mathrm{v}_{j-1})\big) *_j a + \sum_{n=1}^{r} \big(\mathbf{F}_j^*(\mathrm{w}_{j-1}^n)\big) *_j b_n.$$

This means that the decomposition from level j to level $j-1$ is the DFTs of convolutions of the level-j framelet coefficients with masks, and that the reconstruction from level $j-1$ to level j is the sum of convolutions of the adjoint DFTs of level-$(j-1)$ coefficients with masks. Since the convolution is the sum of point-wise multiplications, the computational steps of the framelet transforms are in proportion to those of DFTs on the simplex.

The FFT on T^2 uses, up to log factors, $\mathscr{O}(N)$ operations for an input sequence of size N. If for $j \geq 1$, the ratio N_j/N_{j-1} of the numbers of the nodes of the quadrature rules Q_{N_j} and $Q_{N_{j-1}}$ is equivalent to a constant C, $C > 1$, the computational steps of the framelet transforms (both the decomposition and reconstruction) between levels $0, 1, \ldots, J$, $J \geq 1$, are $\mathscr{O}((r+1)N_J)$ for the sequence v_J of the framelet coefficients of size N_J, and the redundancy rate of the framelet transforms is also $\mathscr{O}((r+1)N_J)$. The framelets with the quadrature rules using triangular Kronecker lattices, as shown in Sect. 5, satisfy that $N_j/N_{j-1} \sim 4$. Thus, the framelet transforms between levels 0 to J have computational steps in proportion to 2^{2J}.

7 Discussion

In the paper, we only consider the framelet transforms for one framelet system with starting level 0. The results can be generalized to a sequence of framelet systems as [12, 18], which will allow one more flexibility in constructing framelets.

The decomposition holds for framelets with any quadrature rules on the simplex. In order to achieve the tightness of the framelets and thus exact reconstruction for functions on the simplex by framelets, the quadrature rules are required to be exact for polynomials, see Sects. 3 and 4. However, polynomial-exact rules are generally difficult to construct on the simplex, see [9, Chapter 3].

Triangular Kronecker lattices with equal weights used in Sect. 5 are low-discrepancy [2], but not exact for polynomials. In this case, the reconstruction will incur errors. To overcome this, the masks and quadrature rules shall be constructed to satisfy the condition

$$\overline{\widehat{a}\left(\frac{\lambda_\ell}{2^j}\right)}\widehat{a}\left(\frac{\lambda_{\ell'}}{2^j}\right)\mathscr{U}_{\ell,\ell'}(Q_{N_{j-1}})+\sum_{n=1}^{r}\overline{\widehat{b_n}\left(\frac{\lambda_\ell}{2^j}\right)}\widehat{b_n}\left(\frac{\lambda_{\ell'}}{2^j}\right)\mathscr{U}_{\ell,\ell'}(Q_{N_j})=\mathscr{U}_{\ell,\ell'}(Q_{N_j}),$$

for $j \geq 1$ and for $\ell, \ell' \in \mathbb{N}_0$ satisfying $\overline{\widehat{\alpha}\left(\frac{\lambda_\ell}{2^j}\right)}\widehat{\alpha}\left(\frac{\lambda_{\ell'}}{2^j}\right) \neq 0$, where $\mathscr{U}_{\ell,\ell'}(Q_{N_j}) := \sum_{k=0}^{N_j} \omega_{j,k} P_\ell(\boldsymbol{x}_{j,k}) \overline{P_{\ell'}(\boldsymbol{x}_{j,k})}$ is the numerical integration of $P_\ell \overline{P_{\ell'}}$ over T^2 by quadrature rule Q_{N_j}, see [18, Theorem 2.4]. This condition requires that the quadrature rules for framelets have good properties for numerical integration over the simplex. Besides the triangular Kronecker lattices used in the paper, one may consider other quadrature rules with low discrepancy on the simplex, for example, the analogues to quasi-Monte Carlo (QMC) points in the cube and spheres, see [3, 4, 7, 16].

To implement the fast algorithms for the DFTs in (21) and (22), we need fast transforms for the bases P_ℓ. For example, we can represent the bases $P_{\ell,m}$ in (12) by trigonometric polynomials and apply the FFT on $\mathbb{R}$ to achieve fast algorithms for the DFTs on T^2.

Acknowledgements The authors thank the anonymous referees for their valuable comments. We are grateful to Kinjal Basu, Yuan Xu and Xiaosheng Zhuang for their helpful discussions.

References

1. Aktaş, R., Xu, Y.: Sobolev orthogonal polynomials on a simplex. Int. Math. Res. Not. IMRN **13**, 3087–3131 (2013)
2. Basu, K., Owen, A.B.: Low discrepancy constructions in the triangle. SIAM J. Numer. Anal. **53**(2), 743–761 (2015)
3. Brauchart, J.S., Saff, E.B., Sloan, I.H., Womersley, R.S.: QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere. Math. Comput. **83**(290), 2821–2851 (2014)
4. Brauchart, J.S., Dick, J., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Covering of spheres by spherical caps and worst-case error for equal weight cubature in Sobolev spaces. J. Math. Anal. Appl. **431**(2), 782–811 (2015)
5. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia, PA (1992)
6. de Goes, F., Desbrun, M., Tong, Y.: Vector field processing on triangle meshes. In: ACM SIGGRAPH 2016 Courses, p. 27. ACM, New York (2016)
7. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
8. Dong, B.: Sparse representation on graphs by tight wavelet frames and applications. Appl. Comput. Harmon. Anal. **42**(3), 452–479 (2017)
9. Dunkl, C.F., Xu, Y.: Orthogonal Polynomials of Several Variables. Encyclopedia of Mathematics and its Applications, 2nd edn. Cambridge University Press, Cambridge (2014)
10. Dyn, N.: Subdivision schemes in computer-aided geometric design. In: Advances in Numerical Analysis, vol. II (Lancaster, 1990), pp. 36–104. Oxford University Press, New York (1992)
11. Greco, F., Coox, L., Maurin, F., Desmet, W.: NURBS-enhanced maximum-entropy schemes. Comput. Methods Appl. Mech. Eng. **317**, 580–597 (2017)

12. Han, B.: Pairs of frequency-based nonhomogeneous dual wavelet frames in the distribution space. Appl. Comput. Harmon. Anal. **29**(3), 330–353 (2010)
13. Li, H., Xu, Y.: Discrete Fourier analysis on fundamental domain and simplex of A_d lattice in d-variables. J. Fourier Anal. Appl. **16**(3), 383–433 (2010)
14. Li, H., Sun, J., Xu, Y.: Discrete Fourier analysis, cubature, and interpolation on a hexagon and a triangle. SIAM J. Numer. Anal. **46**(4), 1653–1681 (2008)
15. Maggioni, M., Mhaskar, H.N.: Diffusion polynomial frames on metric measure spaces. Appl. Comput. Harmon. Anal. **24**(3), 329–353 (2008)
16. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford Science Publications. The Clarendon Press/Oxford University Press, New York (1994)
17. Sun, J.: Multivariate Fourier series over a class of non tensor-product partition domains. J. Comput. Math. **21**(1), 53–62 (2003)
18. Wang, Y.G., Zhuang, X.: Tight framelets and fast framelet filter bank transforms on manifolds. Appl. Comput. Harmon. Anal. (to appear)

Solving Partial Differential Equations with Multiscale Radial Basis Functions

Holger Wendland

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The goal of this paper is to review, discuss and extend the current theory on multiscale radial basis functions for solving elliptic partial differential equations. Multiscale radial basis functions provide approximation spaces using different scales and shifts of a compactly supported, positive definite function in an orderly fashion. In this paper, both collocation and Galerkin approximation are described and analysed. To this end, first symmetric and non-symmetric recovery is discussed. Then, error estimates for both schemes are derived, though special emphasis is given to Galerkin approximation, since the current situation here is not as clear as in the case of collocation. We will distinguish between stationary and non-stationary multiscale approximation spaces and multilevel approximation schemes. For Galerkin approximation, we will establish error bounds in the stationary setting based upon Cea's lemma showing that the approximation spaces are indeed rich enough. Unfortunately, convergence of a simple residual correction algorithm, which is often applied in this context to compute the approximation, can only be shown for a non-stationary multiscale approximation space.

1 Introduction

Radial basis functions are by now a well-established tool in multivariate approximation theory with many areas of applications such as image registration, meshfree methods for partial differential equations, fluid-structure interaction, learning theory and many more. There are several books (for example [3, 12, 39]) and survey articles [2, 16, 35] available.

H. Wendland (✉)
Department of Mathematics, University of Bayreuth, Bayreuth, Germany
e-mail: holger.wendland@uni-bayreuth.de

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_55

Multiscale radial basis functions and multilevel algorithms for the approximation of functions have been introduced in [15, 34], right after compactly supported radial basis functions were invented in [37, 43]. However, though earlier attempts at proving convergence have been made in [20, 30] they did not apply to the algorithms developed in [15]. Further numerical observations were made in [4, 11, 13, 38], this time already for the solution of partial differential equations.

The first sophisticated proofs of convergence were given in [24, 25, 28] for multilevel interpolation on the sphere and, based upon this, in [41] for multilevel interpolation on bounded domains. Further improvements for interpolation were then given in [9, 23, 27, 36]. A recent overview is in [42].

After that, proofs of convergence were also given for multilevel methods for solving partial differential equations. For example, proofs for methods based upon collocation are given in [5, 6, 8, 26], while [7] discusses Galerkin approximation.

Finally, multilevel schemes like those discussed here have been used in computer graphics [32], in the context of neural networks [14] and in the context of machine learning [44].

The goal of this paper is to review, discuss and extend results on multiscale radial basis functions and multilevel algorithms for the numerical solution of elliptic partial differential equations. We will address both major techniques in this context, collocation and Galerkin projection. The next section is devoted to laying the mathematical ground for this. It discusses single-scale or one-step discretisation techniques based on radial basis functions in a rather general form. Special emphasis is given to the difference between symmetric and non-symmetric recovery. In Sect. 3, our main section, we will first give a general definition of multiscale RBF approximation spaces and state a general multilevel algorithm. After that, we will discuss results for collocation. The main part of this section is devoted to results on Galerkin projections. Here, a new convergence proof is given and compared to previous results.

2 One-Level Approximation

Before we can discuss multiscale approximation spaces and multilevel methods based on radial basis functions, it is necessary to have a look at the standard one-level radial basis function method for solving partial differential equations. We will do this in the context of optimal recovery but also discuss unsymmetric recovery.

The general setup is as follows. We are given a Hilbert space $\mathscr{H}$ consisting of functions $f : \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^d$ is a bounded domain. We want to recover a function $u \in \mathscr{H}$, for which we only have discrete data $\lambda_1(u), \dots, \lambda_N(u)$, given by linearly independent functionals $\lambda_1, \dots, \lambda_N \in \mathscr{H}^*$, where $\mathscr{H}^*$ denotes the dual space to $\mathscr{H}$.

Definition 1 The *optimal or symmetric recovery* $s_\Lambda \in \mathscr{H}$ of a function $u \in \mathscr{H}$ from the data $f_1 := \lambda_1(u), \ldots, f_N := \lambda_N(u)$ given by $\lambda_1, \ldots, \lambda_N \in \mathscr{H}^*$ is the unique element $s_\Lambda \in \mathscr{H}$ which solves

$$\min \left\{ \|s\|_{\mathscr{H}} : s \in \mathscr{H} \text{ with } \lambda_j(s) = f_j, 1 \le j \le N \right\}.$$

It is well-known (see for example [39]) that the solution s_Λ is indeed unique. Moreover, it can be computed directly if the Riesz representers of the functionals are known.

Theorem 1 *Let $\lambda_1, \ldots, \lambda_N \in \mathscr{H}^*$ be linearly independent. Let $v_1, \ldots, v_N \in \mathscr{H}$ denote their Riesz representers, respectively, i.e. we have $\lambda_j(v) = \langle v, v_j \rangle_{\mathscr{H}}$ for $1 \le j \le N$ and $v \in \mathscr{H}$. Then, the optimal recovery s_Λ of u is given by*

$$s_\Lambda = \sum_{j=1}^{N} \alpha_j v_j,$$

where the coefficients are determined by the linear system $A_\Lambda \boldsymbol{\alpha} = \boldsymbol{f}$ with $A_\Lambda \in \mathbb{R}^{N \times N}$ having entries $a_{ij} = \langle v_i, v_j \rangle_{\mathscr{H}} = \lambda_i(v_j) = \lambda_j(v_i)$ and $\boldsymbol{f} \in \mathbb{R}^N$ having entries $f_i := \lambda_i(u)$.

While the above theorem seems to be satisfactory, we will soon see that it might be reasonable to change the set-up a little. Instead of having one set of linearly independent functionals $\lambda_1, \ldots, \lambda_N \in \mathscr{H}^*$ with Riesz representers $v_1, \ldots, v_N$, let us now assume that we have a second set of linearly independent functionals $\mu_1, \ldots, \mu_N \in \mathscr{H}^*$ with Riesz representers $w_1, \ldots, w_N \in \mathscr{H}$. We will only assume that each family of functionals is linearly independent but not that the λ_j functionals are independent of the μ_j functionals. As a matter of fact, we will allow $\lambda_j = \mu_j$ so that we are back in the situation of optimal recovery.

Definition 2 The unsymmetric recovery $s_{\Lambda,M} \in \mathscr{H}$ of a function $u \in \mathscr{H}$ from the data $f_1 := \lambda_1(u), \ldots, f_N := \lambda_N(u)$ using the functionals $\mu_1, \ldots, \mu_N \in \mathscr{H}^*$ with Riesz representers $w_1, \ldots, w_N$ is defined to be the element

$$s_{\Lambda,M} = \sum_{j=1}^{N} \alpha_j w_j$$

where $\boldsymbol{\alpha} \in \mathbb{R}^N$ is the solution of $A_{\Lambda,M} \boldsymbol{\alpha} = \boldsymbol{f}$, where $A_{\Lambda,M}$ has entries $a_{ij} = \lambda_i(w_j) = \langle v_i, w_j \rangle_{\mathscr{H}}$, provided that $A_{\Lambda,M}$ is invertible, and $\boldsymbol{f} \in \mathbb{R}^N$ is given by $f_j := \lambda_j(u)$.

In the case of $\lambda_j = \mu_j$ for $1 \le j \le N$ the unsymmetric recovery becomes the symmetric recovery. From this point of view, the unsymmetric recovery is a generalisation of the symmetric case.

In contrast to symmetric or optimal recovery, where the collocation matrix A_Λ is positive definite by definition and hence invertible, the situation is not so clear in the case of unsymmetric collocation, where the matrix $A_{\Lambda,M}$ might not even be symmetric any more and might even become singular. We will discuss this in more details later on.

We will use all of this in the specific situation of $\mathscr{H}$ being a reproducing kernel Hilbert space, since this allows us to state the Riesz representers of functionals explicitly.

Definition 3 A Hilbert space $\mathscr{H}$ consisting of continuous functions $f : \Omega \to \mathbb{R}$ with $\Omega \subseteq \mathbb{R}^d$ is called a *reproducing kernel Hilbert space*, if there is a unique function $\Phi : \Omega \times \Omega \to \mathbb{R}$ with the properties

- $\Phi(\cdot,\boldsymbol{x}) \in \mathscr{H}$ for all $\boldsymbol{x} \in \Omega$,
- $f(\boldsymbol{x}) = \langle f, \Phi(\cdot,\boldsymbol{x})\rangle_{\mathscr{H}}$ for all $f \in \mathscr{H}$ and all $\boldsymbol{x} \in \Omega$.

The function Φ is called the *reproducing kernel* of $\mathscr{H}$.

The second property shows that the point evaluation functional $\delta_{\boldsymbol{x}}$ has $\Phi(\cdot,\boldsymbol{x})$ as its Riesz representer, i.e. the Riesz representer is given by applying this functional to the second argument of the kernel. This is also true for any other functional $\lambda \in \mathscr{H}^*$. To see this, we first note that the kernel is necessarily symmetric, since

$$\Phi(\boldsymbol{x},\boldsymbol{y}) = \langle \Phi(\cdot,\boldsymbol{y}), \Phi(\cdot,\boldsymbol{x})\rangle_{\mathscr{H}} = \langle \Phi(\cdot,\boldsymbol{x}), \Phi(\cdot,\boldsymbol{y})\rangle_{\mathscr{H}} = \Phi(\boldsymbol{y},\boldsymbol{x}).$$

Hence, if $v_\lambda \in \mathscr{H}$ denotes the Riesz representer of $\lambda \in \mathscr{H}^*$ then

$$\lambda(\Phi(\boldsymbol{x},\cdot)) = \langle \Phi(\boldsymbol{x},\cdot), v_\lambda\rangle_{\mathscr{H}} = \langle v_\lambda, \Phi(\cdot,\boldsymbol{x})\rangle_{\mathscr{H}} = v_\lambda(\boldsymbol{x}).$$

As a consequence, we can express the system matrices A_Λ and $A_{\Lambda,M}$ completely in terms of the kernel. In the following we will use the notation $\lambda^{\boldsymbol{x}}\Phi(\cdot,\boldsymbol{x})$ to indicate that the functional λ acts with respect to the second variable of the kernel.

Corollary 1 *Let $\mathscr{H}$ be a reproducing kernel Hilbert space with reproducing kernel Φ. Let $\Lambda = \{\lambda_1, \dots, \lambda_N\} \subseteq \mathscr{H}^*$ and $M = \{\mu_1, \dots, \mu_N\} \subseteq \mathscr{H}^*$ denote two sets of linearly independent functionals. Then, the system matrices associated to symmetric and unsymmetric collocation are given by*

$$A_\Lambda = \left(\lambda_i^{\boldsymbol{x}}\lambda_j^{\boldsymbol{y}}\Phi(\boldsymbol{x},\boldsymbol{y})\right), \qquad A_{\Lambda,M} = \left(\lambda_i^{\boldsymbol{x}}\mu_j^{\boldsymbol{y}}\Phi(\boldsymbol{x},\boldsymbol{y})\right).$$

Let us shortly discuss what this means when it comes to solving a simple elliptic boundary value problem with either collocation or a Galerkin approach. In both cases, we will assume that the kernel $\Phi : \Omega \times \Omega \to \mathbb{R}$ is actually defined on $\mathbb{R}^d \times \mathbb{R}^d$ and is translation invariant, i.e. there is an even function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that

$$\Phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}-\boldsymbol{y}) = \phi(\boldsymbol{y}-\boldsymbol{x}), \qquad \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d.$$

We start with collocation and look at the toy problem

$$-\Delta u = f \text{ in } \Omega, \qquad u = g \text{ on } \partial\Omega.$$

In the case of collocation, we want to enforce these equations on discrete points. Hence, we can pick data sites $X_1 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \Omega$ and $X_2 = \{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\} \subseteq \partial\Omega$ and define the test functionals λ_j via

$$\lambda_j(u) = \begin{cases} -\Delta u(\boldsymbol{x}_j) & \text{for } 1 \le j \le n, \\ u(\boldsymbol{x}_j) & \text{for } n+1 \le j \le N. \end{cases}$$

Then, the symmetric recovery will form approximants as

$$s_\Lambda(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \lambda_j^{\boldsymbol{y}} \Phi(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{j=1}^{n} \alpha_j \Delta\phi(\boldsymbol{x} - \boldsymbol{x}_j) + \sum_{j=n+1}^{N} \alpha_j \phi(\boldsymbol{x} - \boldsymbol{x}_j).$$

The associated system matrix has a natural block structure of the form

$$A_\Lambda = \begin{pmatrix} \Delta^2\phi(\boldsymbol{x}_i - \boldsymbol{x}_j) & -\Delta\phi(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ -\Delta\phi(\boldsymbol{x}_i - \boldsymbol{x}_j) & \phi(\boldsymbol{x}_i - \boldsymbol{x}_j) \end{pmatrix},$$

and the right-hand side $\boldsymbol{f} \in \mathbb{R}^N$ has entries $f_j = f(\boldsymbol{x}_j)$ for $1 \le j \le n$ and $f_j = g(\boldsymbol{x}_j)$ for $n+1 \le j \le N$.

This matrix is symmetric and positive definite and this type of symmetric collocation has been studied extensively, for example in [10, 17–19, 29, 40], giving explicit error and stability results on the collocation process.

However, since it requires twice the application of the differential operator to compute the system matrix, one often also sees the unsymmetric approach, where the functionals $\mu_j(u) = \delta_j(u) = u(\boldsymbol{x}_j)$ and their Riesz representers $w_j = \Phi(\cdot, \boldsymbol{x}_j) = \phi(\cdot - \boldsymbol{x}_j)$ are used to form the approximation space. In this situation the unsymmetric recovery becomes

$$s_{\Lambda,M}(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \phi(\boldsymbol{x} - \boldsymbol{x}_j)$$

with the system matrix

$$A_{\Lambda,M} = \begin{pmatrix} -\Delta\phi(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ \phi(\boldsymbol{x}_i - \boldsymbol{x}_j) \end{pmatrix},$$

which clearly requires only one application of the differential operator. However, it is also apparent that the system matrix is no longer symmetric and, as a matter of fact, might even become singular [22].

We can conclude from this simple example that when it comes to collocation, the symmetric approach is the mathematically sound one and this is the one we will further investigate later on.

For our second example, we slightly change the toy problem. This time, we look at

$$-\Delta u + u = f \text{ in } \Omega, \qquad \frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \partial\Omega,$$

where $\boldsymbol{n}$ denotes the unit outer normal vector on $\partial\Omega$. The reason for this change is that the weak formulation for this new toy problem becomes

$$a(u,v) := \int_\Omega (\nabla u \cdot \nabla v + uv)\, dx = \int_\Omega f v dx, \qquad u, v \in H^1(\Omega),$$

so that the bilinear form is coercive for $H^1(\Omega)$ and not only for $H_0^1(\Omega)$.

To discretise this weak problem using our symmetric or unsymmetric recovery strategies, we choose again discretisation points $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \Omega$ and define functionals

$$\lambda_j(u) := a(u, \Phi(\cdot, \boldsymbol{x}_j)), \qquad 1 \le j \le N.$$

Interestingly, these functionals employ the Riesz representers of our second family of functionals μ_j defined by $\mu_j(u) = u(\boldsymbol{x}_j)$ so that we can also write

$$\lambda_j(u) = a(u, \mu_j^{\boldsymbol{y}} \Phi(\cdot, \boldsymbol{y})).$$

Since a is $H^1(\Omega)$-coercive, this shows that this time the unsymmetric approach is the better one. For the unsymmetric approach, we form approximants as

$$s_{\Lambda,M}(\boldsymbol{x}) = \sum_{j=1}^N \alpha_j \mu_j^{\boldsymbol{y}} \Phi(\cdot, \boldsymbol{y}) = \sum_{j=1}^N \alpha_j \Phi(\cdot, \boldsymbol{x}_j) \tag{1}$$

and determine the coefficients α_j using the system matrix

$$A_{\Lambda,M} = (\lambda_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{y}} \Phi(\boldsymbol{x}, \boldsymbol{y})) = (\lambda_i^{\boldsymbol{x}} \Phi(\boldsymbol{x}, \boldsymbol{x}_j)) = (a(\Phi(\cdot, \boldsymbol{x}_j), \Phi(\cdot, \boldsymbol{x}_i))).$$

This matrix is now positive definite and symmetric and the reconstruction (1) is the Galerkin approximation to u from $V_X = \text{span}\{\Phi(\cdot, \boldsymbol{x}_1), \ldots, \Phi(\cdot, \boldsymbol{x}_N)\}$.

Corollary 2 *In the case of functionals $\lambda_j(u) = a(u, \Phi(\cdot, \boldsymbol{x}_j))$ and $\mu_j(u) = u(\boldsymbol{x}_j)$, the unsymmetric approach gives the Galerkin best approximation as the recovery. In particular, by Cea's Lemma we have the error estimate*

$$\|u - s_{\Lambda,M}\|_{H^1(\Omega)} \le C \inf_{s \in V_X} \|u - s\|_{H^1(\Omega)}.$$

Nonetheless, it is interesting to see that we could also look at the symmetric approach. Here, the approximant would be of the form

$$s_\Lambda(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \lambda_j^y \Phi(\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^{N} \alpha_j a(\Phi(\boldsymbol{x},\cdot), \Phi(\cdot,\boldsymbol{x}_j)).$$

The coefficients are again determined by solving a linear system with the system matrix A_Λ having entries

$$a_{ij} = \lambda_i^x \lambda_j^y \Phi(\boldsymbol{x},\boldsymbol{y}) = a_x(\Phi(\boldsymbol{x},\boldsymbol{x}_i), a_y(\Phi(\boldsymbol{x},\boldsymbol{y}), \Phi(\boldsymbol{y},\boldsymbol{x}_j))),$$

where the index at the bilinear form a indicates the variable with respect to which a is applied.

In the case of a translation invariant and even kernel $\Phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}-\boldsymbol{y})$ this can be reformulated as

$$\begin{aligned} a_{ij} = &- \int_\Omega \int_\Omega \nabla\phi(\boldsymbol{y}-\boldsymbol{x}_j) \cdot H\phi(\boldsymbol{x}-\boldsymbol{y}) \nabla\phi(\boldsymbol{x}-\boldsymbol{x}_i) d\boldsymbol{x} d\boldsymbol{y} \\ &+ \int_\Omega \int_\Omega \phi(\boldsymbol{y}-\boldsymbol{x}_j) \nabla\phi(\boldsymbol{x}-\boldsymbol{y}) \cdot \nabla\phi(\boldsymbol{x}-\boldsymbol{x}_i) d\boldsymbol{x} d\boldsymbol{y} \\ &+ \int_\Omega \int_\Omega \phi(\boldsymbol{x}-\boldsymbol{x}_i) \nabla\phi(\boldsymbol{x}-\boldsymbol{y}) \cdot \nabla\phi(\boldsymbol{y}-\boldsymbol{x}_j) d\boldsymbol{x} d\boldsymbol{y} \\ &+ \int_\Omega \int_\Omega \phi(\boldsymbol{x}-\boldsymbol{y}) \phi(\boldsymbol{y}-\boldsymbol{x}_j) \phi(\boldsymbol{x}-\boldsymbol{x}_i) d\boldsymbol{x} d\boldsymbol{y}, \end{aligned}$$

where $H\phi$ denotes the Hessian matrix having entries $(H\phi)_{ij} = \partial_i \partial_j \phi$. Clearly, this approach has the disadvantage of requiring second order derivatives of ϕ and also requiring the computation of double integrals. Hence, it is not surprising that so far this approach has not been investigated.

From both examples we can summarise our findings as follows. When it comes to collocation then the symmetric approach is the right choice, when it comes to Galerkin approximation then the unsymmetric approach is the right choice.

3 Multilevel Approximation

We will now describe the general setup of multiscale radial basis functions. To this end, we have to understand how scaling affects a reproducing kernel of a reproducing kernel Hilbert space. Since we are mainly interested in solving elliptic PDEs we will, from now on, concentrate on Sobolev spaces $H^\sigma(\Omega)$, which are known to be reproducing kernel Hilbert spaces if the Sobolev embedding theorem holds, i.e. if $\sigma > d/2$. If Ω has a Lipschitz boundary then it is well-known that each function can be extended to become a function in $H^\sigma(\mathbb{R}^d)$. Thus, we will concentrate on reproducing kernels of Sobolev spaces $H^\sigma(\mathbb{R}^d)$. The functions $u \in H^\sigma(\mathbb{R}^d)$ can be described using Fourier transforms. To be more precise, if we define the Fourier transform of a function $u \in L_1(\mathbb{R}^d)$ by

$$\widehat{u}(\boldsymbol{\omega}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} u(\boldsymbol{x}) e^{-i\boldsymbol{x}^T\boldsymbol{\omega}} d\boldsymbol{x}$$

and extend this definition in the usual way to functions $u \in L_2(\mathbb{R}^d)$, then we have the following definition.

Definition 4 For $\sigma \geq 0$, the Sobolev space $H^\sigma(\mathbb{R}^d)$ consists of all functions $u \in L_2(\mathbb{R}^d)$ with

$$\|u\|^2_{H^\sigma(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\widehat{u}(\boldsymbol{\omega})|^2 (1 + \|\boldsymbol{\omega}\|_2^{2\sigma}) d\boldsymbol{\omega} < \infty, \tag{2}$$

Though the reproducing kernel of a reproducing kernel Hilbert space is unique, it is possible to equip this space with another reproducing kernel by altering the inner product of the space. This is frequently used in the theory of radial basis functions. The example, which is relevant for us here, is as follows. Assume that $\phi : \mathbb{R}^d \to \mathbb{R}$ is an integrable function having a Fourier transform $\widehat{\phi}$ satisfying

$$c_1(1 + \|\boldsymbol{\omega}\|_2^{2\sigma})^{-1} \leq \widehat{\phi}(\boldsymbol{\omega}) \leq c_2(1 + \|\boldsymbol{\omega}\|_2^{2\sigma})^{-1}, \qquad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{3}$$

with certain fixed constants $c_1, c_2 > 0$ and $\sigma > d/2$. Then, $\Phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x} - \boldsymbol{y})$ is also a reproducing kernel of $H^\sigma(\mathbb{R}^d)$ but with respect to the inner product

$$\langle f, g \rangle_\phi := \int_{\mathbb{R}^d} \frac{\widehat{f}(\boldsymbol{\omega})\overline{\widehat{g}(\boldsymbol{\omega})}}{\widehat{\phi}(\boldsymbol{\omega})} d\boldsymbol{\omega}, \qquad f, g \in H^\sigma(\mathbb{R}^d).$$

The induced norm $\|\cdot\|_\phi^2 := \langle \cdot, \cdot \rangle_\phi$ is obviously equivalent to the norm defined by (2), where the equivalence constants are determined by c_1 and c_2 from (3).

Important for us is yet another matter. On $H^\sigma(\mathbb{R}^d)$ with $\sigma \geq 0$, we can introduce a scaled norm as follows.

Definition 5 Let $\sigma \geq 0$ be given. Then for each $\delta > 0$, the Sobolev space $H^\sigma(\mathbb{R}^d)$ can be equipped with a scaled norm defined by

$$\|u\|^2_{H^\sigma_\delta} := \int_{\mathbb{R}^d} |\widehat{u}(\boldsymbol{\omega})|^2 (1 + (\delta^2 \|\boldsymbol{\omega}\|_2^2)^\sigma) d\boldsymbol{\omega}.$$

Obviously, this indeed defines a norm on $H^\sigma(\mathbb{R}^d)$ but this time the norm equivalence constants depend on δ. To be more precise, it is straight-forward to see that we have for each $\sigma > 0$ and each $\delta \in (0, 1]$,

$$\|u\|_{H^\sigma_\delta} \leq \|u\|_{H^\sigma(\mathbb{R}^d)} \leq \delta^{-\sigma} \|u\|_{H^\sigma_\delta}, \qquad u \in H^\sigma(\mathbb{R}^d). \tag{4}$$

If we now come back to the case $\sigma > d/2$ and scale a function ϕ having a Fourier transform satisfying (3) by setting $\phi_\delta := \delta^{-d}\phi(\cdot/\delta)$ then its Fourier transform is given by $\widehat{\phi_\delta} = \widehat{\phi}(\delta\cdot)$ such that ϕ_δ generates a norm which is equivalent to the H^σ_δ norm, i.e. it satisfies

$$c_1\|u\|_{\phi_\delta} \le \|u\|_{H^\sigma_\delta} \le c_2\|u\|_{\phi_\delta}, \qquad u \in H^\sigma(\mathbb{R}^d).$$

This can be rephrased as follows.

Lemma 1 *Let ϕ define a reproducing kernel Φ for $H^\sigma(\mathbb{R}^d)$, $\sigma > d/2$, i.e. let $\widehat{\phi}$ satisfy (3). Then, ϕ_δ defines also a reproducing kernel Φ_δ of $H^\sigma(\mathbb{R}^d)$. For $0 < \delta \le 1$, the norms satisfy*

$$c_1\|u\|_{\phi_\delta} \le \|u\|_{H^\sigma_\delta} \le \|u\|_{H^\sigma(\mathbb{R}^d)} \le \delta^{-\sigma}\|u\|_{H^\sigma_\delta} \le c_2\delta^{-\sigma}\|u\|_{\phi_\delta}, \qquad u \in H^\sigma(\mathbb{R}^d).$$

This simple observation allows us now to introduce a sequence of reproducing kernels and our multiscale approximation spaces.

Definition 6 Let $\sigma > d/2$ and let $\phi \in L_1(\mathbb{R}^d)$ satisfy (3) such that $\Phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}-\boldsymbol{y})$ is a reproducing kernel of $H^\sigma(\mathbb{R}^d)$. Let $1 \ge \delta_1 \ge \delta_2 \ge \cdots \ge \delta_m \ge \cdots$ be a non-increasing sequence of scales and define the scaled kernels

$$\Phi_j(\boldsymbol{x},\boldsymbol{y}) := \delta_j^{-d}\phi\left(\left(\boldsymbol{x}-\boldsymbol{y}\right)/\delta_j\right), \qquad \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d. \tag{5}$$

Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain. For each $j \in \mathbb{N}$ let $M_j = \{\mu_1^{(j)},\dots,\mu_{N_j}^{(j)}\} \subseteq H^\sigma(\Omega)^*$ be a set of linearly independent functionals. Then, a *single-scale approximation space* of level j is defined as

$$W_j := \operatorname{span}\left\{\mu^{\boldsymbol{y}}\Phi_j(\cdot,\boldsymbol{y}) : \mu \in M_j\right\}$$

and the *multiscale approximation space* of level j is defined as

$$V_j := W_1 + \cdots + W_j.$$

Often, the functionals $\mu_k^{(j)} \in M_j$ have a one point support $\boldsymbol{x}_k^{(j)}$ meaning that $\mu_k^{(j)}(f) = 0$ for all $f \in C(\Omega)$ with $\boldsymbol{x}_k^{(j)} \not\in \operatorname{supp}(f)$. Later on, we will see that this is indeed the case for the discretisations discussed in this paper. In such a case the fill distance of the set $X_j = \{\boldsymbol{x}_1^{(j)},\dots,\boldsymbol{x}_{N_j}^{(j)}\} \subseteq \Omega$ defined by

$$h_j := h_{X_j,\Omega} := \sup_{\boldsymbol{x}\in\Omega}\ \min_{\boldsymbol{x}_k^{(j)}\in X_j} \|\boldsymbol{x}-\boldsymbol{x}_k^{(j)}\|_2$$

will play a crucial role and we will distinguish our multiscale approximation spaces in the following way.

Definition 7 If each functional $\mu_k \in M_j$ has a one point support $\boldsymbol{x}_k^{(j)} \in X_j \subseteq \Omega$, then the multiscale approximation spaces V_j are called

- *stationary*, if there is a constant $\nu > 0$ such that $\delta_j = \nu h_j$.
- *non-stationary*, if $h_j/\delta_j \to 0$ for $j \to \infty$.

One of the reasons for looking at multiscale approximation spaces is that in the stationary setting the single-scale approximation spaces are usually not rich enough to provide good approximations, see [3, 33, 39].

Throughout this paper, we will work with kernels $\Phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}-\boldsymbol{y})$, where $\phi : \mathbb{R}^d \to \mathbb{R}$ has compact support. Without restriction, we will assume that the support of ϕ is given by the closed unit ball about zero. If in this case the functionals also have a one point support then the single-scale approximation spaces W_j are built from *local* basis functions $(\mu_k^{(j)})^y \Phi(\cdot,\boldsymbol{y})$ having the closed ball of radius δ_j about $\boldsymbol{x}_k^{(j)}$ as support. In this situation we will call W_j also a local approximation space.

We will discuss such spaces for both collocation and Galerkin methods. Before that we will give a simple algorithm to determine an approximation from the space V_j to a given function $f \in H^\sigma(\Omega)$. This algorithm is based on an iterative residual correction and can be formulated in general form using yet another sequence of functionals.

Hence, for each $j \in \mathbb{N}$ let $\Lambda_j := \{\lambda_1^{(j)}, \dots, \lambda_j^{(N_j)}\} \subseteq H^\sigma(\Omega)^*$ be another set of linearly independent functionals, then, the residual correction algorithm can be described as in Algorithm 1. Note that the sets M_j of functionals are used implicitly to define the approximation spaces W_j and hence V_j.

Obviously, the crucial point of Algorithm 1 is to find the approximant $s_j \in W_j$, where W_j is the space built with the trial functionals $\mu_k^{(j)}$, using the test functionals $\lambda_k^{(j)}$. This can be done in various ways. We will mainly be interested in generalised interpolation, where the approximant satisfies $\lambda_k^{(j)}(s_j) = \lambda_k^{(j)}(e_{j-1})$, $1 \le k \le N_j$, i.e. for each level j we have to invert a system matrix as described in Corollary 1. Taking the comments after Corollary 1 into account, we will do this only for combinations of test and trial functionals which guarantee an invertible system matrix. We could, however, also use other techniques like least-squares to determine the approximant s_j, if the system matrix is not invertible.

Algorithm 1 Multilevel approximation

```
Input: Right-hand side $f$, number of levels $n$, sets of functionals $\Lambda_j$.
Output: Approximate solution $u_n \in V_n = W_1 + \cdots + W_n$
Set $u_0 = 0$, $e_0 = f$.
for $j = 1, 2, \dots, n$ do
    Determine a single-scale approximant $s_j \in W_j$ to $e_{j-1}$ such that $\lambda(s_j)$ is close to $\lambda(e_{j-1})$ for
    all $\lambda \in \Lambda_j$.
    Set $u_j = u_{j-1} + s_j$.
    Set $e_j = e_{j-1} - s_j$.
end for
```

3.1 Multilevel Collocation

We now want to discuss the ideas of using multiscale approximation spaces to solve partial differential equations with collocation. In this context, Algorithm 1 is an appropriate tool.

Following the ideas above, we will employ symmetric recovery to determine the single-scale approximant s_j. Let us hence look again at a typical elliptic problem

$$Lu = f \text{ in } \Omega, \qquad u = g \text{ on } \partial\Omega, \tag{6}$$

with given right-hand sides f, g and given elliptic operator L.

To describe the multiscale spaces and the multilevel algorithm, we slightly change the notation. To define the functionals Λ_j on level j, we choose discrete point sets $X_j = Y_j \cup Z_j \subseteq \overline{\Omega}$ consisting of interior points $Y_j \subseteq \Omega$ and boundary points $Z_j \subseteq \partial\Omega$ to represent the functionals in the interior and on the boundary separately. Then, a typical functional $\lambda_k \in \Lambda_j$ takes the form

$$\lambda_k(u) := \begin{cases} Lu(\boldsymbol{x}_k) & \text{if } \boldsymbol{x}_k \in Y_j \\ u(\boldsymbol{x}_k) & \text{if } \boldsymbol{x}_k \in Z_j. \end{cases}$$

Furthermore, for a practical realisation of Algorithm 1 it is more important to record the residuals $f_j = f_{j-1} - Ls_j$ and $g_j = g_{j-1} - s_j$ rather than the error $e_j = e_{j-1} - s_j$. These residuals have only to be computed on the discrete point sets starting with level $j + 1$ and, if the point sets X_j are nested, then the residuals have only to be computed on the finest level n.

Taking this into account, the generic multilevel algorithm becomes the multilevel collocation algorithm, given in Algorithm 2.

Algorithm 2 Multilevel collocation algorithm

```
Input: Right-hand sides $f$ and $g$, number of levels $n$.
Output: Approximate solution $u_n \in V_n = W_1 + \cdots + W_n$.
Set $u_0 = 0, f_0 = f, g_0 = g$
for $j = 1, 2, 3 \ldots$ do
    Determine the single-scale correction $s_j$ to $f_{j-1}$ and $g_{j-1}$ with
        $Ls_j(\boldsymbol{y}) = f_{j-1}(\boldsymbol{y}), \quad \boldsymbol{y} \in Y_j,$
        $s_j(\boldsymbol{z}) = g_{j-1}(\boldsymbol{z}), \quad \boldsymbol{z} \in Z_j.$
    Update the global approximation and the residuals:
        $u_j = u_{j-1} + s_j$
        $f_j = f_{j-1} - Ls_j$
        $g_j = g_{j-1} - s_j$
end for
```

This multilevel scheme has been thoroughly investigated in [8]. Hence, we will only report here on the main convergence result since we want to compare the results to those of the Galerkin multilevel scheme which we will introduce next.

To state the convergence result, we have to recall the fill distance, which is usually employed when it comes to measuring convergence in the area of scattered data approximation.

For point sets $Y_j \subseteq \Omega$ and $Z_j \subseteq \partial\Omega$, the fill distance on Ω and $\partial\Omega$, respectively, are defined as

$$h_{Y_j,\Omega} := \sup_{\boldsymbol{x}\in\Omega} \min_{\boldsymbol{x}_j\in Y_j} \|\boldsymbol{x}-\boldsymbol{x}_j\|_2, \qquad h_{Z_j,\partial\Omega} := \sup_{\boldsymbol{x}\in\partial\Omega} \min_{\boldsymbol{x}_j\in Z_j} \operatorname{dist}(\boldsymbol{x},\boldsymbol{x}_j),$$

where dist denotes the intrinsic distance function on $\partial\Omega$, i.e. the length of the shortest connecting curve.

Theorem 2 *Let $\Omega \subseteq \mathbb{R}^d$ have a $C^{k,s}$-boundary for $s \in [0,1)$, $k \in \mathbb{N}_0$ with $k > d/2$. Assume that $u \in W_2^\sigma(\Omega)$ solves (6) with $\sigma = k + s$. Let $Y_1, Y_2, \ldots$ be a sequence of point sets in Ω and let $Z_1, Z_2, \ldots$ be a sequence of point sets in $\partial\Omega$ having fill distances $h_{Y_j,\Omega}$ and $h_{Z_j,\partial\Omega}$, respectively. Define $h_j := \max\{h_{X_j,\Omega}, h_{Y_j,\partial\Omega}\}$ and assume that there are constants $\mu \in (0,1)$ and $\gamma \in (0,1]$ such that*

$$\gamma\mu h_j \leq h_{j+1} \leq \mu h_j \tag{7}$$

for $j = 1, 2, \ldots$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a continuous, compactly supported function with a Fourier transform satisfying (3). Define the scaled kernels Φ_j by (5), where the scales δ_j satisfy

$$\delta_j = \left(\frac{h_j}{\mu}\right)^{1-\frac{2}{\sigma}}. \tag{8}$$

Then, provided that $h_1 \leq \mu$ is sufficiently small, there exist constants $C, C_1 > 0$, independent of μ, u, j such that

$$\|u - u_n\|_{L_2(\Omega)} \leq C_1 (C\mu^{\sigma-2})^n \|u\|_{H^\sigma(\Omega)}, \qquad n = 1, 2, 3, \ldots. \tag{9}$$

Hence, if the constant $\mu \in (0,1)$ has been chosen sufficiently small, so that $\alpha := C\mu^{\sigma-2} < 1$, the multiscale approximation u_n converges linearly in the number of levels to u.

As mentioned above, a proof of this result can be found in [8]. Here, we want to point out a few additional things.

First of all, since δ_j is not proportional to h_j, but follows the rule (8), we have a non-stationary multiscale approximation space. This has the disadvantage that the collocation matrices become denser and denser from level to level. Another disadvantage is that the condition number also grows from level to level. Nonetheless, the improvement when compared to the non-stationary single-scale approximation is significant. However, it would be desirable to have also convergence of the multilevel algorithm in the case of a stationary multiscale approximation space. Unfortunately, numerical results from [11] indicate that in this case the simple multilevel algorithm does not converge. Suggestions of improving the algorithm have been given in [11, 13] but more research in this direction seems to be necessary.

Finally, note that (9) actually also gives convergence orders in terms of the fill distance. To see this, assume for simplicity that we have $h_{j+1} = \mu h_j$ for all j and $h_1 = \mu$. Then, we obviously have $h_n = \mu^n$ or $\mu = h_n^{1/n}$. Inserting this into (9) gives the following result.

Corollary 3 *Under the assumption of Theorem 2 let $\gamma = 1$, i.e. $h_{j+1} = \mu h_j$. Let $\epsilon > 0$ and choose $\mu \in (0, 1)$ so small that $C\mu^\epsilon < 1$. Then, we have the error bound*

$$\|u - u_n\|_{L_2(\Omega)} \leq C_1 h_n^{\sigma-2-\epsilon} \|u\|_{H^\sigma(\Omega)}.$$

Proof As mentioned above, we have $\mu^n = h_n$. Moreover, from (9) and the fact that $C\mu^\epsilon < 1$, we see that

$$\begin{aligned}\|u - u_n\|_{L_2(\Omega)} &\leq C_1 (C\mu^\epsilon \mu^{\sigma-2-\epsilon})^n \|u\|_{H^\sigma}(\Omega)\\ &< C_1 \mu^{(\sigma-2-\epsilon)n} \|u\|_{H^\sigma(\Omega)}\\ &= C_1 h_n^{\sigma-2-\epsilon} \|u\|_{H^\sigma(\Omega)}.\end{aligned}$$ □

It is interesting to see that, with ϵ going to zero, this yields the same approximation order $\sigma - 2$ which the one-level approach with a fixed support radius would yield.

3.2 Multilevel Galerkin Approximation

The idea of using a multilevel scheme in the context of Galerkin approximation has been suggested in [38] and has then further been investigated in [7]. However, there have been different observations regarding the convergence of the multilevel algorithm and we will take a closer look at this now.

The idea is as follows. Again assume that we want to solve a strictly elliptic PDE with natural boundary conditions of the form

$$-\mathrm{div}(A\nabla u) + u = f \quad \text{in } \Omega, \qquad \boldsymbol{n} \cdot A\nabla u = 0 \quad \text{on } \partial\Omega$$

with a positive definite matrix $A \in \mathbb{R}^{d \times d}$. Then the weak formulation means to find a function $u \in H^1(\Omega)$ such that

$$a(u, v) := \int_\Omega [(A\nabla u) \cdot \nabla v + uv]\, d\boldsymbol{x} = F(v) := \int_\Omega f v d\boldsymbol{x}, v \in H^1(\Omega). \tag{10}$$

Solving this problem in a multiscale radial basis function setting means that we pick increasingly finer data sets $X_1, X_2, \ldots \subseteq \Omega$ and a non-increasing sequence of scales $\delta_1 \geq \delta_2 \geq \ldots$ to define approximation spaces

$$W_j := \operatorname{span}\{\Phi_j(\cdot, \boldsymbol{x}) : \boldsymbol{x} \in X_j\},$$
$$V_j := W_1 + \cdots + W_j.$$

Then, the classical Galerkin approach leads to the following definition.

Definition 8 The approximate Galerkin solution to the weak problem (10) is defined as the function $u_n^* \in V_n$ satisfying

$$a(u_n^*, v) = F(v), \qquad v \in V_n.$$

Since the bilinear form a is $H^1(\Omega)$-coercive, Cea's lemma tells us now that u_n^* approximates the true solution $u \in H^1(\Omega)$ approximately as good as the best $H^1(\Omega)$-approximation to u from V_n. This gives in particular the following result.

Theorem 3 *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a continuous, compactly supported function with Fourier transform satisfying (3) with $\sigma > d/2$. Assume that the solution u of (10) satisfies $u \in H^\sigma(\Omega)$. Let the scaled kernel Φ_j of (5) be defined with $\delta_j = h_j/\mu$, where h_j is the fill distance of X_j in Ω and $\mu \in (0,1)$ is fixed. Finally, assume that $h_{j+1} = \mu h_j$. Then, for each $\epsilon > 0$ there is a $\mu_0 = \mu_0(\epsilon) \in (0,1)$ such that the approximate Galerkin solution $u_n^* \in V_n$ satisfies the error bound*

$$\|u - u_n^*\|_{H^1(\Omega)} \leq C h_n^{\sigma-1-\epsilon} \|u\|_{H^\sigma(\Omega)},$$

provided $\mu \leq \mu_0(\epsilon)$.

Proof By Cea's lemma we have

$$\|u - u_n^*\|_{H^1(\Omega)} \leq C \inf_{v \in V_n} \|u - v\|_{H^1(\Omega)}.$$

The latter best approximation error can be bounded by a multilevel interpolant, which gives the stated error bound, see [41] and the extensions in [42]. □

The question remains how to efficiently compute the approximate Galerkin solution of the multiscale approximation space. A natural choice is to use our multilevel residual correction algorithm. In this situation it takes the specific form given in Algorithm 3.

It is important to note that $u_n \in V_n$ produced by this algorithm is not the approximate Galerkin solution u_n^*. Nonetheless, the sequences $\{u_j\}$, $\{s_j\}$ and $\{e_j\}$ produced by Algorithm 3 satisfy the following relations.

Proposition 1 *The sequences $\{u_j\}$, $\{s_j\}$ and $\{e_j\}$ from Algorithm 3 have the following properties.*

Algorithm 3 Multilevel Galerkin approximation

```
Input: Right-hand side $f$, number of levels $n$
Output: Approximate solution $u_n \in V_n := W_1 + \cdots + W_n$
Set $u_0 = 0, e_0 = f$.
for $j = 1, 2, \ldots, n$ do
    Determine a single-scale approximant $s_j \in W_j$ to $e_{j-1}$ with
        $a(s_j, v) = F(v) - a(u_{j-1}, v), \qquad v \in W_j.$
    Set $u_j = u_{j-1} + s_j$.
    Set $e_j = e_{j-1} - s_j$.
end for
```

1. *As before, we have* $u_n = s_1 + \cdots + s_n$ *and* $e_n = e_{n-1} - s_n = u - u_n$.
2. *The function* $u_j \in V_j$ *satisfies* $a(u_j, v) = F(v)$ *for all* $v \in W_j$. *Hence,*

$$\|u - u_j\|_{H^1(\Omega)} \le C \inf_{v \in W_j} \|u - u_{j-1} - v\|_{H^1(\Omega)}.$$

3. *The function* s_j *is the approximate Galerkin solution to* e_{j-1}, *i.e* $a(s_j, v) = a(e_{j-1}, v)$ *for all* $v \in W_j$. *Hence,*

$$\|e_{j-1} - s_j\|_{H^1(\Omega)} \le C \inf_{v \in W_j} \|e_{j-1} - v\|_{H^1(\Omega)}.$$

4. *The sequence* $\{e_j\}$ *is monotone decreasing in the energy norm, i.e.* $a(e_j, e_j) \le a(e_{j-1}, e_{j-1})$ *and hence it possesses the stability property*

$$\|e_j\|_{H^1(\Omega)} \le C\|u\|_{H^1(\Omega)}, \qquad j \in \mathbb{N}_0.$$

In each case, the constant $C > 0$ *is independent of the sequences and the level.*

Proof The first property follows as usual by induction. For the second property we note that we have $F(v) = a(u_{j-1} + s_j, v) = a(u_j, v)$ for all $v \in W_j$ by construction and thus $a(u - u_j, v) = 0$ for all $v \in W_j$. Hence, if we denote the energy norm by $\|u\|_a^2 := a(u, u)$ we have for an arbitrary $v \in W_j$ that

$$\begin{aligned} \|u - u_j\|_a^2 &= a(u - u_j, u - u_j) = a(u - u_j, u - u_{j-1} - s_j) \\ &= a(u - u_j, u - u_{j-1} - v) \le \|u - u_j\|_a \|u - u_{j-1} - v\|_a. \end{aligned}$$

Dividing by $\|u - u_j\|_a$ and using the fact that $v \in W_j$ was arbitrary gives

$$\|u - u_j\|_a \le \inf_{v \in W_j} \|u - u_{j-1} - v\|_a. \tag{11}$$

Using now the norm equivalence of $\|\cdot\|_a$ with $\|\cdot\|_{H^1(\Omega)}$ yields the second property.

The third statement is just the second statement using the identities $e_j = u - u_j = u - u_{j-1} - s_j = e_{j-1} - s_j$.

Finally, choosing $v = 0$ in (11) yields $\|e_j\|_a \le \|e_{j-1}\|_a$. This shows in particular $\|e_j\|_a \le \|e_0\|_a = \|u\|_a$ and the norm equivalence between the energy and the $H^1(\Omega)$-norm finally gives $\|e_j\|_{H^1(\Omega)} \le C\|u\|_{H^1(\Omega)}$. □

Thus, while u_n^* is the Galerkin approximation to u from the multiscale space V_n, the function $u_n \in V_n$ only satisfies the Galerkin orthogonality for W_n in the above sense.

It is an open question, whether u_n nonetheless converges to u or not and if it converges in which sense. This is of particular interest in the stationary multiscale setting. Here, [38] contains numerical evidence indicating that the multilevel algorithm seems not to converge. However, in [7] a proof for convergence was given and a numerical example seems to corroborate this. In both cases the numerical evidence has to be considered carefully since the numerical results depend on numerical integration adding additional errors to the scheme.

Unfortunately, in the proof of [7] an approximation result for the single-scale approximation was used, which was too optimistic for the given situation. We will now give a convergence proof of the multilevel scheme in a weaker form. As mentioned above, in [7] the stationary situation of choosing $\delta_j = \mu h_j$ was investigated. However, if the too optimistic assumption on the single-scale approximation is replaced, their proof only works in the situation of a non-stationary setting. To be more precise, we will see that $\delta_j = \mu h_j^{1/3}$ is required to achieve convergence with this proof. This, of course, means that the support radii also grow from level to level leading to denser matrices.

To prove this result, we need to recap some results from the one-level approximation. The first auxiliary result is a sampling inequality which is usually used in this context. We will employ it in a form coming from [31].

Lemma 2 *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. Let $\sigma > d/2$. Let $X \subseteq \Omega$ be a finite point set with sufficiently small fill distance $h_{X,\Omega}$. Then, there is a constant $C > 0$, independent of X, such that for all $f \in H^\sigma(\Omega)$ vanishing on X, we have*

$$\|f\|_{H^\mu(\Omega)} \le C h_{X,\Omega}^{\sigma-\mu} \|f\|_{H^\sigma(\Omega)}.$$

for $0 \le \mu \le \sigma$.

The second auxiliary result comes from [41], to be more precise, it is a summary of Lemmas 4 and 5 in [41]. It gives an approximation result for band-limited functions from weighted Sobolev spaces. Recall that a function $f \in L_2(\mathbb{R}^d)$ is band-limited if there is a $\tau > 0$ such that $\widehat{f}$ is compactly supported in $B_\tau(\mathbf{0})$, the ball about zero with radius τ. We will denote the collection of all such functions by $\mathscr{B}_\tau$. We also need the *separation distance*

$$q_X := \min_{j \ne k} \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2$$

of a point set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.

Lemma 3 *Let $\sigma \geq \beta > d/2$, $\delta \in (0,1]$ and $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathbb{R}^d$ with $q_X \leq \delta$. Then, there exists a constant $\kappa > 0$, independent of X and δ such that to each $f \in H^\beta(\mathbb{R}^d)$ there is a band-limited function $f_{\kappa/q_X} \in \mathscr{B}_{\kappa/q_X}$ with $f_{\kappa/q_X}|X = f|X$ and*

$$\|f - f_{\kappa/q_X}\|_{H^\beta_\delta} \leq 5\|f\|_{H^\beta_\delta}, \tag{12}$$

$$\|f_{\kappa/q_X}\|_{H^\sigma_\delta} \leq C(\kappa\delta/q_X)^{\sigma-\beta}\|f\|_{H^\beta_\delta}. \tag{13}$$

Note that function f_{κ/q_X} might also depend on δ, though this is not explicitly stated.

With this result, we can bound the $H^1(\Omega)$-error of an interpolant for functions of $H^2(\Omega)$. In its proof, as in what follows, we need a universal, linear and bounded extension operator, which is described in our final auxiliary result, see [1].

Lemma 4 *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary. Then, there is a linear, extension operator $E : H^\sigma(\Omega) \to H^\sigma(\mathbb{R}^d)$ satisfying*

1. *$Ef|\Omega = f$,*
2. *$\|Ef\|_{H^\sigma(\mathbb{R}^d)} \leq C_\sigma\|f\|_{H^\sigma(\Omega)}$*

for all $f \in H^\sigma(\Omega)$ and all $\sigma \geq 0$.

Now we are in the position to formulate the single-scale approximation result that we require for proving convergence of the non-stationary multilevel scheme.

Proposition 2 *Let $d \leq 3$. Let $\Omega \subseteq \mathbb{R}^d$ be bounded with a Lipschitz boundary. Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \Omega$ be quasi-uniform in the sense that $h_{X,\Omega} \leq c_{qu}q_X$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfy (3) and denote the interpolant to $f \in H^2(\Omega)$ on X using $\phi_\delta = \delta^{-d}\phi(\cdot/\delta)$ by $I_{X,\delta}f$. Then, there is a constant $C > 0$ such that*

$$\|f - I_{X,\delta}f\|_{H^1(\Omega)} \leq C\frac{h_{X,\Omega}}{\delta^2}\|f\|_{H^2(\Omega)}. \tag{14}$$

Proof Since $d \leq 3$, the Sobolev embedding theorem guarantees $H^2(\Omega) \subseteq C(\Omega)$. Since $f - I_{X,\delta}f$ vanishes on X, the sampling inequality from Lemma 2 yields

$$\|f - I_{X,\delta}f\|_{H^1(\Omega)} \leq Ch_{X,\Omega}\|f - I_{X,\delta}f\|_{H^2(\Omega)}.$$

Next, we choose $\widetilde{f} := (Ef)_{\kappa/q_X}$ from Lemma 3 and split the latter term as

$$\begin{aligned}\|f - I_{X,\delta}f\|_{H^2(\Omega)} &= \|Ef - I_{X,\delta}(Ef)\|_{H^2(\Omega)} \\ &\leq \|Ef - \widetilde{f}\|_{H^2(\Omega)} + \|\widetilde{f} - I_{X,\delta}\widetilde{f}\|_{H^2(\Omega)},\end{aligned}$$

where we used the fact that $f|X = Ef|X = \widetilde{f}|X$ and hence the interpolants to each of these functions are the same.

The first term in the above inequality can be bounded by

$$\begin{aligned}\|Ef-\widetilde{f}\|_{H^2(\Omega)} &\le \|Ef-\widetilde{f}\|_{H^2(\mathbb{R}^d)} \le \delta^{-2}\|Ef-\widetilde{f}\|_{H^2_\delta} \le 5\delta^{-2}\|Ef\|_{H^2_\delta}\\ &\le 5\delta^{-2}\|Ef\|_{H^2(\mathbb{R}^d)} \le C\delta^{-2}\|f\|_{H^2(\Omega)},\end{aligned}$$

using (4), (12) and the continuity of the extension operator. Next, using the sampling inequality from Lemma 2, the norm equivalence from Lemmas 1, 3 and 4 yields

$$\begin{aligned}\|\widetilde{f}-I_{X,\delta}\widetilde{f}\|_{H^2(\Omega)} &\le Ch_{X,\Omega}^{\sigma-2}\|\widetilde{f}-I_{X,\delta}\widetilde{f}\|_{H^\sigma(\mathbb{R}^d)} \le h_{X,\Omega}^{\sigma-2}\delta^{-\sigma}\|\widetilde{f}-I_{X,\delta}\widetilde{f}\|_{H^\sigma_\delta}\\ &\le Ch_{X,\Omega}^{\sigma-2}\delta^{-\sigma}\|\widetilde{f}\|_{H^\sigma_\delta} \le Ch_{X,\Omega}^{\sigma-2}\delta^{-\sigma}\left(\frac{\kappa\delta}{q_X}\right)^{\sigma-2}\|\widetilde{f}\|_{H^2_\delta}\\ &\le C\left(\frac{h_{X,\Omega}}{q_X}\right)^{\sigma-2}\delta^{-2}\|\widetilde{f}\|_{H^2_\delta} \le C\delta^{-2}\|\widetilde{f}\|_{H^2(\mathbb{R}^d)}\\ &\le C\delta^{-2}\|Ef\|_{H^2(\mathbb{R}^d)} \le C\delta^{-2}\|f\|_{H^2(\Omega)}.\end{aligned}$$

Taking this all together yields the stated result. □

A similar result can be found in [21, Lemma 4.6], for more general Sobolev spaces and without the condition that X has to be quasi-uniform. But the result there is not given for the interpolant but only guarantees the existence of an element $s \in V_n$ satisfying the inequality.

Unfortunately, in our proof below we will need $L_2(\Omega)$-estimates for our approximate Galerkin solution. To derive these, we will assume that we can apply the Aubin-Nitsche trick, i.e. the solution of the dual problem satisfies Friedrich's inequality. To be more precise:

Assumption 1 *The solution $w \in H^1(\Omega)$ of the dual problem $a(v, w) = F(v) = \langle f, v\rangle_{L_2(\Omega)}$ for all $v \in H^1(\Omega)$ satisfies $w \in H^2(\Omega)$ and $\|w\|_{H^2(\Omega)} \le C\|f\|_{L_2(\Omega)}$.*

It is well-known, that Assumption 1 is, for example, satisfied if Ω has a smooth boundary or if Ω is convex with a Lipschitz boundary.

While the previous result (14) should replace the too optimistic bound from [7, Theorem 3.1], which does not hold in the case of scaled radial basis functions, the following result is the correct version to replace the too optimistic [7, Lemma 3.2].

Proposition 3 *Let $u \in H^1(\Omega)$ be the solution of (10) and let $s \in W_\delta = \operatorname{span}\{\phi_\delta(\cdot - \boldsymbol{x}) : \boldsymbol{x} \in X\}$ be the approximate Galerkin approximation, where $X \subseteq \Omega$ is a discrete, quasi-uniform set with $q_X \le \delta \le 1$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfies (3). Let Assumption 1 be satisfied. Then,*

$$\|u-s\|_{L_2(\Omega)} \le C\frac{h_{X,\Omega}}{\delta^2}\|u-s\|_{H^1(\Omega)}. \tag{15}$$

Proof The proof is given by the Aubin-Nitsche trick. We choose $w \in H^1(\Omega)$ as the solution of $a(v,w) = \langle u-s, v\rangle_{L_2(\Omega)}$, $v \in H^1(\Omega)$ and know by Assumption 1 that $w \in H^2(\Omega)$ with $\|w\|_{H^2(\Omega)} \le C\|u-s\|_{L_2(\Omega)}$. Then, we have for $v = I_{X,\delta}w \in W_\delta$ that

$$\begin{aligned}\|u-s\|_{L_2(\Omega)}^2 &= \langle u-s, u-s\rangle_{L_2(\Omega)}\\ &= a(u-s,w) = a(u-s,w-v)\\ &\le C\|u-s\|_{H^1(\Omega)}\|w-v\|_{H^1(\Omega)}\\ &\le C\frac{h_{X,\Omega}}{\delta^2}\|w\|_{H^2(\Omega)}\|u-s\|_{H^1(\Omega)}\\ &\le C\frac{h_{X,\Omega}}{\delta^2}\|u-s\|_{L_2(\Omega)}\|u-s\|_{H^1(\Omega)}.\end{aligned}$$

Dividing now by $\|u-s\|_{L_2(\Omega)}$ gives the desired result. □

The estimate given in (15) has two consequences. Obviously, we can use the bound on the $H^1(\Omega)$-norm to derive

$$\|u-s\|_{L_2(\Omega)} \le C\frac{h^2}{\delta^4}\|u\|_{H^2(\Omega)}.$$

We will not need this inequality but will employ the following one. The Galerkin orthogonality and Lemma 1 yield

$$\|u-s\|_{H^1(\Omega)} \le C\|u\|_{H^1(\Omega)} \le C\|Eu\|_{H^1(\mathbb{R}^d)} \le C\delta^{-1}\|Eu\|_{H^1_\delta},$$

where E denotes the extension operator again, such that we also have

$$\|u-s\|_{L_2(\Omega)} \le \frac{h_{X,\Omega}}{\delta^3}\|Eu\|_{H^1_\delta}. \tag{16}$$

We are now in the position to formulate and prove convergence of the non-stationary multilevel Galerkin method.

Theorem 4 *Let $\Omega \subseteq \mathbb{R}^d$ with $d \le 3$ be bounded. Let $u \in H^1(\Omega)$ denote the solution of (10). Let Assumption 1 be satisfied. Let $\phi \in L_1(\mathbb{R}^d)$ be continuous and compactly supported with Fourier transform satisfying (3). Let $X_1, X_2, \ldots \subseteq \Omega$ be a sequence of quasi-uniform discrete sets with fill distances $h_j := h_{X_j,\Omega}$ satisfying $\mu\gamma h_j \le h_{j+1} \le \mu h_j$ with some fixed $\gamma \in (0,1]$ and $\mu \in (0,1)$. Let $\{\delta_j\}$ be a non-increasing sequence of support radii satisfying*

$$\delta_j = \frac{h_j^{1/3}}{\mu^{1/9}}.$$

Finally, let $u_n \in V_n$ be the approximate multilevel Galerkin approximation to u. Then, there are constants $C, C_1 > 0$ such that

$$\|u - u_n\|_{L_2(\Omega)} \leq C_1 (C\mu^{1/3})^n \|u\|_{H^1(\Omega)}. \tag{17}$$

In particular, there is a $\mu_0 \in (0, 1]$ such that the method converges for all $\mu \leq \mu_0$ linearly in the number of levels.

Proof For $j = 1, 2, \ldots$ let $e_j = u - u_j = e_{j-1} - s_j$ denote the error. We want to establish the recurrence

$$\|Ee_j\|_{H^1_{\delta_{j+1}}} \leq C\mu^{1/3} \|Ee_{j-1}\|_{H^1_{\delta_j}}. \tag{18}$$

We start by splitting

$$\begin{aligned}\|Ee_j\|^2_{H^1_{\delta_{j+1}}} &= \int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 (1 + \delta^2_{j+1} \|\boldsymbol{\omega}\|_2^2) d\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} + \delta^2_{j+1} \int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 \|\boldsymbol{\omega}\|_2^2 d\boldsymbol{\omega}. \end{aligned} \tag{19}$$

For the first integral on the right-hand side we notice that $e_j = e_{j-1} - s_j$, where s_j is the approximate Galerkin approximation to e_{j-1} from W_j such that (16) yields, with $u = e_{j-1}$, $s = s_j$ and $\delta = \delta_j$,

$$\begin{aligned}\int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} = \|Ee_j\|^2_{L_2(\mathbb{R}^d)} &\leq C\|e_j\|^2_{L_2(\Omega)} \leq C \left(\frac{h_j}{\delta_j^3}\right)^2 \|Ee_{j-1}\|^2_{H^1_{\delta_j}} \\ &\leq C\mu^{2/3} \|Ee_{j-1}\|^2_{H^1_{\delta_j}}.\end{aligned}$$

For the second integral in (19) we have the estimate

$$\begin{aligned}\delta^2_{j+1} \int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 \|\boldsymbol{\omega}\|_2^2 d\boldsymbol{\omega} &\leq \delta^2_{j+1} \int_{\mathbb{R}^d} |\widehat{Ee_j}(\boldsymbol{\omega})|^2 (1 + \|\boldsymbol{\omega}\|_2^2) d\boldsymbol{\omega} \\ &\leq C\delta^2_{j+1} \|e_j\|^2_{H^1(\Omega)} \leq C\delta^2_{j+1} \|e_{j-1}\|^2_{H^1(\Omega)} \\ &\leq C\delta^2_{j+1} \|Ee_{j-1}\|^2_{H^1(\mathbb{R}^d)} \\ &\leq C \left(\frac{\delta_{j+1}}{\delta_j}\right)^2 \|Ee_{j-1}\|^2_{H^1_{\delta_j}} \\ &\leq C\mu^{2/3} \|Ee_{j-1}\|^2_{H^1_{\delta_j}}.\end{aligned}$$

Taking this together and taking the root gives indeed (18). Finally, Proposition 3 yields

$$\begin{aligned}\|u-u_n\|_{L_2(\Omega)} &= \|e_n\|_{L_2(\Omega)} \le C\frac{h_n}{\delta_n^2}\|e_n\|_{H^1(\Omega)}\\ &\le C\frac{h_n}{\delta_n^2\delta_{n+1}}\|Ee_n\|_{H^1_{\delta_{n+1}}} \le C\|Ee_n\|_{H^1_{\delta_{n+1}}},\end{aligned}$$

where, in the last step, we have used that

$$\frac{h_n}{\delta_n^2\delta_{n+1}} = \frac{h_n}{\delta_n^3}\frac{\delta_n}{\delta_{n+1}} \le \mu^{1/3}\left(\frac{1}{\gamma\mu}\right)^{1/3} = \gamma^{-1/3}.$$

Applying the recursion (18) now n times proves the final result. □

It is interesting to note that the convergence is given in the $L_2(\Omega)$-norm rather than the more natural $H^1(\Omega)$-norm. Moreover, even if we needed Friedrich's inequality to derive the error estimate, the norm on the right-hand side of (17) is the $H^1(\Omega)$-norm instead of the $H^2(\Omega)$-norm. This might have two consequences. On the one hand, it might be possible to derive this result without relying on Friedrich's inequality. This would give a true $H^1(\Omega)$ theory. Then, of course, it is natural that the convergence has to be measured in a weaker norm, i.e. the $L_2(\Omega)$-norm. On the other hand, it might be possible to improve the result so that the final error estimate will employ the $H^2(\Omega)$-norm on the right-hand side of (17). Both of these possible consequences are subject to further research.

In the same spirit is the following observation. We can express the convergence again in terms of the fill distance of the finest level, as we have done in Corollary 3 for collocation. Here, assuming again $h_{j+1} = \mu h_j$, we see that for each $\epsilon \in (0, 1/3)$, we can find a μ_0 such that

$$\|u-u_n\|_{L_2(\Omega)} \le Ch_n^{\frac{1}{3}-\epsilon}\|u\|_{H^1(\Omega)}.$$

This is rather weak since it gives less than linear convergence. It also means that the proof does not guarantee convergence in the $H^1(\Omega)$-norm. But the above discussion is a starting point for investigating this further.

In the case of the stationary multiscale approximation spaces, the situation is as follows. By Theorem 3 we know that the spaces V_n are rich enough to provide good approximations. Moreover, by Proposition 1 we know that the approximation generated by the residual correction algorithm (Algorithm 3) is stable, but there is no convergence proof yet. It might very well be that the algorithm does not capture enough of the information provided by the multiscale space. In [38] it was suggested to use an iterative projection method, which is known to converge to the Galerkin approximation $u_n^* \in V_n$. Unfortunately, the convergence seems to be slow. Hence, an improved algorithm might be required.

References

1. Brenner, S., Scott, L.: The Mathematical Theory of Finite Element Methods, 3rd edn. Springer, New York (1994)
2. Buhmann, M.D.: Radial basis functions. In: Acta Numerica 2000, vol. 9, pp. 1–38. Cambridge University Press, Cambridge (2000)
3. Buhmann, M.D.: Radial Basis Functions. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2003)
4. Chen, C.S., Ganesh, M., Golberg, M.A., Cheng, A.H.D.: Multilevel compact radial functions based computational schemes for some elliptic problems. Comput. Math. Appl. **43**, 359–378 (2002)
5. Chernih, A., Le Gia, Q.T.: Multiscale methods with compactly supported radial basis functions for elliptic partial differential equations on bounded domains. ANZIAM J. (E) **54**, C137–C152 (2012)
6. Chernih, A., Le Gia, Q.T.: Multiscale methods with compactly supported radial basis functions for the Stokes problem on bounded domains. Adv. Comput. Math. **42**, 1187–1208 (2013)
7. Chernih, A., Le Gia, Q.T.: Multiscale methods with compactly supported radial basis functions for Galerkin approximation of elliptic PDEs. IMA J. Numer. Anal. **34**, 569–591 (2014)
8. Farrell, P., Wendland, H.: RBF multiscale collocation for second order elliptic boundary value problems. SIAM J. Numer. Anal. **51**, 2403–2425 (2013)
9. Farrell, P., Gillow, K., Wendland, H.: Multilevel interpolation of divergence-free vector fields. IMA J. Numer. Anal. **37**, 332–353 (2017)
10. Fasshauer, G.E.: Solving partial differential equations by collocation with radial basis functions. In: Méhauté, A.L., Rabut, C., Schumaker, L.L. (eds.) Surface Fitting and Multiresolution Methods, pp. 131–138. Vanderbilt University Press, Nashville (1997)
11. Fasshauer, G.E.: Solving differential equations with radial basis functions: multilevel methods and smoothing. Adv. Comput. Math. **11**, 139–159 (1999)
12. Fasshauer, G.E.: Meshfree Approximation Methods with MATLAB. World Scientific, Singapore (2007)
13. Fasshauer, G.E., Jerome, J.W.: Multistep approximation algorithms: improved convergence rates through postconditioning with smoothing kernels. Adv. Comput. Math. **10**, 1–27 (1999)
14. Ferrari, S., Maggioni, M., Borhese, N.A.: Multiscale approximation with hierarchical radial basis functions networks. IEEE Trans. Neural Netw. **15**, 178–188 (2004)
15. Floater, M.S., Iske, A.: Multistep scattered data interpolation using compactly supported radial basis functions. J. Comput. Appl. Math. **73**, 65–78 (1996)
16. Fornberg, B., Flyer, N.: Solving PDEs with radial basis functions. In: Iserles, A. (ed.) Acta Numerica, vol. 24, pp. 215–258. Cambridge University Press, Cambridge (2015)
17. Franke, C., Schaback, R.: Convergence order estimates of meshless collocation methods using radial basis functions. Adv. Comput. Math. **8**, 381–399 (1998)
18. Franke, C., Schaback, R.: Solving partial differential equations by collocation using radial basis functions. Appl. Math. Comput. **93**, 73–82 (1998)
19. Giesl, P., Wendland, H.: Meshless collocation: Error estimates with application to dynamical systems. SIAM J. Numer. Anal. **45**, 1723–1741 (2007)
20. Hales, S.J., Levesley, J.: Error estimates for multilevel approximation using polyharmonic splines. Numer. Algoritm. **30**, 1–10 (2002)
21. Heuer, N., Tran, T.: A mixed method for Dirichlet problems with radial basis functions. Comput. Math. Appl. **66**, 2045–2055 (2013)
22. Hon, Y.C., Schaback, R.: On unsymmetric collocation by radial basis functions. J. Appl. Math. Comput. **119**, 177–186 (2001)
23. Le Gia, Q.T., Wendland, H.: Data compression on the sphere using multiscale radial basis functions. Adv. Comput. Math. **40**, 923–943 (2014)
24. Le Gia, Q.T., Sloan, I., Wendland, H.: Multiscale analysis in Sobolev spaces on the sphere. SIAM J. Numer. Anal. **48**, 2065–2090 (2010)

25. Le Gia, Q.T., Sloan, I., Wendland, H.: Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. Appl. Comput. Harmon. Anal. **32**, 401–412 (2012)
26. Le Gia, Q.T., Sloan, I., Wendland, H.: Multiscale RBF collocation for solving PDEs on spheres. Numer. Math. **121**, 99–125 (2012)
27. Le Gia, Q.T., Sloan, I.H., Wendland, H.: Zooming from global to local: a multiscale RBF approach. Adv. Comput. Math. **43**, 581–606 (2017)
28. Li, M., Cao, F.: Multiscale interpolation on the sphere: convergence rate and inverse theorem. Appl. Math. Comput. **263**, 134–150 (2015)
29. Morton, T.M., Neamtu, M.: Error bounds for solving pseudodifferential equatons on spheres by collocation with zonal kernels. J. Approx. Theory **114**, 242–268 (2002)
30. Narcowich, F.J., Schaback, R., Ward, J.D.: Multilevel interpolation and approximation. Appl. Comput. Harmon. Anal. **7**, 243–261 (1999)
31. Narcowich, F.J., Ward, J.D., Wendland, H.: Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. Constr. Approx. **24**, 175–186 (2006)
32. Ohtake, Y., Belyaev, A., Alexa, M., Turk, G., Seidel, H.P.: Multi-level partition of unity implicits. ACM Trans. Graph. **22**, 463–470 (2003)
33. Ron, A.: The L_2-approximation orders of principal shift-invariant spaces generated by a radial basis function. In: Braess, D., et al. (eds.) Numerical Methods in Approximation Theory. vol. 9: Proceedings of the Conference Held in Oberwolfach, Germany, 24–30 Nov 1991. International Series of Numerican Mathematics, vol. 105, pp. 245–268. Birkhäuser, Basel (1992)
34. Schaback, R.: Creating surfaces from scattered data using radial basis functions. In: Dæhlen, M., Lyche, T., Schumaker, L.L. (eds.) Mathematical Methods for Curves and Surfaces, pp. 477–496. Vanderbilt University Press, Nashville (1995)
35. Schaback, R., Wendland, H.: Kernel techniques: from machine learning to meshless methods. In: Iserles, A. (ed.) Acta Numerica, vol. 15, pp. 543–639. Cambridge University Press, Cambridge (2006)
36. Townsend, A., Wendland, H.: Multiscale analysis in Sobolev spaces on bounded domains with zero boundary values. IMA J. Numer. Anal. **33**, 1095–1114 (2013)
37. Wendland, H.: Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. **4**, 389–396 (1995)
38. Wendland, H.: Numerical solutions of variational problems by radial basis functions. In: Chui, C.K., Schumaker, L.L. (eds.) Approximation Theory IX, vol. 2: Computational Aspects, pp. 361–368. Vanderbilt University Press, Nashville (1998)
39. Wendland, H.: Scattered Data Approximation. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2005)
40. Wendland, H.: On the stability of meshless symmetric collocation for boundary value problems. BIT **47**, 455–468 (2007)
41. Wendland, H.: Multiscale analysis in Sobolev spaces on bounded domains. Numer. Math. **116**, 493–517 (2010)
42. Wendland, H.: Multiscale radial basis functions. In: Pesenson, I., Gia, Q.T.L., Mayeli, A., Mhaskar, H., Zhou, D.X. (eds.) Frames and Other Bases in Abstract and Function Spaces – Novel Methods in Harmonic Analysis, vol. 1, pp. 265–299. Birkhäuser, Cham (2017)
43. Wu, Z.: Compactly supported positive definite radial functions. Adv. Comput. Math. **4**, 283–292 (1995)
44. Xu, B., Lu, S., Zhong, M.: Multiscale support vector regression method in Sobolev spaces on bounded domains. Appl. Anal. **94**, 548–569 (2015)

Tractability of Approximation for Some Weighted Spaces of Hybrid Smoothness

Arthur G. Werschulz

Dedicated to our friend and colleague Ian H. Sloan on the occasion of his 80th birthday.

Abstract A great deal of work has studied the tractability of approximating (in the L_2-norm) functions belonging to weighted unanchored Sobolev spaces of dominating mixed smoothness of order 1 over the unit d-cube. In this paper, we generalize these results. Let r and s be non-negative integers, with $r \leq s$. We consider the approximation of complex-valued functions over the torus $\mathbb{T}^d = [0, 2\pi]^d$ from weighted spaces $H^{s,1}_{\Gamma}(\mathbb{T}^d)$ of hybrid smoothness, measuring error in the $H^r(\mathbb{T}^d)$-norm. Here we have isotropic smoothness of order s, the derivatives of order s having dominating mixed smoothness of order 1. If $r = s = 0$, then $H^{0,1}(\mathbb{T}^d)$ is a well-known weighted unachored Sobolev space of dominating smoothness of order 1, whereas we have a generalization for other values of r and s. Besides its independent interest, this problem arises (with $r = 1$) in Galerkin methods for solving second-order elliptic problems. Suppose that continuous linear information is admissible. We show that this new approximation problem is topologically equivalent to the problem of approximating $H^{s-r,1}_{\Gamma}(\mathbb{T}^d)$ in the $L_2(\mathbb{T}^d)$-norm, the equivalence being independent of d. It then follows that our new problem attains a given level of tractability if and only if approximating $H^{s-r,1}_{\Gamma}(\mathbb{T}^d)$ in the $L_2(\mathbb{T}^d)$-norm has the same level of tractability. We further compare the tractability of our problem to that of $L_2(\mathbb{T}^d)$-approximation for $H^{0,1}_{\Gamma}(\mathbb{T}^d)$. We then analyze the tractability of our problem for various families of weights.

A. G. Werschulz (✉)
Department of Computer and Information Science, Fordham University, New York, NY, USA

Department of Computer Science, Columbia University, New York, NY, USA
e-mail: agw@cis.fordham.edu; agw@cs.columbia.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_56

1 Introduction

Much recent research in information-based complexity has dealt with the issue of tractability. To what extent is it computationally feasible to solve this problem? To get an idea of the scope of this area, see the three-monograph series [7–9]. Most of the work in this area has dealt with the integration problem (which was the initial impetus for studying tractability in the first place) and the approximation problem (mainly in L_p-norms, with most of the work for the case $p = 2$). This paper deals with the latter.[1]

It has long been known that the L_2-approximation problem for the unit ball of $H^s(I^d)$ over the unit d-cube I^d has nth minimal error $\Theta(n^{-s/d})$, so that $\Theta(\varepsilon^{-d/s})$ information evaluations are necessary and sufficient for an ε-approximation. Were it not for the Θ-factors, this would imply that this problem suffers from what Richard Bellman [1] called "the curse of dimensionality", i.e., an exponential dependence on the dimension d. It turns out that things are not quite as bad as this. To avoid some technical difficulties, we'll use spaces $H^s(\mathbb{T}^d)$ defined over the d-torus $\mathbb{T}^d$, rather than over the d-dimensional unit cube I^d. Kühn et al. [5] showed that the Θ-factors decay polynomially in d, and that this problem does not suffer from the curse of dimensionality.

However, we would much prefer something stronger; in particular, we would like to have polynomial tractability, with nth minimal error at most $Cd^q\varepsilon^{-p}$ for C p, and q independent of ε and d or (better yet) strong polynomial tractability, with nth minimal error at most $C\varepsilon^{-p}$ for C and p independent of ε and d. However, the results in [5] imply that the aforementioned problem is not polynomially tractable. So if we want a better tractability result, we need to change the space of functions being approximated.

Now the spaces $H^s(\mathbb{T}^d)$ are isotropic—all variables are equally important. This has led many authors to use anisotropic spaces. In particular, we have used weighted spaces that (algebraically) are tensor products of $H^1(I)$, with the weight family Γ entering into the norm. These are weighted versions of spaces having *mixed smoothness*, as per [6]. In [11], we were able to find conditions on certain weights families Γ that were necessary and sufficient for the $L_2(\mathbb{T}^d)$-approximation problem to be (strongly) polynomially tractable.

We would like to extend these results to weighted spaces of *hybrid* smoothness, see [10]. These are weighted versions of the spaces $H^{s_1,s_2}(\mathbb{T}^d)$, the members of which being periodic functions having isotropic smoothness of order s_1 and dominating mixed smoothness of order s_2.

[1] This introduction is merely an overview. Precise definitions are given in Sect. 2.

In this paper, we make a first step in such a study. We will consider spaces $H_\Gamma^{s,1}(\mathbb{T}^d)$. Functions belonging to this space have Sobolev derivatives of order s, said derivatives themselves having one derivative in each coordinate direction. The weights only apply to the anisotropic part of the $H^{s,1}(\mathbb{T}^d)$-norm. We measure error in the $H^r(\mathbb{T}^d)$-sense. Here r and s are non-negative integers, with $r \le s$.

We have an ulterior motive for studying these particular spaces. Suppose we are trying to solve the elliptic problem $-\Delta u + qu = f$ over $\mathbb{T}^d$, with f, q in the unit ball of $H^{0,1}(\mathbb{T}^d)$. Suppose further that we have an elliptic regularity result, saying that $u \in H_\Gamma^{2,1}(\mathbb{T}^d)$ for $f, q \in H_\Gamma^{0,1}(\mathbb{T}^d)$. Then the error of a Galerkin method using an optimal test/trial space will roughly be the minimal error for the $H^1(\mathbb{T}^d)$-approximation problem over $H^{2,1}(\mathbb{T}^d)$. This explains our interest in the $H^r(\mathbb{T}^d)$-approximation problem for $H_\Gamma^{s,1}(\mathbb{T}^d)$ with $r = 1$ and $s = 2$. In this paper, we study the general case (with $r \le s$), which is as easy to handle as the special case $r = 1$ and $s = 2$. In addition, we expect the results of this paper to hold for negative r; this is important because the case $r = -1$ occurs in non-regular second-order elliptic problems, see (e.g.) [3] for further discussion.

The overall structure of this paper is as follows. In Sect. 2, we precisely define the terminology surrounding the problem we're trying to solve. The results we seek depend on spectral information of a particular linear operator on $H_\Gamma^{s,1}(\mathbb{T}^d)$, which we give in Sect. 3. Finally, Sect. 4 gives the tractability results for our approximation problem:

1. If $\mathrm{App}_{\Gamma,0,0}$ has a given level of tractability, then $\mathrm{App}_{\Gamma,r,s}$ has at least the same level of tractability, and the exponent(s) for $\mathrm{App}_{\Gamma,r,s}$ are bounded from above by those for $\mathrm{App}_{\Gamma,0,0}$.
2. Under certain boundedness conditions, $\mathrm{App}_{\Gamma,r,s}$ has a given level of tractability iff $\mathrm{App}_{\Gamma,0,0}$ has at least the same level of tractability. We give estimates relating the exponents for these two problems.
3. For the unweighted case, $\mathrm{App}_{\Gamma,r,s}$ is quasi-polynomially tractable, with exponent $2/\ln 2 \doteq 2.88539$.
4. For bounded product weights:
   (a) $\mathrm{App}_{\Gamma,r,s}$ is always quasi-polynomially tractable. We give an estimate of the exponent.
   (b) We give conditions on the weights that are necessary and sufficient to guarantee (strong) polynomial tractability, along with estimates of the exponents.
5. For bounded finite-order and finite-diameter weights, $\mathrm{App}_{\Gamma,r,s}$ is always polynomially tractable. We give estimates for the exponents.

2 Problem Definition

In this section, we define the approximation problem to be studied and recall some basic concepts of information-based complexity.

First, we establish some notational conventions. We let $\mathbb{N}$ denote the strictly positive integers, with $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denoting the natural numbers. (As usual, we let $\mathbb{Z}$ denote the integers.) Next, we let $\mathbb{T}$ denote the torus $[0, 2\pi]$, so that $\mathbb{T}^d$ is the d-torus. We identify opposite points on the d-torus, so that for any $f\colon \mathbb{T}^d \to \mathbb{C}$, we have $f(\boldsymbol{x}) = f(\boldsymbol{y})$ whenever $\boldsymbol{x} - \boldsymbol{y} \in 2\pi\,\mathbb{Z}^d$. In this sense, functions on the d-torus are periodic. We denote points in $\mathbb{R}^d$ by boldface italic letters, and points in $\mathbb{Z}^d$ (including multi-indices) by boldface roman letters. The unit ball of a normed space X is denoted by $\mathscr{B}X$. Any product over the empty set is defined to be the appropriate multiplicative identity.

We now describe some Sobolev spaces, see (e.g.) [4, 5, 10, 13] for further discussion. Let $L_2(\mathbb{T}^d)$ denote the space of complex-valued square-integrable functions over $\mathbb{T}^d$ and let $r \in \mathbb{N}_0$. Then

$$H^r(\mathbb{T}^d) = \left\{f \in L_2(\mathbb{T}^d) : \ D^{\mathbf{m}} f \in L_2(\mathbb{T}^d) \ \text{ for } \ |\mathbf{m}| \le r\right\},$$

is the (classical) isotropic Sobolev space of order r, which is a Hilbert space under the usual inner product

$$\langle f, g\rangle_{H^r(\mathbb{T}^d)} = \sum_{|\mathbf{m}|\le r} \langle D^{\mathbf{m}} f, D^{\mathbf{m}} g\rangle_{L_2(\mathbb{T}^d)}.$$

Here, for $\mathbf{m} = (m_1, m_2, \dots, m_d) \in \mathbb{N}_0^d$, we write

$$D^{\mathbf{m}} = \prod_{j=1}^{d} \frac{\partial^{m_j}}{\partial x_j^{m_j}} \qquad \text{and} \qquad \boldsymbol{z}^{\mathbf{m}} = \prod_{j=1}^{d} z_j^{m_j} \quad \forall\, \boldsymbol{z} = (z_1, \dots, z_d) \in \mathbb{C}^d,$$

as well as $|\mathbf{m}| = \sum_{j=1}^d m_j$. Here, the partial derivative $\partial/\partial x_j$ is in the distributional sense.

For $s \in \mathbb{N}_0$, we define the space[2]

$$H^{s,1}(\mathbb{T}^d) = \{\, v \in H^s(\mathbb{T}^d) : \partial_{\mathrm{u}} v \in H^s(\mathbb{T}^d) \text{ for all } \mathrm{u} \subseteq [d]\,\}$$

of hybrid smoothness, which is a Hilbert space under the inner product

$$\langle v, w\rangle_{H^{s,1}(\mathbb{T}^d)} = \sum_{\mathrm{u}\subseteq[d]} \langle \partial_{\mathrm{u}} v, \partial_{\mathrm{u}} w\rangle_{H^s(\mathbb{T}^d)} \qquad \forall\, v, w \in H^{s,1}_{\Gamma}(\mathbb{T}^d).$$

[2] The superscript 1 in $H^{s,1}(\mathbb{T}^d)$ means that we are taking dominating mixed derivatives of order 1.

Here, we write

$$\partial_{\mathfrak{u}} = \prod_{i \in \mathfrak{u}} \frac{\partial}{\partial x_i} \qquad \forall\, \mathfrak{u} \subseteq [d],$$

where $[d] := \{1, 2, \ldots, d\}$.

Our final Sobolev space is a weighted version of the space $H^{s,1}(\mathbb{T}^d)$. Let

$$\Gamma = \{\, \gamma_{d,\mathfrak{u}} \geq 0 : \mathfrak{u} \subseteq [d], d \in \mathbb{N}\,\}$$

be a given set of non-negative weights $\gamma_{d,\mathfrak{u}}$, with $\gamma_{\mathfrak{u},\emptyset} = 1$ for all $d \in \mathbb{N}$. Then we let $H^{s,1}_{\Gamma}(\mathbb{T}^d)$ be $H^{s,1}(\mathbb{T}^d)$, but under the inner product

$$\langle v, w\rangle_{H^{s,1}_{\Gamma}(\mathbb{T}^d)} = \sum_{\substack{\mathfrak{u} \subseteq [d] \\ \gamma_{d,\mathfrak{u}} > 0}} \gamma_{d,\mathfrak{u}}^{-1} \langle \partial_{\mathfrak{u}} v, \partial_{\mathfrak{u}} w\rangle_{H^{s}(\mathbb{T}^d)} \qquad \forall\, v, w \in H^{s,1}_{\Gamma}(\mathbb{T}^d). \tag{1}$$

Clearly $H^{s,1}_{\Gamma}(\mathbb{T}^d)$ is a Hilbert space under this inner product.

We now describe the problem we wish to solve. Let $r, s \in \mathbb{N}_0$, with $r \leq s$. Our goal is to approximate functions from $\mathscr{B}H^{s,1}_{\Gamma}(\mathbb{T}^d)$ in the $H^r(\mathbb{T}^d)$-norm. This approximation problem is described by the embedding operator $\mathrm{App}_{d,\Gamma,r,s}\colon H^{s,1}_{\Gamma}(\mathbb{T}^d) \to H^r(\mathbb{T}^d)$, which is defined as

$$\mathrm{App}_{d,\Gamma,r,s} f = f \qquad \forall f \in H^{s,1}_{\Gamma}(\mathbb{T}^d).$$

Remark 1 We note some special cases of this problem:

1. Suppose that $r = s = 0$. Then $\mathrm{App}_{d,\Gamma,r,s} = \mathrm{App}_{d,\Gamma,0,0}$, and our problem is that of approximating functions from $\mathscr{B}H^{0,1}_{\Gamma}(\mathbb{T}^d)$ in the $L_2(\mathbb{T}^d)$-norm. This problem is analogous to the problem that was extensively covered in [11], the main difference being that [11] dealt with functions defined over the unit cube, rather than the unit torus.
2. Let $\Gamma(\emptyset)$ be given by

$$\gamma_{d,\mathfrak{u}} = \begin{cases} 1 & \text{if } \mathfrak{u} = \emptyset, \\ 0 & \text{otherwise.} \end{cases} \qquad \forall\, \mathfrak{u} \subseteq [d], d \in \mathbb{N}.$$

Allowing a slight abuse of language, we call $\Gamma(\emptyset)$ *empty weights*. Then $\mathrm{App}_{d,\Gamma,r,s} = \mathrm{App}_{d,\Gamma(\emptyset),r,s}$, and our problem is that of approximating functions from $\mathscr{B}H^{s}(\mathbb{T}^d)$ in the $H^r(\mathbb{T}^d)$-norm. This problem was studied for the case $r = 0$ in [5, 13] and for arbitrary $r \geq 0$ in [10].

3. Let $\Gamma(\textsc{unw})$ be defined as

$$\gamma_{d,\mathfrak{u}} = 1 \qquad \forall\, \mathfrak{u} \subseteq [d].$$

Then $\mathrm{App}_{d,\Gamma,r,s} = \mathrm{App}_{d,\Gamma(\textsc{unw}),r,s}$ and we are trying to solve the *unweighted* case. Our problem is now that of approximating functions from $\mathscr{B}H^{s,1}(\mathbb{T}^d)$ in the $H^r(\mathbb{T}^d)$-norm. A non-periodic version of this problem (over the unit cube, rather than the torus) was discussed in [11, Section 4.1.1].

4. If $\Gamma(\Pi)$ is a set of weights defined by

$$\gamma_{d,\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_{d,j} \qquad \forall\, \mathfrak{u} \subseteq [d], d \in \mathbb{N}, \tag{2}$$

where

$$\gamma_{d,1} \geq \gamma_{d,2} \geq \cdots \geq \gamma_{d,d} > 0 \qquad \forall\, d \in \mathbb{N}, \tag{3}$$

the $\Gamma(\Pi)$ is said to be a set of *product weights*. We may refer to the set $\{\, \gamma_{d,j} : j \in [d], d \in \mathbb{N}\,\}$ as being the *weightlets* for $\Gamma(\Pi)$.

5. We say that $\Gamma(\textsc{fow})$ is a family of *finite-order weights* if there exists $\omega \in \mathbb{N}_0$ such that

$$\gamma_{d,\mathfrak{u}} = 0 \qquad \text{for all } d \in \mathbb{N} \text{ and } \mathfrak{u} \text{ such that } |\mathfrak{u}| > \omega.$$

The smallest ω for which this holds is said to be the *order* of $\Gamma(\textsc{fow})$. As a special case, we say that $\Gamma(\textsc{fdw})$ is a family of *finite-diameter weights* if

$$\gamma_{d,\mathfrak{u}} = 0 \qquad \text{for all } d \in \mathbb{N} \text{ and all } \mathfrak{u} \text{ with } \operatorname{diam}(\mathfrak{u}) \geq q.$$

The smallest q for which this holds is said to be the *diameter* of $\Gamma(\textsc{fdw})$. □

Remark 2 We can slightly simplify the sum appearing in (1), as in [12]. If we adopt the convention that $0/0 = 0$, we can write

$$\langle v, w\rangle_{H^{s,1}_\Gamma(\mathbb{T}^d)} = \sum_{\mathfrak{u}\subseteq[d]} \gamma^{-1}_{d,\mathfrak{u}} \langle \partial_\mathfrak{u} v, \partial_\mathfrak{u} w\rangle_{H^s(\mathbb{T}^d)} \qquad \forall\, v, w \in H^{s,1}_\Gamma(\mathbb{T}^d),$$

provided that we require

$$\partial_\mathfrak{u} w = 0 \text{ for any } \mathfrak{u} \subseteq [d] \text{ such that } \gamma_{d,\mathfrak{u}} = 0 \qquad \forall\, w \in H^{s,1}_\Gamma(\mathbb{T}^d).$$

Of course, if $\partial_\mathfrak{u} v = 0$, then $\partial_\mathfrak{v} w = 0$ for any superset $\mathfrak{v}$ of $\mathfrak{u}$. This imposes the natural condition

$$\gamma_{d,\mathfrak{u}} = 0 \implies \gamma_{d,\mathfrak{v}} = 0 \qquad \text{for any } \mathfrak{v} \subseteq [d] \text{ for which } \mathfrak{v} \supseteq \mathfrak{u}. \tag{4}$$

In the remainder of this paper, we shall assume that (4) *holds.* Now suppose that $\gamma_{d,j} = 0$ for some $j \in [d]$ and $d \in \mathbb{N}$. Using (4), we see that $\gamma_{d,\mathfrak{u}} = 0$ for any $\mathfrak{u}$ containing j as an element, and so the variable x_j plays no part in the problem $\mathrm{App}_{d,\Gamma,r,s}$. So there is no essential loss of generality in assuming that

$$\gamma_{d,j} > 0 \qquad \forall j \in [d], d \in \mathbb{N}. \tag{5}$$

In the remainder of this paper, we shall also assume that (5) *holds for product weights.* □

An approximation is given by an algorithm $A_{d,\Gamma,r,s,n}$ using at most n linear functionals on $H_\Gamma^{s,1}(\mathbb{T}^d)$. That is, there exist continuous linear functionals $L_1, L_2 \ldots, L_n$ on $H_\Gamma^{s,1}(\mathbb{T}^d)$ and a function $\phi_n \colon \mathbb{R}^n \to H^r(\mathbb{T}^d)$ such that

$$A_{d,\Gamma,r,s,n}(f) = \phi_n\left(L_1(f), L_2(f), \ldots, L_n(f)\right) \qquad \forall f \in \mathscr{B}H_\Gamma^{s,1}(\mathbb{T}^d).$$

The worst case *error* of $A_{d,\Gamma,n,r,s}$ is given by

$$e(A_{d,\Gamma,r,s,n}) = \sup_{f \in \mathscr{B}H_\Gamma^{s,1}(\mathbb{T}^d)} \|f - A_{d,\Gamma,r,s,n}f\|_{H^r(\mathbb{T}^d)}.$$

For simplicity's sake, we measure the cost of an algorithm by the number of information evaluations it uses.

Let $\varepsilon > 0$ be a given error tolerance. An algorithm yields an *ε-approximation* if its error is at most ε. We define the *information complexity* $n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s})$ as the minimal number of linear functionals defined on $H_\Gamma^{s,1}(\mathbb{T}^d)$ needed to find an algorithm whose error as most ε.

As in [11], we have $\| \mathrm{App}_{d,\Gamma,r,s} \| = 1$. Hence it follows that

$$e(0, \mathrm{App}_{d,\Gamma,r,s}) = e(A_{d,r,s,0}, \mathrm{App}_{d,\Gamma,r,s}) = 1,$$

where $A_{d,r,s,0}$ is the *zero algorithm* defined by

$$A_{d,\Gamma,r,s,0}f \equiv 0 \qquad \forall f \in H_\Gamma^{s,1}(\mathbb{T}^d).$$

Thus $n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) = 0$ for $\varepsilon \geq 1$. So in the remainder of this paper, we assume that $\varepsilon \in (0, 1)$, since the problem is trivial otherwise.

It is well-known that there exist algorithms with arbitrarily small error iff the operator $\mathrm{App}_{d,\Gamma,r,s}$ is compact. Since we want to find ε-approximations for any $\varepsilon \in (0, 1)$, we shall assume that $\mathrm{App}_{d,\Gamma,r,s}$ is compact in the remainder of this paper. This compactness holds if either $r < s$ (by Rellich's Theorem, see e.g. [2, p. 219]) or if at least one of the weights $\gamma_{d,\mathfrak{u}}$ is positive (from the results in [11]).

Let $\{(\lambda_{d,n}, e_{d,n})\}_{n\in\mathbb{N}}$ denote the eigensystem of $W_{d,\Gamma,r,s} = \mathrm{App}^*_{d,\Gamma,r,s}\,\mathrm{App}_{d,\Gamma,r,s}$, with $H^{s,1}_{\Gamma}(\mathbb{T}^d)$-orthonormal eigenvectors $e_{d,n}$ and with the eigenvalues $\lambda_{d,n}$ forming a non-increasing sequence

$$\lambda_{d,1} = 1 \geq \lambda_{d,2} \geq \cdots > 0,$$

Then the algorithm

$$A_n(f) = \sum_{i=1}^{n} \langle f, e_{d,i}\rangle_{H^{s,1}_{\Gamma}(\mathbb{T}^d)} e_{d,i} = \sum_{i=1}^{n} \lambda_{d,i}^{-1} \langle f, e_{d,i}\rangle_{L_2(\mathbb{T}^d)} e_{d,i} \qquad \forall f \in \mathscr{B}H^{s,1}_{\Gamma}(\mathbb{T}^d)$$

minimizes the worst case error among *all* algorithms using n linear functionals on $H^{s,1}_{\Gamma}(\mathbb{T}^d)$, with error

$$e(A_n) = \sqrt{\lambda_{d,n+1}},$$

so that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) = \inf\{\, n \in \mathbb{N}_0 : \lambda_{d,n} > \varepsilon^2 \,\}. \tag{6}$$

We are now ready to describe various levels of tractability for the approximation problem $\mathrm{App}_{\Gamma,r,s} = \{\mathrm{App}_{d,\Gamma,r,s}\}_{d\in\mathbb{N}}$. This problem can satisfy any of the following tractability criteria, listed in decreasing order of desirability, see [7] for further discussion.

1. The problem is *strongly* (polynomial) *tractable* if there exists $p \geq 0$ such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq C \left(\frac{1}{\varepsilon}\right)^p \qquad \forall\, \varepsilon \in (0,1), d \in \mathbb{N}. \tag{7}$$

   When this holds, we define

$$p(\mathrm{App}_{\Gamma,r,s}) = \inf\{p \geq 0 : (7) \text{ holds}\}$$

   to be the *exponent of strong tractability*.

2. The problem is (polynomially) *tractable* if there exist non-negative numbers C, p, and q such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq C \left(\frac{1}{\varepsilon}\right)^p d^q \qquad \forall\, \varepsilon \in (0,1), d \in \mathbb{N}. \tag{8}$$

   Numbers $p = p(\mathrm{App}_{\Gamma,r,s})$ and $q = q(\mathrm{App}_{\Gamma,r,s})$ such that (8) holds are called ε- and *d-exponents of tractability*; these need not be uniquely defined.

3. The problem is *quasi-polynomially tractable* if there exist $C \geq 0$ and $t \geq 0$ such that

$$n(\varepsilon, S_d) \leq C \exp\bigl(t(1+\ln\varepsilon^{-1})(1+\ln d)\bigr) \qquad \forall\, \varepsilon \in (0,1), \forall\, d \in \mathbb{N}. \tag{9}$$

   The infimum of all t such that (9) holds is said to be the *exponent of quasi-polynomial tractability*, denoted t^{qpoly}.
4. Let t_1 and t_2 be non-negative numbers. The problem is (t_1, t_2)-*weakly tractable* if non-negative numbers, with

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\ln\, n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s})}{\varepsilon^{-t_1}+d^{t_2}} > 0. \tag{10}$$

   The problem is said to be *weakly tractable* if it is $(1,1)$-weakly tractable, and *uniformly weakly tractable* if it is (t_1, t_2)-weakly tractable for all positive t_1 and t_2. For more details, see [10].
5. The problem is *intractable* if it is not (t_1, t_2)-weakly tractable for any non-negative t_1 and t_2.
6. The problem suffers from the *curse of dimensionality* if there exists $c > 1$ such that[3]

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \geq c^d \qquad \forall\, d \in \mathbb{N}. \tag{11}$$

3 Spectral Results

If we want to follow the prescription for determining minimal error algorithms for our problem, we clearly need to know the eigenvalues and eigenvectors of $W_{d,\Gamma,r,s}$. That's what we'll be doing in this section.

First, a bit more notation. Let $\mathrm{i} = \sqrt{-1}$. For $\mathbf{k} = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{T}^d$, let $\mathbf{k}\cdot\boldsymbol{x} = \sum_{j=1}^d k_j x_j$. Define

$$e_{d,\mathbf{k}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp(\mathrm{i}\,\mathbf{k}\cdot\boldsymbol{x}) \qquad \forall\, \boldsymbol{x} \in \mathbb{T}^d.$$

For any $f \in H^r(\mathbb{T}^d)$, we have

$$D^{\mathbf{m}} f = \sum_{\mathbf{k}\in\mathbb{Z}^d} (\mathrm{i}\,\mathbf{k})^{\mathbf{m}} c_{d,\mathbf{k}}(f)\, e_{d,\mathbf{k}} \qquad \text{for } |\mathbf{m}| \leq r,$$

[3] We follow [5, (5.3)] in using $1 + c$ with $c > 0$ rather than $c > 1$.

where

$$c_{d,\mathbf{k}}(f) = \int_{\mathbb{T}^d} f(\boldsymbol{x}) \, \exp(-\mathrm{i}\,\mathbf{k}\cdot\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

is the **k**th Fourier coefficient of f and convergence is in the $L_2(\mathbb{T}^d)$-sense.

Theorem 1 *For* $\mathbf{k} \in \mathbb{Z}^d$, *let*

$$\lambda_{d,\mathbf{k},\Gamma,r,s} = \frac{\beta_{d,r,\mathbf{k}}}{\beta_{d,s,\mathbf{k}}}\,\alpha_{d,\mathbf{k},\Gamma}, \tag{12}$$

where

$$\alpha_{d,\mathbf{k},\Gamma} = \Bigg(\sum_{\substack{\mathfrak{u}\subseteq[d] \\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}^{-1} \prod_{j\in\mathfrak{u}} k_j^2 \Bigg)^{-1} \tag{13}$$

and

$$\beta_{d,r,\mathbf{k}} = \sum_{|\mathbf{m}|\le r} \mathbf{k}^{2\mathbf{m}}. \tag{14}$$

Then the following hold:

1. *The vectors* $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ *form an orthogonal basis for* $H_\Gamma^{s,1}(\mathbb{T}^d)$, *with*

$$\|e_{d,\mathbf{k}}\|^2_{H_\Gamma^{s,1}(\mathbb{T}^d)} = \alpha_{d,\mathbf{k},\Gamma}^{-1}\,\beta_{d,s,\mathbf{k}} \qquad \forall\,\mathbf{k}\in\mathbb{Z}^d. \tag{15}$$

2. *The vectors* $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ *form an orthogonal basis for* $H^r(\mathbb{T}^d)$, *with*

$$\|e_{d,\mathbf{k}}\|^2_{H^r(\mathbb{T}^d)} = \beta_{d,r,\mathbf{k}} \qquad \forall\,\mathbf{k}\in\mathbb{Z}^d. \tag{16}$$

3. *The eigensystem of* $W_{d,\Gamma,r,s}$ *is given by* $\{(\lambda_{d,\mathbf{k},\Gamma,r,s}, e_{d,\mathbf{k}})\}_{\mathbf{k}\in\mathbb{Z}^d}$, *so that*

$$W_{d,\Gamma,r,s}e_{d,\mathbf{k}} = \lambda_{d,\mathbf{k},\Gamma,r,s}\,e_{d,\mathbf{k}} \qquad \forall\,\mathbf{k}\in\mathbb{Z}^d. \tag{17}$$

4. *The information complexity is given by*

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) = \left|\{\,\mathbf{k}\in\mathbb{Z}^d : \lambda_{d,\mathbf{k},\Gamma,r,s} > \varepsilon^2\,\}\right|. \tag{18}$$

Proof For part 1, we need to show that that $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ is an orthogonal basis for $H^{s,1}_{\Gamma}(\mathbb{T}^d)$. Let $v \in H^{s,1}_{\Gamma}(\mathbb{T}^d)$. Now for any $\mathbf{k} \in \mathbb{Z}^d$ and any $\mathfrak{u} \subseteq [d]$, we have

$$\partial_{\mathfrak{u}} D^{\mathbf{m}} e_{d,\mathbf{k}} = \prod_{j\in\mathfrak{u}} (-\mathrm{i}\, k_j) \prod_{j=1}^{d} (-\mathrm{i}\, k_j)^{m_j}\, e_{d,\mathbf{k}} = (-\mathrm{i})^{|\mathfrak{u}|+|\mathbf{m}|} \mathbf{k}^{\mathbf{m}} e_{d,\mathbf{k}}$$

and so we may integrate by parts and use periodicity to see that

$$\begin{aligned}
\langle \partial_{\mathfrak{u}} D^{\mathbf{m}} v, \partial_{\mathfrak{u}} D^{\mathbf{m}} e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} &= (-1)^{|\mathfrak{u}|+|\mathbf{m}|} \langle v, \partial_{\mathfrak{u}}^2 D^{2\mathbf{m}} e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} \\
&= (-1)^{|\mathfrak{u}|+|\mathbf{m}|} (-\mathrm{i})^{2(|\mathfrak{u}|+\mathbf{m}|)} \left(\prod_{j\in\mathfrak{u}} k_j^2 \right) \mathbf{k}^{2|\mathbf{m}|} \langle v, e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} \\
&= \left(\prod_{j\in\mathfrak{u}} k_j^2 \right) \mathbf{k}^{2\mathbf{m}} \langle v, e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)}.
\end{aligned}$$

Hence for any $\mathbf{k} \in \mathbb{Z}^d$, we have

$$\begin{aligned}
\langle v, e_{d,\mathbf{k}} \rangle_{H^{s,1}_{\Gamma}(\mathbb{T}^d)} &= \sum_{\substack{\mathfrak{u}\subseteq[d] \\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}^{-1} \langle \partial_{\mathfrak{u}} v, \partial_{\mathfrak{u}} e_{d,\mathbf{k}} \rangle_{H^s(\mathbb{T}^d)} \\
&= \sum_{\substack{\mathfrak{u}\subseteq[d] \\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}^{-1} \sum_{|\mathbf{m}|\le s} \langle \partial_{\mathfrak{u}} D^{\mathbf{m}} v, \partial_{\mathfrak{u}} D^{\mathbf{m}} e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} \\
&= \left(\sum_{\substack{\mathfrak{u}\subseteq[d] \\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}^{-1} \prod_{j\in\mathfrak{u}} k_j^2 \right) \left(\sum_{|\mathbf{m}|\le s} \mathbf{k}^{2|\mathbf{m}|} \right) \langle v, e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} \\
&= \alpha_{d,\mathbf{k},\Gamma}^{-1}\, \beta_{d,s,\mathbf{k}} \langle v, e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)}.
\end{aligned} \tag{19}$$

In particular, we see that

$$\langle e_{d,\mathbf{p}}, e_{d,\mathbf{k}} \rangle_{H^{s,1}_{\Gamma}(\mathbb{T}^d)} = \alpha_{d,\mathbf{k},\Gamma}^{-1}\, \beta_{d,s,\mathbf{k}}\, \delta_{\mathbf{p},\mathbf{k}} \qquad \forall\, \mathbf{k}, \mathbf{p} \in \mathbb{Z}^d, \tag{20}$$

with $\delta_{\mathbf{k},\mathbf{p}}$ being the Kronecker delta. Hence $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ is an $H^{s,1}_{\Gamma}(\mathbb{T}^d)$-orthogonal set, the norm of whose elements being given by (15). To see that this set is a basis, we need only show that this set is $H^{s,1}_{\Gamma}(\mathbb{T}^d)$-complete. So let $v \in H^{s,1}_{\Gamma}(\mathbb{T}^d)$ satisfy $\langle v, e_{d,\mathbf{k}} \rangle_{H^{s,1}_{\Gamma}(\mathbb{T}^d)} = 0$ for all $\mathbf{k} \in \mathbb{Z}^d$. Once again using (19), it follows that $\langle v, e_{d,\mathbf{k}} \rangle_{L_2(\mathbb{T}^d)} = 0$ for all $\mathbf{k} \in \mathbb{Z}^d$. Since $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ is an orthogonal basis for $L_2(\mathbb{T}^d)$, it follows that $v = 0$. Hence $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ is $H^{s,1}_{\Gamma}(\mathbb{T}^d)$-complete, as required.

Setting $\Gamma = \Gamma(\emptyset)$ in part 1, we immediately have part 2.

To see that part 3 holds, note that

$$\langle W_{d,\Gamma,r,s} e_{d,\mathbf{k}}, e_{d,\mathbf{p}} \rangle_{H_\Gamma^{s,1}(\mathbb{T}^d)} = \langle e_{d,\mathbf{k}}, e_{d,\mathbf{p}} \rangle_{H^r(\mathbb{T}^d)} = \beta_{d,r,\mathbf{k}}\, \delta_{\mathbf{k},\mathbf{p}}, \qquad \forall\, \mathbf{k}, \mathbf{p} \in \mathbb{Z}^d,$$

the second equality following from part 2. Since $\{e_{d,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ is an orthogonal basis for $H_\Gamma^{s,1}(\mathbb{T}^d)$, it follows that $W_{d,\Gamma,r,s} e_{d,\mathbf{k}}$ must be a multiple of $e_{d,\mathbf{k}}$, which means that $e_{d,\mathbf{k}}$ is an eigenvector of $W_{d,\Gamma,r,s}$. Thus $W_{d,\Gamma,r,s} e_{d,\mathbf{k}} = \lambda_{d,\mathbf{k},\Gamma,r,s} e_{d,\mathbf{k}}$ for some $\lambda_{d,\mathbf{k},\Gamma,r,s} > 0$, with

$$\lambda_{d,\mathbf{k},\Gamma,r,s} = \frac{\|e_{d,\mathbf{k}}\|^2_{H^r(\mathbb{T}^d)}}{\|e_{d,\mathbf{k}}\|^2_{H_\Gamma^{s,1}(\mathbb{T}^d)}}, \tag{21}$$

as usual. Part 3 follows once we use (15) and (16) in (21).

Finally, part 4 follows immediately from (6), along with the remaining parts of this theorem.

As a special case, let $s = r$. Then the problem $\mathrm{App}_{d,\Gamma,r,r}$ is equivalent to the problem $\mathrm{App}_{d,\Gamma,0,0}$:

Corollary 1 *The following results hold for the problem* $\mathrm{App}_{d,\Gamma,r,r}$*:*

1. *The operators $W_{d,\Gamma,r,r}$ and $W_{d,\Gamma,0,0}$ both have $\{(e_{d,\mathbf{k}}, \alpha_{d,\mathbf{k},\Gamma})\}_{\mathbf{k}\in\mathbb{Z}^d}$ as their eigensystems.*
2. *Minimal errors, minimal error algorithms, and levels of tractability are the same for our problem* $\mathrm{App}_{d,\Gamma,r,r}$ *and for the problem* $\mathrm{App}_{d,\Gamma,0,0}$. □

Just as we have reduced the problem $\mathrm{App}_{d,\Gamma,r,r}$ to the problem $\mathrm{App}_{d,\Gamma,0,0}$, we can also reduce the problem $\mathrm{App}_{d,\Gamma,r,s}$ to the simpler problem $\mathrm{App}_{d,\Gamma,0,s-r}$. Let

$$\eta_{d,\mathbf{k}} = 1 + \sum_{j=1}^{d} k_j^2. \tag{22}$$

We then have

Theorem 2 *Let $r, s \in \mathbb{N}_0$, with $s \geq r$.*

1. *The eigenvectors of $W_{d,\Gamma,r,s}$ are given by $\{\, e_{d,\mathbf{k}} : \mathbf{k} \in \mathbb{Z}^d \,\}$.*
2. *The eigenvalues of $W_{d,\Gamma,r,s}$ satisfy the inequality*

$$\begin{aligned}\frac{1}{r!(s-r)!}\,\lambda_{d,\mathbf{k},\Gamma,0,s-r} \leq \frac{1}{r!}\,\frac{\alpha_{d,\mathbf{k},\Gamma}}{\eta_{d,\mathbf{k}}^{s-r}} \leq \lambda_{d,\mathbf{k},\Gamma,r,s} \leq s!\,\frac{\alpha_{d,\mathbf{k},\Gamma}}{\eta_{d,\mathbf{k}}^{s-r}} \\ \leq s!\,\lambda_{d,\mathbf{k},\Gamma,0,s-r}\end{aligned} \tag{23}$$

for all $\mathbf{k} \in \mathbb{Z}^d$.

Proof Let $d \in \mathbb{N}$ and $\mathbf{k} \in \mathbb{Z}^d$. As in [5], we may use the multinomial theorem to see that

$$\beta_{d,\ell,\mathbf{k}} \le \eta_{d,\mathbf{k}}^{\ell} \le \ell!\, \beta_{d,\ell,\mathbf{k}}. \qquad \forall \ell \in \mathbb{N}_0.$$

We then have

$$\frac{1}{r!\,(s-r)!\,\beta_{d,s-r,\mathbf{k}}} \le \frac{1}{r!\,\eta_{d,\mathbf{k}}^{s-r}} \le \frac{\beta_{d,r,\mathbf{k}}}{\beta_{d,s,\mathbf{k}}} \le \frac{s!}{\eta_{d,\mathbf{k}}^{s-r}} \le \frac{s!}{\beta_{d,s-r,\mathbf{k}}}.$$

This result now follows from Theorem 1 and (18). □

From Theorem 2, we see that minimal errors for our problem $\mathrm{App}_{d,\Gamma,r,s}$ and for the simpler problem $\mathrm{App}_{d,\Gamma,0,s-r}$ are essentially the same.

4 Tractability Results

We now compare the tractability of our problem $\mathrm{App}_{\Gamma,r,s} = \{\mathrm{App}_{d,\Gamma,r,s}\}_{d\in\mathbb{N}}$ with the problem $\mathrm{App}_{\Gamma,0,0} = \{\mathrm{App}_{d,\Gamma,0,0}\}_{d\in\mathbb{N}}$. The papers [11, 12] studied this latter problem, except for functions defined over the unit cube instead of the unit torus.

4.1 General Weights

We first give tractability results that hold for any weights, regardless of their structure (or lack thereof), depending only some boundedness conditions. Our main result is that our approximation problem $\mathrm{App}_{\Gamma,r,s}$ has the same level of tractability as the problem $\mathrm{App}_{\Gamma,0,0}$, which is the periodic version of the problem studied in [12]. In what follows, we let

$$M_d = \max\left\{1, \max_{j\in[d]} \gamma_{d,\{j\}}\right\} \qquad \text{and} \qquad m_d = \min_{\substack{\mathfrak{u}\subseteq[d]\\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}. \tag{24}$$

Clearly both M_d and m_d are positive numbers.

First, we compare the information complexity of these problems.

Theorem 3 *For all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$, we have*

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \ge n\left(\left(r! M_d^{s-r}\right)^{1/(2(s-r+1))} \varepsilon^{1/(s-r+1)}, \mathrm{App}_{d,\Gamma,0,0}\right), \tag{25}$$

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le n\left(\left(\frac{m_d^{s-r}}{s!}\right)^{1/(2(s-r+1))} \varepsilon^{1/(s-r+1)}, \mathrm{App}_{d,\Gamma,0,0}\right), \tag{26}$$

and

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le n(\varepsilon, \mathrm{App}_{d,\Gamma,0,0}). \tag{27}$$

Proof We first show that (25) holds. Let $\mathbf{k} \in \mathbb{Z}^d$. From (5), (13), (18), and (22), it follows that

$$\alpha_{d,\mathbf{k},\Gamma}^{-1} \ge 1 + \sum_{j=1}^{d} \gamma_{d,j}^{-1} k_j^2 \ge 1 + \min_{j\in[d]} \gamma_{d,j}^{-1} \sum_{j=1}^{d} k_j^2 \ge M_d^{-1}\left(1 + \sum_{j=1}^{d} k_j^2\right) = M_d^{-1}\,\eta_{d,\mathbf{k}}.$$

Since $\eta_{d,\mathbf{k}}^{-1} \ge M_d^{-1}\alpha_{d,\mathbf{k},\Gamma}$, we may use Theorem 2 to see that

$$\lambda_{d,\mathbf{k},\Gamma,r,s} \ge \frac{1}{r!}\frac{\alpha_{d,\mathbf{k},\Gamma}}{\eta_{d,\mathbf{k}}^{s-r}} \ge \frac{1}{r!M_d^{s-r}}\alpha_{d,\mathbf{k},\Gamma}^{s-r+1}.$$

Using part 4 of Theorem 1 and the previous estimate, we now have

$$\begin{aligned} n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) &= \left|\{\,\mathbf{k} \in \mathbb{Z}^d : \lambda_{d,\mathbf{k},\Gamma,r,s} > \varepsilon^2\,\}\right| \\ &\ge \left|\left\{\,\mathbf{k} \in \mathbb{Z}^d : \frac{1}{r!M_d^{s-r}}\alpha_{d,\mathbf{k},\Gamma}^{s-r+1} > \varepsilon^2\,\right\}\right| \\ &= \left|\left\{\,\mathbf{k} \in \mathbb{Z}^d : \alpha_{d,\mathbf{k},\Gamma} > \left(r!M_d^{s-r}\varepsilon^2\right)^{1/(s-r+1)}\,\right\}\right| \\ &= n\left(\left(r!M_d^{s-r}\right)^{1/(2(s-r+1))}\varepsilon^{1/(s-r+1)}, \mathrm{App}_{d,\Gamma,0,0}\right), \end{aligned}$$

as required.

The proof of (26) is similar to that of (25), except that we start with the bound

$$\alpha_{d,\mathbf{k},\Gamma}^{-1} = \sum_{\substack{\mathfrak{u}\subseteq[d]\\ \gamma_{d,\mathfrak{u}}>0}} \gamma_{d,\mathfrak{u}}^{-1} \prod_{j\in\mathfrak{u}} k_j^2 \le m_d^{-1} \sum_{\substack{\mathfrak{u}\subseteq[d]\\ \gamma_{d,\mathfrak{u}}>0}} \prod_{j\in\mathfrak{u}} k_j^2 \le m_d^{-1}\beta_{d,1,\mathbf{k}} = m_d^{-1}\eta_{d,k}.$$

Finally, (27) follows from (18) and Theorem 1. □

We now show that the level of tractability of our problem $\mathrm{App}_{\Gamma,r,s}$ is often the same as that of the problem $\mathrm{App}_{\Gamma,0,0}$.

Theorem 4 *If* $\mathrm{App}_{\Gamma,0,0}$ *has a given level of tractability, then* $\mathrm{App}_{\Gamma,r,s}$ *has at least the same level of tractability, and the exponent(s) for* $\mathrm{App}_{\Gamma,r,s}$ *are bounded from above by those for* $\mathrm{App}_{\Gamma,0,0}$. *Moreover, recalling the definition* (24) *of* M_d, *we have* *the following:*

1. *If*

$$M := \sup_{d\in\mathbb{N}} M_d < \infty \tag{28}$$

then the following hold:

a. $\mathrm{App}_{\Gamma,r,s}$ *is strongly polynomially tractable iff* $\mathrm{App}_{\Gamma,0,0}$ *is strongly polynomially tractable, in which case the exponents of strong tractability satisfy the inequality*

$$\frac{1}{s-r+1}\, p(\mathrm{App}_{\Gamma,0,0}) \le p(\mathrm{App}_{\Gamma,r,s}) \le p(\mathrm{App}_{\Gamma,0,0}). \tag{29}$$

b. $\mathrm{App}_{\Gamma,r,s}$ *is quasi-polynomially tractable iff* $\mathrm{App}_{\Gamma,0,0}$ *is quasi-polynomially tractable, in which case the exponents of strong quasi-polynomial tractability satisfy the inequality*

$$\begin{aligned}\frac{1}{\max\left\{s-r, \frac{1}{2}\ln(r!M^{s-r})\right\}+1}\, t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0}) \le t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,r,s})& \\ \le t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0}).&\end{aligned} \tag{30}$$

2. *If*

$$\sup_{d\in\mathbb{N}} d^{-q} M_d < \infty \tag{31}$$

for some $q \ge 0$*, then* $\mathrm{App}_{\Gamma,r,s}$ *is polynomially tractable iff* $\mathrm{App}_{\Gamma,0,0}$ *is polynomially tractable.*

Proof The first statement in the theorem follows immediately from (27).

For part 1, suppose that (28) holds.

We first prove part 1(a). From the first statement in the theorem, it suffices to show that if $\mathrm{App}_{\Gamma,r,s}$ is strongly polynomially tractable, then the same is true for $\mathrm{App}_{\Gamma,0,0}$, and that the first inequality in (29) holds. So let $\mathrm{App}_{\Gamma,r,s}$ be strongly polynomially tractable, so that for any $p > p(\mathrm{App}_{\Gamma,r,s})$, there exists $C > 0$ such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le C\varepsilon^{-p} \qquad \forall\, \varepsilon \in (0,1), d \in \mathbb{N}.$$

Set

$$\varepsilon_d = \left(r!M_d^{s-r}\right)^{1/(2(s-r+1))} \varepsilon^{1/(s-r+1)}, \tag{32}$$

so that

$$\varepsilon^{-1} = (r!M_d^{s-r})^{1/2}\varepsilon_d^{-(s-r+1)}.$$

Using (25) and (32), we see that

$$\begin{aligned} n(\varepsilon_d, \mathrm{App}_{d,\Gamma,0,0}) &\le n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le C\varepsilon^{-p} = C\left(M_d^{s-r}r!\right)^{p/2}\varepsilon_d^{-(s-r+1)p} \\ &\le C\left(M^{s-r}r!\right)^{p/2}\varepsilon_d^{-(s-r+1)p}. \end{aligned}$$

Varying $\varepsilon > 0$, we see that ε_d can assume arbitrary positive values here. Since p may be chosen arbitrarily close to $p(\mathrm{App}_{\Gamma,r,s})$, we see that $\mathrm{App}_{\Gamma,0,0}$ is strongly polynomially tractable, and that (29) holds, as required.

We now prove part 1(b). It suffices to show that if $\mathrm{App}_{\Gamma,r,s}$ is strongly quasi-polynomially tractable, then so is $\mathrm{App}_{\Gamma,0,0}$, and that the first inequality in (30) holds. So suppose that $\mathrm{App}_{\Gamma,0,0}$ is quasi-polynomially tractable. Then for any $t > t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0})$, there exists $C > 0$ such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le C\exp\left(t(1+\ln\varepsilon^{-1})(1+\ln d)\right) \qquad \forall\, \varepsilon \in (0,1), d \in \mathbb{N}.$$

Once again, define ε_d by (32) and use (25) to see that

$$\begin{aligned} n(\varepsilon_d, \mathrm{App}_{d,\Gamma,0,0}) &\le n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \le C\exp\left(t(1+\ln\varepsilon^{-1})(1+\ln d)\right) \\ &= C\exp\left[t\left(1+(s-r+1)\ln\varepsilon_d^{-1} + \tfrac{1}{2}\ln(r!M_d^{s-r})\right)(1+\ln d)\right] \\ &\le C\exp\left[t\left(1+(s-r+1)\ln\varepsilon_d^{-1} + \tfrac{1}{2}\ln(r!M^{s-r})\right)(1+\ln d)\right] \end{aligned} \tag{33}$$

Define $g\colon [0,\infty) \to [0,\infty)$ as

$$g(\xi) = \frac{1+(s-r+1)\xi + \frac{1}{2}\ln(r!M^{s-r})}{1+\xi} \qquad \forall\, \xi \ge 0.$$

We find that

$$\sup_{\xi\ge 0} g(\xi) = \max\left\{g(0), \lim_{\xi\to\infty} g(\xi)\right\} = \max\left\{s-r, \tfrac{1}{2}\ln(r!M^{s-r})\right\} + 1.$$

From (33), we now see that

$$n(\varepsilon_d, \mathrm{App}_{d,\Gamma,0,0}) \le C\exp\left(t_1(1+\ln\varepsilon_d^{-1})(1+\ln d)\right),$$

where

$$t_1 = t\sup_{d\in\mathbb{N}} g(\ln\varepsilon_d^{-1}) = t\left(\max\left\{s-r, \tfrac{1}{2}\ln(r!M^{s-r})\right\} + 1\right). \tag{34}$$

Arguing as in the strongly polynomial case, we see that $\mathrm{App}_{\Gamma,0,0}$ is quasi-polynomially tractable, with

$$t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,r,s}) \le \left(\max\left\{s-r, \tfrac{1}{2}\ln(r!M^{s-r})\right\}+1\right) t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0}),$$

as required.

For part 2, suppose that (31) holds, so that $M := \sup_{d\in\mathbb{N}} d^{-q}M_d < \infty$. Suppose also that $\mathrm{App}_{\Gamma,r,s}$ is polynomially tractable, so that there exist positive C, ℓ, and p such that such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,0,0}) \le Cd^{\ell}\varepsilon^{-p} \qquad \forall\, d\in\mathbb{N}, \varepsilon\in(0,1).$$

Once again defining ε_d as in (32), we have

$$\begin{aligned} n(\varepsilon_d, \mathrm{App}_{d,\Gamma,0,0}) &\le C\,d^{\ell}\varepsilon^{-p} = C\,d^{\ell}(r!M_d^{s-r})^{p/2}\varepsilon_d^{-(s-r+1)p} \\ &\le C\,d^{\ell}(r!M^{s-r})^{p/2}\varepsilon_d^{-(s-r+1)p}. \end{aligned}$$

Hence $\mathrm{App}_{\Gamma,0,0}$ is polynomially tractable. □

Remark 3 The non-trivial results in Theorem 4 hold when the boundedness conditions (28) or (31) are satisfied. Suppose that we allow unbounded weights. Although the tractability of $\mathrm{App}_{\Gamma,r,s}$ is no worse than the tractability of $\mathrm{App}_{\Gamma,0,0}$, we can say nothing in the opposite direction in this case. As an extreme example, we show a choice of (unbounded) weights such that $\mathrm{App}_{\Gamma,r,s}$ to be strongly polynomially tractable, but for which $\mathrm{App}_{\Gamma,0,0}$ suffers from the curse of dimensionality.

Define our weight set Γ as

$$\gamma_{d,\mathfrak{u}} = \begin{cases} 1 & \text{if } \mathfrak{u}=\emptyset, \\ (1+c)^{2d} & \text{if } \mathfrak{u}=\{1\}, \\ 0 & \text{otherwise.} \end{cases}$$

This is actually a sequence of univariate problems, for which

$$\alpha_{1,k,\Gamma} = \left(1+(1+c)^{-2d}\right)k^2 \qquad \text{and} \qquad \eta_{1,k} = 1+k^2.$$

From Theorem 2, we see that the eigenvalues of $W_{1,\Gamma,r,s}$ satisfy

$$\lambda_{1,k} \le s!\,\frac{\alpha_{1,k,\Gamma}}{\eta_{1,k}^{s-r}} = \frac{s!}{\left(1+(1+c)^{-2d}(1+k^2)\right)(1+k^2)^{s-r}} \le \frac{s!}{(1+k^2)^{s-r}}.$$

Hence we may use (23) to see that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq \left|\left\{ k \in \mathbb{Z} : \frac{s!}{(1+k^2)^{s-r}} > \varepsilon^2 \right\}\right| = 2\left\lfloor \sqrt{\left(\frac{s!}{\varepsilon^2}\right) - 1} \right\rfloor - 1$$
$$= \Theta\left(\varepsilon^{1/(s-r)}\right),$$

and so $\mathrm{App}_{\Gamma,r,s}$ is strongly polynomially tractable, provided that $r < s$. On the other hand, we have

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,0,0})` = |\{ k \in \mathbb{Z} : \alpha_{1,k,\Gamma} > \varepsilon^2 \}| = |\{ k \in \mathbb{Z} : 1 + (1+c)^{-2d} k^2 > \varepsilon^2 \}|$$
$$= 2\left\lfloor (1+c)^d \sqrt{\varepsilon^{-2} - 1} \right\rfloor - 1 = \Theta\left((1+c)^d \varepsilon^{-1}\right),$$

and so $\mathrm{App}_{\Gamma,0,0}$ suffers from the curse of dimensionality. □

Remark 4 If we are willing to live with an upper bound that depends on d, we can improve the ε-exponent in Theorem 4. (This is an example of the tradeoff of exponents, as described several places in [7].) To be specific, suppose that $\mathrm{App}_{\Gamma,0,0}$ is strongly polynomially tractable. Then for any $p > p(\mathrm{App}_{\Gamma,0,0})$, there exists $C > 0$ such that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,0,0}) \leq C\varepsilon^{-p} \qquad \forall\, \varepsilon \in (0,1), d \in \mathbb{N}.$$

Choosing such a p, d, and ε, let

$$\varepsilon_d = \left(\frac{m_d^{s-r}}{s!}\right)^{1/(2(s-r+1))} \varepsilon^{1/(s-r+1)},$$

where m_d is defined by (24). Using (26), the previous inequality tells us that

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq n(\varepsilon_d, \mathrm{App}_{\Gamma,0,0}) \leq C\left(\frac{s!}{m_d^{s-r}}\right)^{p/(2(s-r+1))} \varepsilon^{-p/(s-r+1)}. \tag{35}$$

Let $m = \inf_{d\in\mathbb{N}} m_d$. There are two cases to consider:

1. Suppose that $m > 0$. Then the $H_\Gamma^{s,1}(\mathbb{T}^d)$-norms are equivalent to the $H^{s,1}(\mathbb{T}^d)$-norms, with equivalence factors independent of d. As we shall see in Sect. 4.2, the problems $\mathrm{App}_{\Gamma(\mathrm{UNW}),r,s}$ and $\mathrm{App}_{\Gamma(\mathrm{UNW}),0,0}$ are both quasi-polynomially tractable, each having exponent $2/\ln 2 \doteq 2.88539$. Hence the same is true for the problems $\mathrm{App}_{\Gamma,r,s}$ and $\mathrm{App}_{\Gamma,0,0}$. Thus part 1(a) of Theorem 4 never comes into play when $m > 0$, and so the estimate (35) does not apply.

2. Suppose that $m = 0$. Then the bound (35) truly depends on d. To cite two examples:

   - Suppose that $m_d \geq C_\alpha d^{-\alpha}$ for some $\alpha > 0$ and $C_\alpha > 0$. Using (35), and letting

     $$C_1 = C\left(\frac{s!}{C_\alpha^{s-r}}\right)^{p/(2(s-r+1))},$$

     we see that

     $$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq C_1\, d^{\alpha p(s-r)/(2(s-r+1))}\, \varepsilon^{-p/(s-r+1)}.$$

     Since p can be chosen arbitrarily close to $p(\mathrm{App}_{\Gamma,0,0})$, this is a polynomially-tractable upper bound on $n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s})$, with

     $$d\text{-exponent: } \frac{\alpha(s-r)\,p(\mathrm{App}_{\Gamma,0,0})}{2(s-r+1)} \qquad \text{and} \qquad \varepsilon^{-1}\text{-exponent: } \frac{p(\mathrm{App}_{\Gamma,0,0})}{s-r+1}.$$

   - Suppose that for *any* $\alpha > 0$, there exists $C_\alpha >$ such that $m_d \geq C_\alpha d^{-\alpha}$. (For instance, this holds if m_d is bounded from below by a power of log d.) We now see that the results of the previous case hold for positive α, no matter how small. Hence we find that $\mathrm{App}_{\Gamma,r,s}$ is polynomially tractable for such Γ, with

     $$d\text{-exponent: } 0 \qquad \text{and} \qquad \varepsilon^{-1}\text{-exponent: } \frac{1}{s-r+1} p(\mathrm{App}_{\Gamma,0,0}).$$

     This is close to, but not identical to, a strong polynomial bound for which

     $$p(\mathrm{App}_{\Gamma,r,s}) = \frac{1}{s-r+1}\, p(\mathrm{App}_{\Gamma,0,0}). \tag{36}$$

     We might describe such a bound as being *almost strongly polynomial.* □

Remark 5 From Remark 4 we see that the left-hand inequality in (29) cannot be improved. However, this fact does not imply that there are problems for which (36) holds. To see that such problems do exist, suppose we choose our weights as

$$\gamma_{d,\mathfrak{u}} = \begin{cases} 1 & \text{if } \mathfrak{u} = \emptyset \text{ or } \mathfrak{u} = \{1\}, \\ 0 & \text{otherwise.} \end{cases}$$

We claim that (36) holds for this problem. Indeed, the eigenvalues of $W_{d,\Gamma,0,0}$ are given by $1/(1+k^2)$ for $k \in \mathbb{Z}$, so that Theorem 2 tells us that the eigenvalues of $W_{d,\Gamma,r,s}$ are bounded from below by $1/\left(r!(1+k^2)^{s-r+1}\right)$ and from above by

$s!/(1+k^2)^{s-r+1}$. It now follows that $n(\varepsilon, \mathrm{App}_{\Gamma,0,0}) = \Theta(\varepsilon^{-1})$ and $n(\varepsilon, \mathrm{App}_{\Gamma,r,s}) = \Theta(\varepsilon^{-1/(s-r+1)})$. Since $p(\mathrm{App}_{\Gamma,0,0}) = 1$ and $p(\mathrm{App}_{\Gamma,r,s}) = 1/(s-r+1)$, we see that (36) holds, as claimed. □

Remark 6 Note that Theorem 4 doesn't mention (r_1, r_2)-weak tractability. That's because (r_1, r_2)-weak tractability simply never arises. To see this, we distinguish between two cases:

1. Suppose that we allow Γ to contain an unbounded sequence of weights. Using Remark 3, we can find a case in which $\mathrm{App}_{\Gamma,0,0}$ suffers from the curse of dimensionality, but $\mathrm{App}_{\Gamma,r,s}$ is strongly polynomially tractable.
2. The alternative is to suppose that the weights are uniformly bounded, with $M = \sup_{d\in\mathbb{N}} \max_{\mathfrak{u}\subseteq[d]} \gamma_{d,\mathfrak{u}} < \infty$. We claim that $\mathrm{App}_{\Gamma,0,0}$ is always (at least) quasi-polynomially tractable in this case, so that the same is true for $\mathrm{App}_{\Gamma,r,s}$ by part 1(b) of Theorem 4.

   Indeed, to see that $\mathrm{App}_{\Gamma,0,0}$ with weights bounded by M is always (at least) quasi-polynomially tractable, note that this problem is no harder than the problem $\mathrm{App}_{\Gamma,0,0}$ for which $\gamma_{d,\mathfrak{u}} \equiv M$. From Theorem 1, we see that the eigenvalues of this latter problem are given by

$$\lambda_{d,\mathbf{k},\Gamma,0,0} = \alpha_{d,\mathbf{k},\Gamma} = M\prod_{j=1}^{d} \frac{1}{1+k_j^2}.$$

   As in Remark 4, this latter problem is quasi-polynomially tractable. Hence $\mathrm{App}_{\Gamma,0,0}$ is at least quasi-polynomially tractable, as claimed. □

The right-hand inequality in Theorem 3 may be summarized as saying that our approximation problem $\mathrm{App}_{d,\Gamma,r,s}$ is no harder than the approximation problem $\mathrm{App}_{d,\Gamma,0,0}$ studied in [11]. The left-hand inequality tells us that $\mathrm{App}_{d,\Gamma,r,s}$ may be easier than $\mathrm{App}_{d,\Gamma,0,0}$. Despite this gap, we find that these two problems sometimes share the same level of tractability, as we shall see in what follows.

4.2 The Unweighted Case

If we specify the structure of the weights, we can get more detailed results. We first look at the unweighted case $\Gamma = \Gamma(\text{UNW})$, see item 1 in Remark 1. Our main result is that this problem is quasi-polynomially tractable.

Theorem 5 *Suppose that $\Gamma = \Gamma(\text{UNW})$. Let*

$$\tau^* = \frac{1}{\ln 2} \doteq 1.44270 \tag{37}$$

and

$$c_1 = \left(\sum_{j=-\infty}^{\infty} \left(\frac{1}{1+k^2} \right)^{\tau^*} \right)^{1/\tau^*} \doteq 2.09722.$$

Then

$$n(\varepsilon, \mathrm{App}_{d,\Gamma(\mathrm{UNW}),r,s}) \le c_1 \exp\left(2\tau^*(1+\ln d)(1+\ln \varepsilon^{-1})\right),$$

and so $\mathrm{App}_{\Gamma(\mathrm{UNW}),r,s}$ *is quasi-polynomially tractable. Moreover, its exponent is*

$$t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma(\mathrm{UNW}),r,s}) = 2\tau^* = \frac{2}{\ln 2} \doteq 2.88539.$$

Proof From [7, Theorem 23.2], we have

$$t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma(\mathrm{UNW}),r,s}) = 2 \inf\{\, \tau > 0 : C_\tau < \infty \,\},$$

where

$$C_\tau = \sup_{d \in \mathbb{N}} C_{\tau,d},$$

with

$$C_{\tau,d} = \frac{1}{d^2} \left(\sum_{\mathbf{k} \in \mathbb{Z}^d} \lambda_{d,\mathbf{k},\Gamma_1,r,s}^{\tau(1+\ln d)} \right)^{1/\tau}.$$

Moreover,

$$n(\varepsilon, \mathrm{App}_{\Gamma(\mathrm{UNW}),r,s}) \le C_\tau^\tau \exp\left(2\tau(1+\ln \varepsilon^{-1})(1+\ln d)\right)$$

for any $\tau > 0$ such that $C_\tau < \infty$. It suffices to show that τ^* is the minimal τ for which $C_\tau < \infty$.

Choose $\tau > 0$ such that $C_\tau < \infty$; we must show that $\tau \ge \tau^*$. For any $p > 0$, Theorem 2 tells us that

$$\sum_{\mathbf{k} \in \mathbb{Z}^d} \lambda_{d,\mathbf{k},\Gamma(\mathrm{UNW}),r,s}^p \ge \left(\frac{1}{(s-r)!} \right)^p \sum_{\mathbf{k} \in \{0,1\}^d} \left(\frac{\alpha_{d,\mathbf{k},\Gamma(\mathrm{UNW})}}{\eta_{d,\mathbf{k}}^{s-r}} \right)^p.$$

But for $\mathbf{k} \in \{0,1\}^d$, we have

$$\eta_{d,\mathbf{k}} = 1 + \sum_{j=1}^{d} k_j^2 \le 1 + d$$

and

$$\alpha_{d,\mathbf{k},\Gamma(\text{UNW})} = \prod_{j=1}^{d} \frac{1}{1+k_j^2} = \left(\tfrac{1}{2}\right)^{|\{j\in[d]:k_j=1\}|}.$$

Hence for any $p \geq 0$, we have

$$\begin{aligned}\sum_{\mathbf{k}\in\mathbb{Z}^d} \lambda^p_{d,\mathbf{k},\Gamma(\text{UNW}),r,s} &\geq \left(\frac{1}{(s-r)!(1+d)^{s-r}}\right)^p \sum_{\mathbf{k}\in\{0,1\}^d} \left(\tfrac{1}{2}\right)^{p|\{j\in[d]:k_j=1\}|} \\ &= \left(\frac{1}{(s-r)!(1+d)^{s-r}}\right)^p \sum_{j=0}^{d} \binom{d}{j} \left(\tfrac{1}{2}\right)^{pj} \\ &= \left(\frac{1}{(s-r)!(1+d)^{s-r}}\right)^p \left[1+\left(\tfrac{1}{2}\right)^p\right]^d.\end{aligned}$$

Let $p = \tau(1+\ln d)$ and take logarithms. Then

$$\ln\left[\sum_{\mathbf{k}\in\mathbb{Z}^d} \lambda^{\tau(1+\ln d)}_{d,\mathbf{k},\Gamma(\text{UNW}),r,s}\right] \geq d\ln\left[1+\left(\tfrac{1}{2}\right)^{\tau(1+\ln d)}\right] - \tau(1+\ln d)\ln\left((s-r)!(1+d)^{s-r}\right).$$

Since $\ln(1+\delta) \geq \delta - \frac{1}{2}\delta^2$ for $\delta \geq 0$, we have

$$\ln\left[1+\left(\tfrac{1}{2}\right)^{\tau(1+\ln d)}\right] \geq \left(\tfrac{1}{2}\right)^{\tau(1+\ln d)}\left(1-\left(\tfrac{1}{2}\right)^{\tau(1+\ln d)+1}\right).$$

Since $d \in \mathbb{N}$ and $\tau > 0$, we have $\left(\frac{1}{2}\right)^{\tau(1+\ln d)+1} \leq \left(\frac{1}{2}\right)^{\tau+1} \leq \frac{1}{2}$, and so

$$\ln\left[1+\left(\tfrac{1}{2}\right)^{\tau(1+\ln d)}\right] \geq \tfrac{1}{2}\left(\tfrac{1}{2}\right)^{\tau(1+\ln d)} = 2^{-(\tau+1)}d^{-\tau\ln 2}.$$

Without loss of generality, let $d \geq 2$, so that

$$\begin{aligned}&\tau(1+\ln d)\ln\left((s-r)!(1+d)^{s-r}\right) \\ &\qquad \leq \tau\left(1+\frac{1}{\ln 2}\right)^2 \ln^2 d + \tau\left(1+\frac{1}{\ln 2}\right)[\ln(s-r)! + (s-r)\ln d].\end{aligned}$$

Thus

$$\ln\left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\lambda_{d,\mathbf{k},\Gamma(\text{UNW}),r,s}^{\tau(1+\ln d)}\right] \ge 2^{-(\tau+1)}d^{1-\tau\ln 2} - \tau\left(1+\frac{1}{\ln 2}\right)^2\ln^2 d - \tau\left(1+\frac{1}{\ln 2}\right)[\ln(s-r)! + (s-r)\ln d],$$

and so

$$\begin{aligned}\ln C_{\tau,d} &= \tau^{-1}\ln\left[\sum_{\mathbf{k}\in\mathbb{Z}^d}\lambda_{d,\mathbf{k},\Gamma(\text{UNW}),r,s}^{\tau(1+\ln d)}\right] - 2\ln d\\ &\ge \tau^{-1}2^{-(\tau+1)}d^{1-\tau\ln 2}\\ &\quad - \left[\left(1+\frac{1}{\ln 2}\right)^2\ln^2 d + \left[3+\frac{1}{\ln 2}(s-r)\right]\ln d + \left(1+\frac{1}{\ln 2}\right)\ln(s-r)!\right].\end{aligned}$$

Since $\sup_{d\in\mathbb{N}} C_{\tau,d}$ must be finite, we see that the exponent of d must be non-positive. Hence we must have

$$\tau \ge \tau^* = \frac{1}{\ln 2} \doteq 1.44270,$$

as required.

It remains to show that $C_{\tau^*} < \infty$. From (27), it suffices to show that $C_{\tau^*} < \infty$ for $\text{App}_{\Gamma,0,0}$. Suppose first that $d = 1$. Again using (27), we see that

$$\lambda_{1,k,\Gamma(\text{UNW}),r,s} \le \lambda_{1,k,\Gamma(\text{UNW}),0,0} = \frac{1}{1+k^2},$$

and so

$$C_{\tau^*,1}^{\tau^*} \le c_1 := \sum_{k\in\mathbb{Z}}\lambda_{1,k,\Gamma(\text{UNW}),0,0}^{\tau^*} = \sum_{k=-\infty}^{\infty}\left(\frac{1}{1+k^2}\right)^{\tau^*}.$$

Since the terms in the series are $\Theta(k^{-2\tau^*})$, with $\tau^* \doteq 1.44270$, the series converges; using Mathematica, we find that $c_1 \doteq 2.09722$.

Now suppose that $d \ge 2$. Since

$$\lambda_{d,\mathbf{k},\Gamma(\text{UNW}),r,s} \le \lambda_{d,\mathbf{k},\Gamma(\text{UNW}),0,0} \le \alpha_{d,\mathbf{k},\Gamma(\text{UNW})} = \prod_{j=1}^{d}\frac{1}{1+k_j^2},$$

we have

$$C_{\tau^*,d}^{\tau^*} \le \frac{1}{d^2}\sum_{\mathbf{k}\in\mathbb{Z}^d}\lambda_{d,\mathbf{k},\Gamma(\text{UNW}),0,0}^{\tau^*(1+\ln d)} \le \frac{1}{d^2}\sum_{k_1\in\mathbb{Z}}\sum_{k_2\in\mathbb{Z}}\cdots\sum_{k_d\in\mathbb{Z}}\left(\prod_{j=1}^{d}\frac{1}{1+k_j^2}\right)^{\tau^*(1+\ln d)}$$
$$= \frac{1}{d^2}\left[\sum_{k=-\infty}^{\infty}\left(\frac{1}{1+k^2}\right)^{\tau^*(1+\ln d)}\right]^d. \tag{38}$$

Since $d \ge 2$ and $\tau^* = 1/\ln 2$, we have

$$\begin{aligned}\sum_{k=-\infty}^{\infty}\left(\frac{1}{1+k^2}\right)^{\tau^*(1+\ln d)} &= \sum_{k=-\infty}^{\infty}(d\mathrm{e})^{-\ln(1+k^2)/\ln 2}\\ &= \frac{1}{d\mathrm{e}}\sum_{k=-\infty}^{\infty}(d\mathrm{e})^{-\ln[(1+k^2)/2]/\ln 2}\\ &\le \frac{1}{d\mathrm{e}}\sum_{k=-\infty}^{\infty}(2\mathrm{e})^{-\ln[(1+k^2)/2]/\ln 2}\\ &= \frac{1}{d\mathrm{e}}\sum_{k=-\infty}^{\infty}\left(\frac{2}{1+k^2}\right)^{1+1/\ln 2}.\end{aligned} \tag{39}$$

Since the terms in the series

$$c_2 := \sum_{k=-\infty}^{\infty}\left(\frac{2}{1+k^2}\right)^{1+1/\ln 2} \tag{40}$$

are $\Theta(j^{-2(1+1/\ln 2)})$ and $2(1+1/\ln 2) > 1$, the series converges; again using Mathematica, we find that $c_2 \doteq 7.70707$. Combining (38)–(40), we find that

$$\sup_{d\ge 2} C_{\tau^*,d}^{\tau^*} \le \sup_{d\ge 2}\frac{1}{d^2}\left(\frac{c_2}{d\mathrm{e}}\right)^d = \frac{1}{4}\left(\frac{c_2}{2\mathrm{e}}\right)^2 =: c_3,$$

where $c_3 \doteq 0.502423$, which is finite, completing the proof for the case $d \ge 2$. Combining the results for $d = 1$ and $d \ge 2$, we see that

$$C_{\tau^*} = \sup_{d\ge 2} C_{\tau^*,d} = \max\{c_1, c_3\}^{1/\tau^*} \doteq 1.67089,$$

as needed to prove the theorem. □

Remark 7 Note that the exponent of quasi-polynomial tractability is $2/\ln 2$, independent of the values of r and s. □

4.3 Product Weights

In this section, we look at *product weights* $\Gamma(\Pi)$, which are defined by (2), subject to the condition (3) on the weightlets. As was the case for the space studied in [11, 12], we find that $H^{0,1}_{\Gamma(\Pi)}(\mathbb{T}^d) = \left[H^{0,1}_{\Gamma(\Pi)}(\mathbb{T})\right]^{\otimes d}$ has a tensor product structure for product weights, with

$$\alpha_{d,\mathbf{k},\Gamma} = \prod_{j=1}^{d} \frac{\gamma_{d,j}}{\gamma_{d,j} + k_j^2} \qquad \forall\, \mathbf{k} \in \mathbb{Z}^d.$$

In what follows, we shall assume that the weightlets $\gamma_{d,j}$ are uniformly bounded, i.e., that there exists $M > 0$ such that

$$\gamma_{d,j} \le M \qquad \forall\, j \in [d], d \in \mathbb{N}. \tag{41}$$

Remark 8 What happens if (41) does not hold? If we allow weightlets that are not uniformly bounded, then $\mathrm{App}_{\Gamma,r,s}$ can suffer from the curse of dimensionality. One such instance is given by choosing $\gamma_{d,j} \equiv d$ for all $j \in [d]$ and $d \in \mathbb{N}$. For a given $d \in \mathbb{N}$, let

$$\varepsilon_d = \frac{1}{2\sqrt{(s-r)!(1+d)^{s-r}(1+d^{-1})^d}} \sim \frac{1}{2\sqrt{(s-r)!\,\varepsilon\, d^{s-r}}} \qquad \text{as } d \to \infty.$$

Following the approach in [12, Section 5.2], we can show that $\lambda_{d,\mathbf{k},\Gamma,r,s} > \varepsilon_d^2$ for any $\mathbf{k} \in \{0,1\}^d$. Since $|\{0,1\}^d| = 2^d$, it follows that $n(\varepsilon_d, \mathrm{App}_{\Gamma,r,s}) \ge 2^d$. □

4.3.1 Quasi-Polynomial Tractability

We claim that our approximation problem is always quasi-polynomially tractable for bounded product weights. Indeed, let Π_M denote product weights for which $\gamma_{d,j} \equiv M$. Then $\mathrm{App}_{\Gamma(\Pi),r,s}$ is no harder than $\mathrm{App}_{\Gamma(\Pi_M),r,s}$, since

$$\alpha_{d,\mathbf{k},\Gamma(\Pi)} \le \alpha_{d,\mathbf{k},\Gamma(\Pi_M)} = \prod_{j=1}^{d} \frac{1}{1 + k_j^2/M}.$$

It is now easy to see that $\mathrm{App}_{\Gamma(\Pi_M),r,s}$ is quasi-polynomially tractable, whence $\mathrm{App}_{\Gamma(\Pi)}$ is also quasi-polynomially tractable. The exponents of quasi-polynomial tractability satisfy

$$t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma(\Pi),r,s}) \le t^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma(\Pi_M),r,s}) = \frac{2}{\ln(1+M^{-1})}.$$

Moreover, the bound in this inequality is sharp, being attained by choosing equal weightlets $\Pi = \Pi_M$. To see why this is so, simply reiterate the proof of Theorem 5, replacing k_j^2 by k_j^2/M and $\frac{1}{2}$ by $1 + 1/M$ and making sure to use the upper bound $\eta_{d,\mathbf{k}} \leq 1 + d^2/M$.

4.3.2 Polynomial and Strong Polynomial Tractability

From Theorem 4, we see that since our weights are bounded, our approximation problem $\mathrm{App}_{\Gamma,r,s}$ is (strongly) polynomially tractable iff the same is true for the approximation problem $\mathrm{App}_{\Gamma,0,0}$. We now look at (strong) polynomial tractability in more detail:

Theorem 6 *We have the following results for bounded product weights.*

1. $\mathrm{App}_{\Gamma,r,s}$ *is strongly polynomially tractable iff there exists* $\tau > \frac{1}{2}$ *such that* $A_\tau < \infty$*, where*

$$A_\tau = \sup_{d\in\mathbb{N}} \sum_{j=1}^{d} \gamma_{d,j}^{\tau}.$$

   a. *The exponent of strong polynomial tractability satisfies the inequality*

$$p(\mathrm{App}_{\Gamma,r,s}) \in \left[\max\left\{ 1, \frac{1}{s-r+1} p(\mathrm{App}_{\Gamma,0,0}) \right\}, p(\mathrm{App}_{\Gamma,0,0}) \right].$$

   Hence when $p^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0}) = 1$*, we have*

$$p^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,r,s}) = p^{\mathrm{qpoly}}(\mathrm{App}_{\Gamma,0,0}).$$

   b. *Let*

$$p(\mathrm{App}_{\Gamma,0,0}) = 2\tau^*,$$

   where

$$\tau^* = \inf\{\, \tau > \tfrac{1}{2} : A_\tau < \infty \,\} \geq \tfrac{1}{2}.$$

   Then for all $\tau > \tau^*$*, we have*

$$n(\varepsilon, \mathrm{App}_{d,\Gamma,r,s}) \leq n(\varepsilon, \mathrm{App}_{d,\Gamma,0,0}) \leq \varepsilon^{-2\tau} \exp\left(2\,\zeta(2\tau)\pi^{-2\tau} A_\tau\right),$$

where

$$\zeta(s) = \sum_{j=1}^{\infty} \frac{1}{j^s}$$

denotes the Riemann zeta function.

2. $\mathrm{App}_{\Gamma,r,s}$ *is polynomially tractable iff there exists* $\tau > \frac{1}{2}$ *such that* $B_\tau < \infty$*, where*

$$B_\tau = \limsup_{d\to\infty} \frac{1}{\ln\, d} \sum_{j=1}^{d} \gamma_{d,j}^{\tau}.$$

When this holds, then for any $q_\tau > B_\tau$ *there exists a positive* C_τ *such that*

$$n(\varepsilon, S_d) \leq C_\tau\, d^{q_\tau}\, \varepsilon^{-2\tau} \qquad \forall\, \varepsilon \in (0,1),\ d \in \mathbb{N}.$$

3. *For product weights independent of d, i.e., such that* $\gamma_{d,j} \equiv \gamma_j$ *for all* $d \in \mathbb{N}$*, strong polynomial tractability and polynomial tractability for* $\mathrm{App}_{\Gamma,r,s}$ *are equivalent.*

Proof Follow the proof of [12, Thm. 5.3]. Take account of the following changes:

1. The factor π^2 in [12, Thm. 5.3] does not appear.
2. The expression $(k_j - 1)^2$ in [12, Thm. 5.3] becomes k_j^2.
3. Sums are over $\mathbb{Z}^d$ or $\mathbb{Z}$, rather than over $\mathbb{N}_0^d$ or $\mathbb{N}_0$. □

4.4 Bounded Finite-Order and Finite-Diameter Weights

As seen in Remark 3, if we allow unbounded weights, then we can run into situations in which $\mathrm{App}_{\Gamma,r,s}$ is strongly polynomially tractable, but $\mathrm{App}_{\Gamma,0,0}$ suffers from the curse of dimensionality. So we're only interested in *bounded* finite-order and finite-diameter weights, so that there exists $M > 0$ such that

$$M := \sup_{d\in\mathbb{N}} \sup_{\mathrm{u}\subseteq[d]} \gamma_{d,\mathrm{u}} < \infty.$$

Now Theorem 3 tells us that our problem $\mathrm{App}_{d,\Gamma,r,s}$ is no harder than the problem $\mathrm{App}_{d,\Gamma,0,0}$. So we may follow the approach in the proof of [12, Theorem 5.4], which relies on [11, Theorem 4.1], to see that for any $\tau > 0$, there exist $C_{\tau,\omega} > 0$ such that

$$n(\varepsilon, \mathrm{App}_{\Gamma,r,s}) \leq C_{\tau,\omega} M^{\tau/2} d^{\omega} \varepsilon^{-\tau}. \tag{42}$$

Thus $\mathrm{App}_{\Gamma,r,s}$ is always polynomially tractable for finite-order weights. Finally, since finite-diameter weights are a special case of finite-order weights of order 1, we may substitute $\omega = 1$ in (42) to get a polynomially-tractable upper bound for $\mathrm{App}_{\Gamma,r,s}$ with finite-diameter weights.

Acknowledgements I am happy to thank Erich Novak and Henryk Woźniakowski for their helpful and insightful remarks. Moreover, the referees made suggestions that improved the paper, for which I also extend my thanks.

References

1. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton, NJ (1957)
2. Borthwick, D.: Introduction to Partial Differential Equations. Universitext. Springer International Publishing, Cham (2017)
3. Dahlke, S., Novak, E., Sickel, W.: Optimal approximation of elliptic problems by linear and nonlinear mappings. I. J. Complex. **22**(1), 29–49 (2006)
4. Griebel, M., Knapek, S.: Optimized tensor-product approximation spaces. Constr. Approx. **16**(4), 525–540 (2000)
5. Kühn, T., Sickel, W., Ullrich, T.: Approximation numbers of Sobolev embeddings—sharp constants and tractability. J. Complex. **30**(2), 95–116 (2014)
6. Kühn, T., Sickel, W., Ullrich, T.: Approximation of mixed order Sobolev functions on the d-torus: asymptotics, preasymptotics, and d-dependence. Constr. Approx. **42**(3), 353–398 (2015)
7. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume I: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society (EMS), Zürich (2008)
8. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume II: Standard Information For Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
9. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume III: Standard Information For Operators. EMS Tracts in Mathematics, vol. 18. European Mathematical Society (EMS), Zürich (2012)
10. Siedlecki, P., Weimar, M.: Notes on (s, t)-weak tractability: a refined classification of problems with (sub)exponential information complexity. J. Approx. Theory **200**, 227–258 (2015)
11. Werschulz, A.G., Woźniakowski, H.: Tractability of multivariate approximation over a weighted unanchored Sobolev space. Constr. Approx. **30**(3), 395–421 (2009)
12. Werschulz, A.G., Woźniakowski, H.: Tight tractability results for a model second-order Neumann problem. Found. Comput. Math. **15**(4), 899–929 (2015)
13. Werschulz, A.G., Woźniakowski, H.: A new characterization of (s, t)-weak intractability. J. Complex. **38**, 68–79 (2017)

Efficient Spherical Designs with Good Geometric Properties

Robert S. Womersley

Dedicated to Ian H. Sloan on the occasion of his 80th birthday in acknowledgement of his many fruitful ideas and generosity.

Abstract Spherical t-designs on $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ provide N nodes for an equal weight numerical integration rule which is exact for all spherical polynomials of degree at most t. This paper considers the generation of efficient, where N is comparable to $(1+t)^d/d$, spherical t-designs with good geometric properties as measured by their mesh ratio, the ratio of the covering radius to the packing radius. Results for $\mathbb{S}^2$ include computed spherical t-designs for $t = 1, \ldots, 180$ and symmetric (antipodal) t-designs for degrees up to 325, all with low mesh ratios. These point sets provide excellent points for numerical integration on the sphere. The methods can also be used to computationally explore spherical t-designs for $d = 3$ and higher.

1 Introduction

Consider the d-dimensional unit sphere

$$\mathbb{S}^d = \left\{ \boldsymbol{x} \in \mathbb{R}^{d+1} : |\boldsymbol{x}| = 1 \right\}$$

where the standard Euclidean inner product is $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^{d+1} x_i y_i$ and $|\boldsymbol{x}|^2 = \boldsymbol{x} \cdot \boldsymbol{x}$.

A numerical integration (quadrature) rule for $\mathbb{S}^d$ is a set of N points $\boldsymbol{x}_j \in \mathbb{S}^d, j = 1, \ldots, N$ and associated weights $w_j > 0, j = 1, \ldots, N$ such that

$$Q_N(f) := \sum_{j=1}^{N} w_j f(\boldsymbol{x}_j) \approx I(f) := \int_{\mathbb{S}^d} f(\boldsymbol{x}) d\sigma_d(\boldsymbol{x}). \tag{1}$$

R. S. Womersley (✉)
School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: r.womersley@unsw.edu.au

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_57

Here $\sigma_d(\boldsymbol{x})$ is the normalised Lebesgue measure on $\mathbb{S}^d$ with surface area

$$\omega_d := \frac{2\pi^{(d+1)/2}}{\Gamma((d+1)/2)},$$

where $\Gamma(\cdot)$ is the gamma function.

Let $\mathbb{P}_t(\mathbb{S}^d)$ denote the set of all spherical polynomials on $\mathbb{S}^d$ of degree at most t. A *spherical t-design* is a set of N points $X_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ on $\mathbb{S}^d$ such that equal weight quadrature using these nodes is exact for all spherical polynomials of degree at most t, that is

$$\frac{1}{N}\sum_{j=1}^{N} p(\boldsymbol{x}_j) = \int_{\mathbb{S}^d} p(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}), \quad \forall p \in \mathbb{P}_t(\mathbb{S}^d). \tag{2}$$

Spherical t-designs were introduced by Delsarte, Goethals and Seidel [24] who provided several characterizations and established lower bounds on the number of points N required for a spherical t-design. Seymour and Zaslavsky[55] showed that spherical t-designs exist on $\mathbb{S}^d$ for all N sufficiently large. Bondarenko, Radchenko and Viazovska [8] established that there exists a C_d such that spherical t-designs on $\mathbb{S}^d$ exist for all $N \geq C_d\, t^d$, which is the optimal order. The papers [5, 19, 21] provide a sample of many on spherical designs and algebraic combinatorics on spheres.

An alternative approach, not investigated in this paper, is to relax the condition $w_j = 1/N$ that the quadrature weights are equal so that $|w_j/(1/N) - 1| \leq \epsilon$ for $j = 1, \ldots, N$ and $0 \leq \epsilon < 1$, but keeping the condition that the quadrature rule is exact for polynomials of degree t (see [57, 69] for example).

The aim of this paper is not to find spherical t-designs with the minimal number of points, nor to provide proofs that a particular configuration is a spherical t-design. Rather the aim is to find sequences of point sets which are at least computationally spherical t-designs, have a low number of points and are geometrically well-distributed on the sphere. Such point sets provide excellent nodes for numerical integration on the sphere, as well as hyperinterpolation [39, 56, 59] and fully discrete needlet approximation [65]. These methods have a requirement that the quadrature rules are exact for certain degree polynomials. More generally, [40] provides a summary of numerical integration on $\mathbb{S}^2$ with geomathematical applications in mind.

1.1 Spherical Harmonics and Jacobi Polynomials

A *spherical harmonic* of degree ℓ on $\mathbb{S}^d$ is the restriction to $\mathbb{S}^d$ of a homogeneous and harmonic polynomial of total degree ℓ defined on $\mathbb{R}^{d+1}$. Let $\mathbb{H}_\ell$ denote the set of all spherical harmonics of exact degree ℓ on $\mathbb{S}^d$. The dimension of the linear space $\mathbb{H}_\ell$ is

$$Z(d, \ell) := (2\ell + d - 1)\frac{\Gamma(\ell + d - 1)}{\Gamma(d)\Gamma(\ell + 1)} \asymp (\ell + 1)^{d-1}, \tag{3}$$

where $a_\ell \asymp b_\ell$ means $c\, b_\ell \le a_\ell \le c'\, b_\ell$ for some positive constants c, c', and the asymptotic estimate uses [47, Eq. 5.11.12].

Each pair $\mathbb{H}_\ell$, $\mathbb{H}_{\ell'}$ for $\ell \neq \ell' \geq 0$ is $\mathbb{L}_2$-orthogonal, $\mathbb{P}_L(\mathbb{S}^d) = \bigoplus_{\ell=0}^{L} \mathbb{H}_\ell$ and the infinite direct sum $\bigoplus_{\ell=0}^{\infty} \mathbb{H}_\ell$ is dense in $\mathbb{L}_p(\mathbb{S}^d)$, $p \geq 2$, see e.g. [64, Ch.1]. The linear span of $\mathbb{H}_\ell$, $\ell = 0, 1, \ldots, L$, forms the space $\mathbb{P}_L(\mathbb{S}^d)$ of spherical polynomials of degree at most L. The dimension of $\mathbb{P}_L(\mathbb{S}^d)$ is

$$D(d, L) := \dim \mathbb{P}_L(\mathbb{S}^d) = \sum_{\ell=0}^{L} Z(d, \ell) = Z(d+1, L). \tag{4}$$

Let $P_\ell^{(\alpha,\beta)}(z)$, $-1 \leq z \leq 1$, be the Jacobi polynomial of degree ℓ for $\alpha, \beta > -1$. The Jacobi polynomials form an orthogonal polynomial system with respect to the Jacobi weight $w_{\alpha,\beta}(z) := (1-z)^\alpha (1+z)^\beta$, $-1 \leq z \leq 1$. We denote the normalised Legendre (or ultraspherical/Gegenbauer) polynomials by

$$P_\ell^{(d+1)}(z) := \frac{P_\ell^{\left(\frac{d-2}{2}, \frac{d-2}{2}\right)}(z)}{P_\ell^{\left(\frac{d-2}{2}, \frac{d-2}{2}\right)}(1)},$$

where, from [61, (4.1.1)],

$$P_\ell^{(\alpha,\beta)}(1) = \frac{\Gamma(\ell+\alpha+1)}{\Gamma(\ell+1)\Gamma(\alpha+1)}, \tag{5}$$

and [61, Theorem 7.32.2, p. 168],

$$\left|P_\ell^{(d+1)}(z)\right| \leq 1, \qquad -1 \leq z \leq 1. \tag{6}$$

The derivative of the Jacobi polynomial satisfies [61]

$$\frac{\mathrm{d}\, P_\ell^{(\alpha,\beta)}(z)}{\mathrm{d}\, z} = \frac{\ell+\alpha+\beta+1}{2}\, P_{\ell-1}^{(\alpha+1,\beta+1)}(z), \tag{7}$$

so

$$\frac{\mathrm{d}\, P_\ell^{(d+1)}(z)}{\mathrm{d}\, z} = \frac{(\ell+d-1)(\ell+d/2)}{d} P_{\ell-1}^{(d+3)}(z). \tag{8}$$

Also if ℓ is odd then the polynomials $P_\ell^{(d+1)}$ are odd and if ℓ is even the polynomials $P_\ell^{(d+1)}$ are even.

A *zonal function* $K : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$ depends only on the inner product of the arguments, i.e. $K(\boldsymbol{x}, \boldsymbol{y}) = \mathfrak{K}(\boldsymbol{x} \cdot \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^d$, for some function $\mathfrak{K} : [-1, 1] \to \mathbb{R}$. Frequent use is made of the zonal function $P_\ell^{(d+1)}(\boldsymbol{x} \cdot \boldsymbol{y})$.

Let $\{Y_{\ell,k} : k = 1, \ldots, Z(d,\ell),\ \ \ell = 0, \ldots, L\}$ be an *orthonormal basis* for $\mathbb{P}_L(\mathbb{S}^d)$. The normalised Legendre polynomial $P_\ell^{(d+1)}(\boldsymbol{x} \cdot \boldsymbol{y})$ satisfies the *addition theorem* (see [3, 61, 64] for example)

$$\sum_{k=1}^{Z(d,\ell)} Y_{\ell,k}(\boldsymbol{x}) Y_{\ell,k}(\boldsymbol{y}) = Z(d,\ell) P_\ell^{(d+1)}(\boldsymbol{x} \cdot \boldsymbol{y}). \tag{9}$$

1.2 Number of Points

Delsarte, Goethals and Seidel [24] showed that an N point t-design on $\mathbb{S}^d$ has $N \geq N^*(d,t)$ where

$$N^*(d,t) := \begin{cases} 2\binom{d+k}{d} & \text{if } t = 2k+1, \\ \binom{d+k}{d} + \binom{d+k-1}{d} & \text{if } t = 2k. \end{cases} \tag{10}$$

On $\mathbb{S}^2$

$$N^*(2,t) := \begin{cases} \frac{(t+1)(t+3)}{4} & \text{if } t \text{ odd}, \\ \frac{(t+2)^2}{4} & \text{if } t \text{ even}. \end{cases} \tag{11}$$

Bannai and Damerell [6, 7] showed that *tight spherical t-designs* which achieve the lower bounds (10) cannot exist except for a few special cases (for example except for $t = 1, 2, 3, 5$ on $\mathbb{S}^2$).

Yudin [68] improved (except for some small values of d, t, see Table 2), the lower bounds (10), by an exponential factor $(4/e)^{d+1}$ as $t \to \infty$, so $N \geq N^+(d,t)$ where

$$N^+(d,t) := 2\frac{\int_0^1 (1-z^2)^{(d-2)/2}\,\mathrm{d}z}{\int_\gamma^1 (1-z^2)^{(d-2)/2}\,\mathrm{d}z} = \frac{\sqrt{\pi}\,\Gamma(d/2)/\Gamma((d+1)/2)}{\int_\gamma^1 (1-z^2)^{(d-2)/2}\,\mathrm{d}z}, \tag{12}$$

and γ is the largest zero of the derivative $\frac{dP_t^{(d+1)}(z)}{dz}$ and hence the largest zero of $P_{t-1}^{(\alpha+1,\alpha+1)}(z)$ where $\alpha = (d-2)/2$. Bounds [2, 61] on the largest zero of $P_n^{(\alpha,\alpha)}(z)$ are

$$\begin{aligned} \cos\left(\frac{j_0(\nu)}{n+\alpha+1/2}\right) &\leq \gamma \\ &\leq \sqrt{\frac{(n-1)(n+2\alpha-1)}{(n+\alpha-3/2)/(n+\alpha-1/2)}} \cos\left(\frac{\pi}{n+1}\right), \end{aligned} \tag{13}$$

where $j_0(\nu)$ is the first positive zero of the Bessel function $J_\nu(x)$.

Numerically there is strong evidence that spherical t-designs with $N = D(2,t) = (t+1)^2$ points exist, [17] and [18] used interval methods to *prove* existence of spherical t-designs with $N = (t+1)^2$ for all values of t up to 100, but there is no proof yet that spherical t-designs with $N \leq D(2,t)$ points exist for all degrees t. Hardin and Sloane [33, 34] provide tables of designs with modest numbers of points, exploiting icosahedral symmetry. They conjecture that for $d = 2$ spherical t-designs exist with $N = t^2/2 + o(t^2)$ for all t. The numerical experiments reported here and available from [66] strongly support this conjecture.

McLaren [45] defined efficiency E for a quadrature rule as the ratio of the number of independent functions for which the rule is exact to the number of arbitrary constants in the rule. For a spherical t-design with N points on $\mathbb{S}^d$ (and equal weights)

$$E = \frac{\dim \mathbb{P}_t(\mathbb{S}^d)}{dN} = \frac{D(d,t)}{dN}. \tag{14}$$

In these terms the aim is to find spherical t-designs with $E \geq 1$. McLaren [45] exploits symmetry (in particular octahedral and icosahedral) to seek rules with optimal efficiency. The aim here is not to maximise efficiency by finding the minimal number of points for a t-design on $\mathbb{S}^d$, but rather a sequence of *efficient* t-designs with $N \asymp \frac{D(d,t)}{d} \asymp \frac{(1+t)^d}{d}$. Such efficient t-designs provide a practical tool for numerical integration and approximation.

1.3 Geometric Quality

The Geodesic distance between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^d$ is

$$\operatorname{dist}(\boldsymbol{x},\boldsymbol{y}) = \cos^{-1}(\boldsymbol{x}\cdot\boldsymbol{y}),$$

while the Euclidean distance is

$$|\boldsymbol{x}-\boldsymbol{y}| = \sqrt{2(1-\boldsymbol{x}\cdot\boldsymbol{y})} = 2\sin(\operatorname{dist}(\boldsymbol{x},\boldsymbol{y})/2).$$

The *spherical cap* with centre $\boldsymbol{z} \in \mathbb{S}^d$ and radius $\eta \in [0,\pi]$ is

$$\mathscr{C}(\boldsymbol{z};\eta) = \left\{\boldsymbol{x} \in \mathbb{S}^d : \operatorname{dist}(\boldsymbol{x},\boldsymbol{z}) \leq \eta\right\}.$$

The *separation distance*

$$\delta(X_N) = \min_{i \neq j} \operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

is twice the packing radius for spherical caps of the same radius and centers in X_N. The best packing problem (or Tammes problem) has a long history [21], starting with [53, 62]. A sequence of point sets $\{X_N\}$ with $N \to \infty$ has the optimal order separation if there exists a constant c_d^{pck} independent of N such that

$$\delta(X_N) \geq c_d^{\mathrm{pck}}\, N^{-1/d}.$$

The separation, and all the zonal functions considered in subsequent sections, are determined by the set of inner products

$$\mathscr{A}(X_N) := \left\{\boldsymbol{x}_i \cdot \boldsymbol{x}_j, i = 1, \ldots, N, j = i+1, \ldots, N\right\} \tag{15}$$

which has been widely used in the study of spherical codes, see [21] for example. Then

$$\max_{z \in \mathscr{A}(X_N)} z = \cos(\delta(X_N)).$$

Point sets are only considered different if the corresponding sets (15) differ, as they are invariant under an orthogonal transformation (rotation) of the point set and permutation (relabelling) of the points.

The *mesh norm* (or fill radius)

$$h(X_N) = \max_{\boldsymbol{x} \in \mathbb{S}^d} \min_{j=1,\ldots,N} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}_j)$$

gives the *covering radius* for covering the sphere with spherical caps of the same radius and centers in X_N. A sequence of point sets $\{X_N\}$ with $N \to \infty$ has the optimal order covering if there exists a constant c_d^{cov} independent of N such that

$$h(X_N) \leq c_d^{\mathrm{cov}}\, N^{-1/d}.$$

The *mesh ratio* is

$$\rho(X_N) = \frac{2h_{X_N}}{\delta_{X_N}} \geq 1.$$

A common assumption in numerical methods is that the mesh ratio is uniformly bounded, that is the point sets are *quasi-uniform*. Minimal Riesz s-energy and best packing points can also produce quasi-uniform point sets [9, 23, 35].

Yudin [67] showed that a spherical t-design with N points has a covering radius of the optimal order $1/t$. Reimer extended this to quadrature rules exact for polynomials of degree t with positive weights. Thus a spherical t-design with $N = O(t^d)$ points provides an optimal order covering.

The union of two spherical t-designs with N points is a spherical t-design with $2N$ points. A spherical design with arbitrarily small separation can be obtained as one N point set is rotated relative to the other. Thus an assumption on the separation of the points of a spherical design is used to derive results, see [37] for example. This simple argument is not possible if N is less than twice a lower bound (10) or (12) on the number of points in a spherical t-design.

Bondarenko, Radchenko and Viazovska [10] have shown that on $\mathbb{S}^d$ well-separated spherical t-designs exist for $N \geq c'_d\, t^d$. This combined with Yudin's result on the covering radius of spherical designs mean that there exist spherical t-designs with $N = O(t^d)$ points and uniformly bounded mesh ratio.

There are many other "geometric" properties that could be used, for example the spherical cap discrepancy, see [31] for example, (using normalised surface measure so $|\mathbb{S}^d| = 1$)

$$\sup_{\boldsymbol{x}\in\mathbb{S}^d,\eta\in[0,\pi]} \left| |\mathscr{C}(\boldsymbol{x},\eta)| - \frac{|X_N \cap \mathscr{C}(\boldsymbol{x},\eta)|}{N} \right|,$$

or a Riesz s-energy, see [12] for example,

$$E_s(X_N) = \sum_{1\leq i<j\leq N} \frac{1}{|\boldsymbol{x}_i - \boldsymbol{x}_j|^s}.$$

In distinguishing between spherical t-designs with the same number N of points we prefer those with lower mesh ratio. Note that some authors, see [9, 35] for example, define the mesh ratio as $\tilde{\rho}(X_N) = h(X_N)/\delta(X_N) \geq 1/2$.

2 Variational Characterizations

Delsarte, Goethals and Seidel [24] showed that $X_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{S}^d$ is a spherical t-design if and only if the Weyl sums satisfy

$$r_{\ell,k}(X_N) := \sum_{j=1}^{N} Y_{\ell,k}(\boldsymbol{x}_j) = 0 \qquad k = 1, \ldots, Z(d,\ell), \quad \ell = 1, \ldots, t, \tag{16}$$

as the integral of all spherical harmonics of degree $\ell \geq 1$ is zero from orthogonality with the constant ($\ell = 0$) polynomial $Y_{0,1} = 1$ which is not included.

In matrix form

$$\boldsymbol{r}(X_N) := \overline{\boldsymbol{Y}}\boldsymbol{e} = \boldsymbol{0}$$

where $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathbb{R}^N$ and $\overline{\boldsymbol{Y}} \in \mathbb{R}^{D(d,t)-1\times N}$ is the spherical harmonic basis matrix excluding the first row.

Let $\psi_t : [-1,1] \to \mathbb{R}$ be a polynomial of degree $t \geq 1$ with

$$\psi_t(z) = \sum_{\ell=1}^{t} a_\ell P_\ell^{(d+1)}(z), \qquad a_\ell > 0 \text{ for } \ell = 1, \ldots, t, \tag{17}$$

so the generalised Legendre coefficients a_ℓ for degrees $\ell = 1, \ldots, t$ are all strictly positive. Clearly any such function ψ_t can be scaled by an arbitrary positive constant without changing these properties.

Consider now an arbitrary set X_N of N points on $\mathbb{S}^d$. Sloan and Womersley [58] considered the variational form

$$V_{t,N,\psi}(X_N) := \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \psi_t(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

which from (6) satisfies

$$0 \leq V_{t,N,\psi}(X_N) \leq \sum_{\ell=1}^{t} a_\ell = \psi_t(1).$$

Moreover the average value is

$$\overline{V}_{t,N,\psi} := \int_{\mathbb{S}^d} \cdots \int_{\mathbb{S}^d} V_{t,N,\psi}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) d\sigma_d(\boldsymbol{x}_1) \cdots d\sigma_d(\boldsymbol{x}_N) = \frac{\psi_t(1)}{N}.$$

As the upper bound and average of $V_{t,N,\psi}(X_N)$ depend on $\psi_t(1)$, we concentrate on functions ψ for which $\psi_t(1)$ does not grow rapidly with t.

From the addition theorem (9), $V_{t,N,\psi}(X_N)$ is a weighted sum of squares with strictly positive coefficients

$$V_{t,N,\psi}(X_N) = \frac{1}{N^2} \sum_{\ell=1}^{t} \frac{a_\ell}{Z(d,\ell)} \sum_{k=1}^{Z(d,\ell)} (r_{\ell,k}(X_N))^2 = \frac{1}{N^2} \boldsymbol{r}(X_N)^T \boldsymbol{D}\, \boldsymbol{r}(X_N), \tag{18}$$

where $\boldsymbol{D}$ is the diagonal matrix with strictly positive diagonal elements $\frac{a_\ell}{Z(d,\ell)}$ for $k = 1, \ldots, Z(d,\ell), \ell = 1, \ldots, t$. Thus, from (16), X_N is a spherical t-design if and only if

$$V_{t,N,\psi}(X_N) = 0.$$

Moreover, if the *global* minimum of $V_{t,N,\psi}(X_N) > 0$ then there are no spherical t-designs on $\mathbb{S}^d$ with N points.

Given a polynomial $\widehat{\psi}_t(z)$ of degree t and strictly positive Legendre coefficients, the zero order term may need to be removed to get $\psi_t(z) = \widehat{\psi}_t(z) - a_0$ where for $\mathbb{S}^d$

and $\alpha = (d-2)/2$,

$$a_0 = \int_{-1}^{1} \widehat{\psi}_t(z) \left(1 - z^2\right)^{\alpha} dz.$$

Three examples of polynomials on $[-1, 1]$ with strictly positive Legendre coefficients for $\mathbb{S}^d$ and zero constant term, with $\alpha = (d-2)/2$ are:

Example 1

$$\psi_{1,t}(z) = z^{t-1} + z^t - a_0 \tag{19}$$

where

$$a_0 = \frac{\Gamma(\alpha + 3/2)}{\sqrt{\pi}} \begin{cases} \frac{\Gamma(t/2)}{\Gamma(\alpha+1+t/2)} & t \text{ odd,} \\ \frac{\Gamma((t+1)/2)}{\Gamma(\alpha+3/2+t/2)} & t \text{ even.} \end{cases} \tag{20}$$

For $d = 2$ this simplifies to $a_0 = 1/t$ if t is odd and $a_0 = 1/(t+1)$ if t is even. This function was used by Grabner and Tichy [31] for symmetric point sets where only even values of t need to be considered, as all odd degree polynomials are integrated exactly.

Example 2

$$\psi_{2,t}(z) = \left(\frac{1+z}{2}\right)^t - a_0 \tag{21}$$

where

$$a_0 = \frac{2}{\sqrt{\pi}} 4^{\alpha} \Gamma(\alpha + 3/2) \frac{\Gamma(\alpha + 1 + t)}{\Gamma(2\alpha + 2 + t)}. \tag{22}$$

For $d = 2$ this simplifies to $a_0 = 1/(1+t)$. This is a scaled version of the function $(1+z)^t$ used by Cohn and Kumar [19] for which a_0 must be scaled by 2^t producing more cancellation errors for large t.

Example 3

$$\psi_{3,t}(z) = P_t^{(\alpha+1,\alpha)}(z) - a_0 \tag{23}$$

where a_0 is given by (22). The expansion in terms of Jacobi polynomials in Szegő [61, Section 4.5] gives

$$\sum_{\ell=0}^{t} Z(d, \ell) P_\ell^{(d+1)}(z) = \frac{1}{a_0} P_t^{(\alpha+1,\alpha)}(z).$$

For S^2 this is equivalent to

$$\sum_{\ell=1}^{t}(2\ell+1)P_{\ell}^{(d+1)}(z) = (t+1)P_t^{(1,0)}(z) - 1$$

used in Sloan and Womersley [58].

3 Quadrature Error

The error for numerical integration depends on the smoothness of the integrand. Classical results are based on the error of best approximation of the integrand f by polynomials [51], (see also [40] for more details on $\mathbb{S}^2$). For $f \in C^{\kappa}(\mathbb{S}^d)$, there exists a constant $c = c(\kappa, f)$ such that the numerical integration error satisfies

$$\left| \int_{\mathbb{S}^d} f(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) - \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}_j) \right| \leq c\, t^{-\kappa}.$$

If $N = O(t^d)$ then the right-hand-side becomes $N^{-\kappa/d}$. Thus for functions with reasonable smoothness it pays to increase the degree of precision t.

Similar results are presented in [13], building on the work of [36, 38], for functions f in a Sobolev space $\mathbb{H}^s(\mathbb{S}^d)$, $s > d/2$. The *worst-case-error* for equal weight (quasi Monte-Carlo) numerical integration using an arbitrary point set X_N is

$$WCE(X_N, s, d) := \sup_{f \in \mathbb{H}^s(\mathbb{S}^d), ||f||_{\mathbb{H}^s(\mathbb{S}^d)} \leq 1} \left| \int_{\mathbb{S}^d} f(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) - \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}_j) \right|. \qquad (24)$$

From this it immediately follows that the error for numerical integration satisfies

$$\left| \int_{\mathbb{S}^d} f(\boldsymbol{x}) \mathrm{d}\sigma_d(\boldsymbol{x}) - \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}_j) \right| \leq WCE(X_N, s, d)\ \|f\|_{\mathbb{H}^s(\mathbb{S}^d)}.$$

Spherical t-designs X_N with $N = O(t^d)$ points satisfy the optimal order rate of decay of the worst case error, for any $s > d/2$, namely

$$WCE(X_N, s, d) = O\left(N^{-s/d}\right), \qquad N \to \infty.$$

Thus spherical t-designs with $N = O(t^d)$ points are ideally suited to the numerical integration of smooth functions.

4 Computational Issues

The aim is to find a spherical t-design with N points on $\mathbb{S}^d$ by finding a point set X_N achieving the global minimum of zero for the variational function $V_{t,N,\psi}(X_N)$. This section considers several computational issues: the evaluation of $V_{t,N,\psi}(X_N)$ either as a double sum or using its representation (18) as a sum of squares; the parametrisation of the point set X_N; the number of points N as a function of t and d; the choice of optimization algorithm which requires evaluation of derivatives with respect to the chosen parameters; exploiting the sum of squares structure which requires evaluating the spherical harmonics and their derivatives; and imposing structure on the point set, for example symmetric (antipodal) point sets. An underlying issue is that optimization problems with points on the sphere typically have many different local minima with different characteristics. Here we are seeking both a global minimizer with value 0 and one with good geometric properties as measured by the mesh ratio.

The calculations were performed using Matlab, on a Linux computational cluster using nodes with up to 16 cores. In all cases analytic expressions for the derivatives with respect to the chosen parametrisation were used.

4.1 Evaluating Criteria

Although the variational functions are nonnegative, there is significant cancellation between the (constant) diagonal elements $\psi_t(1)$ and all the off-diagonal elements with varying signs as

$$V_{t,N,\psi}(X_N) = \frac{1}{N}\psi_t(1) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \psi_t(\boldsymbol{x}_i \cdot \boldsymbol{x}_j).$$

Accurate calculation of such sums is difficult, see [41] for example, especially getting reproducible results on multi-core architecture with dynamic scheduling of parallel non-associative floating point operations [25]. Example 1 has $\psi_{1,t}(1) = 2$ and Example 2 has $\psi_{2,t}(1) = 1$, both independent of t, while Example 3 has

$$\psi_{3,t}(1) = \frac{\Gamma(t+\alpha+2)}{\Gamma(t+1)\Gamma(\alpha+2)} - 1,$$

which grows with the degree t (for $d = 2$, $\psi_{3,t}(1) = t$). These functions are illustrated in Fig. 1. As the variational objectives can be scaled by an arbitrary positive constant, you could instead have used $\psi_{3,t}\frac{\Gamma(t+1)\Gamma(\alpha+2)}{\Gamma(t+\alpha+2)}$. Ratios of gamma functions, as in the expressions for a_0, should not be evaluated directly, but rather simplified for small values of d or evaluated using the log-gamma function. The

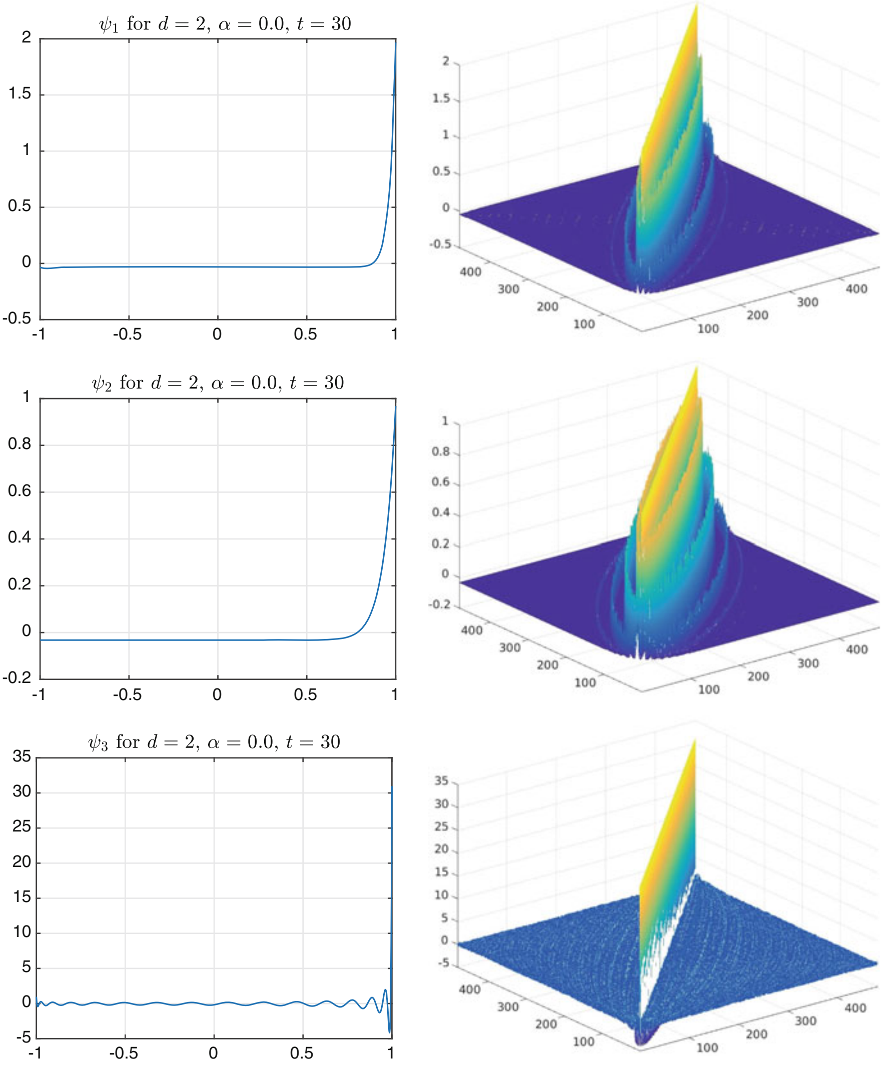


Fig. 1 For $d = 2$, $t = 30$, a spherical t-design with $N = 482$, the functions $\psi_{k,t}$ and arrays $\psi_{k,t}(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$ for $k = 1, 2, 3$

derivatives, essential for large scale non-linear optimization algorithms, are readily calculated using

$$\nabla_{\boldsymbol{x}_k} V_{t,N,\psi}(X_N) = 2 \sum_{\substack{i=1\\ i \neq k}}^{N} \psi_t'(\boldsymbol{x}_i \cdot \boldsymbol{x}_k)\boldsymbol{x}_i$$

and the Jacobian of the (normalised) spherical parametrisation (see Sect. 4.2).

Because of the interest in the use of spherical harmonics for the representation of the Earth's gravitational field there has been considerable work, see [42, 43] and [28, Section 7.24.2] for example, on the evaluation of high degree spherical harmonics for $\mathbb{S}^2$. For $(x, y, z)^T \in \mathbb{S}^2$ the real spherical harmonics [54, Chapter 3, Section 18] are usually expressed in terms of the coordinates $z = \cos(\theta)$ and ϕ. In terms of the coordinates $(x, \phi_2) = (\cos(\phi_1), \phi_2)$, see (28) below, they are the $Z(2, \ell) = 2\ell + 1$ functions

$$\begin{aligned} Y_{\ell,\ell+1-k}(x, \phi_2) &:= \hat{c}_{\ell,k}(1 - x^2)^{k/2} S_\ell^{(k)}(x) \sin(k\phi_2), \quad k = 1, \ldots, \ell, \\ Y_{\ell,\ell+1}(x, \phi_2) &:= \hat{c}_{\ell,0} S_\ell^{(0)}(x), \\ Y_{\ell,\ell+1+k}(x, \phi_2) &:= \hat{c}_{\ell,k}(1 - x^2)^{k/2} S_\ell^{(k)}(x) \cos(k\phi_2), \quad k = 1, \ldots, \ell. \end{aligned} \tag{25}$$

where $S_\ell^{(k)}(x) = \sqrt{\frac{(\ell-k)!}{(\ell+k)!}} P_\ell^k(x)$ are versions of the Schmidt semi-normalised associated Legendre functions for which stable three-term recurrences exist for high (about 2700) degrees and orders. The normalization constants $\hat{c}_{\ell,0}$, $\hat{c}_{\ell,k}$ are, for normalised surface measure,

$$\hat{c}_{\ell,0} = \sqrt{2\ell + 1}, \quad \hat{c}_{\ell,k} = \sqrt{2}\sqrt{2\ell + 1}, \quad k = 1, \ldots, \ell,$$

For $\mathbb{S}^2$ these expressions can be used to directly evaluate the Weyl sums (16), and hence their sum of squares, and their derivatives.

4.2 Spherical Parametrisations

There are many ways to organise a spherical parametrisation of $\mathbb{S}^d$. For $\phi_i \in [0, \pi]$ for $i = 1, \ldots, d - 1$ and $\phi_d \in [0, 2\pi)$ define $\boldsymbol{x} \in \mathbb{S}^d$ by

$$x_1 = \cos(\phi_1) \tag{26}$$

$$x_i = \prod_{k=1}^{i-1} \sin(\phi_k) \cos(\phi_i), \quad i = 2, \ldots, d \tag{27}$$

$$x_{d+1} = \prod_{k=1}^{d} \sin(\phi_k) \tag{28}$$

The inverse transformation used is, for $i = 1, \ldots, d - 1$

$$\phi_i = \begin{cases} 0 & \text{if } x_k = 0, \quad k = i, \ldots, d + 1, \\ \cos^{-1}\left(x_i / \sqrt{\sum_{k=i}^{d+1} x_k^2} \right) & \text{otherwise;} \end{cases} \tag{29}$$

$$\phi_d = \tan^{-1}\left(x_{d+1}/x_d\right). \tag{30}$$

The last component can be calculated using the four quadrant `atan2` function and periodicity to get $\phi_d \in [0, 2\pi)$. Spherical parametrisations introduce potential singularities when $\phi_i = 0$ or $\phi_i = \pi$ for any $i = 1, \ldots, d-1$.

As all the functions considered are zonal, they are invariant under an orthogonal transformation (rotation). Thus the point sets are normalised so that the $d+1$ by N matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_N]$ has

$$\begin{aligned}
&\boldsymbol{X}_{i,j} = 0 \quad \text{for } i = j+1, \ldots, d+1, \quad j = 1, \ldots, \min(d, N) \\
&\boldsymbol{X}_{i,i} \geq 0 \quad \text{for } i = 1, \ldots, \min(d, N).
\end{aligned}$$

The first normalised point is $\boldsymbol{x}_1 = \boldsymbol{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{d+1}$. Such a rotation can easily be calculated using the QR factorization of $\boldsymbol{X}$ combined with sign changes to the rows Q. The corresponding normalised spherical parametrisation has

$$\Phi_{i,j} = 0 \quad \text{for } i = j, \ldots, d, \quad j = 1, \ldots, \min(d, N),$$

where the jth column of Φ corresponds to the point $\boldsymbol{x}_j$, $j = 1, \ldots, N$. The optimisation variables are then $\Phi_{i,j}, i = 1, \ldots, \min(j-1, d)$, $j = 2, \ldots, N$, stored as the vector $\boldsymbol{\phi} \in \mathbb{R}^n$ where

$$n = \begin{cases} \frac{N(N-1)}{2} & \text{for } N \leq d, \\ Nd - \frac{d(d+1)}{2} & \text{for } N > d, \end{cases} \tag{31}$$

so

$$\begin{aligned}
&\boldsymbol{\phi}_p = \Phi_{i,j}, \quad i = 1, \ldots, \min(j-1, d), \quad j = 2, \ldots, N, \\
&p = \begin{cases} \frac{\min(j-1,d)(\min(j-1,d)-1)}{2} + i & \text{for } j = 2, \ldots, \min(d, N) \\ \frac{d(d-1)}{2} + (j-d-1)d + i & \text{for } j = d+1, \ldots, N, \quad N > d. \end{cases}
\end{aligned}$$

It is far easier to work with a spherical parametrisation with bound constraints than to impose the quadratic constraints $\boldsymbol{x}_j \cdot \boldsymbol{x}_j = 1, j = 1, \ldots, N$, especially for large N. As the optimization criteria have the effect of moving the points apart, the use of the normalised point sets reduces difficulties with singularities at the boundaries corresponding to $\Phi_{i,j} = 0$ or $\Phi_{i,j} = \pi$, $i = 1, \ldots, d-1$.

For $\mathbb{S}^2$, these normalised point sets may be rotated (the variable components reordered) using

$$Q = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

to get the commonly [27, 57, 66] used normalization with the first point at the north pole and the second on the prime meridian.

A symmetric (or antipodal) point set ($\boldsymbol{x} \in X_N \iff -\boldsymbol{x} \in X_N$) must have N even, so can be represented as $\boldsymbol{X} = [\overline{\boldsymbol{X}}\ -\overline{\boldsymbol{X}}]$ where the $d+1$ by $N/2$ array of points $\overline{\boldsymbol{X}}$ is normalised as above.

If only zonal function functions depending just on the inner products $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ are used then you could use the variables $\boldsymbol{Z}_{i,j} = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$, so

$$\boldsymbol{Z} \in \mathbb{R}^{N \times N}, \quad \boldsymbol{Z}^T = \boldsymbol{Z}, \quad \boldsymbol{Z} \succeq 0, \quad \operatorname{diag}(\boldsymbol{Z}) = \boldsymbol{e}, \quad \operatorname{rank}(\boldsymbol{Z}) = d+1,$$

where $\boldsymbol{e} = (1, \dots, 1)^T \in \mathbb{R}^N$ and $\boldsymbol{Z} \succeq 0$ indicates $\boldsymbol{Z}$ is positive semi-definite. The major difficulties with such a parametrisation are the number $N(N-1)/2$ of variables and the rank condition. Semi-definite programming relaxations (without the rank condition) have been used to get bounds on problems involving points on the sphere (see, for example, [4]).

4.3 Degrees of Freedom for $\mathbb{S}^d$

Using a normalised spherical parametrisation of N points on $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ there are $n = Nd - d(d+1)/2$ variables (assuming $N \geq d$). The number of conditions for a t-design is

$$m = \sum_{\ell=1}^{t} Z(d, \ell) = D(d,t) - 1 = Z(d+1, t) - 1.$$

Using the simple criterion that the number of variables n is at least the number of conditions m, gives the number of points as

$$\widehat{N}(d,t) := \left\lceil \frac{1}{d} \left(Z(d+1,t) + \frac{d(d+1)}{2} - 1 \right) \right\rceil. \tag{32}$$

For $\mathbb{S}^2$ there are $n = 2N - 3$ variables and $m = (t+1)^2 - 1$ conditions giving

$$\widehat{N}(2,t) := \lceil (t+1)^2)/2 \rceil + 1.$$

Grabner and Sloan [30] obtained separation results for N point spherical t-designs when $N \leq \tau\ 2N^*$ and $\tau < 1$. For $d = 2$, $\widehat{N}$ is less than twice the lower bound N^* as

$$\widehat{N}(2,t) = 2N^*(2,t) - t,$$

but the difference is only a lower order term. The values for $\widehat{N}(2,t)$, $N^*(2,t)$ and the Yudin lower bound $N^+(2,t)$ are available in Tables 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16.

The idea of exploiting symmetry to reduce the number of conditions that a quadrature rule should satisfy at least goes back to Sobolev [60]. For a symmetric point set (both $\boldsymbol{x}_j, -\boldsymbol{x}_j$ in X_N) then all odd degree polynomials $Y_{\ell,k}$ or $P_\ell^{(d+1)}$ are automatically integrated exactly by an equal weight quadrature rule. Thus, for t odd, the number of conditions to be satisfied is

$$m = \sum_{\ell=1}^{(t-1)/2} Z(d, 2\ell) = \frac{\Gamma(t+d)}{\Gamma(d+1)\Gamma(t)} - 1. \tag{33}$$

The number of free variables in a normalised symmetric point set $\boldsymbol{X} = [\overline{\boldsymbol{X}} \quad -\overline{\boldsymbol{X}}]$ (assuming $N/2 \geq d$) is

$$n = \left(\frac{Nd}{2} - \frac{d(d+1)}{2}\right). \tag{34}$$

Again the simple requirement that $n \geq m$ gives the number of points as

$$\overline{N}(d,t) := 2\left\lceil \frac{1}{d}\left(\frac{\Gamma(t+d)}{\Gamma(d+1)\Gamma(t)} - 1 + \frac{d(d+1)}{2}\right)\right\rceil. \tag{35}$$

For $d = 2$ this simplifies, again for t odd, to

$$\overline{N}(2,t) := 2\left\lceil \frac{t^2+t+4}{4} \right\rceil.$$

$\overline{N}(2,t)$ is slightly less than $\widehat{N}(2,t)$, comparable to twice the lower bound $N^*(2,t)$ as

$$\overline{N}(2,t) = 2N^*(2,t) - \frac{3}{2}t + \begin{cases} \frac{3}{2} & \text{if} \quad \bmod\,(t,4) = 1, \\ \frac{1}{2} & \text{if} \quad \bmod\,(t,4) = 3. \end{cases}$$

However $\overline{N}(2,t)$ is not less than $\tau\ 2N^*(2,t)$, $\tau < 1$, as required by Grabner and Sloan [30].

The leading term of both $\widehat{N}(d,t)$ and $\overline{N}(d,t)$ is $D(d,t)/d$, see Table 1, where $D(d,t)$ defined in (4) is the dimension of $\mathbb{P}_t(\mathbb{S}^d)$. From (14), a spherical t-design

Table 1 The lower bound $N^*(d,t)$, the number of points $\overline{N}(d,t)$ (symmetric point set) and $\widehat{N}(d,T)$ to match the number of conditions and the dimension of $\mathbb{P}_t(\mathbb{S}^d)$ for $d = 2, 3, 4, 5$

d	$N^*(d,t)$	$\overline{N}(d,t)$	$\widehat{N}(d,t)$	$D(d,t)$
2	$\frac{t^2}{4} + t + O(1)$	$\frac{t^2}{2} + \frac{t}{2} + O(1)$	$\frac{t^2}{2} + t + O(1)$	$t^2 + 2t + 1$
3	$\frac{t^3}{24} + \frac{3t^2}{8} + O(t)$	$\frac{t^3}{9} + \frac{t^2}{3} + O(t)$	$\frac{t^3}{9} + \frac{t^2}{2} + O(t)$	$\frac{t^3}{3} + \frac{3t^2}{2} + O(t)$
4	$\frac{t^4}{192} + \frac{t^3}{12} + O(t^2)$	$\frac{t^4}{48} + \frac{t^3}{8} + O(t^2)$	$\frac{t^4}{48} + \frac{t^3}{6} + O(t^2)$	$\frac{t^4}{12} + \frac{2t^3}{3} + O(t^2)$
5	$\frac{t^5}{1920} + \frac{5t^4}{384} + O(t^3)$	$\frac{t^5}{300} + \frac{t^4}{30} + O(t^3)$	$\frac{t^5}{300} + \frac{t^4}{24} + O(t^3)$	$\frac{t^5}{60} + \frac{5t^4}{24} + O(t^3)$

with $\widehat{N}(d,t)$ or $\overline{N}(d,t)$ points has efficiency $E \approx 1$. Also the leading term of both $\overline{N}(d,t)$ and $\widehat{N}(d,t)$ is $2^d/d$ times the leading term of the lower bound $N^*(d,t)$.

4.4 Optimization Algorithms

As with many optimization problems on the sphere there are many distinct (not related by an orthogonal transformation or permutation) point sets giving local minima of the optimization objective. For example, Erber and Hockney [26] and Calef et al. [16] studied the minimal energy problem for the sphere and the large number of stable configurations.

Gräf and Potts [32] develop optimization methods on general Riemannian manifolds, in particular $\mathbb{S}^2$, and both Newton-like and conjugate gradients methods. Using a fast method for spherical Fourier coefficients at non-equidistant points they obtain approximate spherical designs for high degrees.

While mathematically it is straight forward to conclude that if $V_{t,N,\psi}(X_N) = 0$ then X_N is a spherical t-design, deciding when a quantity is zero with the limits of standard double precision floating point arithmetic with machine precision $\epsilon = 2.2 \times 10^{-16}$ is less clear (should 10^{-14} be regarded as zero?). Extended precision libraries and packages like Maple or Mathematica can help. A point set X_N with $V_{t,N,\psi}(X_N) \approx \epsilon$ does not give a mathematical proof that is X_N is a spherical t-design, but X_N may still be computationally useful in applications.

On the other hand showing that the global minimum of $V_{t,N,\psi}(X_N)$ is strictly positive, so no spherical t-design with N points exist, is an intrinsically hard problem. Semi-definite programming [63] provides an approach [50] to the global optimization of polynomial sum of squares for modest degrees.

For $d = 2$ a variety of gradient based bound constrained optimization methods, for example the limited memory algorithm [15, 46], were tried both to minimise the variational forms $V_{t,N,\psi}(X_N)$. Classically, see [48] for example, methods can exploit the sum of squares structure $\boldsymbol{r}(X_N)^T\boldsymbol{r}(X_N)$. In both cases it is important to provide derivatives of the objective with respect to the parameters. Using the normalised spherical parametrisation $\boldsymbol{\phi}$ of X_N, the Jacobian of the residual $\boldsymbol{r}(\boldsymbol{\phi})$ is $\boldsymbol{A} : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ where $n = dN - d(d+1)/2$ and $m = D(d,t) - 1$

$$\boldsymbol{A}_{i,j}(\boldsymbol{\phi}) = \frac{\partial r_i(\boldsymbol{\phi})}{\partial \phi_j}, \qquad i = 1,\dots,m, \quad j = 1,\dots,n,$$

where $i = (\ell-1)Z(d+1,\ell-1) + k$, for $k = 1,\dots,Z(d,\ell)$, $\ell = 1,\dots,t$.

For symmetric point sets with $N = \overline{N}(d,t)$ points, the number of variables n is given by (34) and the number of conditions m by (33) corresponding to even degree spherical harmonics.

The well-known structure of a nonlinear least squares problem, see [48] for example, gives, ignoring the $1/N^2$ scaling in (18),

$$f(\boldsymbol{\phi}) = \boldsymbol{r}(\boldsymbol{\phi})^T\, \boldsymbol{D}\, \boldsymbol{r}(\boldsymbol{\phi}), \tag{36}$$

$$\nabla f(\boldsymbol{\phi}) = 2\boldsymbol{A}(\boldsymbol{\phi})^T \boldsymbol{D}\, \boldsymbol{r}(\boldsymbol{\phi}), \tag{37}$$

$$\nabla^2 f(\boldsymbol{\phi}) = 2\boldsymbol{A}(\boldsymbol{\phi})^T \boldsymbol{D}\boldsymbol{A}(\boldsymbol{\phi}) + 2\sum_{i=1}^{m} r_i(\boldsymbol{\phi}) D_{ii} \nabla^2 r_i(\boldsymbol{\phi}). \tag{38}$$

If $\boldsymbol{\phi}^*$ has $\boldsymbol{r}(\boldsymbol{\phi}^*) = \boldsymbol{0}$ and $\boldsymbol{A}(\boldsymbol{\phi}^*)$ has rank n, the Hessian $\nabla^2 f(\boldsymbol{\phi}^*) = 2\boldsymbol{A}(\boldsymbol{\phi}^*)^T \boldsymbol{D}\boldsymbol{A}(\boldsymbol{\phi}^*)$ is positive definite and $\boldsymbol{\phi}^*$ is a strict global minimizer. Here this is only possible when $n = m$, for example when $d = 2$ and t is odd, see Tables 2, 3, 4, 5, 6, 7, 8 and 9, and in the symmetric case when $t \bmod 4 = 3$, see Tables 10, 11, 12, 13, 14, 15 and 16. For $d = 2$ the other values of t have $n = m + 1$, so there is generically a one parameter family of solutions even when the Jacobian has full rank. When $d = 3$, the choice $N = \widehat{N}(3, t)$ gives $n = m$, $n = m + 1$ or $n = m + 3$ depending on the value of t, see Table 18. Thus a Levenberg-Marquadt or trust region method, see [48] for example, in which the search direction satisfies

$$\left(\boldsymbol{A}^T \boldsymbol{D}\boldsymbol{A} + \nu \boldsymbol{I}\right) \boldsymbol{d} = \boldsymbol{A}^T \boldsymbol{D}\boldsymbol{r}$$

was used. When $n > m$ the Hessian of the variational form $V_{t,N,\psi}(X_N)$ evaluated using one of the three example functions (19), (21) or (23) will also be singular at the solution. These disadvantages could have been reduced by choosing the number of points N so that $n < m$, but then there may not be solutions with $V_{t,N,\psi}(X_N) = 0$, that is spherical t-designs may not exist for that number of points.

Many local solutions were found as well as (computationally) global solutions which differed depending on the starting point and the algorithm parameters (for example the initial Levenberg-Marquadt parameter ν, initial trust region, line search parameters etc.). Even when $n = m$ there are often multiple spherical designs for same t, N, which are strict global minimisers, but have different inner product sets $\mathscr{A}(X_N)$ in (15) and different mesh ratios.

4.5 Structure of Point Sets

There are a number of issues with the spherical designs studied here.

- There is no proof that spherical t-designs on $\mathbb{S}^d$ with $N = t^d/d + O(t^{d-1})$ points exist for all t (that is the constant in the Bondarenko et al. result [8] is $C_d = 1/d$ (or lower), as suggested by [33] for $\mathbb{S}^2$).

- The point sets are not nested, that is the points of a spherical t-design are not necessarily a subset of the points of a t'-design for some $t' > t$.
- The point sets do not lie on bands of equal ϕ_1 (latitude on $\mathbb{S}^2$) making them less amenable for FFT based methods.
- The point sets are obtained by extensive calculation, rather than generated by a simple algorithms as for generalized spiral or equal area points on $\mathbb{S}^2$ [52]. Once calculated the point sets are easy to use.

An example of a point set on $\mathbb{S}^2$ that satisfies the last three issues are the HELAPix points[29], which provide a hierarchical, equal area (so exact for constants), iso-latitude set of points widely used in cosmology.

5 Tables of Results

5.1 Spherical t-Designs with no Imposed Symmetry for $\mathbb{S}^2$

From Tables 2, 3, 4, 5, 6, 7, 8 and 9 the variational criteria based on the three functions $\psi_{1,t}$, $\psi_{2,t}$ and $\psi_{3,t}$ all have values close to the double precision machine precision of $\epsilon = 2.2 \times 10^{-16}$ for all degrees $t = 1, \dots, 180$. Despite being theoretically non-negative, rounding error sometimes gives negative values, but still close to machine precision. The potential values using $\psi_{3,t}$ are slightly larger due to the larger value of $\psi_{3,t}(1)$. The tables also give the unscaled sum of squares

$$f_t(X_N) = \boldsymbol{r}(X_N)^T \boldsymbol{r}(X_N), \tag{39}$$

which are plotted in Fig. 2. These tables also list both the Delsarte, Goethals and Seidel lower bounds $N^*(2,t)$ and the Yudin lower bound $N^+(2,t)$, plus the actual number of points N. The number of points $N = \widehat{N}(2,t)$, apart from $t = 3, 5, 7, 9, 11, 13, 15$ when $N = \widehat{N}(2,t) - 1$. There may well be spherical t-designs with smaller values of N and special symmetries, see [33] for example. For all these point sets the mesh ratios $\rho(X_N)$ are less than 1.81, see Fig. 2. All these point sets are available from [66].

5.2 Symmetric Spherical t-Designs for $\mathbb{S}^2$

For $\mathbb{S}^2$ a t-design with a sightly smaller number of points $\overline{N}(2,t)$ can be found by constraining the point sets to be symmetric (antipodal). A major computational advantage of working with symmetric point sets is the reduction (approximately half), for a given degree t, in the number of optimization variables n and the number

Table 2 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees $t = 1$–24, except $t = 3, 5, 7, 9, 11, 13, 15$ when $N = \widehat{N}(2,t) - 1$

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
1	2	2	3	3	3	3.7e−17	1.9e−17	1.2e−17	3.0e−32	2.0944	1.5708	1.50
2	4	4	6	9	8	3.5e−17	3.5e−17	4.3e−17	1.1e−31	1.5708	0.9553	1.22
3	6	6	8	13	15	3.1e−17	−4.8e−18	−1.0e−17	6.2e−30	1.2310	0.9553	1.55
4	9	9	14	25	24	−2.1e−18	−5.5e−18	4.5e−17	8.6e−30	0.8630	0.6913	1.60
5	12	12	18	33	35	−3.4e−17	6.9e−19	−1.1e−16	5.2e−27	0.8039	0.5749	1.43
6	16	16	26	49	48	2.0e−17	5.0e−18	6.0e−17	7.4e−29	0.6227	0.4911	1.58
7	20	20	32	61	63	1.1e−17	4.5e−18	6.1e−17	1.5e−28	0.5953	0.4357	1.46
8	25	25	42	81	80	−9.6e−18	−2.3e−18	−2.8e−17	2.1e−28	0.4845	0.3958	1.63
9	30	31	50	97	99	1.0e−17	−1.3e−18	−2.4e−17	7.0e−28	0.4555	0.3608	1.58
10	36	37	62	121	120	1.6e−17	2.3e−18	7.7e−17	8.1e−28	0.3945	0.3308	1.68
11	42	43	72	141	143	8.3e−18	8.8e−18	1.2e−16	2.1e−27	0.3750	0.2989	1.59
12	49	50	86	169	168	−9.7e−18	−5.1e−18	−7.6e−17	2.9e−27	0.3241	0.2761	1.70
13	56	58	98	193	195	−7.4e−18	4.3e−18	−6.9e−17	1.1e−26	0.3028	0.2567	1.70
14	64	66	114	225	224	7.0e−18	1.5e−18	−6.4e−18	1.2e−26	0.2838	0.2402	1.69
15	72	75	128	253	255	−4.2e−18	−1.6e−18	−1.2e−16	2.8e−26	0.2644	0.2279	1.72
16	81	84	146	289	288	3.8e−18	6.3e−18	−5.0e−17	3.0e−26	0.2568	0.2115	1.65
17	90	94	163	323	323	2.0e−17	1.1e−17	1.9e−16	4.7e−26	0.2333	0.2070	1.77
18	100	104	182	361	360	−4.1e−18	−5.1e−18	−1.6e−16	5.7e−26	0.2243	0.1880	1.68
19	110	115	201	399	399	1.5e−17	7.7e−18	1.1e−16	7.7e−26	0.2086	0.1843	1.77
20	121	127	222	441	440	1.3e−17	1.3e−17	2.0e−17	1.3e−25	0.2105	0.1697	1.61
21	132	139	243	483	483	1.6e−17	−3.0e−18	−4.3e−17	1.4e−25	0.1900	0.1677	1.77
22	144	151	266	529	528	−7.6e−19	−1.0e−17	8.6e−17	1.8e−25	0.1887	0.1574	1.67
23	156	164	289	575	575	−6.8e−18	1.1e−18	−3.7e−17	2.6e−25	0.1759	0.1554	1.77
24	169	178	314	625	624	2.4e−18	−9.3e−18	−1.2e−16	3.2e−25	0.1730	0.1451	1.68

Table 3 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees $t = 25$–48

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
25	182	192	339	675	675	3.1e−18	1.1e−17	−1.2e−16	4.3e−25	0.1628	0.1407	1.73
26	196	207	366	729	728	6.1e−18	−2.0e−17	−8.6e−17	4.3e−25	0.1533	0.1333	1.74
27	210	222	393	783	783	3.3e−18	1.7e−17	−1.4e−16	4.9e−25	0.1485	0.1305	1.76
28	225	238	422	841	840	−1.8e−17	1.9e−17	−2.5e−16	5.8e−25	0.1490	0.1252	1.68
29	240	254	451	899	899	−2.0e−17	1.4e−17	−3.3e−16	7.2e−25	0.1405	0.1214	1.73
30	256	271	482	961	960	−3.5e−17	4.2e−18	−8.6e−17	8.8e−25	0.1381	0.1152	1.67
31	272	289	513	1023	1023	−3.3e−17	−4.2e−19	−1.7e−16	1.4e−24	0.1313	0.1132	1.72
32	289	307	546	1089	1088	−5.1e−17	−5.6e−18	−2.2e−16	1.7e−24	0.1315	0.1098	1.67
33	306	325	579	1155	1155	−5.0e−17	2.7e−17	−2.1e−16	1.7e−24	0.1292	0.1054	1.63
34	324	344	614	1225	1224	3.1e−17	2.8e−17	−3.0e−16	2.2e−24	0.1235	0.1030	1.67
35	342	364	649	1295	1295	3.0e−17	−1.6e−18	−3.5e−16	2.4e−24	0.1139	0.1005	1.76
36	361	384	686	1369	1368	5.1e−17	−9.6e−19	−3.9e−16	3.4e−24	0.1170	0.0970	1.66
37	380	405	723	1443	1443	5.1e−17	−6.6e−18	−3.1e−16	3.1e−24	0.1113	0.0962	1.73
38	400	426	762	1521	1520	2.8e−17	2.6e−17	−3.6e−16	4.0e−24	0.1079	0.0925	1.71
39	420	448	801	1599	1599	3.0e−17	−6.2e−17	−3.6e−16	4.1e−24	0.1079	0.0933	1.73
40	441	470	842	1681	1680	1.0e−16	5.5e−17	−4.6e−16	5.0e−24	0.1068	0.0875	1.64
41	462	493	883	1763	1763	1.1e−16	2.8e−17	−2.6e−16	5.5e−24	0.0998	0.0858	1.72
42	484	516	926	1849	1848	−1.7e−19	−2.1e−17	−3.8e−16	6.8e−24	0.1007	0.0829	1.65
43	506	540	969	1935	1935	−7.3e−19	−4.7e−17	−4.8e−16	7.0e−24	0.0964	0.0819	1.70
44	529	565	1014	2025	2024	2.9e−17	2.6e−17	−4.0e−16	9.0e−24	0.0980	0.0805	1.64
45	552	590	1059	2115	2115	2.4e−17	−2.5e−17	−3.7e−16	9.4e−24	0.0911	0.0787	1.73
46	576	615	1106	2209	2208	7.6e−17	6.9e−17	−3.5e−16	1.2e−23	0.0949	0.0763	1.61
47	600	642	1153	2303	2303	8.4e−17	−3.7e−17	−2.7e−16	1.3e−23	0.0898	0.0751	1.67
48	625	668	1202	2401	2400	−3.1e−17	−4.2e−17	−3.8e−16	1.6e−23	0.0869	0.0748	1.72

Table 4 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees $t = 49$–72

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
49	650	696	1251	2499	2499	−2.3e−17	1.2e−16	−3.7e−16	1.5e−23	0.0839	0.0725	1.73
50	676	723	1302	2601	2600	3.5e−17	2.9e−17	−1.8e−16	2.2e−23	0.0858	0.0712	1.66
51	702	752	1353	2703	2703	2.6e−17	−1.1e−16	−3.0e−16	2.0e−23	0.0838	0.0694	1.66
52	729	781	1406	2809	2808	1.0e−16	6.5e−17	−3.4e−16	2.3e−23	0.0809	0.0676	1.67
53	756	810	1459	2915	2915	1.1e−16	−5.3e−17	−3.4e−16	2.8e−23	0.0768	0.0674	1.75
54	784	840	1514	3025	3024	−2.6e−17	2.5e−17	−2.5e−16	2.9e−23	0.0783	0.0656	1.68
55	812	870	1569	3135	3135	−3.0e−17	5.1e−17	−3.2e−16	3.1e−23	0.0741	0.0650	1.75
56	841	902	1626	3249	3248	−9.2e−17	−5.7e−17	−2.8e−16	3.5e−23	0.0778	0.0629	1.62
57	870	933	1683	3363	3363	−9.0e−17	−1.2e−16	−2.0e−16	3.8e−23	0.0717	0.0631	1.76
58	900	965	1742	3481	3480	−2.1e−16	−1.7e−16	−2.3e−16	4.6e−23	0.0732	0.0615	1.68
59	930	998	1801	3599	3599	−2.0e−16	−8.5e−17	−2.3e−16	5.2e−23	0.0708	0.0608	1.72
60	961	1031	1862	3721	3720	−9.6e−17	3.7e−18	−1.9e−16	5.8e−23	0.0718	0.0592	1.65
61	992	1065	1923	3843	3843	−9.4e−17	5.3e−17	−4.3e−17	6.4e−23	0.0663	0.0584	1.76
62	1024	1099	1986	3969	3968	2.6e−17	2.4e−17	−1.9e−16	6.7e−23	0.0699	0.0577	1.65
63	1056	1134	2049	4095	4095	2.6e−17	−4.8e−18	−1.5e−16	7.1e−23	0.0680	0.0564	1.66
64	1089	1170	2114	4225	4224	3.3e−17	2.7e−17	6.5e−17	7.4e−23	0.0662	0.0562	1.70
65	1122	1206	2179	4355	4355	2.7e−17	−1.7e−16	9.7e−18	8.9e−23	0.0647	0.0552	1.70
66	1156	1242	2246	4489	4488	−1.0e−16	−1.0e−16	−1.0e−16	8.9e−23	0.0616	0.0537	1.74
67	1190	1279	2313	4623	4623	−1.1e−16	−2.2e−16	−1.7e−16	1.1e−22	0.0609	0.0534	1.75
68	1225	1317	2382	4761	4760	2.5e−16	2.2e−16	−1.5e−17	1.1e−22	0.0620	0.0523	1.69
69	1260	1355	2451	4899	4899	2.5e−16	−1.3e−16	−8.2e−17	1.2e−22	0.0590	0.0516	1.75
70	1296	1394	2522	5041	5040	1.0e−16	4.1e−18	−8.3e−17	1.4e−22	0.0595	0.0513	1.73
71	1332	1433	2593	5183	5183	1.0e−16	−1.7e−16	−3.2e−17	1.5e−22	0.0587	0.0496	1.69
72	1369	1473	2666	5329	5328	1.7e−16	1.7e−16	−6.6e−17	1.7e−22	0.0603	0.0494	1.64

Table 5 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2, t)$ and degrees $t = 73$–96

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
73	1406	1513	2739	5475	5475	1.7e−16	8.5e−17	−2.5e−16	1.6e−22	0.0567	0.0488	1.72
74	1444	1554	2814	5625	5624	1.8e−16	9.0e−17	−2.0e−16	1.9e−22	0.0582	0.0476	1.64
75	1482	1595	2889	5775	5775	1.8e−16	−1.7e−16	−1.6e−16	2.0e−22	0.0577	0.0475	1.64
76	1521	1637	2966	5929	5928	2.8e−16	1.8e−16	−3.8e−16	2.2e−22	0.0547	0.0471	1.72
77	1560	1680	3043	6083	6083	2.8e−16	−2.1e−16	−3.4e−16	2.3e−22	0.0546	0.0459	1.68
78	1600	1723	3122	6241	6240	3.2e−16	2.3e−16	−2.2e−16	2.6e−22	0.0554	0.0455	1.64
79	1640	1766	3201	6399	6399	3.3e−16	2.2e−16	−1.3e−16	2.7e−22	0.0538	0.0449	1.67
80	1681	1810	3282	6561	6560	2.6e−16	2.8e−16	−1.7e−16	3.0e−22	0.0525	0.0450	1.71
81	1722	1855	3363	6723	6723	2.6e−16	2.0e−16	−4.6e−16	3.4e−22	0.0537	0.0436	1.62
82	1764	1900	3446	6889	6888	3.1e−16	2.2e−16	−4.2e−16	3.6e−22	0.0532	0.0431	1.62
83	1806	1946	3529	7055	7055	3.2e−16	−2.9e−16	−3.2e−16	3.6e−22	0.0505	0.0426	1.69
84	1849	1992	3614	7225	7224	−3.9e−17	1.3e−17	−3.9e−16	3.9e−22	0.0516	0.0424	1.64
85	1892	2039	3699	7395	7395	−5.1e−17	1.7e−17	−3.8e−16	4.1e−22	0.0479	0.0418	1.75
86	1936	2087	3786	7569	7568	−1.6e−16	−1.4e−16	−4.4e−16	4.8e−22	0.0488	0.0427	1.75
87	1980	2135	3873	7743	7743	−1.6e−16	−1.4e−16	−4.5e−16	4.9e−22	0.0488	0.0414	1.69
88	2025	2183	3962	7921	7920	2.5e−16	2.3e−16	−4.2e−16	5.0e−22	0.0493	0.0404	1.64
89	2070	2232	4051	8099	8099	2.5e−16	−3.2e−16	−4.9e−16	5.4e−22	0.0454	0.0403	1.78
90	2116	2282	4142	8281	8280	−9.9e−17	−1.6e−17	−4.1e−16	6.0e−22	0.0489	0.0394	1.61
91	2162	2332	4233	8463	8463	−1.1e−16	2.3e−16	−4.0e−16	6.2e−22	0.0473	0.0393	1.66
92	2209	2383	4326	8649	8648	−2.2e−16	−2.9e−16	−5.0e−16	6.7e−22	0.0481	0.0389	1.61
93	2256	2434	4419	8835	8835	−2.1e−16	8.6e−17	−5.1e−16	7.1e−22	0.0467	0.0382	1.64
94	2304	2486	4514	9025	9024	4.3e−17	1.4e−16	−5.6e−16	7.5e−22	0.0463	0.0383	1.65
95	2352	2538	4609	9215	9215	3.5e−17	−1.0e−16	−6.4e−16	7.9e−22	0.0462	0.0377	1.63
96	2401	2591	4706	9409	9408	−1.7e−16	−5.9e−17	−6.8e−16	9.0e−22	0.0458	0.0369	1.61

Table 6 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2, t)$ and degrees $t = 97$–120

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
97	2450	2644	4803	9603	9603	−1.8e−16	−1.2e−16	−6.2e−16	8.8e−22	0.0429	0.0367	1.71
98	2500	2698	4902	9801	9800	−1.6e−16	−1.6e−16	−5.3e−16	1.0e−21	0.0453	0.0362	1.60
99	2550	2753	5001	9999	9999	−1.6e−16	−1.9e−16	−5.5e−16	1.0e−21	0.0424	0.0370	1.75
100	2601	2808	5102	10, 201	10, 200	1.4e−16	2.6e−16	−6.4e−16	1.1e−21	0.0432	0.0357	1.65
101	2652	2863	5203	10, 403	10, 403	1.5e−16	−2.8e−16	−6.2e−16	1.2e−21	0.0433	0.0351	1.62
102	2704	2919	5306	10, 609	10, 608	−1.1e−16	6.7e−18	−5.1e−16	1.3e−21	0.0424	0.0350	1.65
103	2756	2976	5409	10, 815	10, 815	−1.2e−16	8.7e−17	−6.7e−16	1.3e−21	0.0428	0.0345	1.61
104	2809	3033	5514	11, 025	11, 024	−2.7e−16	−1.3e−16	−7.0e−16	1.5e−21	0.0424	0.0343	1.62
105	2862	3091	5619	11, 235	11, 235	−2.8e−16	4.7e−18	−6.4e−16	1.5e−21	0.0395	0.0345	1.75
106	2916	3149	5726	11, 449	11, 448	1.9e−16	1.9e−16	−6.1e−16	1.6e−21	0.0410	0.0335	1.63
107	2970	3208	5833	11, 663	11, 663	1.9e−16	−1.5e−16	−6.8e−16	1.6e−21	0.0393	0.0337	1.72
108	3025	3267	5942	11, 881	11, 880	−4.3e−16	−3.7e−16	−7.0e−16	1.8e−21	0.0385	0.0340	1.77
109	3080	3327	6051	12, 099	12, 099	−4.3e−16	2.6e−16	−6.8e−16	1.9e−21	0.0381	0.0333	1.75
110	3136	3388	6162	12, 321	12, 320	1.9e−16	3.9e−17	−7.8e−16	2.0e−21	0.0396	0.0324	1.64
111	3192	3449	6273	12, 543	12, 543	1.9e−16	−1.5e−16	−8.9e−16	2.1e−21	0.0376	0.0321	1.70
112	3249	3510	6386	12, 769	12, 768	6.0e−17	1.6e−16	−8.2e−16	2.2e−21	0.0379	0.0322	1.70
113	3306	3573	6499	12, 995	12, 995	6.1e−17	−1.5e−16	−7.3e−16	2.3e−21	0.0373	0.0315	1.69
114	3364	3635	6614	13, 225	13, 224	2.2e−16	2.1e−16	−8.1e−16	2.8e−21	0.0381	0.0312	1.64
115	3422	3698	6729	13, 455	13, 455	2.2e−16	1.5e−16	−6.7e−16	2.5e−21	0.0367	0.0310	1.69
116	3481	3762	6846	13, 689	13, 688	−4.3e−16	−2.6e−16	−7.3e−16	2.8e−21	0.0368	0.0308	1.67
117	3540	3826	6963	13, 923	13, 923	−4.3e−16	6.8e−17	−7.9e−16	2.8e−21	0.0355	0.0304	1.71
118	3600	3891	7082	14, 161	14, 160	−4.5e−16	−2.8e−16	−7.5e−16	3.0e−21	0.0360	0.0306	1.70
119	3660	3957	7201	14, 399	14, 399	−4.5e−16	2.0e−16	−7.5e−16	3.3e−21	0.0358	0.0308	1.72
120	3721	4023	7322	14, 641	14, 640	−5.2e−16	−4.7e−16	−8.2e−16	3.4e−21	0.0348	0.0297	1.70

Table 7 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees $t = 121$–144

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
121	3782	4089	7443	14,883	14,883	−5.2e−16	−1.5e−16	−9.0e−16	3.3e−21	0.0338	0.0295	1.75
122	3844	4156	7566	15,129	15,128	1.8e−16	1.7e−16	−8.9e−16	3.6e−21	0.0356	0.0295	1.66
123	3906	4224	7689	15,375	15,375	1.8e−16	1.3e−16	−6.8e−16	4.0e−21	0.0352	0.0292	1.66
124	3969	4292	7814	15,625	15,624	−9.1e−17	6.3e−17	−8.4e−16	4.1e−21	0.0348	0.0298	1.71
125	4032	4360	7939	15,875	15,875	−8.7e−17	−1.9e−16	−1.0e−15	4.2e−21	0.0337	0.0288	1.70
126	4096	4430	8066	16,129	16,128	−8.7e−17	−1.9e−16	−8.1e−16	4.4e−21	0.0332	0.0283	1.71
127	4160	4499	8193	16,383	16,383	−7.6e−17	−6.8e−18	−8.1e−16	4.5e−21	0.0321	0.0283	1.76
128	4225	4570	8322	16,641	16,640	6.9e−16	6.6e−16	−8.6e−16	4.9e−21	0.0330	0.0283	1.72
129	4290	4641	8451	16,899	16,899	7.1e−16	−6.4e−16	−9.5e−16	5.1e−21	0.0341	0.0279	1.63
130	4356	4712	8582	17,161	17,160	2.8e−16	2.7e−16	−9.7e−16	5.4e−21	0.0322	0.0278	1.73
131	4422	4784	8713	17,423	17,423	2.8e−16	−1.6e−16	−8.1e−16	5.6e−21	0.0325	0.0276	1.70
132	4489	4856	8846	17,689	17,688	2.8e−16	1.3e−16	−1.0e−15	5.8e−21	0.0324	0.0278	1.72
133	4556	4929	8979	17,955	17,955	2.8e−16	8.4e−17	−9.2e−16	6.1e−21	0.0335	0.0269	1.61
134	4624	5003	9114	18,225	18,224	4.5e−16	4.1e−16	−1.0e−15	6.4e−21	0.0321	0.0267	1.67
135	4692	5077	9249	18,495	18,495	4.5e−16	−2.8e−16	−1.0e−15	7.0e−21	0.0309	0.0270	1.75
136	4761	5152	9386	18,769	18,768	5.4e−16	5.0e−16	−1.0e−15	7.3e−21	0.0306	0.0261	1.70
137	4830	5227	9523	19,043	19,043	5.5e−16	2.7e−16	−8.8e−16	7.8e−21	0.0323	0.0260	1.61
138	4900	5303	9662	19,321	19,320	2.5e−16	1.0e−16	−8.8e−16	8.1e−21	0.0301	0.0267	1.78
139	4970	5379	9801	19,599	19,599	2.5e−16	−3.5e−16	−9.6e−16	8.6e−21	0.0303	0.0271	1.79
140	5041	5456	9942	19,881	19,880	7.5e−17	6.8e−17	−9.5e−16	9.3e−21	0.0298	0.0263	1.77
141	5112	5533	10,083	20,163	20,163	7.1e−17	−3.8e−16	−9.5e−16	8.9e−21	0.0299	0.0257	1.72
142	5184	5611	10,226	20,449	20,448	2.2e−16	2.2e−16	−9.4e−16	9.4e−21	0.0300	0.0255	1.70
143	5256	5689	10,369	20,735	20,735	2.1e−16	6.1e−16	−1.1e−15	9.4e−21	0.0293	0.0250	1.71
144	5329	5768	10,514	21,025	21,024	−6.5e−16	−6.1e−16	−1.1e−15	9.9e−21	0.0285	0.0253	1.77

Table 8 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees t = 145–168

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
145	5402	5848	10,659	21,315	21,315	−6.5e−16	1.3e−16	−1.1e−15	1.0e−20	0.0283	0.0250	1.77
146	5476	5928	10,806	21,609	21,608	6.6e−16	6.2e−16	−1.1e−15	1.1e−20	0.0285	0.0250	1.76
147	5550	6009	10,953	21,903	21,903	6.6e−16	−4.2e−16	−1.1e−15	1.1e−20	0.0287	0.0251	1.75
148	5625	6090	11,102	22,201	22,200	−3.8e−16	−2.3e−16	−1.1e−15	1.2e−20	0.0282	0.0241	1.71
149	5700	6172	11,251	22,499	22,499	−3.8e−16	−3.6e−16	−1.2e−15	1.1e−20	0.0276	0.0245	1.77
150	5776	6254	11,402	22,801	22,800	1.4e−16	1.3e−16	−1.2e−15	1.3e−20	0.0279	0.0243	1.74
151	5852	6337	11,553	23,103	23,103	1.3e−16	5.8e−16	−1.2e−15	1.4e−20	0.0275	0.0239	1.74
152	5929	6420	11,706	23,409	23,408	−8.5e−16	−6.8e−16	−1.4e−15	1.3e−20	0.0274	0.0238	1.73
153	6006	6504	11,859	23,715	23,715	−8.4e−16	4.8e−16	−1.1e−15	1.4e−20	0.0268	0.0237	1.77
154	6084	6589	12,014	24,025	24,024	−4.5e−16	−4.8e−16	−1.2e−15	1.5e−20	0.0271	0.0235	1.74
155	6162	6674	12,169	24,335	24,335	−4.5e−16	2.6e−16	−1.0e−15	1.6e−20	0.0268	0.0240	1.79
156	6241	6759	12,326	24,649	24,648	4.6e−16	3.5e−16	−1.2e−15	1.6e−20	0.0269	0.0234	1.74
157	6320	6845	12,483	24,963	24,963	4.6e−16	−2.7e−16	−1.2e−15	1.7e−20	0.0269	0.0238	1.77
158	6400	6932	12,642	25,281	25,280	3.7e−17	1.0e−16	−1.3e−15	1.7e−20	0.0262	0.0228	1.74
159	6480	7019	12,801	25,599	25,599	2.5e−17	−5.4e−16	−1.3e−15	1.7e−20	0.0262	0.0227	1.73
160	6561	7107	12,962	25,921	25,920	−8.0e−16	−7.6e−16	−1.2e−15	1.9e−20	0.0265	0.0230	1.74
161	6642	7195	13,123	26,243	26,243	−8.0e−16	2.2e−16	−1.2e−15	1.9e−20	0.0256	0.0225	1.75
162	6724	7284	13,286	26,569	26,568	1.2e−16	2.0e−16	−1.3e−15	2.1e−20	0.0256	0.0228	1.78
163	6806	7373	13,449	26,895	26,895	1.1e−16	−6.1e−16	−1.3e−15	2.2e−20	0.0257	0.0221	1.72
164	6889	7463	13,614	27,225	27,224	−5.7e−16	−6.3e−16	−1.5e−15	2.1e−20	0.0257	0.0223	1.74
165	6972	7553	13,779	27,555	27,555	−5.7e−16	−3.7e−16	−1.5e−15	2.2e−20	0.0251	0.0225	1.79
166	7056	7644	13,946	27,889	27,888	−9.6e−16	−8.1e−16	−1.3e−15	2.3e−20	0.0251	0.0217	1.73
167	7140	7736	14,113	28,223	28,223	−9.5e−16	5.6e−16	−1.3e−15	2.3e−20	0.0252	0.0221	1.75
168	7225	7828	14,282	28,561	28,560	−1.8e−16	−1.8e−16	−1.4e−15	2.5e−20	0.0248	0.0214	1.73

Table 9 Spherical t-designs on $\mathbb{S}^2$ with no symmetry restrictions, $N = \widehat{N}(2,t)$ and degrees $t = 169$–180

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
169	7310	7921	14,451	28,899	28,899	−1.8e−16	−7.2e−16	−1.3e−15	2.6e−20	0.0249	0.0225	1.81
170	7396	8014	14,622	29,241	29,240	−1.3e−16	−1.3e−16	−1.4e−15	2.7e−20	0.0263	0.0210	1.60
171	7482	8108	14,793	29,583	29,583	−1.3e−16	6.2e−16	−1.4e−15	2.8e−20	0.0261	0.0209	1.60
172	7569	8202	14,966	29,929	29,928	3.5e−16	3.5e−16	−1.4e−15	2.9e−20	0.0242	0.0212	1.75
173	7656	8297	15,139	30,275	30,275	3.4e−16	2.9e−16	−1.3e−15	3.0e−20	0.0243	0.0212	1.74
174	7744	8392	15,314	30,625	30,624	1.4e−16	1.6e−16	−1.4e−15	3.0e−20	0.0244	0.0211	1.72
175	7832	8488	15,489	30,975	30,975	1.3e−16	−1.5e−17	−1.5e−15	3.3e−20	0.0241	0.0208	1.72
176	7921	8584	15,666	31,329	31,328	1.2e−16	−6.0e−18	−1.4e−15	3.5e−20	0.0238	0.0206	1.73
177	8010	8681	15,843	31,683	31,683	1.3e−16	2.2e−16	−1.3e−15	3.4e−20	0.0238	0.0212	1.78
178	8100	8779	16,022	32,041	32,040	−2.3e−16	−1.4e−16	−1.2e−15	3.6e−20	0.0233	0.0202	1.74
179	8190	8877	16,201	32,399	32,399	−2.3e−16	2.2e−16	−1.2e−15	3.6e−20	0.0239	0.0202	1.69
180	8281	8976	16,382	32,761	32,760	−7.8e−16	−6.7e−16	−1.3e−15	3.7e−20	0.0236	0.0202	1.72

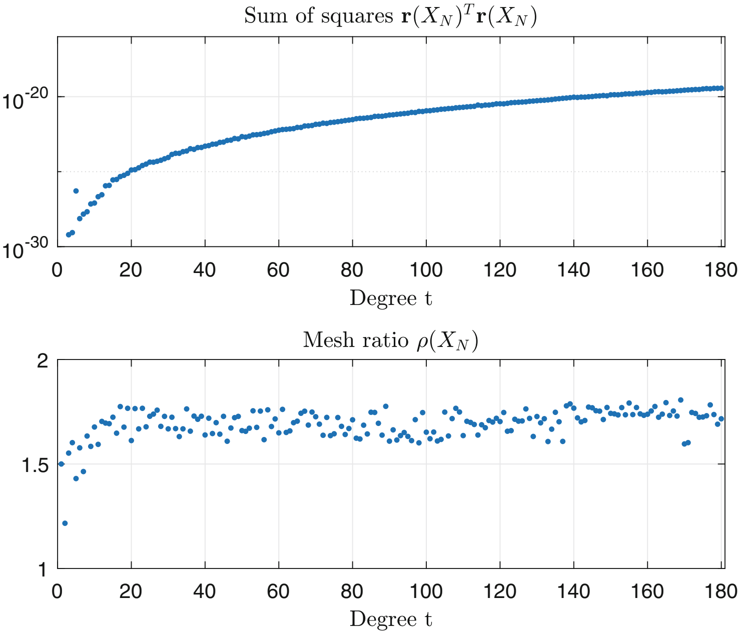


Fig. 2 Sum of squares of Weyl sums and mesh ratios for spherical t-designs on $\mathbb{S}^2$

of terms m in the Weyl sums and hence the unscaled sum of squares (39). Tables 10, 11, 12, 13, 14, 15 and 16 list the characteristics of the calculated t-designs for $t = 1, 3, 5, \ldots, 325$, as a symmetric $2k$-design is automatically a $2k + 1$-design. These tables have $t = \overline{N}(2, t)$ except for $t = 1, 7, 11$. These point sets, again available from [66], provide excellent sets of points for numerical integration on $\mathbb{S}^2$ with mesh ratios all less than 1.78 for degrees up to 325, as illustrated in Fig. 3.

5.3 Designs for d = 3

For $d = 3$, $Z(3, \ell) = (\ell + 1)^2$, so the dimension of the space of polynomials of degree at most t in $\mathbb{S}^3$ is $D(3, t) = Z(4, t) = (t+1)(t+2)(2t+3)/6$. Comparing the number of variables with the number of conditions, with no symmetry restrictions, gives

$$\widehat{N}(3, t) = \left\lceil \frac{2t^3 + 9t^2 + 13t + 36}{18} \right\rceil,$$

Table 10 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees $t = 1$–45, except for $t = 1$ when $N = \overline{N}(2,t) - 2$ and $t = 7, 11$ when $N = \overline{N}(2,t) + 2$

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
1	2	2	2	0	0	0.0e+00	0.0e+00	0.0e+00	0.0e+00	3.1416	1.5708	1.00
3	6	6	6	3	5	3.5e−17	0.0e+00	0.0e+00	6.1e−31	1.5708	0.9553	1.22
5	12	12	12	9	14	6.9e−18	1.5e−17	−8.5e−17	5.2e−30	1.1071	0.6524	1.18
7	20	20	32	29	27	2.1e−17	4.0e−18	1.9e−16	3.4e−28	0.5863	0.4480	1.53
9	30	31	48	45	44	4.2e−18	−4.1e−18	−1.9e−16	1.2e−25	0.4611	0.3860	1.67
11	42	43	70	67	65	−2.4e−17	−1.9e−18	−1.4e−16	8.6e−27	0.3794	0.3017	1.59
13	56	58	94	91	90	−8.4e−18	4.8e−18	−1.3e−16	1.6e−25	0.3146	0.2639	1.68
15	72	75	120	117	119	7.9e−18	8.0e−18	8.6e−17	3.9e−26	0.2900	0.2352	1.62
17	90	94	156	153	152	8.7e−18	1.0e−17	−6.0e−17	9.1e−26	0.2457	0.2039	1.66
19	110	115	192	189	189	−5.4e−18	−9.2e−19	−1.3e−16	7.8e−25	0.2248	0.1899	1.69
21	132	139	234	231	230	2.2e−17	−7.2e−19	1.2e−16	3.9e−25	0.2009	0.1689	1.68
23	156	164	278	275	275	−3.3e−18	2.1e−18	8.0e−17	7.5e−25	0.1822	0.1548	1.70
25	182	192	328	325	324	−9.5e−18	1.1e−17	−2.4e−16	9.0e−25	0.1722	0.1421	1.65
27	210	222	380	377	377	−4.6e−19	2.0e−17	1.9e−17	2.4e−24	0.1567	0.1328	1.70
29	240	254	438	435	434	5.8e−18	2.2e−17	−9.2e−17	2.6e−24	0.1448	0.1229	1.70
31	272	289	498	495	495	−3.0e−17	−3.2e−18	−3.7e−16	4.6e−24	0.1376	0.1151	1.67
33	306	325	564	561	560	−4.0e−17	2.3e−17	−3.5e−16	5.4e−24	0.1292	0.1075	1.66
35	342	364	632	629	629	2.1e−17	−3.8e−18	−4.2e−16	8.3e−24	0.1197	0.1011	1.69
37	380	405	706	703	702	3.8e−17	−5.0e−18	−2.6e−16	9.9e−24	0.1165	0.0968	1.66
39	420	448	782	779	779	2.6e−17	−6.2e−17	−3.2e−16	1.5e−23	0.1082	0.0910	1.68
41	462	493	864	861	860	9.7e−17	2.7e−17	−2.9e−16	2.2e−23	0.1025	0.0863	1.68
43	506	540	948	945	945	−7.1e−18	−4.5e−17	−2.6e−16	2.2e−23	0.0988	0.0824	1.67
45	552	590	1038	1035	1034	2.5e−17	−2.8e−17	−2.7e−16	3.2e−23	0.0936	0.0793	1.69

Table 11 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees t = 47–93

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
47	600	642	1130	1127	1127	7.9e−17	−4.6e−17	−2.3e−16	4.3e−23	0.0889	0.0758	1.71
49	650	696	1228	1225	1224	−3.5e−17	1.2e−16	−2.3e−16	5.0e−23	0.0856	0.0727	1.70
51	702	752	1328	1325	1325	2.8e−17	−1.2e−16	−2.0e−16	6.4e−23	0.0823	0.0706	1.72
53	756	810	1434	1431	1430	8.9e−17	−5.9e−17	−4.3e−16	9.1e−23	0.0799	0.0679	1.70
55	812	870	1542	1539	1539	−1.1e−17	5.5e−17	−3.2e−16	1.0e−22	0.0772	0.0664	1.72
57	870	933	1656	1653	1652	−7.2e−17	−1.2e−16	4.8e−19	1.2e−22	0.0742	0.0630	1.70
59	930	998	1772	1769	1769	−1.9e−16	−1.1e−16	−2.5e−16	1.5e−22	0.0722	0.0605	1.68
61	992	1065	1894	1891	1890	−6.1e−17	5.8e−17	−1.4e−16	2.1e−22	0.0701	0.0583	1.66
63	1056	1134	2018	2015	2015	2.8e−17	−3.4e−18	−9.0e−17	2.5e−22	0.0674	0.0568	1.68
65	1122	1206	2148	2145	2144	3.0e−17	−1.7e−16	−1.1e−16	2.6e−22	0.0664	0.0544	1.64
67	1190	1279	2280	2277	2277	−1.0e−16	−2.2e−16	9.8e−17	3.0e−22	0.0634	0.0541	1.71
69	1260	1355	2418	2415	2414	2.4e−16	−1.1e−16	−2.2e−16	3.5e−22	0.0616	0.0521	1.69
71	1332	1433	2558	2555	2555	7.1e−17	−1.7e−16	−1.8e−16	4.2e−22	0.0596	0.0514	1.72
73	1406	1513	2704	2701	2700	1.7e−16	6.8e−17	−2.9e−16	5.8e−22	0.0575	0.0495	1.72
75	1482	1595	2852	2849	2849	1.5e−16	−1.7e−16	−3.0e−16	6.2e−22	0.0569	0.0480	1.69
77	1560	1680	3006	3003	3002	2.6e−16	−2.0e−16	−1.0e−16	7.1e−22	0.0555	0.0463	1.67
79	1640	1766	3162	3159	3159	2.9e−16	2.2e−16	−2.1e−16	9.8e−22	0.0533	0.0452	1.70
81	1722	1855	3324	3321	3320	2.8e−16	2.1e−16	−2.1e−16	1.2e−21	0.0533	0.0446	1.67
83	1806	1946	3488	3485	3485	2.8e−16	−3.0e−16	−2.1e−16	1.2e−21	0.0506	0.0436	1.72
85	1892	2039	3658	3655	3654	−1.2e−17	9.3e−18	−5.1e−16	1.5e−21	0.0499	0.0419	1.68
87	1980	2135	3830	3827	3827	−1.8e−16	−1.5e−16	−4.6e−16	1.5e−21	0.0490	0.0416	1.70
89	2070	2232	4008	4005	4004	2.5e−16	−3.2e−16	−2.9e−16	1.7e−21	0.0473	0.0405	1.71
91	2162	2332	4188	4185	4185	−6.1e−17	2.3e−16	−3.3e−16	2.0e−21	0.0466	0.0395	1.70
93	2256	2434	4374	4371	4370	−2.6e−16	8.5e−17	−5.5e−16	2.2e−21	0.0460	0.0389	1.69

Table 12 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees t = 95–141

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
95	2352	2538	4562	4559	4559	8.8e−17	−1.4e−16	−6.0e−16	2.5e−21	0.0449	0.0380	1.69
97	2450	2644	4756	4753	4752	−1.3e−16	−1.6e−16	−5.7e−16	2.9e−21	0.0439	0.0371	1.69
99	2550	2753	4952	4949	4949	−1.6e−16	−2.3e−16	−5.8e−16	3.1e−21	0.0425	0.0365	1.72
101	2652	2863	5154	5151	5150	2.0e−16	−3.2e−16	−5.6e−16	3.3e−21	0.0422	0.0353	1.67
103	2756	2976	5358	5355	5355	−6.0e−17	1.3e−16	−6.2e−16	4.0e−21	0.0410	0.0349	1.70
105	2862	3091	5568	5565	5564	−2.2e−16	4.5e−17	−7.2e−16	4.6e−21	0.0403	0.0342	1.70
107	2970	3208	5780	5777	5777	1.9e−16	−1.5e−16	−8.4e−16	5.3e−21	0.0397	0.0341	1.72
109	3080	3327	5998	5995	5994	−4.1e−16	2.5e−16	−8.6e−16	5.9e−21	0.0386	0.0329	1.71
111	3192	3449	6218	6215	6215	1.3e−16	−1.9e−16	−6.7e−16	6.6e−21	0.0376	0.0324	1.72
113	3306	3573	6444	6441	6440	1.1e−16	−1.5e−16	−7.2e−16	7.2e−21	0.0377	0.0318	1.69
115	3422	3698	6672	6669	6669	2.2e−16	2.0e−16	−6.7e−16	8.8e−21	0.0375	0.0316	1.69
117	3540	3826	6906	6903	6902	−3.7e−16	1.6e−17	−5.6e−16	8.6e−21	0.0358	0.0309	1.73
119	3660	3957	7142	7139	7139	−4.0e−16	2.0e−16	−9.7e−16	9.5e−21	0.0356	0.0303	1.71
121	3782	4089	7384	7381	7380	−5.0e−16	−9.3e−17	−8.7e−16	1.1e−20	0.0351	0.0297	1.69
123	3906	4224	7628	7625	7625	1.7e−16	1.3e−16	−7.7e−16	1.2e−20	0.0349	0.0292	1.67
125	4032	4360	7878	7875	7874	−3.1e−17	−2.5e−16	−6.0e−16	1.4e−20	0.0339	0.0289	1.71
127	4160	4499	8130	8127	8127	−1.1e−16	−7.2e−18	−9.6e−16	1.5e−20	0.0335	0.0288	1.72
129	4290	4641	8388	8385	8384	6.9e−16	−6.4e−16	−1.0e−15	1.5e−20	0.0328	0.0277	1.69
131	4422	4784	8648	8645	8645	2.8e−16	−1.6e−16	−9.6e−16	1.7e−20	0.0320	0.0274	1.71
133	4556	4929	8914	8911	8910	2.3e−16	1.3e−16	−9.0e−16	2.0e−20	0.0317	0.0269	1.70
135	4692	5077	9182	9179	9179	4.3e−16	−2.8e−16	−1.0e−15	2.0e−20	0.0318	0.0267	1.68
137	4830	5227	9456	9453	9452	5.3e−16	2.7e−16	−9.0e−16	2.4e−20	0.0308	0.0261	1.70
139	4970	5379	9732	9729	9729	2.0e−16	−3.0e−16	−8.7e−16	2.4e−20	0.0303	0.0259	1.71
141	5112	5533	10,014	10,011	10,010	6.9e−17	−3.8e−16	−1.1e−15	2.8e−20	0.0300	0.0254	1.69

Table 13 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2, t)$ and odd degrees t = 143–189

t	$N^*(2, t)$	$N^+(2, t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
143	5256	5689	10,298	10,295	10,295	2.2e−16	6.0e−16	−1.0e−15	3.0e−20	0.0302	0.0255	1.69
145	5402	5848	10,588	10,585	10,584	−6.3e−16	1.4e−16	−6.7e−16	3.3e−20	0.0289	0.0248	1.72
147	5550	6009	10,880	10,877	10,877	6.5e−16	−4.2e−16	−1.2e−15	3.5e−20	0.0286	0.0246	1.72
149	5700	6172	11,178	11,175	11,174	−3.3e−16	−3.6e−16	−8.7e−16	3.7e−20	0.0287	0.0241	1.68
151	5852	6337	11,478	11,475	11,475	1.3e−16	5.5e−16	−1.0e−15	4.2e−20	0.0283	0.0239	1.69
153	6006	6504	11,784	11,781	11,780	−7.8e−16	4.8e−16	−1.1e−15	4.5e−20	0.0274	0.0235	1.71
155	6162	6674	12,092	12,089	12,089	−4.8e−16	2.7e−16	−1.0e−15	4.8e−20	0.0275	0.0234	1.70
157	6320	6845	12,406	12,403	12,402	4.3e−16	−2.4e−16	−1.1e−15	5.1e−20	0.0269	0.0232	1.72
159	6480	7019	12,722	12,719	12,719	8.6e−17	−5.6e−16	−1.3e−15	5.5e−20	0.0267	0.0229	1.71
161	6642	7195	13,044	13,041	13,040	−7.8e−16	2.3e−16	−1.2e−15	5.9e−20	0.0267	0.0225	1.69
163	6806	7373	13,368	13,365	13,365	1.8e−16	−6.5e−16	−1.3e−15	6.4e−20	0.0262	0.0222	1.69
165	6972	7553	13,698	13,695	13,694	−6.3e−16	−3.5e−16	−1.3e−15	7.0e−20	0.0257	0.0217	1.69
167	7140	7736	14,030	14,027	14,027	−9.0e−16	5.6e−16	−1.3e−15	7.1e−20	0.0255	0.0216	1.69
169	7310	7921	14,368	14,365	14,364	−1.8e−16	−7.8e−16	−1.2e−15	8.4e−20	0.0253	0.0214	1.69
171	7482	8108	14,708	14,705	14,705	−1.3e−16	6.2e−16	−1.2e−15	8.6e−20	0.0248	0.0210	1.70
173	7656	8297	15,054	15,051	15,050	3.9e−16	3.6e−16	−1.6e−15	9.3e−20	0.0246	0.0208	1.69
175	7832	8488	15,402	15,399	15,399	1.9e−16	−7.7e−17	−1.3e−15	1.0e−19	0.0247	0.0204	1.65
177	8010	8681	15,756	15,753	15,752	3.5e−17	2.2e−16	−1.3e−15	1.1e−19	0.0242	0.0205	1.69
179	8190	8877	16,112	16,109	16,109	−2.0e−16	2.8e−16	−1.3e−15	1.1e−19	0.0236	0.0202	1.71
181	8372	9075	16,474	16,471	16,470	−7.4e−16	−4.6e−17	−1.4e−15	1.2e−19	0.0234	0.0197	1.69
183	8556	9275	16,838	16,835	16,835	1.2e−16	4.5e−17	−1.4e−15	1.3e−19	0.0240	0.0198	1.65
185	8742	9477	17,208	17,205	17,204	−2.8e−17	−2.3e−16	−1.3e−15	1.3e−19	0.0224	0.0193	1.73
187	8930	9681	17,580	17,577	17,577	−1.1e−16	−7.5e−16	−1.8e−15	8.2e−20	0.0232	0.0192	1.65
189	9120	9888	17,958	17,955	17,954	−1.8e−16	−5.4e−16	−1.2e−15	1.6e−19	0.0222	0.0190	1.71

Table 14 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees $t = 191$–237

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
191	9312	10,096	18,338	18,335	18,335	2.6e−16	−2.5e−16	−1.3e−15	1.6e−19	0.0221	0.0188	1.70
193	9506	10,307	18,724	18,721	18,720	1.2e−15	−9.4e−16	−1.4e−15	1.8e−19	0.0223	0.0187	1.68
195	9702	10,520	19,112	19,109	19,109	−7.0e−16	5.6e−16	−1.4e−15	1.9e−19	0.0219	0.0187	1.71
197	9900	10,736	19,506	19,503	19,502	6.4e−16	−1.2e−16	−1.2e−15	2.0e−19	0.0218	0.0182	1.66
199	10,100	10,953	19,902	19,899	19,899	−6.3e−16	−3.4e−16	−1.3e−15	2.1e−19	0.0217	0.0179	1.65
201	10,302	11,173	20,304	20,301	20,300	1.2e−15	−2.6e−16	−1.2e−15	2.3e−19	0.0210	0.0178	1.70
203	10,506	11,394	20,708	20,705	20,705	5.0e−16	−3.9e−16	−1.2e−15	2.4e−19	0.0209	0.0178	1.70
205	10,712	11,618	21,118	21,115	21,114	−6.3e−17	−5.8e−17	−1.4e−15	2.5e−19	0.0206	0.0176	1.70
207	10,920	11,844	21,530	21,527	21,527	2.5e−16	−5.6e−16	−1.5e−15	2.6e−19	0.0202	0.0174	1.72
209	11,130	12,073	21,948	21,945	21,944	−1.1e−16	−1.0e−15	−1.2e−15	2.9e−19	0.0201	0.0172	1.71
211	11,342	12,303	22,368	22,365	22,365	9.1e−16	−8.9e−16	−1.4e−15	3.1e−19	0.0201	0.0171	1.70
213	11,556	12,536	22,794	22,791	22,790	−5.1e−16	−5.9e−16	−1.3e−15	3.3e−19	0.0195	0.0168	1.71
215	11,772	12,771	23,222	23,219	23,219	−3.4e−16	6.3e−16	−1.4e−15	3.5e−19	0.0201	0.0168	1.67
217	11,990	13,008	23,656	23,653	23,652	9.8e−17	2.0e−16	−1.5e−15	3.6e−19	0.0196	0.0166	1.69
219	12,210	13,247	24,092	24,089	24,089	2.3e−17	−3.1e−16	−1.4e−15	3.9e−19	0.0194	0.0166	1.71
221	12,432	13,488	24,534	24,531	24,530	−1.4e−15	2.6e−16	−1.4e−15	4.0e−19	0.0194	0.0163	1.68
223	12,656	13,732	24,978	24,975	24,975	−8.1e−17	−1.6e−16	−1.3e−15	4.4e−19	0.0192	0.0164	1.71
225	12,882	13,978	25,428	25,425	25,424	−4.9e−16	−5.3e−16	−1.3e−15	4.5e−19	0.0188	0.0160	1.70
227	13,110	14,226	25,880	25,877	25,877	1.5e−16	6.4e−16	−1.4e−15	4.7e−19	0.0185	0.0158	1.72
229	13,340	14,476	26,338	26,335	26,334	−1.0e−15	1.8e−16	−1.5e−15	5.1e−19	0.0183	0.0156	1.71
231	13,572	14,728	26,798	26,795	26,795	3.6e−17	−5.2e−16	−1.4e−15	5.3e−19	0.0181	0.0157	1.74
233	13,806	14,982	27,264	27,261	27,260	−6.2e−16	−2.4e−16	−1.5e−15	5.7e−19	0.0185	0.0155	1.68
235	14,042	15,239	27,732	27,729	27,729	−9.6e−16	7.0e−17	−1.4e−15	5.9e−19	0.0182	0.0153	1.69
237	14,280	15,498	28,206	28,203	28,202	1.1e−15	3.9e−16	−1.5e−15	6.1e−19	0.0178	0.0154	1.73

Table 15 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees $t = 239$–285

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)p$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
239	14,520	15,759	28,682	28,679	28,679	−1.5e−15	−3.9e−16	−1.5e−15	6.4e−19	0.0176	0.0152	1.72
241	14,762	16,022	29,164	29,161	29,160	−1.5e−15	3.8e−16	−1.4e−15	6.7e−19	0.0180	0.0153	1.70
243	15,006	16,287	29,648	29,645	29,645	8.7e−16	1.9e−16	−1.4e−15	7.2e−19	0.0171	0.0148	1.73
245	15,252	16,555	30,138	30,135	30,134	8.6e−16	−4.1e−16	−1.5e−15	7.4e−19	0.0174	0.0146	1.67
247	15,500	16,825	30,630	30,627	30,627	6.9e−16	−4.2e−16	−1.5e−15	7.9e−19	0.0173	0.0146	1.69
249	15,750	17,097	31,128	31,125	31,124	−5.1e−16	−2.6e−16	−1.6e−15	8.2e−19	0.0168	0.0146	1.74
251	16,002	17,371	31,628	31,625	31,625	−1.1e−15	4.7e−16	−1.6e−15	8.5e−19	0.0167	0.0143	1.71
253	16,256	17,647	32,134	32,131	32,130	9.8e−16	4.7e−16	−1.4e−15	9.2e−19	0.0170	0.0141	1.66
255	16,512	17,925	32,642	32,639	32,639	−2.5e−16	−8.6e−18	−1.8e−15	9.5e−19	0.0165	0.0142	1.72
257	16,770	18,206	33,156	33,153	33,152	−1.6e−15	6.6e−16	−1.7e−15	1.0e−18	0.0166	0.0141	1.70
259	17,030	18,489	33,672	33,669	33,669	−5.0e−16	−6.6e−16	−1.5e−15	1.1e−18	0.0163	0.0142	1.74
261	17,292	18,774	34,194	34,191	34,190	−9.0e−17	−9.6e−16	−1.6e−15	1.1e−18	0.0165	0.0139	1.69
263	17,556	19,061	34,718	34,715	34,715	7.3e−16	1.5e−15	−1.6e−15	1.1e−18	0.0160	0.0136	1.70
265	17,822	19,350	35,248	35,245	35,244	9.8e−16	−7.8e−16	−1.5e−15	1.2e−18	0.0162	0.0138	1.70
267	18,090	19,642	35,780	35,777	35,777	1.7e−16	7.5e−16	−1.7e−15	1.3e−18	0.0154	0.0134	1.74
269	18,360	19,935	36,318	36,315	36,314	−9.0e−16	−1.2e−15	−1.7e−15	1.3e−18	0.0160	0.0134	1.68
271	18,632	20,231	36,858	36,855	36,855	1.3e−15	1.5e−15	−1.9e−15	1.3e−18	0.0157	0.0137	1.75
273	18,906	20,529	37,404	37,401	37,400	−1.7e−15	5.4e−16	−1.7e−15	1.4e−18	0.0152	0.0132	1.73
275	19,182	20,830	37,952	37,949	37,949	1.6e−15	1.7e−16	−1.6e−15	1.5e−18	0.0152	0.0132	1.74
277	19,460	21,132	38,506	38,503	38,502	−4.1e−17	5.6e−16	−1.8e−15	1.6e−18	0.0153	0.0130	1.70
279	19,740	21,437	39,062	39,059	39,059	−9.5e−16	−5.9e−16	−1.9e−15	1.6e−18	0.0152	0.0132	1.75
281	20,022	21,743	39,624	39,621	39,620	−2.9e−16	−1.4e−15	−1.9e−15	1.7e−18	0.0150	0.0129	1.72
283	20,306	22,052	40,188	40,185	40,185	1.7e−15	1.3e−15	−1.9e−15	1.8e−18	0.0150	0.0126	1.69
285	20,592	22,363	40,758	40,755	40,754	−8.5e−16	−1.6e−15	−2.1e−15	1.8e−18	0.0148	0.0127	1.71

Table 16 Symmetric spherical t-designs on $\mathbb{S}^2$ with $N = \overline{N}(2,t)$ and odd degrees $t = 287$–325

t	$N^*(2,t)$	$N^+(2,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$f_t(X_N)p$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
287	20,880	22,677	41,330	41,327	41,327	6.9e−16	−9.4e−16	−1.5e−15	2.0e−18	0.0149	0.0126	1.70
289	21,170	22,992	41,908	41,905	41,904	−1.1e−15	−6.1e−16	−1.8e−15	2.0e−18	0.0148	0.0126	1.71
291	21,462	23,310	42,488	42,485	42,485	−4.0e−17	−1.5e−15	−1.9e−15	2.2e−18	0.0149	0.0124	1.66
293	21,756	23,630	43,074	43,071	43,070	9.5e−16	5.9e−16	−1.9e−15	2.2e−18	0.0142	0.0123	1.74
295	22,052	23,952	43,662	43,659	43,659	1.7e−15	1.2e−15	−1.5e−15	2.3e−18	0.0143	0.0124	1.74
297	22,350	24,276	44,256	44,253	44,252	−9.0e−17	7.9e−16	−1.6e−15	2.4e−18	0.0143	0.0121	1.70
299	22,650	24,602	44,852	44,849	44,849	−8.3e−16	1.4e−15	−1.7e−15	2.5e−18	0.0141	0.0121	1.72
301	22,952	24,931	45,454	45,451	45,450	5.8e−16	3.6e−17	−1.8e−15	2.6e−18	0.0140	0.0120	1.70
303	23,256	25,262	46,058	46,055	46,055	3.0e−16	1.0e−15	−1.8e−15	2.7e−18	0.0140	0.0118	1.70
305	23,562	25,595	46,668	46,665	46,664	5.8e−16	−1.2e−15	−1.9e−15	2.7e−18	0.0140	0.0118	1.68
307	23,870	25,930	47,280	47,277	47,277	−1.0e−15	−1.4e−15	−1.7e−15	2.9e−18	0.0139	0.0117	1.69
309	24,180	26,267	47,898	47,895	47,894	−1.8e−15	1.4e−15	−1.6e−15	2.9e−18	0.0136	0.0118	1.73
311	24,492	26,607	48,518	48,515	48,515	5.9e−16	2.3e−16	−2.0e−15	3.2e−18	0.0135	0.0116	1.72
313	24,806	26,948	49,144	49,141	49,140	1.6e−15	−9.6e−16	−1.8e−15	3.3e−18	0.0139	0.0115	1.66
315	25,122	27,292	49,772	49,769	49,769	−3.8e−17	1.0e−15	−2.0e−15	3.4e−18	0.0133	0.0114	1.72
317	25,440	27,638	50,406	50,403	50,402	−1.2e−15	9.1e−18	−1.8e−15	3.5e−18	0.0134	0.0114	1.69
319	25,760	27,986	51,042	51,039	51,039	7.8e−16	1.2e−15	−1.7e−15	3.6e−18	0.0134	0.0112	1.67
321	26,082	28,337	51,684	51,681	51,680	−2.0e−15	1.2e−15	−1.7e−15	3.7e−18	0.0133	0.0112	1.69
323	26,406	28,689	52,328	52,325	52,325	−3.2e−16	−1.2e−15	−1.7e−15	3.9e−18	0.0131	0.0115	1.76
325	26,732	29,044	52,978	52,975	52,974	1.2e−15	−1.7e−15	−1.7e−15	4.0e−18	0.0124	0.0110	1.77

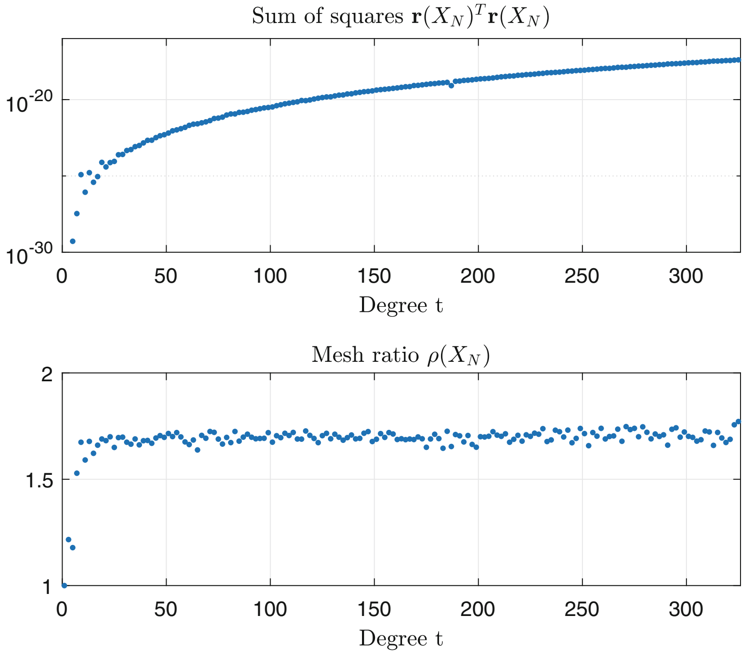


Fig. 3 Sum of squares of Weyl sums and mesh ratios for symmetric spherical t-designs on $\mathbb{S}^2$

while for symmetric spherical designs on $\mathbb{S}^3$

$$\overline{N}(3,t) = 2\left\lceil \frac{t^3 + 3t^2 + 2t + 30}{18} \right\rceil.$$

There are six regular convex polytopes with $N = 5, 8, 16, 24, 120$ and 600 vertices on S^3 [22] (the 5-cell, 16-cell, 8-cell, 24-cell, 600-cell and 120-cell respectively) giving spherical t-designs for $t = 2, 3, 5, 7, 9, 11$ and 11. The energy of regular sets on $\mathbb{S}^3$ with $N = 2, 3, 4, 5, 6, 8, 10, 12, 13, 24, 48$ has been studied by [1]. The $N = 24$ vertices of the D4 root system [20] provides a one-parameter family of 5-designs on $\mathbb{S}^3$. The Cartesian coordinates of the regular point sets are known, and these can be numerically verified to be spherical designs. The three variational criteria using (19), (21) and (23) are given for these point sets in Table 17. Figure 4 clearly illustrates the difference between the widely studied [11, 21, 24] inner product set $\mathscr{A}(X_N)$ for a regular point set (the 600-cell with $N = 120$) and a computed spherical 13-design with $N = 340$.

Table 17 Regular spherical t-designs on $\mathbb{S}^3$ for degrees $t = 1, 2, 3, 5, 7, 11$

t	$N^*(3,t)$	$N^+(3,t)$	$\widehat{N}(3,t)$	N	*sym*	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
1	2	2	4	2	1	5.6e−17	0.0e+00	0.0e+00	3.1416	1.5708	1.00
2	5	5	7	5	0	8.4e−17	4.4e−17	1.7e−16	1.8235	1.3181	1.45
3	8	8	12	8	1	1.5e−17	−1.1e−16	−2.6e−17	1.5708	1.0472	1.33
3	8	8	12	16	1	−1.4e−17	−9.6e−17	−1.5e−16	1.0472	1.0472	2.00
5	20	19	32	24	1	−4.7e−17	1.3e−16	−6.8e−17	1.0472	0.7854	1.50
7	40	40	70	48	1	2.8e−17	−1.2e−16	1.8e−16	0.7854	0.6086	1.55
11	112	117	219	120	1	−1.7e−17	−5.6e−17	−5.7e−16	0.6283	0.3881	1.24
11	112	117	219	600	1	2.9e−17	−3.7e−17	−8.1e−17	0.2709	0.3881	2.87

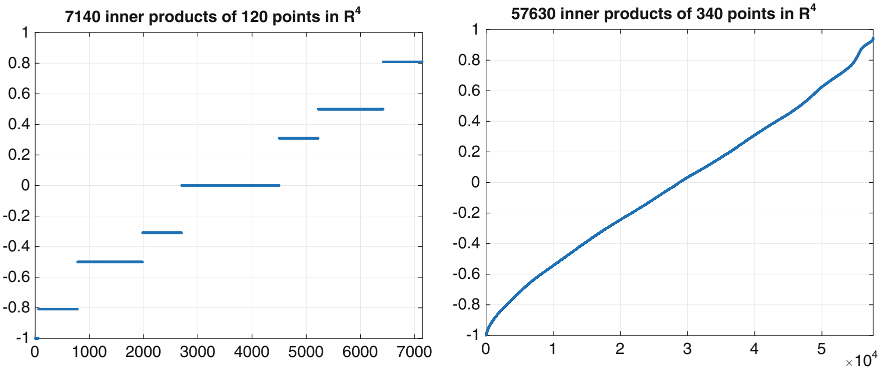


Fig. 4 Inner product sets $\mathscr{A}(X_N)$ for 600-cell with $N = 120$ and 13-design with $N = 340$ on $\mathbb{S}^3$

The results of some initial experiments in minimising the three variational criteria are given in Tables 18 and 19. These point sets, including any updates, are again available from [66]. For $d > 2$, it is more difficult to quickly generate a point set with a good mesh ratio to serve as an initial point for the optimization algorithms. One strategy is to randomly generate starting points, but this both makes the optimization problem harder and tends to produce nearby point sets which are local minimisers and have poor mesh ratios as the random initial points may have small separation [14]. Another possibility is the generalisation of equal area points to $d > 2$ by Leopardi [44]. For a given t and N there are still many different point sets with objective values close to 0 and different mesh ratios. To fully explore spherical t-designs for $d > 2$, a stable implementation of the spherical harmonics is needed, so that least squares minimisation can be fully utilised.

Acknowledgements This research includes extensive computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW Sydney.

Table 18 Computed spherical t-designs on $\mathbb{S}^3$ for degrees t = 1,...,20, with $N = \widehat{N}(3,t)$

t	$N^*(3,t)$	$N^+(3,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
1	2	2	4	6	4	6.2e−17	0.0e+00	0.0e+00	1.5708	1.5708	2.00
2	5	5	7	15	13	6.3e−17	5.6e−17	1.0e−16	1.3585	1.1683	1.72
3	8	8	12	30	29	8.1e−18	−9.1e−17	−7.9e−17	1.2311	1.0016	1.63
4	14	13	20	54	54	1.4e−19	5.6e−17	−2.4e−17	0.9414	0.8338	1.77
5	20	19	32	90	90	−3.0e−17	1.3e−16	2.6e−17	0.7816	0.7041	1.80
6	30	28	49	141	139	1.1e−17	1.0e−17	−9.2e−18	0.6883	0.6086	1.77
7	40	40	70	204	203	2.1e−17	−1.2e−16	−1.8e−16	0.5765	0.5454	1.89
8	55	54	97	285	284	−3.2e−17	1.6e−16	−1.5e−17	0.5028	0.4837	1.92
9	70	71	130	384	384	−1.4e−17	7.1e−17	−5.9e−17	0.4688	0.4467	1.91
10	91	92	171	507	505	1.2e−17	6.1e−17	−1.8e−16	0.4404	0.4082	1.85
11	112	117	219	651	649	4.1e−18	−5.2e−17	−2.9e−16	0.3809	0.3748	1.97
12	140	145	275	819	818	3.4e−17	−4.2e−17	−2.2e−16	0.3467	0.3409	1.97
13	168	178	340	1014	1014	3.7e−17	5.6e−17	−3.9e−17	0.3328	0.3225	1.94
14	204	216	415	1239	1239	4.9e−18	−2.0e−17	−1.5e−16	0.3111	0.2982	1.92
15	240	258	501	1497	1495	2.3e−18	1.6e−16	−9.1e−18	0.2909	0.2898	1.99
16	285	306	597	1785	1784	−1.5e−17	−7.6e−17	−1.1e−16	0.2673	0.2616	1.96
17	330	360	705	2109	2108	−3.0e−17	7.9e−17	−4.3e−17	0.2535	0.2507	1.98
18	385	419	825	2469	2469	1.2e−16	1.2e−16	1.1e−15	0.2386	0.2383	2.00
19	440	485	959	2871	2869	3.7e−17	5.5e−18	−6.0e−17	0.2283	0.2267	1.99
20	506	557	1106	3312	3310	1.9e−17	1.0e−16	1.2e−16	0.2163	0.2162	2.00

Table 19 Computed symmetric spherical t-designs on $\mathbb{S}^3$ for degrees t = 1,3,. . . ,31, with $N = \overline{N}(3,t)$

t	$N^*(3,t)$	$N^+(3,t)$	N	n	m	$V_{t,N,\psi_1}(X_N)$	$V_{t,N,\psi_2}(X_N)$	$V_{t,N,\psi_3}(X_N)$	$\delta(X_N)$	$h(X_N)$	$\rho(X_N)$
1	2	2	4	6	4	6.2e−17	0.0e+00	0.0e+00	1.5708	1.5708	2.00
3	8	8	10	30	29	−2.4e−17	−1.1e−16	−2.6e−16	1.3181	0.9776	1.48
5	20	19	28	90	90	−2.0e−17	1.3e−16	9.2e−17	0.8334	0.7303	1.75
7	40	40	60	204	203	5.7e−18	−1.3e−16	−1.1e−16	0.6324	0.5656	1.79
9	70	71	114	384	384	−2.7e−17	6.4e−17	−1.4e−16	0.4863	0.4548	1.87
11	112	117	194	651	649	1.7e−17	−4.7e−17	−2.2e−16	0.4126	0.3860	1.87
13	168	178	308	1014	1014	3.3e−17	5.5e−17	−1.1e−16	0.3454	0.3220	1.87
15	240	258	458	1497	1495	1.8e−18	1.6e−16	−2.7e−17	0.2914	0.2877	1.98
17	330	360	650	2109	2108	−2.9e−17	8.0e−17	−7.4e−17	0.2649	0.2584	1.95
19	440	485	890	2871	2869	3.5e−17	5.4e−18	−5.5e−17	0.2388	0.2380	1.99
21	572	636	1184	3795	3794	1.0e−17	1.4e−16	4.6e−17	0.2139	0.2113	1.98
23	728	816	1538	4899	4899	1.0e−17	−1.9e−16	−2.5e−16	0.1999	0.1951	1.95
25	910	1027	1954	6201	6200	−1.1e−17	5.3e−17	−2.0e−17	0.1795	0.1778	1.98
27	1120	1272	2440	7713	7713	−7.2e−18	8.9e−17	6.9e−17	0.1586	0.1678	2.12
29	1360	1553	3000	9456	9454	1.2e−15	3.4e−16	3.5e−14	0.1528	0.1565	2.05
31	1632	1872	3642	11,439	11,439	3.0e−16	−2.0e−17	9.5e−15	0.1438	0.1474	2.05

References

1. Agboola, D., Knol, A.L., Gill, P.M.W., Loos, P.F.: Uniform electron gases. III. Low density gases on three dimensional spheres. J. Chem. Phys. **143**(8), 084114-1–6 (2015)
2. Area, I., Dimitrov, D.K., Godoy, E., Ronveaux, A.: Zeros of Gegenbauer and Hermite polynomials and connection coefficients. Math. Comput. **73**(248), 1937–1951 (electronic) (2004)
3. Atkinson, K., Han, W.: Spherical Harmonics and Approximations on the Unit Sphere: An Introduction. Lecture Notes in Mathematics, vol. 2044. Springer, Heidelberg (2012)
4. Bachoc, C., Vallentin, F.: New upper bounds for kissing numbers from semidefinite programming. J. Am. Math. Soc. **21**(3), 909–924 (2008)
5. Bannai, E., Bannai, E.: A survey on spherical designs and algebraic combinatorics on spheres. Eur. J. Comb. **30**(6), 1392–1425 (2009)
6. Bannai, E., Damerell, R.M.: Tight spherical designs. I. J. Math. Soc. Japan **31**(1), 199–207 (1979)
7. Bannai, E., Damerell, R.M.: Tight spherical designs. II. J. Lond. Math. Soc. (2) **21**(1), 13–30 (1980)
8. Bondarenko, A., Radchenko, D., Viazovska, M.: Optimal asymptotic bounds for spherical designs. Ann. Math. (2) **178**(2), 443–452 (2013)
9. Bondarenko, A.V., Hardin, D.P., Saff, E.B.: Mesh ratios for best-packing and limits of minimal energy configurations. Acta Math. Hungar. **142**(1), 118–131 (2014)
10. Bondarenko, A., Radchenko, D., Viazovska, M.: Well-separated spherical designs. Constr. Approx. **41**(1), 93–112 (2015)
11. Boyvalenkov, P.G., Delchev, K.: On maximal antipodal spherical codes with few distances. Electron Notes Discrete Math. **57**, 85–90 (2017)
12. Boyvalenkov, P.G., Dragnev, P.D., Hardin, D.P., Saff, E.B., Stoyanova, M.M.: Universal upper and lower bounds on energy of spherical designs. Dolomites Res. Notes Approx. **8**(Special Issue), 51–65 (2015)
13. Brauchart, J.S., Saff, E.B., Sloan, I.H., Womersley, R.S.: QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere. Math. Comput. **83**(290), 2821–2851 (2014)
14. Brauchart, J.S., Reznikov, A.B., Saff, E.B., Sloan, I.H., Wang, Y.G., Womersley, R.S.: Random point sets on the sphere – hole radii, covering and separation. Exp. Math. **27**(1) , 62–81 (2018)
15. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.Y.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
16. Calef, M., Griffiths, W., Schulz, A.: Estimating the number of stable configurations for the generalized Thomson problem. J. Stat. Phys. **160**(1), 239–253 (2015)
17. Chen, X., Womersley, R.S.: Existence of solutions to systems of underdetermined equations and spherical designs. SIAM J. Numer. Anal. **44**(6), 2326–2341 (electronic) (2006)
18. Chen, X., Frommer, A., Lang, B.: Computational existence proofs for spherical t-designs. Numer. Math. **117**(2), 289–305 (2011)
19. Cohn, H., Kumar, A.: Universally optimal distribution of points on spheres. J. Am. Math. Soc. **20**(1), 99–148 (2007)
20. Cohn, H., Conway, J.H., Elkies, N.D., Kumar, A.: The D_4 root system is not universally optimal. Exp. Math. **16**(3), 313–320 (2007)
21. Conway, J.H., Sloane, N.J.A.: Sphere packings, lattices and groups. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 290, 3rd edn. Springer, New York (1999). With additional contributions by E. Bannai, R. E. Borcherds, J. Leech, S. P. Norton, A. M. Odlyzko, R. A. Parker, L. Queen and B. B. Venkov
22. Coxeter, H.S.M.: Regular Polytopes, 3rd edn. Dover Publications, Inc., New York (1973)
23. Damelin, S.B., Maymeskul, V.: On point energies, separation radius and mesh norm for s-extremal configurations on compact sets in $\mathbb{R}^n$. J. Complexity **21**(6), 845–863 (2005)
24. Delsarte, P., Goethals, J.M., Seidel, J.J.: Spherical codes and designs. Geom. Dedicata **6**(3), 363–388 (1977)

25. Demmel, J., Nguyen, H.D.: Parallel reproducible summation. IEEE Trans. Comput. **64**(7), 2060–2070 (2015)
26. Erber, T., Hockney, G.M.: Complex systems: equilibrium configurations of N equal charges on a sphere ($2 \leq N \leq 112$). In: Advances in Chemical Physics, vol. XCVIII, pp. 495–594. Wiley, New York (1997)
27. Fliege, J., Maier, U.: The distribution of points on the sphere and corresponding cubature formulae. IMA J. Numer. Anal. **19**(2), 317–334 (1999)
28. Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F., Ulerich, R.: GNU Scientific Library. https:://www.gnu.org/software.gsl/. Accessed 2016
29. Górski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M.: HEALPix: a framework for high resolution discretisation and fast analysis of data distributed on the sphere. Astrophys. J. **622**, 759–771 (2005)
30. Grabner, P., Sloan, I.H.: Lower bounds and separation for spherical designs. In: Uniform Distribution Theory and Applications, pp. 2887–2890. Mathematisches Forschungsinstitut Oberwolfach Report 49/2013 (2013)
31. Grabner, P.J., Tichy, R.F.: Spherical designs, discrepancy and numerical integration. Math. Comput. **60**(201), 327–336 (1993)
32. Gräf, M., Potts, D.: On the computation of spherical designs by a new optimization approach based on fast spherical Fourier transforms. Numer. Math. **119**(4), 699–724 (2011)
33. Hardin, R.H., Sloane, N.J.A.: McLaren's improved snub cube and other new spherical designs in three dimensions. Discret. Comput. Geom. **15**(4), 429–441 (1996)
34. Hardin, R.H., Sloane, N.J.A.: Spherical designs. http://neilsloane.com/sphdesigns/. Accessed 2017
35. Hardin, D.P., Saff, E.B., Whitehouse, J.T.: Quasi-uniformity of minimal weighted energy points on compact metric spaces. J. Complexity **28**(2), 177–191 (2012)
36. Hesse, K.: A lower bound for the worst-case cubature error on spheres of arbitrary dimension. Numer. Math. **103**(3), 413–433 (2006)
37. Hesse, K., Leopardi, P.: The Coulomb energy of spherical designs on S^2. Adv. Comput. Math. **28**(4), 331–354 (2008)
38. Hesse, K., Sloan, I.H.: Cubature over the sphere S^2 in Sobolev spaces of arbitrary order. J. Approx. Theory **141**(2), 118–133 (2006)
39. Hesse, K., Sloan, I.H.: Hyperinterpolation on the sphere. In: Frontiers in Interpolation and Approximation. Pure Appl. Math. (Boca Raton), vol. 282, pp. 213–248. Chapman & Hall/CRC, Boca Raton (2007)
40. Hesse, K., Sloan, I.H., Womersley, R.S.: Numerical integration on the sphere. In: Freeden, W., Nashed, M.Z., Sonar, T. (eds.) Handbook of Geomathematics, 1st edn., pp. 1187–1220. Springer, Heidelberg (2010)
41. Higham, N.J.: Accuracy and Stability of Numerical Algorithms. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1996)
42. Holmes, S.A., Featherstone, W.E.: A unified approach to the Clenshaw summation and the recursive computation of very high degree and order normalised associated Legendre functions. J. Geodesy **76**, 279–299 (2002)
43. Jekeli, C., Lee, J.K., Kwon, A.H.: On the computation and approximation of ultra-high degree spherical harmonic series. J. Geodesy **81**, 603–615 (2007)
44. Leopardi, P.: A partition of the unit sphere into regions of equal area and small diameter. Electron. Trans. Numer. Anal. **25**, 309–327 (electronic) (2006)
45. McLaren, A.D.: Optimal numerical integration on a sphere. Math. Comput. **17**, 361–383 (1963)
46. Morales, J.L., Nocedal, J.: Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". ACM Trans. Math. Softw. **38**(1), Art. 7, 4 (2011)
47. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.9 of 2014-08-29. Online companion to [49]
48. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)

49. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W. (eds.): NIST Handbook of Mathematical Functions. Cambridge University Press, New York (2010). Print companion to [47]
50. Parrilo, P.A., Sturmfels, B.: Minimizing polynomial functions. In: Algorithmic and Quantitative Real Algebraic Geometry (Piscataway, NJ, 2001). DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, vol. 60, pp. 83–99. American Mathematical Society, Providence (2003)
51. Ragozin, D.L.: Constructive polynomial approximation on spheres and projective spaces. Trans. Am. Math. Soc. **162**, 157–170 (1971)
52. Rakhmanov, E.A., Saff, E.B., Zhou, Y.M.: Minimal discrete energy on the sphere. Math. Res. Lett. **1**(6), 647–662 (1994)
53. Rankin, R.A.: The closest packing of spherical caps in n dimensions. Proc. Glasg. Math. Assoc. **2**, 139–144 (1955)
54. Sansone, G.: Orthogonal Functions. Interscience Publishers, Inc., New York (1959). Revised English ed. Translated from the Italian by A. H. Diamond; with a foreword by E. Hille. Pure and Applied Mathematics, vol. IX, Interscience Publishers, Ltd., London
55. Seymour, P.D., Zaslavsky, T.: Averaging sets: a generalization of mean values and spherical designs. Adv. Math. **52**(3), 213–240 (1984)
56. Sloan, I.H.: Polynomial interpolation and hyperinterpolation over general regions. J. Approx. Theory **83**(2), 238–254 (1995)
57. Sloan, I.H., Womersley, R.S.: Extremal systems of points and numerical integration on the sphere. Adv. Comput. Math. **21**(1–2), 107–125 (2004)
58. Sloan, I.H., Womersley, R.S.: A variational characterisation of spherical designs. J. Approx. Theory **159**(2), 308–318 (2009)
59. Sloan, I.H., Womersley, R.S.: Filtered hyperinterpolation: a constructive polynomial approximation on the sphere. Int. J. Geomath. **3**(1), 95–117 (2012)
60. Sobolev, S.L.: Cubature formulas on the sphere which are invariant under transformations of finite rotation groups. Dokl. Akad. Nauk SSSR **146**, 310–313 (1962)
61. Szegő, G.: Orthogonal Polynomials, 4th edn. American Mathematical Society, Providence (1975). American Mathematical Society, Colloquium Publications, vol. XXIII
62. Tammes, P.M.L.: On the origin of number and arrangement of places of exit on the surface of pollen grains. Recueil des Travaux Botanique Neerlandais **27**, 1–84 (1930)
63. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Rev. **38**(1), 49–95 (1996)
64. Wang, K., Li, L.: Harmonic Analysis and Approximation on the Unit Sphere. Science Press, Beijing (2006)
65. Wang, Y.G., Le Gia, Q.T., Sloan, I.H., Womersley, R.S.: Fully discrete needlet approximation on the sphere. Appl. Comput. Harmon. Anal. **43**, 292–316 (2017)
66. Womersley, R.S.: Efficient spherical designs with good geometric properties. http://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/ (2017)
67. Yudin, V.A.: Covering a sphere and extremal properties of orthogonal polynomials. Discret. Math. Appl. **5**(4), 371–379 (1995)
68. Yudin, V.A.: Lower bounds for spherical designs. Izv. Ross. Akad. Nauk Ser. Mat. **61**(3), 213–223 (1997)
69. Zhou, Y., Chen, X.: Spherical t_ϵ designs and approximation on the sphere. Math. Comput. (2018). http://dx.doi.org/10.1090/mcom/3306

Optimal Points for Cubature Rules and Polynomial Interpolation on a Square

Yuan Xu

Dedicated to Ian H. Sloan on the occasion of his 80th birthday.

Abstract The nodes of certain minimal cubature rule are real common zeros of a set of orthogonal polynomials of degree n. They often consist of a well distributed set of points and interpolation polynomials based on them have desired convergence behavior. We report what is known and the theory behind by explaining the situation when the domain of integrals is a square.

1 Introduction

A numerical integration rule is a finite linear combination of point evaluations that approximates an integral. The degree of precision of such a rule is the highest total degree of polynomials that are evaluated exactly. For a fixed degree of precision, the minimal rule uses the smallest number of point evaluations. Finding a minimal rule is a difficult problem and the most challenging part lies in identifying the set of nodes used in the rule, which is often a desirable set of points for polynomial interpolation. For integration on subsets of the real line, a Gaussian quadrature rule is minimal; its nodes are known to be zeros of orthogonal polynomials and polynomial interpolation based on the nodes has desired convergence behavior. The problem is far less understood in higher dimension, where we have fewer answers and many open questions. The purpose of this paper is to explain the situation when the integral domain is a square on the plane, for which we know more than on any other domain.

We can work with any fixed square and will fix our choice as

$$\square := [-1, 1]^2$$

Y. Xu (✉)
Department of Mathematics, University of Oregon, Eugene, OR, USA
e-mail: yuan@uoregon.edu

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_58

throughout the paper. Let Π_n^2 denote the space of polynomials of (total) degree at most n in two real variables, where the total degree means the sum of degrees in both variables. It is known that $\dim \Pi_n^2 = (n+1)(n+2)/2$. Let W be a nonnegative weight function on the square. For the integral with respect to W, a cubature rule of degree of precision m (abbreviated as *degree m* from now on) is a finite sum, defined below, such that

$$\int_{\square} f(x,y)W(x,y)dxdy = \sum_{k=1}^{N} \lambda_k f(x_k,y_k), \qquad \text{for all } f \in \Pi_m^2, \tag{1}$$

and there exists at least one $f \in \Pi_{m+1}^2$ such that the equality (1) fails to hold. The integer N is the number of nodes. The points $(x_k, y_k) \in \mathbb{R}^2$ are called *nodes* and the numbers λ_k are called *weights* of the cubature rule, respectively. We consider only *positive* cubature rules for which λ_k are all positive.

As in the case of one variable, the nodes of a minimal cubature rule are closely related to the zeros of orthogonal polynomials. A polynomial P is an orthogonal polynomial of degree n with respect to the weight function W if $P \in \Pi_n^2$ and

$$\int_{\square} P(x,y)Q(x,y)W(x,y)dxdy = 0 \qquad \text{for all } Q \in \Pi_{n-1}^2.$$

Let $\mathscr{V}_n(W)$ denote the space of orthogonal polynomials of degree n. Then

$$\dim \mathscr{V}_n(W) = n + 1,$$

as can be seen by applying the Gram-Schmidt process to $x^n, x^{n-1}y, \ldots, xy^{n-1}, y^n$. However, the structure of zeros for polynomials of more than one variable can be complicated and what is needed is the common zeros of a family of orthogonal polynomials of degree n. A common zero of a family of polynomials is a point which is a zero for every polynomial in the family. To be more precise, what we often need is to identify a polynomial ideal, I, generated by a family of orthogonal polynomials in $\mathscr{V}_n(W)$, so that its variety, V, is real and zero-dimensional, and the cardinality of V equals the codimension of I. Given the status of real algebraic geometry, this is difficult in general. Only in a few cases can we establish the existence of a minimal, or near minimal, cubature rule and identify its generating polynomial ideal explicitly. The nodes of such a cubature rule are good points for polynomial interpolation. Indeed, using the knowledge on orthogonal polynomial that vanish on the nodes, it is not difficult to construct a polynomial subspace Π_n^*, so that the problem of finding p such that $p(x_i, y_i) = f(x_i, y_i)$ for all nodes (x_i, y_i) of the cubature rule has a unique solution in Π_n^*. Moreover, this interpolation polynomial is easy to compute and has desirable convergence behavior. The above rough description applies to all cubature rules. Restricting to the square allows us to describe the idea and results without becoming overly tangled by notations.

The minimal or near-minimal cubature rules offer highly efficient tools for high-precision computation of integrals. It is unlikely, however, that they will become a major tool for numerical integration any time soon, because we do not know how to construct them in most cases. Moreover, their usage is likely restricted to lower dimension integrals, since they are even less understood in higher dimensions, where the difficulty increases rapidly as the dimension goes up, and, one could also add, truly high-dimensional numerical integration is really a different problem (see, for example, [7]). Nevertheless, with their deep connection to other fields in mathematics and their promise as high dimensional substitute for Gaussian quadrature rules, minimal cubature rules are a fascinating object to study. It is our hope that this paper will help attract researchers into this topic.

The paper is organized as follows. We review the theoretic results in the following section. In Sect. 3, we discuss minimal and near minimal cubature rules for the Chebyshev weight functions on the square, which includes a discussion on the Padua points. In Sect. 4, we discuss more recent extensions of the results in previous section to a family of weight functions that have a singularity on the diagonal of the square. Finally, in Sect. 5, we describe how cubature rules of lower degrees can be established for unit weight function on the square.

2 Cubature Rules and Interpolation

We are interested in integrals with respect to a fixed weight function W over the square, as in (1), and we assume that all moments of W are finite. A typical example of W is the product weight function

$$W_{\alpha,\beta}(x,y) := (1-x^2)^\alpha(1-y^2)^\beta, \qquad \alpha,\beta > -1.$$

This weight function is centrally symmetric, which means that it is symmetric with respect to the origin; more precisely, it satisfies $W(x,y) = W(-x,-y)$. If we replace $(1-x^2)^\alpha$ by $(1-x)^\alpha(1+x)^\gamma$, with $\gamma \neq \alpha$, the resulting weight function will not be centrally symmetric.

Many of the results below hold for cubature rules with respect to integrals on all domains in the plane, not just for the square. We start with the first lower bound for the number of nodes of cubature rules [23].

Theorem 1 *Let n be a positive integer and let $m = 2n-1$ or $2n-2$. If the cubature rule* (1) *is of degree m, then its number of nodes satisfies*

$$N \geq \dim \Pi_{n-1}^2 = \frac{n(n+1)}{2}. \tag{2}$$

A cubature rule of degree m is called Gaussian if the lower bound (2) is attained. In the one-dimensional case, it is well-known that the Gaussian quadrature rule of degree $2n-1$ has $n = \dim \Pi_{n-1}$ nodes, where Π_n denote the space of polynomials of degree at most n in one variable, and the same number of nodes is needed for the quadrature rule of degree $2n-2$.

For $n = 0, 1, 2, \ldots$, let $\{P_k^n : 0 \le k \le n\}$ be a basis of $\mathscr{V}_n(W)$. We denote by $\mathbb{P}_n$ the set of this basis and we also regard $\mathbb{P}_n$ as a column vector

$$\mathbb{P}_n = (P_0^n, P_1^n, \ldots, P_n^n)^{\mathsf{t}},$$

where the superscript t denotes the transpose. The Gaussian cubature rules can be characterized as follows:

Theorem 2 *Let $\mathbb{P}_s$ be a basis of $\mathscr{V}_s(W)$ for $s = n$ and $n-1$. Then*

1. *A Gaussian cubature rule* (1) *of degree* $2n-1$ *exists if, and only if, its nodes are common zeros of the polynomials in* $\mathbb{P}_n$*;*
2. *A Gaussian cubature rule* (1) *of degree* $2n-2$ *exists if, and only if, its nodes are common zeros of the polynomials in*

$$\mathbb{P}_n + \Gamma\, \mathbb{P}_{n-1},$$

where Γ is a real matrix of size $(n+1) \times n$.

For $m = 2n-1$, the characterization is classical and established in [19]; see also [8, 20]. For $m = 2n-2$, the characterization was established in [18, 21, 27]. As in the classical Gaussian quadrature rules, a Gaussian cubature rule, if it exists, can be derived from integrating the Lagrange interpolation based on its nodes [25].

Let $(x_k, y_k) : 1 \le k \le \dim \Pi_{n-1}^2$ be distinct points in $\mathbb{R}^2$. The Lagrange interpolation polynomial, denoted by $L_n f$, is a polynomial of degree n, such that

$$L_n f(x_k, y_k) = f(x_k, y_k), \qquad 1 \le k \le \dim \Pi_{n-1}^2.$$

If (x_k, y_k) are zeros of a Gaussian cubature rule, then the Lagrange interpolation polynomial is uniquely determined. Moreover, let $K_n(\cdot,\cdot)$ be the reproducing kernel of the space $\mathscr{V}_n(W)$, which can be written as

$$K_n((x,y),(x',y')) := \sum_{m=0}^{n} \sum_{k=0}^{m} P_k^m(x,y) P_k^m(x',y'),$$

where $\{P_k^m : 0 \le k \le m\}$ is an orthonormal basis of $\mathscr{V}_m(W)$; then the Lagrange interpolation polynomial based on the nodes (x_k, y_k) of the Gaussian cubature rule can be written as

$$L_n f(x,y) = \sum_{k=0}^{N} f(x_k, y_k) \ell_{k,n}(x,y), \qquad \ell_{k,n} := \frac{K_{n-1}((x,y),(x_k,y_k))}{K_{n-1}((x_k,y_k),(x_k,y_k))},$$

where λ_k are the cubature weights; moreover, $\lambda_k = 1/K_{n-1}((x_k, y_k), (x_k, y_k))$ is clearly positive.

Another characterization, more explicit, of the Gaussian cubature rules can be given in terms of the coefficient matrices of the three-term relations satisfied by the orthogonal polynomials.

For $n = 0, 1, 2, \ldots$, let $\{P_k^n : 0 \leq k \leq n\}$ be an orthonormal basis of $\mathscr{V}_n(W)$. Then there exist matrices $A_{n,i} : (n+1) \times (n+2)$ and $B_{n,i} : (n+1) \times (n+1)$ such that [8],

$$x_i \mathbb{P}_n(x) = A_{n,i}\mathbb{P}_{n+1}(x) + B_{n,i}\mathbb{P}_n(x) + A_{n-1,i}^{\mathrm{t}}\mathbb{P}_{n-1}(x), \quad x = (x_1, x_2), \tag{3}$$

for $i = 1, 2$. The coefficient matrices $B_{n,i}$ are necessarily symmetric. Furthermore, it is known that $B_{n,i} = 0$ if W is centrally symmetric.

Theorem 3 *Let $n \in \mathbb{N}$, Assume that the cubature rule* (1) *is of degree* $2n-1$.

1. *The number of nodes of the cubature rule satisfies*

$$N \geq \dim \Pi_{n-1}^2 + \frac{1}{2}\operatorname{rank}(A_{n-1,1}A_{n-1,2}^{\mathrm{t}} - A_{n-1,2}A_{n-1,1}^{\mathrm{t}}). \tag{4}$$

2. *The cubature is Gaussian if, and only if,* $A_{n-1,1}A_{n-1,2}^{\mathrm{t}} = A_{n-1,2}A_{n-1,1}^{\mathrm{t}}$.
3. *If W is centrally symmetric, then* (4) *becomes*

$$N \geq \dim \Pi_{n-1}^2 + \left\lfloor \frac{n}{2} \right\rfloor = \frac{n(n+1)}{2} + \left\lfloor \frac{n}{2} \right\rfloor =: N_{\min}. \tag{5}$$

In particular, Gaussian cubature rules do not exist for centrally symmetric weight functions.

The lower bound (5) was established by Möller in his thesis (see [17]). The more general lower bound (4) was established in [26], which reduces to (5) when W is centrally symmetric. The non-existence of the Gaussian cubature rule of degree $2n-1$ for centrally symmetric weight functions motivates the consideration of *minimal cubature rules*, defined as the cubature rule(s) with the smallest number of nodes among all cubature rules of the same degree for the same integral. Evidently, the existence of a minimal cubature rule is a tautology of its definition.

Cubature rules of degree $2n-1$ that attain Möller's lower bound $N_{\min}$ in (5) can be characterize in terms of common zeros of orthogonal polynomials as well.

Theorem 4 *Let W be centrally symmetric. A cubature rule of degree* $2n-1$ *attains Möller's lower bound* (5) *if, and only if, its nodes are common zeros of* $(n+1) - \left\lfloor \frac{n}{2} \right\rfloor$ *many orthogonal polynomials of degree n in* $\mathscr{V}_n(W)$.

This theorem was established in [17]. In the language of polynomial ideal and variety, we say that the nodes of the cubature rule are the variety of a polynomial ideal generated by $\left\lfloor \frac{n+1}{2} \right\rfloor + 1$ many orthogonal polynomials of degree n. More general results of this nature were developed in [26], which shows, in particular, that a cubature rule of degree $2n-1$ with $N = N_{\min} + 1$ exists if its nodes are common zeros of $\lfloor \frac{n+1}{2} \rfloor$ many orthogonal polynomials of degree n in $\mathscr{V}_n(W)$.

These cubature rules can also be derived from integrating their corresponding interpolating polynomials. However, since $N_{\min}$ is not equal to the dimension of Π^2_{n-1}, we need to define an appropriate polynomial subspace in order to guarantee that the Lagrange interpolant is unique. Assume that a cubature rule of degree $2n-1$ with $N = N_{\min}$ exists. Let $\sigma = \lfloor \frac{n}{2} \rfloor$ and let $\mathscr{P}_n := \{P_1, \ldots, P_{n-\sigma}\}$ be the set of orthogonal polynomials whose common zeros are the nodes of the cubature rule. We can assume, without loss of generality, that these polynomials are mutually orthogonal and they form an orthonormal subset of $\mathscr{V}_n(W)$. Let $\mathscr{Q}_n := \{Q_1, \ldots, Q_\sigma\}$ be an orthonormal basis of $\mathscr{V}_n(W) \setminus \operatorname{span} \mathscr{P}_n$, so that $\mathscr{P}_n \cup \mathscr{Q}_n$ is an orthonormal basis of $\mathscr{V}_n(W)$. Then it is shown in [26] that there is a unique polynomial in the space

$$\Pi_n^* := \Pi^2_{n-1} \cup \operatorname{span} \mathscr{Q}_n \tag{6}$$

that interpolates a generic function f on the nodes of the minimal cubature rule; that is, there is a unique polynomial $L_n f \in \Pi_n^*$ such that

$$L_n f(x_k, y_k) = f(x_k, y_k), \qquad 1 \le k \le N_{\min},$$

where (x_k, y_k) are zeros of the minimal cubature rule. Furthermore, this polynomial can be written as

$$L_n f(x,y) = \sum_{k=0}^{N} f(x_k, y_k) \ell_{k,n}(x,y), \qquad \ell_{k,n} := \frac{K_n^*((x,y),(x_k,y_k))}{K_n^*((x_k,y_k),(x_k,y_k))}, \tag{7}$$

where

$$K_n^*((x,y),(x',y')) = K_{n-1}((x,y),(x',y')) + \sum_{j=1}^{\sigma} Q_j(x,y) Q_j(x',y'). \tag{8}$$

Integrating $L_n f$ gives a cubature rule with $N_{\min}$ nodes that is exact for all polynomials in Π^2_{2n-1} and, in particular, $\lambda_k = 1/K_n^*((x_k,y_k),(x_k,y_k))$. Furthermore, the above relation between cubature rules and interpolation polynomials hold if $\sigma = \lfloor \frac{n}{2} \rfloor + 1$ and the cubature rule has $N_{\min} + 1$ points.

All our examples are given for cubature rules for centrally symmetric cases. We are interested in cubature rules that either attain or nearly attain the lower bounds, which means Gaussian cubature of degree $2n-2$ or cubature rules of degree $2n-1$ with $N_{\min}$ nodes or $N_{\min} + 1$ nodes. When such a cubature rule exists, the Lagrange interpolation polynomials based on its nodes possesses good, close to optimal, approximation behavior.

Because our main interest lies in the existence of our cubature rules and the convergence behavior of our interpolation polynomials, we shall not state cubature weights, λ_k in (1), nor explicit formulas for the interpolation polynomials throughout this paper. For all cases that we shall encounter below, these cubature weights can be stated explicitly in terms of known quantities and interpolation polynomials can be written down in closed forms, which can be found in the references that we provide.

3 Results for Chebyshev Weight Function

We start with the product Gegenbauer weight function defined on $[-1, 1]^2$ by

$$W_\lambda(x, y) = (1 - x^2)^{\lambda-1/2}(1 - y^2)^{\lambda-1/2}, \qquad \lambda > -1/2.$$

The cases $\lambda = 0$ and $\lambda = 1$ are the Chebyshev weight functions of the first and the second kind, respectively. One mutually orthogonal basis of $\mathscr{V}_n(W)$ is given by

$$P_k^n(x, y) := C_{n-k}^\lambda(x)C_k^\lambda(y), \qquad 0 \le k \le n,$$

where C_n^λ denotes the usual Gegenbauer polynomial of degree n. When $\lambda = 0$, C_n^λ is replaced by T_n, the Chebyshev polynomial of the first kind, and when $\lambda = 1$, $C_n^\lambda = U_n$, the Chebyshev polynomial of the second kind. Setting $x = \cos\theta$, we have

$$T_n(x) = \cos n\theta \quad \text{and} \quad U_n(x) = \frac{\sin(n + 1)\theta}{\sin\theta}.$$

In the following we always assume that $C_k^\lambda(x) = U_k(x) = T_k(x) = 0$ if $k < 0$.

The first examples of minimal cubature rules were given for Chebyshev weight functions soon after [17]. We start with the Gaussian cubature rules for Chebyshev weight function of the second type in [18].

Theorem 5 *For the product Chebyshev weight function W_1 of the second kind, the Gaussian cubature rules of degree $2n - 2$ exist. Their nodes can be explicitly given by*

$$\begin{aligned} &(\cos \tfrac{2i\pi}{n+2}, \cos \tfrac{(2j-1)\pi}{n+1}), \quad 1 \le i \le (n + 1)/2, \quad 1 \le j \le (n + 1)/2, \\ &(\cos \tfrac{(2i-1)\pi}{n+2}, \cos \tfrac{2j\pi}{n+1}), \quad 1 \le i \le n/2 + 1, \quad 1 \le j \le n/2, \end{aligned} \tag{9}$$

which are common zeros of the polynomials

$$U_{n-k}(x)U_k(y) - U_k(x)U_{n-1-k}(y), \qquad 0 \le k \le n.$$

However, W_1 remains the only weight function on the square for which the Gaussian cubature rules of degree $2n - 2$ are known to exist for all n. For other weight functions, for example, the constant weight function $W_{1/2}(x, y) = 1$, the existence is known only for small n; see the discussion in the last section.

For minimal cubature of degree $2n-1$ that attains Möller's lower bound, we are in better position. The first result is again known for Chebyshev weight functions.

Theorem 6 *For the product Chebyshev weight function W_0 of the first kind, the cubature rules of degree $2n-1$ that attain the lower bound* (5) *exist. Moreover, for $n=2m$, their nodes can be explicitly given by*

$$\begin{aligned} &(\cos \tfrac{i\pi}{m}, \cos \tfrac{(2j+1)\pi}{2m}), && 0 \le i \le m, \quad 0 \le j \le m-1, \\ &(\cos \tfrac{(2i+1)\pi}{m}, \cos \tfrac{j\pi}{m}), && 0 \le i \le m, \quad 1 \le j \le m, \end{aligned} \tag{10}$$

which are common zeros of the polynomials

$$T_{2m-k+1}(x)T_{m-1}(y) - T_{k-1}(x)T_{m-k+1}(y), \qquad 1 \le k \le m+1.$$

For $n=2m$, this was first established in [18], using the characterization in [17], and it was later proved by other methods [1, 16]. The case for $n=2m-1$ is established more recently in [31], for which the structure of orthogonal polynomials that vanish on the nodes is more complicated, see the discussion after Theorem 4.2. The analog of the explicit construction in the case $n=2m$ holds for cubature rules of degree $2n-1$, with $n=2m-1$, that have one more node than the lower bound (5) [26]. The nodes of these formulas are by Xu [28]

$$\begin{aligned} &(\cos \tfrac{2i\pi}{2m-1}, \cos \tfrac{2j\pi}{2m-1}), && 0 \le k \le m-1,\ 0 \le j \le m-1, \\ &(\cos \tfrac{(2m-2i-1)\pi}{2m-1}, \cos \tfrac{(2m-j-1)\pi}{2m-1}), && 0 \le i \le m-1,\ 1 \le j \le m-1, \end{aligned} \tag{11}$$

and they are common zeros of the polynomials

$$T_{2m-k}(x)T_{k-1}(y) - T_{k-1}(x)T_{2m-k}(y), \qquad 1 \le k \le m.$$

These points are well distributed. Two examples are depicted in Fig. 1.

The Lagrange interpolation polynomials based on the nodes of these cubature rules were first studied in [28]. Let $L_n f(x,y)$ denote the Lagrange interpolation polynomial based on the nodes (10) for $n=2m$ and on (11) for $n=2m-1$, which belongs to the space Π_n^* defined in (6). Using the Christoffel-Darboux formula in two variables, these interpolation polynomials can be given explicitly. Their convergence behavior is about optimal among all interpolation polynomials on the square. To be more precise, we introduce the following notation.

Let $\|\cdot\|_p$ denote the usual L^p norm of the space $L^p(\square, W_0)$ for $1 \le p < \infty$, and define it as the uniform norm on the square $\square$ when $p=\infty$. For $f \in C(\square)$, let $E_n(f)_\infty$ be the error of best approximation by polynomials from Π_n^2 in the uniform norm; that is,

$$E_n(f)_\infty = \inf_{P \in \Pi_n^2} \|f-P\|_\infty.$$

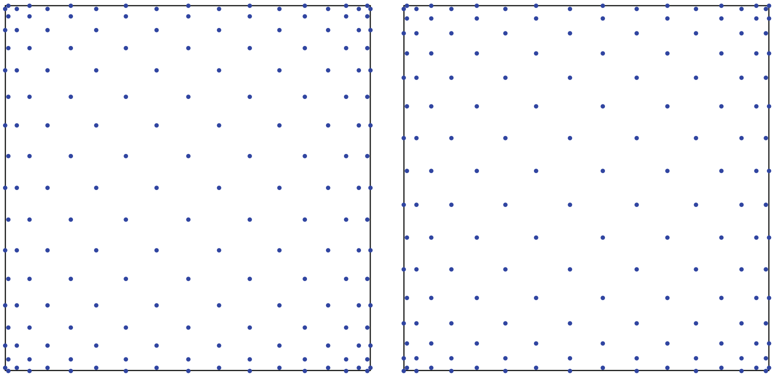


Fig. 1 Left: 180 nodes for minimal cubature rule of degree 35. Right: 162 nodes for near-minimal cubature rule of degree 33

Theorem 7 *Let f be a continuous function on $\square$. Then*

1. *There is a constant $c > 0$, independent of n and f, such that*

$$\|f - L_n f\|_p \le c\, E_n(f)_\infty, \qquad 1 \le p < \infty;$$

2. *The Lebesgue constant $\|L_n\|_\infty := \sup_{\|f\|_\infty \ne 0} \|L_n f\|_\infty$ satisfies*

$$\|L_n\|_\infty = \mathscr{O}((\log n)^2),$$

which is the optimal order among all projection operators from $C(\Omega) \mapsto \Pi_n^*$.

The first item was proved in [28], which shows that $L_n f$ behaves like polynomials of best approximation in L^p norm when $1 \le p < \infty$. The second one was proved more recently in [3], which gives the upper bound of the Lebesgue constant; that this upper bound is optimal was established in [24]. These results indicate that the set of points (10) is optimal for both numerical integration and interpolation. These interpolation polynomials were also considered in [12], and further extended in [13, 14], where points for other Chebyshev weights [18], including $(1-x^2)^{\pm\frac{1}{2}}(1-y^2)^{\mp\frac{1}{2}}$, are considered.

The interpolation polynomial $L_n f$ defined above is of degree n and its set of interpolation points has the cardinality $\dim \Pi_{n-1}^2 + \lfloor n/2 \rfloor$ or one more. One could ask if it is possible to identify another set of points, say X_n, that has the cardinality $\dim \Pi_n^2$ and is just as good, which means that the interpolation polynomials based on X_n should have the same convergence behavior as those in Theorem 7 and the cubature rule with X_n as the set of nodes should be of the degree of precision $2n-1$. If such an X_n exists, the points in X_n need to be common zeros of polynomials of the form

$$\mathbb{P}_{n+1} + \Gamma_1 \mathbb{P}_n + \Gamma_2 \mathbb{P}_{n-1},$$

where Γ_1 and Γ_2 are matrices of sizes $(n+2)\times(n+1)$ and $(n+2)\times n$, respectively. For W_0, such a set indeed exists and known as the Padua points [2, 4]. One version of these points is

$$X_n := \Big\{(\cos \tfrac{2i\pi}{n}, \cos \tfrac{(2j-1)\pi}{n+1}),\ 0 \le i \le \lfloor \tfrac{n}{2} \rfloor\ 1 \le j \le \lfloor \tfrac{n}{2} \rfloor + 1, \\ (\cos \tfrac{(2i-1)\pi}{n}, \cos \tfrac{(2j-2)\pi}{n+1}),\ 1 \le i \le \lfloor \tfrac{n}{2} \rfloor + 1, \quad 1 \le j \le \lfloor \tfrac{n}{2} \rfloor + 2\Big\}, \tag{12}$$

which are common zeros of polynomials Q_k^{n+1}, $0 \le k \le n+1$, defined by

$$Q_0^{n+1}(x,y) = T_{n+1}(x) - T_{n-1}(x), \tag{13}$$

$$Q_k^{n+1}(x,y) = T_{n-k+1}(x)T_k(y) + T_{n-k+1}(y)T_{k-1}(x), \quad 1 \le k \le n+1. \tag{14}$$

Theorem 8 *For $n \in \mathbb{N}$, let X_n be defined as in (12). Then $|X_n| = \dim \Pi_n^2$ and*

1. *There is a cubature rule of degree $2n-1$ with X_n as its set of nodes.*
2. *There is a unique polynomial of degree n that interpolates at the points in X_n, which enjoys the same convergence as that of $L_n f$ given in Theorem 7.*

One interesting property of the Padua points is that they are self-intersection points of a Lissajous curve. For X_n given in (12), the curve is given by Q_0^{n+1} or, in parametric form,

$$(-\cos((n+1)t), -\cos(nt)), \qquad 0 \le t \le (2n+1)\pi,$$

as shown in Fig. 2. The generating curve offers a convenient tool for studying the interpolation polynomial based on Padua points.

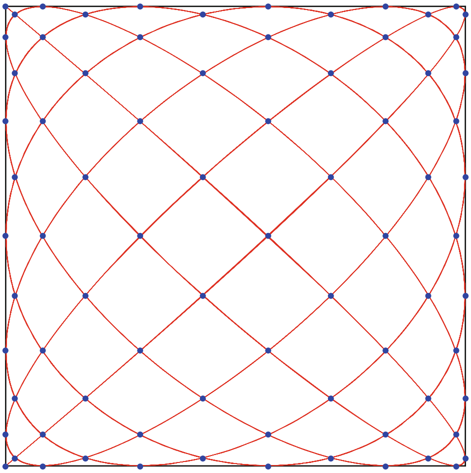


Fig. 2 Seventy-eight Padua points ($n = 11$) and their generating curve

More generally, a Lissajous curve takes of the form $(\cos((n+p)t), \cos(nt))$ with positive integers n and p such that n and $n+p$ are relatively prime. It is known [11] that such a curve has $(n-1)(n+p-1)/2$ self-intersection points inside $[-1,1]^2$. For $p \neq 1$, the number is not equal to the full dimension of Π_m for any m in general. Nonetheless, these points turn out to be good points for cubature rules and for polynomial interpolation, as shown in [9, 10].

4 Results for a Family of Weight Functions

In this section we consider a family of weight functions that include the Chebyshev weight functions as special cases. Let w be a weight function on the interval $[-1,1]$. For $\gamma > -1/2$, we define a weight function

$$\mathscr{W}_\gamma(x,y) := w(\cos(\theta-\phi))w(\cos(\theta+\phi))|x^2-y^2|(1-x^2)^\gamma(1-y^2)^\gamma,$$
$$\text{where } \; x = \cos\theta, \; y = \cos\phi, \quad (x,y) \in [-1,1]^2.$$

When w is the Jacobi weight function $w_{\alpha,\beta}(x) := (1-x)^\alpha(1+x)^\beta$, we denote the weight function $\mathscr{W}_\gamma$ by $W_{\alpha,\beta,\gamma}$. It is not difficult to verify that

$$W_{\alpha,\beta,\gamma}(x,y) := |x+y|^{2\alpha+1}|x-y|^{2\beta+1}(1-x^2)^\gamma(1-y^2)^\gamma. \tag{15}$$

In the special cases of $\alpha = \beta = -\frac{1}{2}$ and $\gamma = \pm\frac{1}{2}$, these are exactly the Chebyshev weight functions. It was proved recently in [29, 31], rather surprisingly, that the results in the previous section can be extended to these weight functions. First, however, we describe a family of mutually orthogonal polynomials. To be more precise, we state this basis only for the weight function $W_{\alpha,\beta,\pm\frac{1}{2}}$.

For $\alpha, \beta > -1$, let $p_n^{(\alpha,\beta)}$ be the normalized Jacobi polynomial of degree n, so that $c_{\alpha,\beta}\int_{-1}^1 |p_n^{(\alpha,\beta)}(x)|^2 w_{\alpha,\beta}(x)dx = 1$ and $p_0^{(\alpha,\beta)}(x) = 1$. For $x = \cos\theta$ and $y = \cos\phi$, we define

$$\begin{aligned} &P_{k,n}^{\alpha,\beta,-\frac{1}{2}}(2xy, x^2+y^2-1) \\ &\quad := p_n^{(\alpha,\beta)}(\cos(\theta-\phi))p_k^{(\alpha,\beta)}(\cos(\theta+\phi)) + p_k^{(\alpha,\beta)}(\cos(\theta-\phi))p_n^{(\alpha,\beta)}(\cos(\theta+\phi)), \\ &P_{k,n}^{\alpha,\beta,\frac{1}{2}}(2xy, x^2+y^2-1) \\ &\quad := \frac{p_{n+1}^{(\alpha,\beta)}(\cos(\theta-\phi))p_k^{(\alpha,\beta)}(\cos(\theta+\phi)) - p_k^{(\alpha,\beta)}(\cos(\theta-\phi))p_{n+1}^{(\alpha,\beta)}(\cos(\theta+\phi))}{2\sin\theta\sin\phi}. \end{aligned}$$

It turns out that $P_{k,n}^{\alpha,\beta,\pm\frac{1}{2}}(u,v)$ itself is a polynomial of degree n in the variables u and v, as can be seen by the elementary trigonometric identities

$$2xy = \cos(\theta-\phi) + \cos(\theta+\phi) \quad \text{and} \quad x^2+y^2-1 = \cos(\theta-\phi)\cos(\theta+\phi),$$

and the fundamental theorem of symmetric polynomials. Furthermore, $P_{k,n}^{\alpha,\beta,\pm\frac{1}{2}}$ (u,v), first studied in [15], are orthogonal polynomials with respect to a weight function on a domain bounded by a parabola and two straight lines and the weight function admit Gaussian cubature rules of all degrees [22]. These polynomials are closely related to the orthogonal polynomials in $\mathscr{V}_n(W_{\alpha,\beta,-\frac{1}{2}})$, as shown in the following proposition established in [30]:

Proposition 1 *Let* $\alpha,\beta > -1$. *A mutually orthogonal basis for* $\mathscr{V}_{2m}(W_{\alpha,\beta,-\frac{1}{2}})$ *is given by*

$$\begin{aligned} {}_1Q_{k,2m}^{\alpha,\beta,\pm\frac{1}{2}}(x,y) &:= P_{k,m}^{\alpha,\beta,\pm\frac{1}{2}}(2xy, x^2+y^2-1), \ 0 \le k \le m, \\ {}_2Q_{k,2m}^{\alpha,\beta,\pm\frac{1}{2}}(x,y) &:= (x^2-y^2)P_{k,m-1}^{\alpha+1,\beta+1,\pm\frac{1}{2}}(2xy, x^2+y^2-1), \ 0 \le k \le m-1, \end{aligned} \tag{16}$$

and a mutually orthogonal basis for $\mathscr{V}_{2m+1}(W_{\alpha,\beta,\pm\frac{1}{2}})$ *is given by*

$$\begin{aligned} {}_1Q_{k,2m+1}^{\alpha,\beta,\pm\frac{1}{2}}(x,y) &:= (x+y)P_{k,m}^{\alpha,\beta+1,\pm\frac{1}{2}}(2xy, x^2+y^2-1), \quad 0 \le k \le m, \\ {}_2Q_{k,2m+1}^{\alpha,\beta,\pm\frac{1}{2}}(x,y) &:= (x-y)P_{k,m-1}^{\alpha+1,\beta,\pm\frac{1}{2}}(2xy, x^2+y^2-1), \quad 0 \le k \le m. \end{aligned} \tag{17}$$

The orthogonal polynomials in (16) of degree $2n$ are symmetric polynomials in x and y, and they are invariant under $(x,y) \mapsto (-x,-y)$. Notice, however, that the product Chebyshev polynomials do not possess such symmetries, even though $W_{-\frac{1}{2},-\frac{1}{2},\pm\frac{1}{2}}$ are the Chebyshev weight functions.

We now return to cubature rules and interpolation and state the following theorem.

Theorem 9 *The minimal cubature rules of degree* $2n-1$ *that attain the lower bound* (5) *exist for the weight function* $\mathscr{W}_{\pm\frac{1}{2}}$ *when* $n = 2m$. *Moreover, the same holds for the weight function* $W_{\alpha,\beta,\pm\frac{1}{2}}$ *when* $n = 2m+1$.

The orthogonal polynomials whose common zeros are nodes of these cubature rules, as described in Theorem 4, can be identified explicitly. Let us consider only $W_{\alpha,\beta,-\frac{1}{2}}$. For $n = 2m$, these polynomials can be chosen as ${}_1Q_{k,2m}^{\alpha,\beta,-\frac{1}{2}}$, $0 \le k \le m$, in (16). For $n = 2m+1$, they can be chosen as ${}_2Q_{k,2m+1}^{\alpha,\beta,-\frac{1}{2}}$, $0 \le k \le m$, in (17), together with one more polynomial

$$\begin{aligned} q_m(x,y) = (x+y)\Big[&p_m^{(\alpha,\beta+1)}(\cos(\theta-\phi))p_m^{(\alpha+1,\beta)}(\cos(\theta+\phi)) \\ &+p_m^{(\alpha,\beta+1)}(\cos(\theta+\phi))p_m^{(\alpha+1,\beta)}(\cos(\theta-\phi))\Big] \end{aligned}$$

in $\mathscr{V}_{2m+1}(W_{\alpha,\beta,-\frac{1}{2}})$, as shown in [31]. For $n = 2m$, the nodes of the minimal cubature rules for $W_{\alpha,\beta,-\frac{1}{2}}$ are not as explicit as those for $n = 2m$. For interpolation, it is often easier to work with the near minimal cubature rule of degree $2n-1$ when $n = 2m+1$, whose number of nodes is just one more than the minimal number $N_{\min}$ in (5). The nodes of the these near minimal rules are common zeros of ${}_2Q_{k,2m+1}^{\alpha,\beta,-\frac{1}{2}}$, $0 \le k \le m$, and a quasi-orthogonal polynomial of the form ${}_1Q_{k,2m+2}^{\alpha,\beta,-\frac{1}{2}} - a_{k,m}{}_1Q_{k,2m}^{\alpha,\beta,-\frac{1}{2}}$, where $a_{k,m}$ are specific constants [31, Theorem 3.5].

The nodes of these cubature rules can be specified. For $\alpha, \beta > -1$ and $1 \le k \le m$, let $\cos\theta_{k,m}^{\alpha,\beta}$ be the zeros of the Jacobi polynomial $P_m^{\alpha,\beta}$ so that

$$0 < \theta_{1,m}^{\alpha,\beta} < \ldots < \theta_{m,m}^{\alpha,\beta} < \pi,$$

and we also define $\theta_{0,m}^{\alpha,\beta} = 0$. We further define

$$s_{j,k}^{\alpha,\beta} := \cos\tfrac{\theta_{j,n}-\theta_{k,n}}{2} \quad \text{and} \quad t_{j,k}^{\alpha,\beta} := \cos\tfrac{\theta_{j,n}+\theta_{k,n}}{2}, \qquad \text{where} \quad \theta_{k,n} = \theta_{k,n}^{\alpha,\beta}.$$

For $n = 2m$, the nodes of the minimal cubature rule of degree $2n-1$ consist of

$$X_{2m}^{\alpha,\beta} := \{(s_{j,k}^{\alpha,\beta}, t_{j,k}^{\alpha,\beta}), (t_{j,k}^{\alpha,\beta}, s_{j,k}^{\alpha,\beta}), (-s_{j,k}^{\alpha,\beta}, -t_{j,k}^{\alpha,\beta}), (-t_{j,k}^{\alpha,\beta}, -s_{j,k}^{\alpha,\beta}) : 1 \le j \le k \le m\}.$$

For $n = 2m+1$, the nodes of the near minimal cubature rule of degree $2n-1$ consist of

$$\begin{aligned} X_{2m+1}^{\alpha,\beta} := \{&(s_{j,k}^{\alpha+1,\beta}, t_{j,k}^{\alpha+1,\beta}), (t_{j,k}^{\alpha+1,\beta}, s_{j,k}^{\alpha+1,\beta}), (-s_{j,k}^{\alpha+1,\beta}, -t_{j,k}^{\alpha+1,\beta}),\\ &(-t_{j,k}^{\alpha+1,\beta}, -s_{j,k}^{\alpha+1,\beta}) : 0 \le j \le k \le m\}. \end{aligned}$$

The weight function $W_{\alpha,\beta,-\frac{1}{2}}$ has a singularity at the diagonal $y = x$ of the square when $\alpha \ne -\frac{1}{2}$, or at the diagonal $y = -x$ of the square when $\beta \ne -\frac{1}{2}$, or at both diagonals when neither α nor β equal to $-\frac{1}{2}$. This is reflected in the distribution of the nodes, which are propelled away from these diagonals. Furthermore, for a fixed m, the points in X_{2m} and X_{2m+1} will be propelled further away for increasing values of α and/or β. In Fig. 3 we depict the nodes of the minimal cubature rules of degree 31 for $W_{\frac{1}{2},\frac{1}{2},-\frac{1}{2}}$, which has singularity on both diagonals, and for $W_{\frac{1}{2},-\frac{1}{2},-\frac{1}{2}}$, which has singularity at the diagonal $y = x$. Writing explicitly, these weight functions are

$$W_{\frac{1}{2},\frac{1}{2},-\frac{1}{2}}(x,y) = \frac{(x-y)^2(x+y)^2}{\sqrt{1-x^2}\sqrt{1-y^2}} \quad \text{and} \quad W_{\frac{1}{2},-\frac{1}{2},-\frac{1}{2}}(x,y) = \frac{(x-y)^2}{\sqrt{1-x^2}\sqrt{1-y^2}}.$$

We also depicted the curves that bound the region that does not contain any nodes, which are given in explicit parametric formulas in [31, Proposition 3.6]. The region without nodes increases in size when α and/or β increase for a fixed m, but they are getting smaller when m increases while α and β are fixed. These figures can be compared to those in Fig. 1 for the case $\alpha = \beta = -\frac{1}{2}$, where the obvious symmetry in $X_n^{\alpha,\beta}$ is not evident.

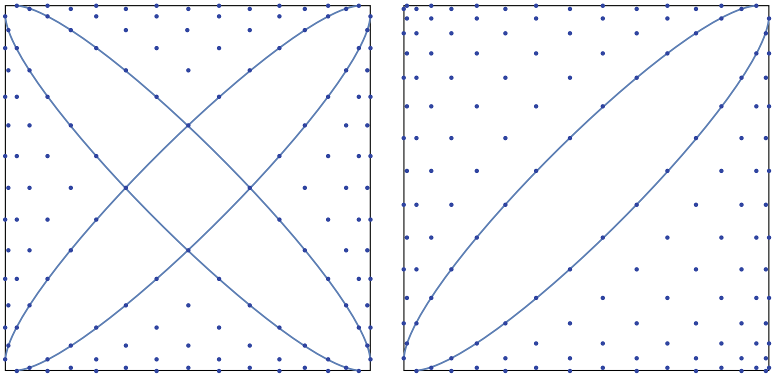


Fig. 3 One hundred forty-four nodes for minimal cubature rule of degree 31 for the weight functions $W_{\frac{1}{2},\frac{1}{2},-\frac{1}{2}}$ (left) and $W_{\frac{1}{2},-\frac{1}{2},-\frac{1}{2}}$ (right)

Let $L_n^{\alpha,\beta} f$ be the interpolation polynomial based on $X_{2m}^{\alpha,\beta}$ when $n = 2m$ and on $X_{2m+1}^{\alpha,\beta}$ when $n = 2m + 1$, as defined in (7). The asymptotics of the Lebesgue constants for these interpolation polynomials can be determined [29, 31].

Theorem 10 *Let $\alpha, \beta \ge -1/2$. The Lebesgue constant of the Lagrange interpolation polynomial $\mathscr{L}_n^{\alpha,\beta} f$ satisfies*

$$\|\mathscr{L}_n^{\alpha,\beta}\|_\infty = \mathscr{O}(1) \begin{cases} n^{2\max\{\alpha,\beta\}+1}, & \max\{\alpha,\beta\} > -1/2, \\ (\log n)^2, & \max\{\alpha,\beta\} = -1/2. \end{cases} \tag{18}$$

It should be mentioned that an explicit formula for the kernel K_n^* in (8) is known, so that the interpolation polynomials $L_n^{\alpha,\beta} f$ can be written down in closed form without solving a large linear system of equations.

5 Minimal Cubature Rules for Constant Weight

The weight functions in the previous two sections contain the Chebyshev weight functions but do not include the weight functions $(1-x^2)^\lambda(1-y^2)^\lambda$ for $\lambda \ne \pm\frac{1}{2}$. In particular, it does not include the constant weight function $W(x, y) = 1$.

For these weight functions, it is possible to establish their existence when n is small. In this section we discuss how these formulas can be constructed. For cubature rules of degree $2n - 2$, we consider the Gaussian cubature rules described in the item 2 of Theorem 2. For cubature rules of degree $2n-1$, we consider minimal cubature rules that attain the lower bound (5). Both these cases can be characterized by non-linear system of equations, which may or may not have solutions. We shall describe these equations and solve them for the constant weight function for small n. The known cases for these cubature rules are listed in [5, 6].

Throughout the rest of this section, we shall assume that $W(x, y) = 1$. Let $\mathscr{V}_n$ be the space of orthogonal polynomials of degree n with respect to the inner product $\langle f, g\rangle = \frac{1}{2}\int_\square f(x, y)g(x, y)dxdy$. Then an orthonormal basis of $\mathscr{V}_n$ is given by

$$P_k^n(x, y) = \widehat{P}_{n-k}(x)\widehat{P}_k(y), \qquad 0 \le k \le n,$$

where $\widehat{P}_n = \sqrt{2n+1}P_n$ and P_n is the classical Legendre polynomial of degree n. In this case, the coefficients $B_{n,i}$ in the three-term relations (3) are zero and the three-term relations take the form

$$\begin{aligned} x\mathbb{P}_n(x, y) &= A_{n,1}\mathbb{P}_{n+1}(x, y) + A_{n-1,1}^{\mathsf{t}}\mathbb{P}_{n-1}(x, y), \\ y\mathbb{P}_n(x, y) &= A_{n,2}\mathbb{P}_{n+1}(x, y) + A_{n-1,2}^{\mathsf{t}}\mathbb{P}_{n-1}(x, y), \end{aligned}$$

where $\mathbb{P}_n = (P_0^n, \ldots, P_n^n)^{\mathsf{t}}$, $A_{n,1}$ and $A_{n,2}$ are given by

$$A_{n,1} = \begin{bmatrix} a_n & & \bigcirc & 0 \\ & \ddots & & \vdots \\ \bigcirc & & a_0 & 0 \end{bmatrix} \quad \text{and} \quad A_{n,2} = \begin{bmatrix} 0 & a_0 & & \bigcirc \\ \vdots & & \ddots & \\ 0 & \bigcirc & & a_n \end{bmatrix},$$

in which

$$a_k := \frac{k+1}{\sqrt{(2k+1)(2k+3)}}, \qquad k = 0, 1, 2, \ldots.$$

5.1 Minimal Cubature Rules of Degree $2n - 2$

By Theorem 2, the nodes of a Gaussian cubature rule of degree $2n - 2$, if it exists, are common zeros of $\mathbb{P}_n + \Gamma_n\mathbb{P}_{n-1}$ for some matrix Γ_n of size $(n+1) \times n$. The latter is characterized in the following theorem [27].

Theorem 11 *The polynomials in $\mathbb{P}_n + \Gamma_n\mathbb{P}_{n-1}$ have $n(n+1)/2$ real, distinct zeros if, and only if, Γ_n satisfies*

$$A_{n-1,1}\Gamma_n = \Gamma_n^{\mathsf{t}}A_{n-1,1}^{\mathsf{t}}, \qquad A_{n-1,2}\Gamma_n = \Gamma_n^{\mathsf{t}}A_{n-1,2}^{\mathsf{t}}, \tag{19}$$

$$\Gamma_n^{\mathsf{t}}(A_{n-1,1}^{\mathsf{t}}A_{n-1,2} - A_{n-1,2}^{\mathsf{t}}A_{n-1,1})\Gamma_n = (A_{n-1,1}A_{n-1,2}^{\mathsf{t}} - A_{n-1,2}A_{n-1,1}^{\mathsf{t}}). \tag{20}$$

The equations in (19) imply that Γ_n can be written in terms of a Hankel matrix $H_n = (h_{i+j})$ of size $(n+1) \times n$,

$$\Gamma = G_nH_nG_{n-1}^{\mathsf{t}}, \quad \text{where} \quad G_n = \operatorname{diag}\{g_{n,0}, g_{n-1,1}, \ldots, g_{1,n-1}, g_{0,n}\} \tag{21}$$

with

$$g_{n-k,k} = \gamma_{n-k}\gamma_k \quad \text{and} \quad \gamma_k = \frac{(2k)!\sqrt{2k+1}}{2^k k!^2}.$$

Thus, solving the system of equations in Theorem 11 is equivalent to solving (20) for the Hankel matrix H_n, which is a nonlinear system of equations and its solution may not exist. Since the matrices in both sides of (20) are skew symmetric, the nonlinear system consists of $n(n-1)/2$ equations and $2n$ variables. The number of variables is equal to the number of equations when $n = 5$.

We found the solution when $n = 3, 4, 5$, which gives Gaussian cubature rules of degree $4, 6, 8$. These cases are known in the literature, see the list in [5]. In the case $n = 3$ and $n = 4$, we were able to solve the system analytically instead of numerically. For $n = 3$, the matrix H_3 takes the form

$$H_3 = \frac{4}{27\sqrt{7}} \begin{bmatrix} -\frac{11}{25} & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & \frac{2}{5} \\ 0 & \frac{2}{5} & 0 \end{bmatrix}.$$

The case H_4 is too cumbersome to write down. In the case $n = 5$, the system is solved numerically, which has multiple solutions but essentially one up to symmetry. This solution, however, has one common zero (or node of the Gaussian cubature rule of degree 8) that lies outside of the square, which agrees with the list in [6].

Solving the system for $n > 5$ *numerically* yields no solution. It is tempting to proclaim that the Gaussian cubature rules of degree $2n-2$ for the constant weight function on the square do not exist for $n \geq 6$, but a proof is still needed.

5.2 Minimal Cubature Rules of Degree $2n-1$

Here we consider minimal cubature rules of degree $2n-1$ that attain the lower bound (5). By Theorem 4, the nodes of such a cubature rule are common zeros of $(n+1) - \lfloor n/2 \rfloor$ many orthogonal polynomials of degree n, which can be written as the elements of $U^{\mathsf{t}}\mathbb{P}_n$, where U is a matrix of size $(n+1) \times (n+1-\lfloor n/2 \rfloor)$ and U has full rank.

Theorem 12 *There exist $(n+1) - \lfloor n/2 \rfloor$ many orthogonal polynomials of degree n, written as $U^{\mathsf{t}}\mathbb{P}_n$, that have $n(n+1) + \lfloor \frac{n}{2} \rfloor$ real, distinct common zeros if, and only if, U satisfies $U^{\mathsf{t}}V = 0$ for a matrix V of size $(n+1) \times \lfloor \frac{n}{2} \rfloor$ that satisfies*

$$A_{n-1,1}(VV^{\mathsf{t}} - I)A_{n-1,2}^{\mathsf{t}} = A_{n-1,2}(VV^{\mathsf{t}} - I)A_{n-1,2}^{\mathsf{t}}, \tag{22}$$

$$VV^{\mathsf{t}}(A_{n-1,1}^{\mathsf{t}}A_{n-1,2} - A_{n-1,2}^{\mathsf{t}}A_{n-1,1})VV^{\mathsf{t}} = 0, \tag{23}$$

where I denotes the identity matrix.

The Eq. (22) implies that the matrix VV^{t} can be written in terms of a Hankel matrix H_n of size $(n+1) \times (n+1)$,

$$VV^{\mathrm{t}} = I + G_n H_n G_n^{\mathrm{t}} := W,$$

where G_n is defined as in (21). Thus, to find the matrix V we need to solve (23) for H_n and make sure that the matrix W is nonnegative definite and has rank $\lfloor \frac{n}{2} \rfloor$, so that it can be factored as VV^{t}. The non-linear system (23) consists of $n(n+1)/2$ equations and has $2n+1$ variables, which may not have a solution.

Comparing with the Gaussian cubature rules of even degree in the previous subsection, however, the situation here is more complicated. We not only need to solve (23), similar to solving (20), for H_n, we also have to make sure that the resulting W is non-negative definite and has rank $\lfloor \frac{n}{2} \rfloor$, which poses an additional constraint that is not so easy to verify.

We found the solutions when $n = 3, 4, 5$ and 6, which gives minimal cubature rules of degree 5, 7, 9, 11. These cases are all known in the literature, see the list in [5] and the references therein. In the case of $n = 4$, there are multiple solutions; for example, one solution has all 12 points inside the square and another one has 2 points outside. In the case $n = 3, 4, 5$, we were able to solve the system analytically instead of numerically. We give Hankel matrices H_n for those cases that have all nodes of the minimal cubature rules inside the square:

$$H_3 = \frac{4}{135} \begin{bmatrix} -\frac{8}{35} & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{4}{35} \end{bmatrix}, \qquad H_4 = \frac{44}{14385} \begin{bmatrix} \frac{94}{231} & 1 & 1 & 1 & -\frac{82}{55} \\ 1 & 1 & 1 & -\frac{82}{55} & 1 \\ 1 & 1 & -\frac{82}{55} & 1 & 1 \\ 1 & -\frac{82}{55} & 1 & 1 & 1 \\ -\frac{82}{55} & 1 & 1 & 1 & \frac{94}{231} \end{bmatrix},$$

and

$$H_5 = \frac{96}{77875} \begin{bmatrix} \frac{1151}{2079} & \frac{10\sqrt{86}}{189} & -\frac{31}{86} & -\frac{1}{9}\sqrt{\frac{43}{2}} & 1 & 0 \\ \frac{10\sqrt{86}}{189} & -\frac{31}{86} & -\frac{1}{9}\sqrt{\frac{43}{2}} & 1 & 0 & 1 \\ -\frac{31}{86} & -\frac{1}{9}\sqrt{\frac{43}{2}} & 1 & 0 & 1 & \frac{1}{9}\sqrt{\frac{43}{2}} \\ -\frac{1}{9}\sqrt{\frac{43}{2}} & 1 & 0 & 1 & \frac{1}{9}\sqrt{\frac{43}{2}} & -\frac{31}{86} \\ 1 & 0 & 1 & \frac{1}{9}\sqrt{\frac{43}{2}} & -\frac{31}{86} & -\frac{10\sqrt{86}}{189} \\ 0 & 1 & \frac{1}{9}\sqrt{\frac{43}{2}} & -\frac{31}{86} & -\frac{10\sqrt{86}}{189} & \frac{1151}{2079} \end{bmatrix}.$$

Once H_n is found, it is easy to verify that W satisfies the desired rank condition and is non-negative definite. We can then find U, or the set of orthogonal polynomials, and then find common zeros. For example, when $n = 5$, we have 4 orthogonal polynomials of degree 5 given by

$$\begin{aligned}
Q_1(x,y) =&\frac{10\sqrt{86}}{189}P_0^5(x,y) + \frac{1081\sqrt{11}}{2835\sqrt{3}}P_1^5(x,y) + P_5^5(x,y),\\
Q_2(x,y) =&\frac{205}{21\sqrt{33}}P_0^5(x,y) + \frac{10\sqrt{86}}{189}P_1^5(x,y) + P_4^5(x,y),\\
Q_3(x,y) =&-\frac{5\sqrt{438}}{27\sqrt{77}}P_0^5(x,y) + \frac{62\sqrt{5}}{81\sqrt{21}}P_1^5(x,y) + P_3^5(x,y),\\
Q_4(x,y) =&-\frac{10\sqrt{5}}{3\sqrt{77}}P_0^5(x,y) - \frac{\sqrt{430}}{9\sqrt{21}}P_1^5(x,y) + P_2^5(x,y),
\end{aligned}$$

which has 17 real common zeros inside the square. Only numerical results are known for the case $n = 6$. We also tried the case $n = 7$, but found no solution numerically.

Acknowledgements The author thanks two anonymous referees for their careful reading and corrections.

The work was supported in part by NSF Grant DMS-1510296.

References

1. Bojanov, B., Petrova, G.: On minimal cubature formulae for product weight function. J. Comput. Appl. Math. **85**, 113–121 (1997)
2. Bos, L., Caliari, M., De Marchi, S., Vianello, M., Xu, Y.: Bivariate Lagrange interpolation at the Padua points: the generating curve approach. J. Approx. Theory **143**, 15–25 (2006)
3. Bos, L., De Marchi, S., Vianello, M.: On the Lebesgue constant for the Xu interpolation formula. J. Approx. Theory **141**, 134–141 (2006)
4. Bos, L., De Marchi, S., Vianello, M., Xu, Y.: Bivariate Lagrange interpolation at the Padua points: the ideal theory approach. Numer. Math. **108**, 43–57 (2007)
5. Cools, R.: Monomial cubature rules since “Stroud”: a compilation – part 2. J. Comput. Appl. Math. **112**, 21–27 (1999)
6. Cools, R., Rabinowitz, P.: Monomial cubature rules since “Stroud”: a compilation. J. Comput. Appl. Math. **48**, 309–326 (1993)
7. Dick, J., Kuo, F., Sloan, I.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
8. Dunkl, C.F., Xu, Y.: Orthogonal Polynomials of Several Variables. Encyclopedia of Mathematics and its Applications, vol. 155. Cambridge University Press, Cambridge (2014)
9. Erb, W.: Bivariate Lagrange interpolation at the node points of Lissajous curves – the degenerate case. Appl. Math. Comput. **289**, 409–425 (2016)
10. Erb, W., Kaethner, C., Ahlborg, M., Buzug, T.M.: Bivariate Lagrange interpolation at the node points of non–degenerate Lissajous curves. Numer. Math. **133**, 685–705 (2016)
11. Fischer, G.: Plane Algebraic Curves, translated by Leslie Kay. American Mathematical Society (AMS), Providence, RI (2001)

12. Harris, L.: Bivariate Lagrange interpolation at the Chebyshev nodes. Proc. Am. Math. Soc. **138**, 4447–4453 (2010)
13. Harris, L.: Bivariate polynomial interpolation at the Geronimus nodes. In: Complex analysis and dynamical systems V. Contemporary Mathematics, vol. 591, pp. 135–147. American Mathematical Society, Providence, RI (2013), Israel Mathematical Conference Proceedings
14. Harris, L.: Lagrange polynomials, reproducing kernels and cubature in two dimensions. J. Approx. Theory, **195**, 43–56 (2015)
15. Koornwinder, T.H.: Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators, I, II. Proc. Kon. Akad. v. Wet., Amsterdam **36**, 48–66 (1974)
16. Li, H., Sun, J., Xu, Y.: Cubature formula and interpolation on the cubic domain. Numer. Math. Theory Methods Appl. **2**, 119–152 (2009)
17. Möller, H.: Kubaturformeln mit minimaler Knotenzahl. Numer. Math. **25**, 185–200 (1976)
18. Morrow, C.R., Patterson, T.N.L.: Construction of algebraic cubature rules using polynomial ideal theory. SIAM J. Numer. Anal. **15**, 953–976 (1978)
19. Mysovskikh, I.P.: Numerical characteristics of orthogonal polynomials in two variables. Vestnik Leningrad Univ. Math. **3**, 323–332 (1976)
20. Mysovskikh, I.P.: Interpolatory Cubature Formulas. Nauka, Moscow (1981)
21. Schmid, H.: On cubature formulae with a minimal number of knots. Numer. Math. **31**, 282–297 (1978)
22. Schmid, H., Xu, Y.: On bivariate Gaussian cubature formula. Proc. Am. Math. Soc. **122**, 833–842 (1994)
23. Stroud, A.H.: Approximate Calculation of Multiple Integrals. Prentice-Hall, Englewood Cliffs, N.J. (1971)
24. Szili, L., Vértesi, P.: On multivariate projection operators. J. Approx. Theory **159**, 154–164 (2009)
25. Xu, Y.: Gaussian cubature and bivariable polynomial interpolation. Math. Comput. **59**, 547–555 (1992)
26. Xu, Y.: Common Zeros of Polynomials in Several Variables and Higher Dimensional Quadrature. Pitman Research Notes in Mathematics Series, Longman, Essex (1994)
27. Xu, Y.: On zeros of multivariate quasi-orthogonal polynomials and Gaussian cubature formulae. SIAM J. Math. Anal. **25**, 991–1001 (1994)
28. Xu, Y.: Lagrange interpolation on Chebyshev points of two variables. J. Approx. Theory **87**, 220–238 (1996)
29. Xu, Y.: Minimal Cubature rules and polynomial interpolation in two variables. J. Approx. Theory **164**, 6–30 (2012)
30. Xu, Y.: Orthogonal polynomials and expansions for a family of weight functions in two variables. Constr. Approx. **36**, 161–190 (2012)
31. Xu, Y.: Minimal Cubature rules and polynomial interpolation in two variables, II. J. Approx. Theory **214**, 49–68 (2017)

Correction to: Multivariate Approximation in Downward Closed Polynomial Spaces

Albert Cohen and Giovanni Migliorati

Correction to:
Chapter 12 in: J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_12

The original version of Chapter 12 was inadvertently published with incorrect details. The theorems 10, 12 and 13 have been updated.

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-319-72456-0_12

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0_59

Index

J. Dick et al. (eds.), *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan*, https://doi.org/10.1007/978-3-319-72456-0