Springer Proceedings in Mathematics & Statistics

Gregory E. Fasshauer
Larry L. Schumaker *Editors*

Approximation Theory XIV: San Antonio 2013

Springer

Springer Proceedings in Mathematics & Statistics

Volume 83

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Gregory E. Fasshauer · Larry L. Schumaker
Editors

Approximation Theory XIV: San Antonio 2013

Springer

Editors
Gregory E. Fasshauer
Department of Applied Mathematics
Illinois Institute of Technology
Chicago, IL
USA

Larry L. Schumaker
Department of Mathematics
Vanderbilt University
Nashville, TN
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-06403-1 ISBN 978-3-319-06404-8 (eBook)
DOI 10.1007/978-3-319-06404-8
Springer Cham Heidelberg New York Dordrecht London

Mathematics Subject Classification (2010): 65D07, 65T60, 47B06, 32E30, 41Axx

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

These proceedings are based on papers presented at the international conference *Approximation Theory XIV*, which was held April 7–10, 2013 in San Antonio, Texas. The conference was the fourteenth in a series of meetings in Approximation Theory held at various locations in the United States, and was attended by 133 participants. Previous conferences in the series were held in Austin, Texas (1973, 1976, 1980, 1992), College Station, Texas (1983, 1986, 1989, 1995), Nashville, Tennessee (1998), St. Louis, Missouri (2001), Gatlinburg, Tennessee (2004), and San Antonio, Texas (2007, 2010).

We are particularly indebted to our plenary speakers: Peter Binev (South Carolina), Annalisa Buffa (Pavia), Michael Floater (Oslo), Kai Hormann (Lugano), Gitta Kutyniok (Berlin), Grady Wright (Boise), and Yuan Xu (Oregon) for their very fine expository talks outlining new research areas. The seventh Vasil A. Popov Prize in Approximation Theory was awarded to Andriy Bondarenko (Kiev), who also presented a plenary lecture. Thanks are also due to the presenters of contributed papers, as well as everyone who attended for making the conference a success.

We are especially grateful to the National Science Foundation for financial support, and also to the Department of Mathematics at Vanderbilt University for its logistical support.

We would also like to express our sincere gratitude to the reviewers who helped select articles for inclusion in this proceedings volume, and also for their suggestions to the authors for improving their papers.

Gregory E. Fasshauer
Larry L. Schumaker

Contents

Contributors

Gerard Awanou Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Ronny Bergmann Department of Mathematics, University Kaiserslautern, Kaiserslautern, Germany

Jean-Luc Bouchot Department of Mathematics, Drexel University, Philadelphia, PA, USA

A. Buffa Istituto di Matematica Applicata e Tecnologie Informatiche 'E. Magenes' del CNR, Pavia, Italy

L. Beirão da Veiga Dipartimento di Matematica, Università di Milano, Milano, Italy

Özge Dalmanoğlu Faculty of Education, Department of Mathematics Education, Başkent University, Ankara, Turkey

Tom Duchamp Department of Mathematics, University of Washington, Seattle, WA, USA

Michael S. Floater Department of Mathematics, University of Oslo, Oslo, Norway

Andrew Gillette Department of Mathematics, University of Arizona, Tucson, AZ, USA

A. Ardeshir Goshtasby Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA

Philipp Grohs Seminar for Applied Mathematics, ETH Zürich, Zürich, Switzerland

Kanghui Guo Missouri State University, Springfield, MO, USA

Kai Hormann Faculty of Informatics, Università della Svizzera italiana, Lugano, Switzerland

Robert Houska Department of Mathematics, University of Houston, Houston, TX, USA

Ertan İbikli Faculty of Science, Department of Mathematics, Ankara University, Ankara, Turkey

Manuel Jaraczewski Helmut Schmidt University, University of the Federal Armed Forces Hamburg, Hamburg, Germany

T. Jordão Departamento de Matemática-ICMC-USP, Universidade de São Paulo, São Carlos, SP, Brazil

Lutz Kämmerer Faculty of Mathematics, Technische Universität Chemnitz, Chemnitz, Germany

Sandra Keiper Department of Mathematics, Technische Universität Berlin, Berlin, Germany

L. Franklin Kemp Collin College, Plano, TX, USA

Alexei Kolesnikov Towson University, Towson, MD, USA

Gitta Kutyniok Department of Mathematics, Technische Universität Berlin, Berlin, Germany

Demetrio Labate Department of Mathematics, University of Houston, Houston, TX, USA

V. A. Menegatto Departamento de Matemática-ICMC-USP, Universidade de São Paulo, São Carlos, SP, Brazil

Jürgen Prestin Institute of Mathematics, University of Lübeck, Lübeck, Germany

Marco Rozgič Helmut Schmidt University, University of the Federal Armed Forces Hamburg, Hamburg, Germany

G. Sangalli Dipartimento di Matematica, Università di Pavia, Pavia, Italy

Martin Schäfer Department of Mathematics, Technische Universität Berlin, Berlin, Germany

Sevilay Kırcı Serenbay Faculty of Education, Department of Mathematics Education, Başkent University, Ankara, Turkey

Jianyun Jimmy Shan Department of Mathematics, University of Illinois, Urbana, IL, USA

Boris Shekhtman Department of Mathematics and Statistics, University of South Florida, Tampa, FL, USA

Lesław Skrzypek Department of Mathematics and Statistics, University of South Florida, Tampa, FL, USA

Tatyana Sorokina Towson University, Towson, MD, USA

Marcus Stiemer Helmut Schmidt University, University of the Federal Armed Forces Hamburg, Hamburg, Germany

Xingping Sun Department of Mathematics, Missouri State University, Springfield, MO, USA

Vesselin Vatchev University of Texas at Brownsville, One West University Boulevard, Brownsville, TX, USA

R. Vázquez Istituto di Matematica Applicata e Tecnologie Informatiche 'E. Magenes' del CNR, Pavia, Italy

Gang Xie Department of Mathematics, East China University of Science and Technology, Shanghai, China

Yuan Xu Department of Mathematics, University of Oregon, Eugene, OR, USA

Qi Ye Department of Mathematics, Syracuse University, Syracuse, NY, USA

Thomas Yu Department of Mathematics, Drexel University, Philadelphia, PA, USA

Isogeometric Method for the Elliptic Monge-Ampère Equation

Gerard Awanou

Abstract We discuss the application of isogeometric analysis to the fully nonlinear elliptic Monge-Ampère equation, an equation nonlinear in the highest order derivatives. The construction of smooth discrete spaces renders isogeometric analysis a natural choice for the discretization of the equation.

Keywords Vanishing viscosity · Monge-Ampère functional · Isogeometric analysis

1 Introduction

We are interested in the numerical resolution of the nonlinear elliptic Monge-Ampère equation

$$\begin{aligned} \det D^2 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where $D^2 v$ denotes the Hessian of a smooth function v, i.e., $D^2 v$ is the matrix with (i, j)th entry $\partial^2 v/(\partial x_i \partial x_j)$. Here Ω is a smooth uniformly convex bounded domain of $\mathbb{R}^2$ which is at least $C^{1,1}$ and $f \in C(\overline{\Omega})$ with $f \geq c_0 > 0$ for a constant c_0. If $f \in C^{0,\alpha}$, $0 < \alpha < 1$, (1) has a classical convex solution in $C^2(\Omega) \cap C(\overline{\Omega})$ and its numerical resolution assuming more regularity on u is understood, e.g., [6, 7, 11]. In the nonsmooth case, various approaches have been proposed, e.g., [16, 17]. For various reasons, it is desirable to use standard discretization techniques, which are valid for both the smooth and the nonsmooth cases. We propose to solve numerically (1) by the discrete version of the sequence of iterates

G. Awanou (✉)
Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, M/C 249, 851 S. Morgan Street, Chicago, IL 60607-7045, USA
e-mail: awanou@uic.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_1,

$$\begin{aligned}(\mathrm{cof}(D^2u_\varepsilon^k+\varepsilon I)) : D^2u_\varepsilon^{k+1} &= \det D^2u_\varepsilon^k + f, \quad \text{in } \Omega \\ u_\varepsilon^{k+1} &= 0, \qquad\qquad\qquad \text{on } \partial\Omega, \end{aligned} \tag{2}$$

where $\varepsilon > 0$, I is the 2×2 identity matrix and we use the notation cof A to denote the matrix of cofactors of A, i.e., for all i, j, $(-1)^{i+j}(\mathrm{cof}\, A)_{ij}$ is the determinant of the matrix obtained from A by deleting its ith row and its jth column. For two $n \times n$ matrices A, B, we recall the Frobenius inner product $A : B = \sum_{i,j=1}^{n} A_{ij}B_{ij}$, where A_{ij} and B_{ij} refer to the entries of the corresponding matrices.

Our recent results [1] indicate that an appropriate space to study a natural variational formulation of (1) is a finite dimensional space of piecewise smooth C^1 functions. For the numerical experiments, we will let V_h be a finite dimensional space of piecewise smooth C^1 functions constructed with the isogeometric analysis paradigm. Numerical results indicate that the proposed iterative regularization (2) is effective for nonsmooth solutions. Formally, the sequence defined by (2) converges to a limit u_ε, and u_ε converges uniformly on compact subsets of Ω to the solution u of (1) as $\varepsilon \to 0$.

For $\varepsilon = 0$, (2) gives the sequence of Newton's method iterates applied to (1). Surprisingly, for the two-dimensional problem, the formal limit u_ε of the sequence u_ε^{k+1} solves the vanishing viscosity approximation of (1)

$$\begin{aligned}\varepsilon\Delta u_\varepsilon + \det D^2u_\varepsilon - f &= 0 \quad \text{in } \Omega \\ u_\varepsilon &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{3}$$

However, discrete versions of Newton's method applied to (3) do not in general perform well for nonsmooth solutions. This led to the development of alternative methods, e.g., the vanishing moment methodology [11]. The key feature in (2) is that the perturbation εI is included to prevent the matrix $D^2u_\varepsilon^k + \varepsilon I$ from being singular.

The difficulty of constructing piecewise polynomial C^1 functions is often cited as a motivation to seek alternative approaches to C^1 conforming approximations of the Monge-Ampère equation. In [1] Lagrange multipliers are used to enforce the C^1 continuity, but the extent to which this constraint is enforced in the computations is comparable to the accuracy of the discretization. With the isogeometric method, the basis functions are also C^1 at the computational level. On the other hand, another advantage of the isogeometric method is the exact representation of a wide range of geometries which we believe would prove useful in applications of the Monge-Ampère equation to geometric optics. Finally, the isogeometric method is widely reported to have better convergence properties than the standard finite element method.

The main difficulty of the numerical resolution of (1) is that Newton's method fails to capture the correct numerical solution when the solution of (1) is not smooth. We proposed in [1] to use a time marching method for solving the discrete equations resulting from a discretization of (1). Moreover in [3] we argued that the correct solution is approximated if one first regularizes the data. However, numerical

experiments reported in [1] and in this paper indicate that regularization of the data may not be necessary.

It is known that the convex solution u of (1) is the unique minimizer of a certain functional J in a set of convex functions S. It is reasonable to expect, although not very easy to make rigorous, that the set S can be approximated by a set of smooth convex functions S_m and minimizers of J in S_m would approximate the minimizer of J in S. We prove that the functional J has a unique minimizer in a ball of C^1 functions centered at a natural interpolant of a smooth solution u. With a sufficiently close initial guess, a minimization algorithm can be used for the computation of the numerical solution. The difficulty of choosing a suitable initial guess may be circumvented by using a global minimization strategy as in [14]. Netherveless our result can be considered a first step toward clarifying whether regularization of the data is necessary for a proven convergence theory of C^1 approximations of (1) in the nonsmooth case.

In this paper the numerical solution u_h is computed as the limit of the sequence $u^k_{\varepsilon,h}$ which solve the discrete variational problem associated with (2). For the case of smooth solutions we use $\varepsilon = 0$ in the resulting discrete problem. See Remark 2. Since (1) is not approximated directly there is a loss of accuracy. Netherveless our algorithm can be considered a step toward the development of fast iterative methods capable of retrieving the correct numerical approximation to (1) in the context of C^1 conforming approximations. Let $u_{\varepsilon,h}$ denote the solution of the discrete problem associated to (3). The existence of $u_{\varepsilon,h}$ and $u^k_{\varepsilon,h}$, the convergence of the sequence $(u^k_{\varepsilon,h})_k$ as $k \to \infty$ as well as the behavior of $u_{\varepsilon,h}$ as $\varepsilon \to 0$ will be addressed in a subsequent paper. These results parallel our recent proof of the convergence of the discrete vanishing moment methodology [2].

This paper falls in the category of papers which do not prove convergence of the discretization of (1) to weak solutions, but give numerical evidence of convergence as well results in the smooth case and/or in particular cases, e.g., [10, 12, 13]. We organize the paper as follows: in the next section we describe the notation used and some preliminaries. In Sect. 3 we prove minimization results at the discrete level. We also derive in Sect. 3 the vanishing viscosity approximation (3) from (2) as well as the discrete variational formulation used in the numerical experiments. In Sect. 4, we recall the isogeometric concept and give numerical results in Sect. 5.

2 Notation and Preliminaries

We denote by $C^k(\Omega)$ the set of all functions having all derivatives of order $\leq k$ continuous on Ω where k is a nonnegative integer or infinity and by $C^0(\overline{\Omega})$, the set of all functions continuous on $\overline{\Omega}$. A function u is said to be uniformly Hölder continuous with exponent α, $0 < \alpha \leq 1$ in Ω if the quantity

$$\sup_{x \neq y} \frac{|u(x) - u(y)|}{|x - y|^\alpha}$$

is finite. The space $C^{k,\alpha}(\Omega)$ consists of functions whose kth order derivatives are uniformly Hölder continuous with exponent α in Ω.

We use the standard notation of Sobolev spaces $W^{k,p}(\Omega)$ with norms $||.||_{k,p}$ and semi-norm $|.|_{k,p}$. In particular, $H^k(\Omega) = W^{k,2}(\Omega)$ and in this case, the norm and seminorms will be denoted, respectively, by $||.||_k$ and semi-norm $|.|_k$. For a function u, we denote by Du its gradient vector and recall that D^2u denotes its Hessian. For a matrix field A, we denote by div A the vector obtained by taking the divergence of each row.

Using the product rule one obtains for sufficiently smooth vector fields v and matrix fields A

$$\text{div}(Av) = (\text{div } A^T) \cdot v + A : (Dv)^T. \tag{4}$$

Moreover, by [8, p. 440]

$$\text{div cof } D^2v = 0. \tag{5}$$

For computation with determinants, the following results are needed.

Lemma 1 *We have*

$$\det D^2v = \frac{1}{2}(\text{cof } D^2v) : D^2v = \frac{1}{2}\,\text{div}\left((\text{cof } D^2v)Dv\right), \tag{6}$$

and for $F(v) = \det D^2v$ *we have*

$$F'(v)(w) = (\text{cof } D^2v) : D^2w = \text{div}\left((\text{cof } D^2v)Dw\right),$$

for v, w *sufficiently smooth.*

Proof For a 2×2 matrix A, one easily verifies that $2 \det A = (\text{cof } A) : A$. It follows that $\det D^2v = 1/2(\text{cof } D^2v) : D^2v$. Using (4) and (5) we obtain $(\text{cof } D^2v) : D^2v = \text{div}\left((\text{cof } D^2v)Dv\right)$ and $(\text{cof } D^2v) : D^2w = \text{div}\left((\text{cof } D^2v)\ Dw\right)$. Finally the expression of the Fréchet derivative is obtained from the definition of Fréchet derivative and the expression $\det D^2v = 1/2(\text{cof } D^2v) : D^2v$. □

Lemma 2 *Let* $v, w \in W^{2,\infty}(\Omega)$ *and* $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$, *then*

$$\left| \int_\Omega (\det D^2v - \det D^2w)\psi \, dx \right| \leq C(|v|_{2,\infty} + |w|_{2,\infty})|v - w|_1|\psi|_1. \tag{7}$$

The above lemma is a simple consequence of the mean value theorem and Cauchy-Schwarz inequalities. For additional details, we refer to [1].

We require our approximation spaces V_h to satisfy the following properties: There exists an interpolation operator Q_h mapping $W^{l+1,p}(\Omega)$ into the space V_h for $1 \leq p \leq \infty$, $0 \leq l \leq d$ with d a constant that depends on V_h and such that

$$||v - Q_hv||_{k,p} \leq C_{ap}h^{l+1-k}||v||_{l+1,p}, \tag{8}$$

for $0 \leq k \leq l$ and

$$||v||_{s,p} \leq C_{inv} h^{l-s+\min(0,\frac{n}{p}-\frac{n}{q})} ||v||_{l,q}, \quad \forall v \in V_h, \tag{9}$$

for $0 \leq l \leq s$, $1 \leq p, q \leq \infty$.

The discussion in [1] is for a space V_h of piecewise polynomials. However, the results quoted here are valid for spaces of piecewise smooth C^1 functions.

We consider the following discretization of (1): find $u_h \in V_h \cap H_0^1(\Omega)$ such that

$$\int_\Omega (\det D^2 u_h) v \, dx = \int_\Omega f v \, dx, \quad \forall v \in V_h \cap H_0^1(\Omega). \tag{10}$$

It can be shown that for $u_h \in H^2(\Omega)$, the left hand side of the above equation is well defined [1]. We recall from [1] that under the assumption that $u \in C^4(\overline{\Omega})$ is a strictly convex function, there exists $\delta > 0$ such that if we define

$$X_h = \left\{ v_h \in V_h, v_h = 0 \quad \text{on } \partial\Omega, ||v_h - Q_h u||_1 < \frac{\delta h^2}{4} \right\},$$

then for h sufficiently small and $v_h \in X_h$, $||v_h - Q_h u||_1 < \delta h^2/2$, v_h is convex with smallest eigenvalue bounded a.e. below by $m'/2$ and above by $3M'/2$. Here m' and M' are respectively lower and upper bounds of the smallest and largest eigenvalues of $D^2 u$ in Ω. The idea of the proof is to use the continuity of the eigenvalues of a matrix as a function of its entries. Thus using (8) with $k = 2$, $p = \infty$ and $l = d$ one obtains that $D^2 Q_h u(x)$ is also positive definite element by element for h sufficiently small. The same argument shows that a C^1 function close to $D^2 Q_h u$ is also piecewise convex and hence convex due to the C^1 continuity. The power of h which appears in the definition of X_h arises from the use of the inverse estimate (9).

We note that by an inverse estimate, for $v_h \in X_h$,

$$||v_h - Q_h u||_{2,\infty} \leq C_{inv} h^{-2} ||v_h - Q_h u||_1 \leq C_{inv} \delta.$$

3 Minimization Results

We first note

Lemma 3 *Let v_n, v, w_n and $w \in W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$ such that $||v_n - v||_{2,\infty} \to 0$ and $||w_n - w||_{2,\infty} \to 0$. Then*

$$\int_\Omega (\det D^2 v_n) w_n \, dx \to \int_\Omega (\det D^2 v) w \, dx \tag{11}$$

$$\int_\Omega f v_n \, dx \to \int_\Omega f v \, dx. \tag{12}$$

Proof Put $\alpha = \int_\Omega (\det D^2 v_n) w_n \, dx - \int_\Omega (\det D^2 v) w \, dx$. We have

$$\alpha = \int_\Omega (\det D^2 v_n - \det D^2 v) w_n \, dx + \int_\Omega (\det D^2 v)(w_n - w) \, dx.$$

Using (7) we obtain

$$|\alpha| \leq C(|v_n|_{2,\infty} + |v|_{2,\infty})|v_n - v|_1 |w_n|_1 + C|v|_{2,\infty}|v|_1 |w_n - w|_1.$$

Since $|v_n - v|_1 \leq C||v_n - v||_{2,\infty}$ and convergent sequences are bounded, (11) follows. We have

$$\left| 3 \int_\Omega f(v_n - v) \, dx \right| \leq C||f||_0 ||v_n - v||_0,$$

and so (12) holds. □

We consider the functional J defined by

$$J(v) = -\int_\Omega v \det D^2 v \, dx + 3 \int_\Omega f v \, dx.$$

We have

Lemma 4 *For $v, w \in W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$*

$$J'(v)(w) = 3 \int_\Omega (f - \det D^2 v) w \, dx.$$

Proof Note that for v, w smooth, vanishing on $\partial\Omega$ and by Lemma 1

$$J'(v)(w) = 3 \int_\Omega f w \, dx - \int_\Omega w \det D^2 v \, dx - \int_\Omega v \operatorname{div}[(\operatorname{cof} D^2 v) Dw] \, dx.$$

But by integration by parts, the symmetry of $D^2 v$ and Lemma 1

$$\begin{aligned} \int_\Omega v \operatorname{div}[(\operatorname{cof} D^2 v) Dw] \, dx &= -\int_\Omega [(\operatorname{cof} D^2 v) Dw] \cdot Dv \, dx = -\int_\Omega [(\operatorname{cof} D^2 v) Dv] \cdot Dw \, dx \\ &= \int_\Omega w \operatorname{div}[(\operatorname{cof} D^2 v) Dv] \, dx = 2 \int_\Omega w \det D^2 v \, dx. \end{aligned}$$

Thus

$$J'(v)(w) = 3 \int_\Omega (f - \det D^2 v) w \, dx.$$

We have proved that for v, w smooth, vanishing on $\partial\Omega$

$$J(v + w) - J(v) = 3 \int_\Omega (f - \det D^2 v) w \, dx + O(|w|_1^2).$$

Since the space of infinitely differentiable functions with compact support is dense in $W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$, the result holds for $v, w \in W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$ by a density argument and using Lemma 3. □

The Euler-Lagrange equation for J is therefore (10).

Remark 1 It has been shown in [4, 19] that a generalized solution of (1) is the unique minimizer of the functional J on the set of convex functions vanishing on the boundary.

Theorem 1 *Let $u \in C^4(\overline{\Omega})$ be the unique strictly convex solution of (1). Then for h sufficiently small, the functional J has a unique minimizer $\hat{u}_h$ in X_h. Moreover, $||u - \hat{u}_h||_1 \to 0$ as $h \to 0$.*

Proof We first note that by (7), the functional J is sequentially continuous in $W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$. For $v_n, v \in W^{2,\infty}(\Omega) \cap H_0^1(\Omega)$ we have

$$J(v_n) - J(v) = 3 \int_\Omega f(v_n - v)\, dx + \int_\Omega (v \det D^2 v - v_n \det D^2 v_n)\, dx.$$

We conclude from Lemma 3 that $J(v_n) \to J(v)$ as $||v_n - v||_{2,\infty} \to 0$. Moreover, using the expression of $J'(v)(w)$ given in Lemma 4, we obtain

$$J''(v)(w)(z) = -3 \int_\Omega w \operatorname{div}[(\operatorname{cof} D^2 v) Dz]\, dx = 3 \int_\Omega [(\operatorname{cof} D^2 v) Dz] \cdot Dw\, dx.$$

We conclude that

$$J''(v)(w)(w) = 3 \int_\Omega [(\operatorname{cof} D^2 v) Dw] \cdot Dw\, dx.$$

That is, J is strictly convex in X_h by definition of X_h. A minimizer, if it exists, is therefore unique.

The argument to prove that J has a minimizer follows the lines of Theorem 5.1 in [9]. We have for some $\theta \in [0, 1]$

$$\begin{aligned} J(v) &= J(0) + J'(0)(v) + \frac{1}{2} J''(\theta v)(v)(v) \\ &= 0 + 3 \int_\Omega f v\, dx + \frac{3}{2} \theta \int_\Omega [(\operatorname{cof} D^2 v) Dv] \cdot Dv\, dx. \end{aligned} \tag{13}$$

We claim that for $v \in X_h$, $v \neq 0$, we have $\theta \neq 0$. Assume otherwise. Then

$$\begin{aligned} 0 &= -\int_\Omega v \det D^2 v\, dx = -\frac{1}{2} \int_\Omega v \operatorname{div}(\operatorname{cof} D^2 v) Dv\, dx \\ &= \frac{1}{2} \int_\Omega [(\operatorname{cof} D^2 v) Dv] \cdot Dv\, dx \geq \frac{m}{2} |v|_1^2, \end{aligned} \tag{14}$$

where m is a lower bound on the smallest eigenvalue of cof $D^2 v$. By the assumption on $v \in X_h$ we have $m > 0$. We obtain the contradiction $v = 0$ and conclude that $\theta \in (0, 1]$.

Next, note that

$$\left| \int_\Omega f v \, dx \right| \leq ||f||_0 ||v||_0 \leq ||f||_0 ||v||_1. \text{ Thus } \int_\Omega f v \, dx \geq -||f||_0 ||v||_1.$$

By (13), we obtain using Poincare's inequality

$$\begin{aligned} J(v) &\geq -3||f||_0||v||_1 + \frac{3}{2}\theta m |v|_1^2 \geq -3||f||_0||v||_1 + C||v||_1^2 \\ &\geq ||v||_1(-3||f||_0 + C||v||_1), \end{aligned} \tag{15}$$

for a constant $C > 0$. Let now $R > 0$ such that

$$X_h \cap \{ v \in V_h \cap H_0^1(\Omega), ||v||_1 \leq R \} \neq \emptyset.$$

Since J is continuous, J is bounded below on the above set. Moreover for $||v||_1 \geq R$, we have

$$J(v) \geq R(-3||f||_0 + CR).$$

We conclude that the functional J is bounded below. We show that its infimum is given by some $\hat{u}_h$ in X_h. Let $v_n \in X_h$ such that $\lim_{n\to\infty} J(v_n) = \inf_{v \in X_h} J(v)$ which has just been proved to be finite. Then the sequence $J(v_n)$ is bounded and by (15), the sequence v_n is also necessary bounded. Let v_{k_n} be a weakly convergent subsequence with limit $\hat{u}_h$. We have

$$\lim_{n\to\infty} J'(\hat{u}_h)(v_{k_n}) = J'(\hat{u}_h)(u_h).$$

Since J is strictly convex in X_h,

$$J(v_{k_n}) \geq J(\hat{u}_h) + J'(\hat{u}_h)(v_{k_n} - \hat{u}_h),$$

and so at the limit $\inf_{v \in X_h} J(v) \geq J(\hat{u}_h)$. This proves that $\hat{u}_h$ minimizes J in X_h.

We now prove that $||u - \hat{u}_h||_1 \to 0$ as $h \to 0$. Note that since $u_h \in X_h$, $||\hat{u}_h - Q_h u||_1 \leq \delta h^2/4$. By (8) and triangle inequality, we obtain the result. □

Remark 2 From the approach taken in [1], we may conclude that (10) has a unique convex solution u_h in X_h which therefore solves the Euler-Lagrange equation for the functional J. Since X_h is open and convex and J convex on X_h, by Theorem 3.9.1 of [15] we have

$$J(v) \geq J(u_h) + J'(u_h)(v - u_h), \quad \forall v \in X_h.$$

Since u_h is a critical point of J in X_h, we get

$$J(v) \geq J(u_h), \quad \forall v \in X_h.$$

We conclude that both u_h and $\hat{u}_h$ are minimizers of J in X_h. By the strict convexity of J in X_h, they are equal. Therefore the unique minimizer of J in X_h solves (10).

We now turn to the regularized problems (2) and (3). The formal limit of u_ε^k as $k \to \infty$ solves

$$\begin{aligned}(\operatorname{cof}(D^2u_\varepsilon + \varepsilon I)) : D^2u_\varepsilon &= \det D^2u_\varepsilon + f \quad \text{in } \Omega \\ u_\varepsilon &= 0 \quad \text{on } \partial\Omega.\end{aligned}$$

But since I and D^2u_ε are 2×2 matrices, we have $\operatorname{cof}(D^2u_\varepsilon + \varepsilon I) = \operatorname{cof} D^2u_\varepsilon + \operatorname{cof} \varepsilon I = \operatorname{cof} D^2u_\varepsilon + \varepsilon I$ and we obtain

$$(\operatorname{cof} D^2u_\varepsilon) : D^2u_\varepsilon + \varepsilon I : D^2u_\varepsilon = \det D^2u_\varepsilon + f.$$

Since $\varepsilon I : D^2u_\varepsilon = \varepsilon \Delta u_\varepsilon$ and by (6) we have $(\operatorname{cof} D^2u_\varepsilon) : D^2u_\varepsilon = 2 \det D^2u_\varepsilon$, we obtain (3).

Next we present the discrete variational formulation used in the numerical experiments. To avoid large errors, we used a damped version of (2). Let $\nu > 0$. We consider the problem

$$\begin{aligned}(\operatorname{cof}(D^2u_\varepsilon^k + \varepsilon I)) : D^2u_\varepsilon^{k+1} &= 2 \det D^2u_\varepsilon^k + \frac{1}{\nu}(-\det D^2u_\varepsilon^k + f) \quad \text{in } \Omega \\ u_\varepsilon^{k+1} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{16}$$

We note that for $\nu = 1$, (16) reduces to (2). Also the formal limit, as $\varepsilon \to 0$ and $k \to \infty$, of u_ε^k solving (16) is a solution of $1/\nu(f - \det D^2u) = 0$.

Let $|x|$ denote the Euclidean norm of $x \in \mathbb{R}^2$. Note that that $D^2(|x|^2/2) = I$ and thus for u_ε^k smooth, $\operatorname{cof}(D^2u_\varepsilon^k + \varepsilon I) = \operatorname{cof} D^2(u_\varepsilon^k + \varepsilon/2|x|^2)$ and thus using (4) and (5) we obtain

$$\begin{aligned}\operatorname{div}\left((\operatorname{cof}(D^2u_\varepsilon^k + \varepsilon I))Du_\varepsilon^{k+1}\right) &= 2 \det D^2u_\varepsilon^k + \frac{1}{\nu}(-\det D^2u_\varepsilon^k + f) \quad \text{in } \Omega \\ u_\varepsilon^{k+1} &= 0 \quad \text{on } \partial\Omega.\end{aligned}$$

This leads to the following discretization: find $u_{\varepsilon,h}^{k+1} \in V_h \cap H_0^1(\Omega)$ such that $\forall v \in V_h \cap H_0^1(\Omega)$

$$\begin{aligned}-\int_\Omega \left((\operatorname{cof}(D^2u_{\varepsilon,h}^k + \varepsilon I))Du_{\varepsilon,h}^{k+1}\right) \cdot Dv\, dx = \int_\Omega \Big(&2 \det D^2u_{\varepsilon,h}^k \\ &+ \frac{1}{\nu}(-\det D^2u_{\varepsilon,h}^k + f)\Big)v\, dx.\end{aligned} \tag{17}$$

For the initial guess $u^0_{\varepsilon,h}$ when $\varepsilon \geq 0$, we take the discrete approximation of the solution of the problem

$$\begin{aligned} \Delta u^0_\varepsilon &= 2\sqrt{f} \quad \text{in } \Omega \\ u^0_\varepsilon &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

While this does not assure that $u^0_{\varepsilon,h} \in X_h$ the above choice appears to work in all our numerical experiments.

Remark 3 For a possible extension of the minimization result in Theorem 1 to the case of nonsmooth solutions, the homogeneous boundary condition is necessary.

4 Isogeometric Analysis

We refer to [20] for a short introduction to isogeometric analysis. Here we give a shorter overview suitable for our needs. Precisely, we are interested in the ability of this approach to generate finite dimensional spaces of piecewise smooth C^1 functions which can be used in the Galerkin method for approximating partial differential equations.

A univariate NURBS of degree p is given by

$$\frac{w_i N_{i,p}(u)}{\sum_{j\in\mathscr{J}} w_j N_{j,p}(u)}, \quad u \in [0,1],$$

with B-splines $N_{i,p}$, weights w_i and an index set $\mathscr{J}$ which encodes its smoothness. The parameter h refers to the maximum distance between the knots $u_i, i \in \mathscr{J}$.

A bivariate NURBS is given by

$$R_{kl}(u,v) = \frac{w_{kl} N_k(u) N_l(v)}{\sum_{i\in\mathscr{I}} \sum_{j\in\mathscr{J}} w_j N_i(u) N_j(v)}, \quad u, v \in [0,1],$$

with index sets $\mathscr{I}$ and $\mathscr{J}$. In the above expression, we omit the degrees pU and pV of the NURBS R_{kl} in the u and v directions.

The domain Ω is described parametrically by a mapping $F : [0,1]^2 \to \Omega$, $F(u,v) = \sum_{i\in\mathscr{I}} \sum_{j\in\mathscr{J}} R_{ij}(u,v) c_{ij}$ with NURBS R_{ij} and control points c_{ij}. We take equally spaced knots u_i, v_j and hence h refers to the size of an element in the parametric domain.

We say that a NURBS R_{kl} has degree p if the univariate NURBS N_k and N_l all have degree p. The NURBS considered in this paper are all of a fixed degree p and C^1.

The basis functions R_{ij} used in the description of the domain are also used in the definition of the finite dimensional space $V_h \subset \text{span}\{R_{ij} \circ F^{-1}\}$. Thus the numerical solution takes the form

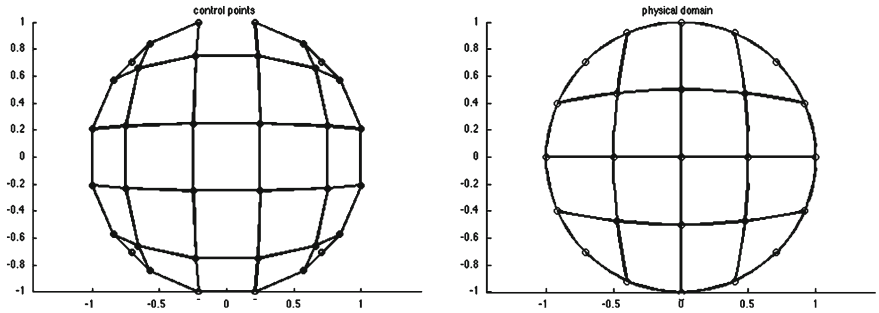


Fig. 1 *Circle* represented exactly. pU = 2, pV = 2

$$T_h(x, y) = \sum_{i \in \mathscr{I}} \sum_{j \in \mathscr{J}} R_{ij}(F^{-1}(x, y))q_{ij},$$

with unknowns q_{ij}.

It can be shown [18] that there exists an interpolation operator Q_h mapping $H^r(\Omega), r \geq p + 1$ into V_h such that if $0 \leq l \leq p + 1, 0 \leq l \leq r \leq p + 1$, we have

$$|u - Q_h u|_l \leq Ch^{r-l}||u||_r,$$

with C independent of h. Thus the approximation property (8) holds for spaces constructed with the isogeometric analysis concept. For the inverse estimates (9), we refer to [5].

5 Numerical Results

The implementation was done by modifying the companion code to [20]. The computational domain is taken as the unit circle: $x^2 + y^2 - 1 = 0$ with an initial triangulation depicted in Fig. 1. The numerical solutions are obtained by computing $u^k_{\varepsilon,h}$ defined by (17). We consider the following test cases.

Test 1 (smooth solution): $u(x, y) = (x^2 + y^2 - 1)e^{x^2+y^2}$ with $f(x, y) = 4e^{2(x^2+y^2)}$ $(x^2 + y^2)^2(2x^2 + 3 + 2y^2)$. Numerical results are given in Table 1. Since pU = 2, pV = 2, the approximation space in the parametric domain contains piecewise polynomials of degree $p = 2$. The analysis in [1] suggests that the rate of convergence for smooth solutions is $O(h^p)$ in the H^1 norm, $O(h^{p+1})$ and $O(h^{p-1})$ in the L^2 and H^2 norms respectively. No regularization or damping was necessary for this case.

Test 2 (No known exact solution): $f = e^{x^2+y^2}$, $g = 0$. As expected the numerical solution displayed in Fig. 2 appears to be a convex function.

Test 3 (solution not in $H^1(\Omega)$): $u(x, y) = -\sqrt{1 - x^2 - y^2}$ with $f(x, y) = 1/(x^2 + y^2 - 1)^2$. With regularization and damping, we were able to avoid the divergence

Table 1 Smooth solution $u(x, y) = (x^2 + y^2 - 1)e^{x^2+y^2}$

h	n_{it}	L^2 Norm	Rate	H^1 norm	Rate	H^2 norm	Rate
$1/2^6$	3	$4.5620\ 10^{-1}$		$1.5565\ 10^{-0}$		$1.1877\ 10^{+1}$	
$1/2^7$	6	$8.4903\ 10^{-3}$	5.75	$1.6442\ 10^{-1}$	3.24	$5.0963\ 10^{-0}$	1.2
$1/2^8$	4	$7.7160\ 10^{-4}$	3.46	$3.9573\ 10^{-2}$	2.05	$2.5880\ 10^{-0}$	0.97
$1/2^9$	4	$9.0321\ 10^{-5}$	3.09	$9.8122\ 10^{-3}$	2.01	$1.3019\ 10^{-0}$	0.99
$1/2^{10}$	4	$1.1077\ 10^{-5}$	3.03	$2.4462\ 10^{-3}$	2.00	$6.5184\ 10^{-1}$	0.99

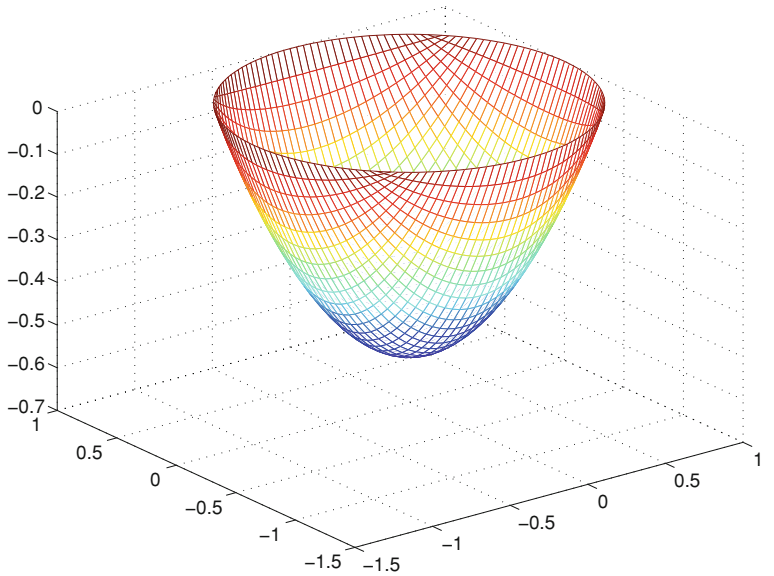


Fig. 2 Convex solution with data $f = e^{x^2+y^2}$, $g = 0$ with $\nu = 2.5, \varepsilon = 0.01, h = 1/32$. No known analytical formula

Table 2 Solution not in $H^1(\Omega)\, u(x, y) = -\sqrt{1 - x^2 - y^2}$ with $\nu = 2.5, \varepsilon = 0.01$

h	n_{it}	L^2 norm	Rate
$1/2^5$	42	$4.0261\ 10^{-1}$	
$1/2^6$	2	$1.7529\ 10^{-1}$	1.20
$1/2^7$	5	$1.3612\ 10^{-1}$	0.36
$1/2^8$	3	$1.0609\ 10^{-1}$	0.36
$1/2^9$	2	$9.6321\ 10^{-2}$	0.14
$1/2^{10}$	4	$7.8179\ 10^{-2}$	0.30

of the discretization. These results should be compared with the ones in [1] where iterative methods with only a linear convergence rate were proposed for nonsmooth solutions of (1). Note that u in this case is highly singular as f vanishes on $\partial\Omega$.

In the tables n_{it} refers to the number of iterations for Newton's method (Table 2).

Acknowledgments The author thanks the two referees for their suggestions and their careful reading of the manuscript. This work began when the author was supported in part by a 2009–2013 Sloan Foundation Fellowship and NSF grant DMS-1319640. Part of this work was completed when the author was in residence at the Mathematical Sciences Research Institute (MSRI) in Berkeley, California, Fall 2013. The MSRI receives major funding from the National Science Foundation under Grant No. 0932078 000.

References

1. Awanou, G.: Pseudo transient continuation and time marching methods for Monge-Ampère type equations (2013). http://arxiv.org/pdf/1301.5891.pdf
2. Awanou, G.: Spline element method for the Monge-Ampère equation (2013). http://arxiv.org/abs/1012.1775
3. Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equation: Aleksandrov solutions (2013). http://arxiv.org/pdf/1310.4568v1.pdf
4. Bakelman, I.J.: Variational problems and elliptic Monge-Ampère equations. J. Differential Geom. **18**(4), 669–699 (1984)
5. Bazilevs, Y., Beirão da Veiga, L., Cottrell, J.A., Hughes, T.J.R., Sangalli, G.: Isogeometric analysis: approximation, stability and error estimates for h-refined meshes. Math. Models Methods Appl. Sci. **16**(7), 1031–1090 (2006)
6. Böhmer, K.: On finite element methods for fully nonlinear elliptic equations of second order. SIAM J. Numer. Anal. **46**(3), 1212–1249 (2008)
7. Brenner, S.C., Gudi, T., Neilan, M., Sung, L.Y.: C^0 penalty methods for the fully nonlinear Monge-Ampère equation. Math. Comp. **80**(276), 1979–1995 (2011)
8. Evans, L.C.: Partial Differential Equations, Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (1998)
9. Faragó, I., Karátson, J.: Numerical Solution of Nonlinear Elliptic Problems Via Preconditioning Operators: Theory and Applications, Advances in Computation: Theory and Practice, vol. 11. Nova Science, Hauppauge (2002)
10. Feng, X., Neilan, M.: Convergence of a Fourth Order Singular Perturbation of the n-Dimensional Radially Symmetric Monge-Ampere Equation (Submitted)
11. Feng, X., Neilan, M.: Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. J. Sci. Comput. **47**(3), 303–327 (2011)
12. Froese, B.D., Oberman, A.M.: Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. J. Comput. Phys. **230**(3), 818–834 (2011)
13. Glowinski, R.: Numerical methods for fully nonlinear elliptic equations. In: ICIAM 07–6th International Congress on Industrial and Applied Mathematics, pp. 155–192. European Mathematical Society, Zürich (2009)
14. Mohammadi, B.: Optimal transport, shape optimization and global minimization. C. R. Math. Acad. Sci. Paris **344**(9), 591–596 (2007)
15. Niculescu, C.P., Persson, L.E.: Convex Functions and Their Applications. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 23. Springer, New York (2006). (A contemporary approach)
16. Oberman, A.M.: Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. Discrete Contin. Dyn. Syst. Ser. B **10**(1), 221–238 (2008)
17. Oliker, V.I., Prussner, L.D.: On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ and its discretizations. I. Numer. Math. **54**(3), 271–293 (1988)
18. Tagliabue, A., Dede, L., Quarteroni, A.: Isogeometric analysis and error estimates for high order partial differential equations in fluid dynamics (Preprint)
19. Tso, K.: On a real Monge-Ampère functional. Invent. Math. **101**(2), 425–448 (1990)
20. Vuong, A.V., Heinrich, C., Simeon, B.: ISOGAT: a 2D tutorial MATLAB code for isogeometric analysis. Comput. Aided Geom. Design **27**(8), 644–655 (2010)

Dual Compatible Splines on Nontensor Product Meshes

L. Beirão da Veiga, A. Buffa, G. Sangalli and R. Vázquez

Abstract In this paper we introduce the concept of dual compatible (DC) splines on nontensor product meshes, study the properties of this class, and discuss their possible use within the isogeometric framework. We show that DC splines are linear independent and that they also enjoy good approximation properties.

Keywords Isogeometric analysis · Spline theory · T-splines · Numerical methods for partial differential equations

1 Introduction

Tensor product multivariate spline spaces are easy to construct and their mathematical properties directly extend from the univariate case. However, the tensor product construction restricts the possibility of local refinement which is a severe limitation for their use within the isogeometric framework, i.e., as discretization spaces for the numerical solution of partial differential equations. This is particularly true in problems that exhibit solutions with layers or singularities. In this paper, we discuss an

L. B. da Veiga (✉)
Dipartimento di Matematica, Università di Milano, via Saldini 50, 20133, Milano, Italy
e-mail: lourenco.beirao@unimi.it

A. Buffa · R. Vázquez
Istituto di Matematica Applicata e Tecnologie Informatiche 'E. Magenes' del CNR, via Ferrata 1, 27100, Pavia, Italy
e-mail: annalisa@imati.cnr.it

R. Vázquez
e-mail: vazquez@imati.cnr.it

G. Sangalli
Dipartimento di Matematica, Università di Pavia, via Ferrata 1, 27100, Pavia, Italy
e-mail: giancarlo.sangalli@unipv.it

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_2,

extension of splines spaces that go beyond the tensor product structure, and therefore allow local mesh refinement.

Three approaches have emerged in the isogeometric community: T-splines, Locally refinable (LR) splines, and hierarchical splines. T-splines have been proposed in [1] for applications to CAGD and have been adopted for isogeometric methods since [2]. Nowadays, they are likely the most popular approach among engineers: for example, they have been used for shell problems [3], fluid–structure interaction problems [4], and contact mechanics simulation [5]. The algorithm for local refinement has evolved since its introduction (in [6]) and while the first approach was not efficient in isogeometric methods (see for example [7]) the more recent developments (e.g., [8]) overcome the initial limitations. The mathematical literature on T-splines is very recent and mainly restricted to the two-dimensional case. It is based on the notion of Analysis-Suitable (AS) T-splines: these are a subset of T-splines, introduced in [9] and extended to arbitrary degree in [10], for which fundamental properties hold. LR-splines [11] and Hierarchical splines [12] have been proposed more recently in the isogeometric literature and represent a valid alternative to T-splines. However, for reasons of space and because of our expertise, we restrict the presentation to T-splines.

This paper is organized as follows. First, we set up our main notation of Sect. 2. Then, we introduce the notion of Dual-Compatible (DC) set of B-splines. This is a set of multivariate B-splines without a global tensor product structure but endowed with a weaker structure that still guarantees some key properties. The main one is that their linear combination spans a space (named DC space) that can be associated with a dual space by a construction of a dual basis. The existence of a "good" dual space implies other mathematical properties that are needed in isogeometric methods: for example, (local) linear independence and partition of unity of the DC set of B-spline functions, and optimal approximation properties of the DC space. The framework we propose here is an extension of the one introduced in [10], and covers arbitrary space dimension.

2 Preliminaries

Given two positive integers p and n, we say that $\Xi := \{\xi_1, \dots, \xi_{n+p+1}\}$ is a p-open knot vector if

$$\xi_1 = \cdots = \xi_{p+1} < \xi_{p+2} \leq \cdots \leq \xi_n < \xi_{n+1} = \cdots = \xi_{n+p+1},$$

where repeated knots are allowed. Without loss of generality, we assume in the following that $\xi_1 = 0$ and $\xi_{n+p+1} = 1$.

We introduce also the vector $Z = \{\zeta_1, \dots, \zeta_N\}$ of knots without repetitions, also called breakpoints, and denote by m_j, the multiplicity of the breakpoint ζ_j, such that

$$\Xi = \{\underbrace{\zeta_1, \dots, \zeta_1}_{m_1 \text{ times}}, \underbrace{\zeta_2, \dots, \zeta_2}_{m_2 \text{ times}}, \dots, \underbrace{\zeta_N, \dots, \zeta_N}_{m_N \text{ times}}\}, \tag{1}$$

with $\sum_{i=1}^{N} m_i = n + p + 1$. We assume $m_j \leq p+1$, for all internal knots. Note that the points in Z form a partition of the unit interval $I = (0, 1)$, i.e., a mesh, and the local mesh size of the element $I_i = (\zeta_i, \zeta_{i+1})$ is called $h_i = \zeta_{i+1} - \zeta_i$, for $i = 1, \ldots, N-1$.

From the knot vector Ξ, B-spline functions of degree p are defined following the well-known Cox-DeBoor recursive formula; we start with piecewise constants ($p = 0$):

$$\widehat{B}_{i,0}(\zeta) = \begin{cases} 1 & \text{if } \xi_i \leq \zeta < \xi_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and for $p \geq 1$ the *B-spline* functions are defined by the recursion

$$\widehat{B}_{i,p}(\zeta) = \frac{\zeta - \xi_i}{\xi_{i+p} - \xi_i} \widehat{B}_{i,p-1}(\zeta) + \frac{\xi_{i+p+1} - \zeta}{\xi_{i+p+1} - \xi_{i+1}} \widehat{B}_{i+1,p-1}(\zeta), \tag{3}$$

where it is here formally assumed that 0/0 = 0.

This gives a set of n B-splines that, among many other properties, are non-negative and form a partition of unity. They also form a basis of the space of *splines*, that is, piecewise polynomials of degree p with $k_j := p - m_j$ continuous derivatives at the points ζ_j, for $j = 1, \ldots, N$. Therefore, $-1 \leq k_j \leq p-1$, and the maximum multiplicity allowed, $m_j = p+1$, gives $k_j = -1$ which stands for a discontinuity at ζ_j.

We denote the *univariate spline space* spanned by the B-splines by

$$S_p(\Xi) = \text{span}\{\widehat{B}_{i,p}, \quad i = 1, \ldots, n\}. \tag{4}$$

Note that the definition of each B-spline $\widehat{B}_{i,p}$ depends only on $p+2$ knots, which are collected in the *local knot vector*

$$\Xi_{i,p} := \{\xi_i, \ldots, \xi_{i+p+1}\}.$$

When needed, we will stress this fact by adopting the notation

$$\widehat{B}_{i,p}(\zeta) = \widehat{B}[\Xi_{i,p}](\zeta). \tag{5}$$

Similarly, the support of each basis function is exactly $\text{supp}(\widehat{B}_{i,p}) = [\xi_i, \xi_{i+p+1}]$. Moreover, given an interval $I_j = (\zeta_j, \zeta_{j+1})$ of the partition, which can also be written as (ξ_i, ξ_{i+1}) for a certain (unique) i, we associate the *support extension* $\tilde{I}_j$ defined as

$$\tilde{I}_j := (\xi_{i-p}, \xi_{i+p+1}), \tag{6}$$

that is the interior of the union of the supports of basis functions whose support intersects I_j.

We concentrate now on the construction of interpolation and projection operators onto the space of splines $S_p(\Xi)$. There are several ways to define projections for splines, and here we only describe the one that will be used in this paper.

We will often make use of the following local quasi-uniformity condition on the knot vector, which is a classical assumption in the mathematical isogeometric literature.

Assumption 1 The partition defined by the knots $\zeta_1, \zeta_2, \ldots, \zeta_N$ is locally quasi-uniform, that is, there exists a constant $\theta \geq 1$ such that the mesh sizes $h_i = \zeta_{i+1} - \zeta_i$ satisfy the relation $\theta^{-1} \leq h_i / h_{i+1} \leq \theta$, for $i = 1, \ldots, N-2$.

Since splines are not in general interpolatory, a common way to define projections is by giving a dual basis, i.e.,

$$\Pi_{p,\Xi} : C^\infty([0,1]) \to S_p(\Xi), \qquad \Pi_{p,\Xi}(f) = \sum_{j=1}^{n} \lambda_{j,p}(f) \widehat{B}_{j,p}, \tag{7}$$

where $\lambda_{j,p}$ are a set of dual functionals verifying

$$\lambda_{j,p}(\widehat{B}_{k,p}) = \delta_{jk}, \tag{8}$$

δ_{jk} being the standard Kronecker symbol. It is trivial to prove that, thanks to this property, the quasi-interpolant $\Pi_{p,\Xi}$ preserves splines, that is,

$$\Pi_{p,\Xi}(f) = f, \quad \forall f \in S_p(\Xi). \tag{9}$$

Here, we adopt the dual basis defined in [13, Sect. 4.6]

$$\lambda_{j,p}(f) = \int_{\xi_j}^{\xi_{j+p+1}} f(s) D^{p+1} \psi_j(s) \, \mathrm{d}s, \tag{10}$$

where $\psi_j(\zeta) = G_j(\zeta)\phi_j(\zeta)$, with

$$\phi_j(\zeta) = \frac{(\zeta - \xi_{j+1}) \cdots (\zeta - \xi_{j+p})}{p!},$$

and

$$G_j(\zeta) = g\left(\frac{2\zeta - \xi_j - \xi_{j+p+1}}{\xi_{j+p+1} - \xi_j}\right),$$

where g is the transition function defined in [13, Theorem 4.37]. In the same reference, it is proved that the functionals $\lambda_{j,p}(\cdot)$ are dual to B-splines in the sense of (8) and stable (see [13, Theorem 4.41]), that is

$$|\lambda_{j,p}(f)| \leq C(\xi_{j+p+1} - \xi_j)^{-1/2} \|f\|_{L^2(\xi_j, \xi_{j+p+1})}, \tag{11}$$

where the constant C grows exponentially with respect to the polynomial degree p with the upperbound

$$C \leq (2p+3)9^p, \tag{12}$$

slightly improved in the literature after the results reported in [13]. Note that these dual functionals are locally defined and only depend on the corresponding local knot vector, that is, adopting a notation as in (5), we can write, when needed:

$$\lambda_{i,p}(f) = \lambda[\Xi_{i,p}](f). \tag{13}$$

The reasons for this choice of the dual basis are mainly historical (in the first paper on the numerical analysis of isogeometric methods [14] the authors used this projection), but also because it verifies the following important stability property:

Proposition 1 *For any non-empty knot span $I_i = (\zeta_i, \zeta_{i+1})$ it holds that*

$$\|\Pi_{p,\Xi}(f)\|_{L^2(I_i)} \leq C\|f\|_{L^2(\tilde{I}_i)}, \tag{14}$$

where the constant C depends only on the degree p, and $\tilde{I}_i$ is the support extension defined in (6). Moreover, if Assumption 1 holds, we also have

$$|\Pi_{p,\Xi}(f)|_{H^1(I_i)} \leq C|f|_{H^1(\tilde{I}_i)}, \tag{15}$$

with the constant C depending only on p and θ, and where H^1 denotes the Sobolev space of order one, endowed with the standard norm and seminorm.

Proof We first show (14). There exists a unique index j such that $I_i = (\zeta_i, \zeta_{i+1}) = (\xi_j, \xi_{j+1})$, and using the definition of B-splines at the beginning of Sect. 2, and in particular their support, it immediately follows that

$$\left\{\ell \in \{1, 2, \ldots, n\} \,:\, \operatorname{supp}(\widehat{B}_{\ell,p}) \cap I_i \neq \emptyset\right\} = \{j-p, j-p+1, \ldots, j\}. \tag{16}$$

Let h_i denotes the length of I_i and $\widetilde{h}_i$ indicates the length of $\tilde{I}_i$. First by definition (7), then recalling that the B-spline basis is positive and a partition of unity, we get

$$\begin{aligned}\|\Pi_{p,\Xi}(f)\|_{L^2(I_i)} &= \Big\| \sum_{\ell=j-p}^{j} \lambda_{\ell,p}(f)\widehat{B}_{\ell,p} \Big\|_{L^2(I_i)} \leq \max_{j-p\leq\ell\leq j} |\lambda_{\ell,p}(f)| \Big\| \sum_{\ell=j-p}^{j} \widehat{B}_{\ell,p} \Big\|_{L^2(I_i)} \\ &= h_i^{1/2} \max_{j-p\leq\ell\leq j} |\lambda_{\ell,p}(f)|.\end{aligned}$$

We now apply bound (11) and obtain

$$\|\Pi_{p,\Xi}(f)\|_{L^2(I_i)} \le Ch_i^{1/2} \max_{j-p\le\ell\le j}(\xi_{\ell+p+1}-\xi_\ell)^{-1/2}\|f\|_{L^2(\xi_\ell,\xi_{\ell+p+1})}$$
$$\le Ch_i^{1/2} \max_{j-p\le\ell\le j}(\xi_{\ell+p+1}-\xi_\ell)^{-1/2}\|f\|_{L^2(\tilde{I}_i)},$$

that yields (14) since clearly $h_i \le (\xi_{\ell+p+1}-\xi_\ell)$, for all ℓ in $\{j-p,\dots,j\}$.

We now show (15). For any real constant c, since the operator $\Pi_{p,\Xi}$ preserves constant functions and using a standard inverse estimate for polynomials on I_i, we get

$$|\Pi_{p,\Xi}(f)|_{H^1(I_i)} = |\Pi_{p,\Xi}(f)-c|_{H^1(I_i)} = |\Pi_{p,\Xi}(f-c)|_{H^1(I_i)}$$
$$\le Ch_i^{-1}\|\Pi_{p,\Xi}(f-c)\|_{L^2(I_i)}.$$

We now apply (14) and a standard approximation estimate for constant functions, yielding

$$|\Pi_{p,\Xi}(f)|_{H^1(I_i)} \le Ch_i^{-1}\|f-c\|_{L^2(\tilde{I}_i)} \le Ch_i^{-1}\widetilde{h}_i|f|_{H^1(\tilde{I}_i)}.$$

Using Assumption 1, it is immediate to check that $\widetilde{h}_i \le Ch_i$ with $C = C(p,\theta)$ so that (15) follows.

The operator $\Pi_{p,\Xi}$ can be modified in order to match boundary conditions. We can define, for all $f \in C^\infty([0,1])$:

$$\tilde{\Pi}_{p,\Xi}(f) = \sum_{j=1}^{n}\tilde{\lambda}_{j,p}(f)\widehat{B}_{j,p} \quad \text{with} \tag{17}$$
$$\tilde{\lambda}_{1,p}(f) = f(0),\ \tilde{\lambda}_{n,p}(f) = f(1),\ \tilde{\lambda}_{j,p}(f) = \lambda_{j,p}(f), \quad j = 2,\dots,n-1.$$

3 Dual Compatible B-Splines

Consider a set of multivariate B-splines

$$\left\{\widehat{B}_{\mathbf{A},\mathbf{p}}, \quad \mathbf{A}\in\mathcal{A}\right\}, \tag{18}$$

where $\mathcal{A}$ is a set of indices. This is a generalization of the tensor product set of multivariate splines where the functions in (18) have the structure

$$\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta}) = \widehat{B}[\Xi_{\mathbf{A},1,p_1}](\zeta_1)\cdots\widehat{B}[\Xi_{\mathbf{A},d,p_d}](\zeta_d) \tag{19}$$

and have in general uncorrelated local knot vectors, that is, two different local knot vectors $\Xi_{\mathbf{A}',\ell,p_\ell}$ and $\Xi_{\mathbf{A}'',\ell,p_\ell}$ in the ℓ-direction are not in general sub-vectors of a global knot vector. This is equivalent to the definition of *point-based splines* in [1]. We assume that there is a one-to-one correspondence between $\mathbf{A}\in\mathcal{A}$ and $\widehat{B}_{\mathbf{A},\mathbf{p}}$.

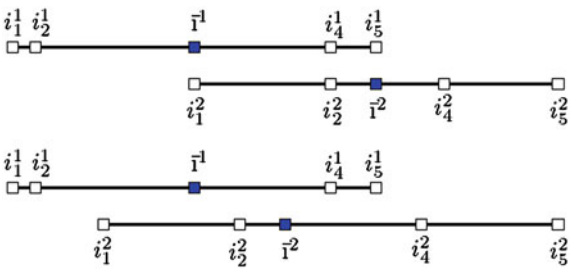


Fig. 1 Overlapping (*left*) and nonoverlapping (*right*) local knot vectors in one dimension

We say that the two p-degree local knot vectors $\Xi' = \{\xi'_1, \ldots, \xi'_{p+2}\}$ and $\Xi'' = \{\xi''_1, \ldots, \xi''_{p+2}\}$ *overlap* if they are subvectors of the same knot vector (that depends on Ξ' and Ξ''), that is there is a knot vector $\Xi = \{\xi_1, \ldots, \xi_k\}$ and k' and k'' such that

$$\begin{aligned} &\forall i = 1, \ldots, p+2, \quad \xi'_i = \xi_{i+k'} \\ &\forall i = 1, \ldots, p+2, \quad \xi''_i = \xi_{i+k''}, \end{aligned} \tag{20}$$

see Fig. 1.

We now define for multivariate B-splines, the notions of *overlap* and *partial overlap* are as follows.

Definition 1 Two B-splines $\widehat{B}_{\mathbf{A}',\mathbf{p}}$ $\widehat{B}_{\mathbf{A}'',\mathbf{p}}$ in (18) overlap if the local knot vectors in each direction overlap. Two B-splines $\widehat{B}_{\mathbf{A}',\mathbf{p}}$ $\widehat{B}_{\mathbf{A}'',\mathbf{p}}$ in (18) partially overlap if, when $\mathbf{A}' \neq \mathbf{A}''$, there exists a direction ℓ such that the local knot vectors $\Xi_{\mathbf{A}',\ell,p_\ell}$ and $\Xi_{\mathbf{A}'',\ell,p_\ell}$ are different and overlap.

From the previous Definition, overlap implies partial overlap. Examples of B-splines overlapping, only partially overlapping, and not partially overlapping are depicted in Fig. 2.

Definition 2 The set (18) is a DC set of B-splines if each pair of B-splines in it partially overlaps. Its span

$$S_{\mathbf{p}}(\mathcal{A}) = \operatorname{span}\left\{\widehat{B}_{\mathbf{A},\mathbf{p}}, \quad \mathbf{A} \in \mathcal{A}\right\}, \tag{21}$$

is denoted as DC spline space.

Note that the partially overlapping condition in Definition 2 needs to be checked only for those B-spline pairs that have nondisjoint support. Indeed, by Definition 1, any two B-splines with disjoint supports are clearly partially overlapping.

A tensor product space is clearly a DC spline space, since every pair of its multivariate B-splines always overlaps by construction. The next proposition shows how the notion of partial overlap is related with the construction of dual basis.

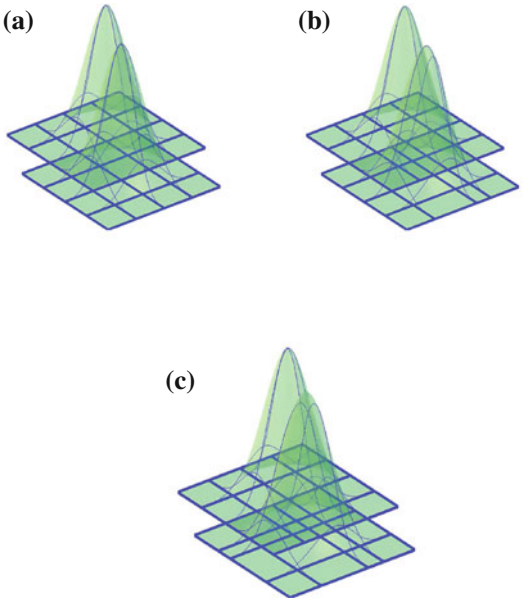


Fig. 2 Example of overlapping, partially overlapping, and not partially overlapping B-splines; *knot lines* are drawn in *blue* **a** Overlapping B-splines, **b** partially overlapping B-splines, **c** not partially overlapping B-splines

Proposition 2 *Assume that* (18) *is a DC set where each* $\widehat{B}_{\mathbf{A},\mathbf{p}}$ *is defined as in* (19)*, i.e., on the local knot vectors* $\Xi_{\mathbf{A},1,p_1},\ldots,\Xi_{\mathbf{A},d,p_d}$*. Consider an associated set of functionals*

$$\left\{\lambda_{\mathbf{A},\mathbf{p}},\quad \mathbf{A}\in\mathcal{A}\right\}, \tag{22}$$

where each $\lambda_{\mathbf{A},\mathbf{p}}$ *is*

$$\lambda_{\mathbf{A},\mathbf{p}}=\lambda[\Xi_{\mathbf{A},1,p_1}]\otimes\cdots\otimes\lambda[\Xi_{\mathbf{A},d,p_d}], \tag{23}$$

and $\lambda[\Xi_{\mathbf{A},\ell,p_\ell}]$ *denotes a univariate functional defined in* (10)*. Then* (22) *is a dual basis for* (18)*.*

Remark 1 The set of dual functionals (10) can be replaced by other choices, see, e.g., [15].

Proof Consider any $\widehat{B}_{\mathbf{A}',\mathbf{p}}$ and $\lambda_{\mathbf{A}'',\mathbf{p}}$, with $\mathbf{A}',\mathbf{A}''\in\mathcal{A}$. We then need to show that

$$\lambda_{\mathbf{A}'',\mathbf{p}}(\widehat{B}_{\mathbf{A}',\mathbf{p}})=\begin{cases}1 & \text{if } \mathbf{A}''=\mathbf{A}',\\ 0 & \text{otherwise.}\end{cases} \tag{24}$$

Clearly, if $\mathbf{A}'=\mathbf{A}''$, then we have $\lambda_{\mathbf{A}'',\mathbf{p}}(\widehat{B}_{\mathbf{A}',\mathbf{p}})=1$ from the definition of dual basis. If $\mathbf{A}'\neq\mathbf{A}''$, thanks to the partial overlap assumption, there is a direction $\bar{\ell}$ such that the local knot vectors $\Xi_{\mathbf{A}',\ell,p_\ell}$ and $\Xi_{\mathbf{A}'',\ell,p_\ell}$ differ and overlap, and then

$$\lambda[\Xi_{\mathbf{A}'',\ell,p_\ell}](\widehat{B}[\Xi_{\mathbf{A}',\ell,p_\ell}])=0,$$

and from (23),

$$\lambda_{\mathbf{A}'',\mathbf{p}}(\widehat{B}_{\mathbf{A}',\mathbf{p}}) = \prod_{\ell=1}^{d} \lambda[\Xi_{\mathbf{A}'',\ell,p_\ell}](\widehat{B}[\Xi_{\mathbf{A}',\ell,p_\ell}]) = 0.$$

The existence of dual functionals implies important properties for a DC set (18) and the related space $S_{\mathbf{p}}(\mathcal{A})$ in (21). We list such properties in the following propositions and remarks.

The first result is the linear independence of set (18), therefore forming a *basis*; they are also a partition of unity.

Proposition 3 *The B-splines in a DC set* (18) *are linearly independent. Moreover, if the constant function belongs to* $S_{\mathbf{p}}(\mathcal{A})$*, they form a partition of unity.*

Proof Assume

$$\sum_{\mathbf{A}\in\mathcal{A}} C_{\mathbf{A}} \widehat{B}_{\mathbf{A},\mathbf{p}} = 0$$

for some coefficients $C_{\mathbf{A}}$. Then for any $\mathbf{A}' \in \mathcal{A}$, applying $\lambda_{\mathbf{A}',\mathbf{p}}$ to the sum, using linearity and (24), we get

$$C_{\mathbf{A}'} = \lambda_{\mathbf{A}',\mathbf{p}}\left(\sum_{\mathbf{A}\in\mathcal{A}} C_{\mathbf{A}} \widehat{B}_{\mathbf{A}',\mathbf{p}}\right) = 0.$$

Similarly, let

$$\sum_{\mathbf{A}\in\mathcal{A}} C_{\mathbf{A}} \widehat{B}_{\mathbf{A},\mathbf{p}} = 1$$

for some coefficients $C_{\mathbf{A}}$. For any $\mathbf{A}' \in \mathcal{A}$, applying $\lambda_{\mathbf{A}',\mathbf{p}}$ as above, we get

$$C_{\mathbf{A}'} = \lambda_{\mathbf{A}',\mathbf{p}}\left(\sum_{\mathbf{A}\in\mathcal{A}} C_{\mathbf{A}} \widehat{B}_{\mathbf{A},\mathbf{p}}\right) = 1.$$

To a B-spline set (18), we can associate a parametric domain

$$\widehat{\Omega} = \bigcup_{\mathbf{A}\in\mathcal{A}} \operatorname{supp}(\widehat{B}_{\mathbf{A},\mathbf{p}})$$

Moreover, we give the following extension of the notion of Bézier mesh.

Definition 3 A parametric Bézier mesh in the parametric domain, denoted by $\widehat{\mathcal{M}}$, is the collection of the the maximal open sets $Q \subset \widehat{\Omega}$ such that for all $\mathbf{A} \in \mathcal{A}$, $\widehat{B}_{\mathbf{A},\mathbf{p}}$ is a polynomial in Q; these Q are denoted (Bézier) elements.

Proposition 4 *In a DC set* (18) *there are at most* $(p_1+1)\cdots(p_d+1)$ *B-splines that are non-null in each element* $Q \in \widehat{\mathcal{M}}$.

Proof Given any point $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d) \in \widehat{\Omega}$, denote by $\mathcal{A}(\boldsymbol{\zeta})$ the subset of $\mathbf{A} \in \mathcal{A}$ such that $\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta}) > 0$. It can be easily checked that $\mathcal{A}(\boldsymbol{\zeta})$ only depends on Q, for all $\boldsymbol{\zeta} \in Q$. Recalling (19) and introducing the notation $\Xi_{\mathbf{A},\ell,p_\ell} = \{\xi_{\ell,1}, \ldots, \xi_{\ell,p_\ell+2}\}$, to each $\mathbf{A} \in \mathcal{A}(\boldsymbol{\zeta})$ we can associate a multi-index $(i_{\mathbf{A},1}, \ldots, i_{\mathbf{A},d})$ such that

$$\forall \ell = 1, \ldots, d, \quad 1 \le i_{\mathbf{A},\ell} \le p_\ell + 1 \text{ and } \xi_{\ell,i_{\mathbf{A},\ell}} \le \zeta_\ell < \xi_{\ell,i_{\mathbf{A},\ell}+1}. \tag{25}$$

From the DC assumption, any two $\widehat{B}_{\mathbf{A}',\mathbf{p}}$ and $\widehat{B}_{\mathbf{A}'',\mathbf{p}}$ with $\mathbf{A}' \neq \mathbf{A}''$ partially overlap, that is, there are different and overlapping $\Xi_{\mathbf{A}',\ell,p_\ell}$ and $\Xi_{\mathbf{A}'',\ell,p_\ell}$; then the indices in (25) fulfill

$$\forall \mathbf{A}', \mathbf{A}'' \in \mathcal{A}(\boldsymbol{\zeta}), \quad \mathbf{A}' \neq \mathbf{A}'' \Rightarrow \exists \ell \text{ such that } i_{\mathbf{A}',\ell} \neq i_{\mathbf{A}'',\ell}. \tag{26}$$

The conclusion follows from (26), since by (25) there are at most $(p_1+1)\cdots(p_d+1)$ distinct multi-indices $(i_{\mathbf{A},1}, \ldots, i_{\mathbf{A},d})$.

Assume that each $\lambda_{\mathbf{A},\mathbf{p}}$ is defined on $L^2(\widehat{\Omega})$. An important consequence of Proposition 2 is that we can build a projection operator $\Pi_{\mathbf{p}} : L^2(\widehat{\Omega}) \to S_{\mathbf{p}}(\mathcal{A})$ by

$$\Pi_{\mathbf{p}}(f)(\boldsymbol{\zeta}) = \sum_{\mathbf{A}\in\mathcal{A}} \lambda_{\mathbf{A},\mathbf{p}}(f)\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta}) \qquad \forall f \in L^2(\widehat{\Omega}),\ \forall \boldsymbol{\zeta} \in \widehat{\Omega}. \tag{27}$$

This allows us to prove the approximation properties of $S_{\mathbf{p}}(\mathcal{A})$. The following result will make use of the notion of support extension $\widetilde{Q}$ associated to an element $Q \subset \widehat{\Omega}$ (or a generic open subset $Q \subset \widehat{\Omega}$) and to the B-spline set (18):

$$\widetilde{Q} = \bigcup_{\substack{\mathbf{A}\in\mathcal{A} \\ \text{supp}(\widehat{B}_{\mathbf{A},\mathbf{p}})\cap Q\neq\emptyset}} \text{supp}(\widehat{B}_{\mathbf{A},\mathbf{p}}).$$

Furthermore, we will denote by $\bar{Q}$, the smallest d-dimensional rectangle in $\widehat{\Omega}$ containing $\widetilde{Q}$. Then the following result holds.

Proposition 5 *Let* (18) *be a DC set of B-splines, then the projection operator* $\Pi_{\mathbf{p}}$ *in* (27) *is (locally) h-uniformly* L^2*-continuous, that is, there exists a constant C only dependent on* $\mathbf{p}$ *such that*

$$||\Pi_{\mathbf{p}}(f)||_{L^2(Q)} \le C||f||_{L^2(\widetilde{Q})} \quad \forall Q \subset \widehat{\Omega},\ \forall f \in L^2(\widehat{\Omega}).$$

Proof Let Q be an element in the parametric domain. Since Proposition 4 and since each $B_{\mathbf{A},\mathbf{p}} \leq 1$ we have that, for any $\boldsymbol{\zeta} \in Q$,

$$\sum_{\mathbf{A}\in\mathcal{A}} \left|\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta})\right| \leq C.$$

Therefore, given any point $\boldsymbol{\zeta} \in Q$, denote by $\mathcal{A}(\boldsymbol{\zeta})$ the subset of $\mathbf{A} \in \mathcal{A}$ such that $\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta}) > 0$, and denote by $Q_{\mathbf{A}}$ the common support of $\widehat{B}_{\mathbf{A},\mathbf{p}}$ and $\lambda_{\mathbf{A},\mathbf{p}}$, by $|Q_{\mathbf{A}}|$ its d-dimensional measure, using (11) it follows that

$$\begin{aligned}\left|\Pi_{\mathbf{p}}(f)(\boldsymbol{\zeta})\right|^2 &= \Big|\sum_{\mathbf{A}\in\mathcal{A}(\boldsymbol{\zeta})} \lambda_{\mathbf{A},\mathbf{p}}(f)\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta})\Big|^2 \leq C \max_{\mathbf{A}\in\mathcal{A}(\boldsymbol{\zeta})} \left|\lambda_{\mathbf{A},\mathbf{p}}(f)\right|^2 \\ &\leq C \max_{\mathbf{A}\in\mathcal{A}(\boldsymbol{\zeta})} |Q_{\mathbf{A}}|^{-1}||f||^2_{L^2(Q_{\mathbf{A}})} \\ &\leq C|Q|^{-1}||f||^2_{L^2(\widetilde{Q})},\end{aligned} \tag{28}$$

where we have used in the last step that $\forall \mathbf{A} \in \mathcal{A}(\boldsymbol{\zeta})$, $Q \subset Q_{\mathbf{A}}$ (and therefore $|Q| \leq |Q_{\mathbf{A}}|$) and that $Q_{\mathbf{A}} \subset \widetilde{Q}$. Since the bound above holds for any $\boldsymbol{\zeta} \in Q$, integrating over Q and applying (28) yields

$$||\Pi_{\mathbf{p}}(f)||^2_{L^2(Q)} \leq C||f||^2_{L^2(\widetilde{Q})}.$$

The continuity of $\Pi_{\mathbf{p}}$ implies the following approximation result in the L^2-norm:

Proposition 6 *Assume that the space of global polynomials of degree* $p = \min_{1\leq\ell\leq d}\{p_\ell\}$ *is included into the space* $S_{\mathbf{p}}(\mathcal{A})$ *and that* $\widehat{\Omega} = [0,1]^d$. *Then there exists a constant* C *only dependent on* $\mathbf{p}$ *such that for* $0 \leq s \leq p+1$

$$||f - \Pi_{\mathbf{p}}(f)||_{L^2(Q)} \leq C(h_{\bar{Q}})^s |f|_{H^s(\bar{Q})} \quad \forall Q \subset \widehat{\Omega},\ \forall f \in H^s(\widehat{\Omega}),$$

where $h_{\bar{Q}}$ *represents the diameter of* $\bar{Q}$.

Proof Let π be any p-degree polynomial. Since $\pi \in S_{\mathbf{p}}(\mathcal{A})$ and $\Pi_{\mathbf{p}}$ is a projection operator, using Proposition 5 it follows that

$$\begin{aligned}||f - \Pi_{\mathbf{p}}(f)||_{L^2(Q)} &= ||f - \pi + \Pi_{\mathbf{p}}(\pi - f)||_{L^2(Q)} \\ &\leq ||f - \pi||_{L^2(Q)} + ||\Pi_{\mathbf{p}}(\pi - f)||_{L^2(Q)} \\ &\leq (1 + C)||f - \pi||_{L^2(\widetilde{Q})} \leq (1 + C)||f - \pi||_{L^2(\bar{Q})}.\end{aligned}$$

The result finally follows by a standard polynomial approximation result.

We conclude this section with a final observation: the notion and construction of Greville sites are easily extended to DC sets of B-splines, and the following representation formula holds:

Proposition 7 *Assume that the linear polynomials belong to the space* $S_{\mathbf{p}}(\mathcal{A})$. *Then we have that*

$$\zeta_\ell = \sum_{\mathbf{A}\in\mathcal{A}} \gamma[\Xi_{\mathbf{A},\ell,p_\ell}]\widehat{B}_{\mathbf{A},\mathbf{p}}(\boldsymbol{\zeta}), \quad \forall \boldsymbol{\zeta} \in \widehat{\Omega},\ 1 \le \ell \le d, \tag{29}$$

where $\gamma[\Xi_{\mathbf{A},\ell,p_\ell}]$ *denotes the average of the* p_ℓ *internal knots of* $\Xi_{\mathbf{A},\ell,p_\ell}$.

Proof The identity (29) easily follows from the expansion of $\Pi_{\mathbf{p}}(\zeta_\ell)$ and the definition of dual functionals which is the same as in the tensor product case, yielding $\lambda_{\mathbf{A},\mathbf{p}}(\zeta_\ell) = \gamma[\Xi_{\mathbf{A},\ell,p_\ell}]$.

References

1. Sederberg, T., Zheng, J., Bakenov, A., Nasri, A.: T-splines and T-NURCCSs. ACM Trans. Graph. **22**(3), 477–484 (2003)
2. Bazilevs, Y., Calo, V., Cottrell, J.A., Evans, J.A., Hughes, T.J.R., Lipton, S., Scott, M., Sederberg, T.: Isogeometric analysis using T-splines. Comput. Methods Appl. Mech. Eng. **199**(5–8), 229–263 (2010)
3. Hosseini, S., Remmers, J.J., Verhoosel, C.V., Borst, R.: An isogeometric solid-like shell element for nonlinear analysis. Int. J. Numer. Methods Eng. **95**(3), 238–256 (2013)
4. Bazilevs, Y., Hsu, M.-C., Scott, M.: Isogeometric fluid–structure interaction analysis with emphasis on non-matching discretizations, and with application to wind turbines. Comput. Methods Appl. Mech. Eng. **249**, 28–41 (2012)
5. Dimitri, R., De Lorenzis, L., Scott, M., Wriggers, P., Taylor, R., Zavarise, G.: Isogeometric large deformation frictionless contact using T-splines. Comput. Methods Appl. Mech. Eng. **269**, 394–414 (2014)
6. Sederberg, T., Cardon, D., Finnigan, G., North, N., Zheng, J., Lyche, T.: T-spline simplification and local refinement. ACM Trans. Graph. **23**(3), 276–283 (2004)
7. Dörfel, M., Jüttler, B., Simeon, B.: Adaptive isogeometric analysis by local *h*-refinement with T-splines. Comput. Methods Appl. Mech. Eng. **199**(5–8), 264–275 (2010)
8. Scott, M., Li, X., Sederberg, T., Hughes, T.J.R.: Local refinement of analysis-suitable T-splines. Comput. Methods Appl. Mech. Eng. **213–216**, 206–222 (2012)
9. Li, X., Zheng, J., Sederberg, T., Hughes, T., Scott, M.: On linear independence of T-spline blending functions. Comput. Aided Geom. Des. **29**(1), 63–76 (2012)
10. Beirão da Veiga, L., Buffa, A., Sangalli, G., Vázquez, R.: Analysis-suitable T-splines of arbitrary degree: definition, linear independence and approximation properties. Math. Models Methods Appl. Sci. **23**(11), 1979–2003 (2013)
11. Dokken, T., Lyche, T., Pettersen, K.F.: Polynomial splines over locally refined box-partitions. Comput. Aided Geom. Design **30**(3), 331–356 (2013)
12. Vuong, A.-V., Giannelli, C., Jüttler, B., Simeon, B.: A hierarchical approach to adaptive local refinement in isogeometric analysis. Comput. Methods Appl. Mech. Eng. **200**(49–52), 3554–3567 (2011)
13. Schumaker, L.L.: Spline Functions: Basic Theory, 3rd edn. Cambridge Mathematical Library, Cambridge University Press, Cambridge (2007)
14. Bazilevs, Y., Beirão da Veiga, L., Cottrell, J.A., Hughes, T.J.R., Sangalli, G.: Isogeometric analysis: approximation, stability and error estimates for *h*-refined meshes. Math. Models Methods Appl. Sci. **16**(7), 1031–1090 (2006)
15. Lee, B.-G., Lyche T., Mørken K.: Some examples of quasi-interpolants constructed from local spline projectors. In: Mathematical Methods for Curves and Surfaces (Oslo, 2000), Innov. Appl. Math., pp. 243–252. Vanderbilt University Press, Nashville, TN (2001)

Multivariate Anisotropic Interpolation on the Torus

Ronny Bergmann and Jürgen Prestin

Abstract We investigate the error of periodic interpolation, when sampling a function on an arbitrary pattern on the torus. We generalize the periodic Strang-Fix conditions to an anisotropic setting and provide an upper bound for the error of interpolation. These conditions and the investigation of the error especially take different levels of smoothness along certain directions into account.

Keywords Anisotropic periodic interpolation · Shift invariant spaces · Lattices · Interpolation error bounds

1 Introduction

Approximation by equidistant translates of a periodic function was first investigated in the univariate case [6, 15]. The multivariate case was developed in [21–23], where the introduction of the periodic Strang-Fix conditions enabled a unified way to the error estimates [16, 17].

Recently, many approaches such as contourlets [7], curvelets [8] or shearlets [10], analyze and decompose multivariate data by focusing on certain anisotropic features. A more general approach are wavelets with composite dilations [11, 13], which inherit an MRA structure similar to the classical wavelets. For periodic functions the multivariate periodic wavelet analysis [1, 9, 14, 18, 19] is a periodic approach to such an anisotropic decomposition. The pattern $\mathscr{P}(\mathbf{M})$ as a basic ingredient to these scaling and wavelet functions models equidistant points with preference of

R. Bergmann (✉)
Department of Mathematics, University Kaiserslautern, Paul-Ehrlich-Straße 31,
D-67663 Kaiserslautern, Germany
e-mail: bergmann@mathematik.uni-kl.de

J. Prestin
Institute of Mathematics, University of Lübeck, Ratzeburger Allee 160,
D-23562 Lübeck, Germany
e-mail: prestin@math.uni-luebeck.de

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_3,

direction, i.e., fixing one direction $\mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{v}\| = 1$, we obtain equidistant points along this direction in the pattern $\mathscr{P}(\mathbf{M})$, though other directions might have other point distances.

This paper presents the interpolation on such patterns $\mathscr{P}(\mathbf{M})$, where $\mathbf{M} \in \mathbb{Z}^{d\times d}$, $d \in \mathbb{N}$, is a regular integer matrix. In order to derive an upper bound for the interpolation error, we introduce function spaces $A^{\alpha}_{\mathbf{M},q}$, where each function is of different directional smoothness due to decay properties of the Fourier coefficients imposed. The periodic Strang-Fix conditions can be generalized to this anisotropic setting, characterizing and quantifying the reproduction capabilities of a fundamental interpolant with respect to a certain set of trigonometric polynomials. Such a fundamental interpolant can then be used for approximation, where the error can be bounded for the functions having certain directional smoothness, i.e., the space $A^{\alpha}_{\mathbf{M},\,q}$.

The rest of the paper is organized as follows: In Sect. 2 we introduce the basic preliminary notations of the pattern $\mathscr{P}(\mathbf{M})$, the corresponding discrete Fourier transform $\mathscr{F}(\mathbf{M})$, and the spaces $A^{\alpha}_{\mathbf{M},\,q}$. Section 3 is devoted to the interpolation problem on the pattern $\mathscr{P}(\mathbf{M})$ and the ellipsoidal periodic Strang-Fix conditions, which generalize the periodic Strang-Fix conditions to an anisotropic setting. For this interpolation, we derive an upper bound with respect to $A^{\alpha}_{\mathbf{M},\,q}$ in Sect. 4. Finally, in Sect. 5 we provide an example that the ellipsoidal Strang-Fix conditions are fulfilled by certain periodized three-directional box splines and their higher dimensional analogs.

2 Preliminaries

2.1 Patterns

Let $d \in \mathbb{N}$. For a regular integral matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$ and two vectors $\mathbf{h}, \mathbf{k} \in \mathbb{Z}^d$ we write $\mathbf{h} \equiv \mathbf{k} \bmod \mathbf{M}$ if there exists a vector $\mathbf{z} \in \mathbb{Z}^d$ such that $\mathbf{h} = \mathbf{k} + \mathbf{M}\mathbf{z}$. The set of congruence classes

$$[\mathbf{h}]_{\mathbf{M}} := \left\{\mathbf{k} \in \mathbb{Z}^d \,;\, \mathbf{k} \equiv \mathbf{h} \bmod \mathbf{M}\right\}, \quad \mathbf{h} \in \mathbb{Z}^d,$$

forms a partition of $\mathbb{Z}^d$ and using the addition $[\mathbf{h}]_{\mathbf{M}} + [\mathbf{k}]_{\mathbf{M}} := [\mathbf{h} + \mathbf{k}]_{\mathbf{M}}$, we obtain the generating group $\left(\mathscr{G}(\mathbf{M}), +\right)$, where the generating set $\mathscr{G}(\mathbf{M})$ is any set of congruence class representatives. If we apply the congruence with respect to the unit matrix

$$\mathbf{E}_d := \left(\delta_{i,\,j}\right)_{i,\,j=1}^{d} \in \mathbb{R}^{d\times d}, \quad \text{where } \delta_{i,\,j} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{else,} \end{cases}$$

denotes the Kronecker delta, to the lattice $\Lambda_{\mathbf{M}} := \mathbf{M}^{-1}\mathbb{Z}^d \subset \mathbb{Q}^d$, we also get congruence classes. Let further $\mathbf{e}_j := \left(\delta_{i,\,j}\right)_{i=1}^{d}$ denote the jth unit vector. We obtain the pattern group $\left(\mathscr{P}(\mathbf{M}), +\right)$ on the corresponding congruence classes $[\mathbf{y}]_{\mathbf{E}_d}, \mathbf{y} \in \Lambda_{\mathbf{M}}$, where the pattern $\mathscr{P}(\mathbf{M})$ is again any set of congruence class representatives of the

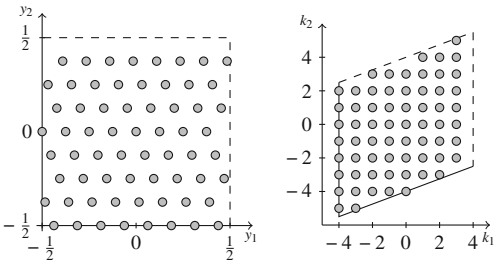


Fig. 1 The pattern $\mathscr{P}_{\mathrm{S}}(\mathbf{M})$ (*left*) and the generating set $\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$ (*right*), where $\mathbf{M} = \left(\begin{smallmatrix} 8 & 3 \\ 0 & 8 \end{smallmatrix}\right)$

congruence classes on the lattice $\Lambda_{\mathbf{M}}$. For any pattern $\mathscr{P}(\mathbf{M})$ we obtain a generating set by $\mathscr{G}(\mathbf{M}) = \mathbf{M}\mathscr{P}(\mathbf{M})$. Using a geometrical argument [4, Lemma II.7], we get $|\mathscr{P}(\mathbf{M})| = |\mathscr{G}(\mathbf{M})| = |\det \mathbf{M}| =: m$. A special pattern $\mathscr{P}_{\mathrm{S}}(\mathbf{M})$ and its corresponding generating set $\mathscr{G}_{\mathrm{S}}(\mathbf{M})$ are given by

$$\mathscr{P}_{\mathrm{S}}(\mathbf{M}) := \left[-\frac{1}{2}, \frac{1}{2}\right)^d \cap \Lambda_{\mathbf{M}} \quad \text{and} \quad \mathscr{G}_{\mathrm{S}}(\mathbf{M}) := \mathbf{M}\left[-\frac{1}{2}, \frac{1}{2}\right)^d \cap \mathbb{Z}^d.$$

We will apply the usual addition, when performing an addition on the set of representatives, i.e., for $\mathbf{x}, \mathbf{y} \in \mathscr{P}(\mathbf{M})$ the expression $\mathbf{x} + \mathbf{y}$ is an abbreviation for choosing the unique element of $[\mathbf{x} + \mathbf{y}]_{\mathbf{E}_d} \cap \mathscr{P}(\mathbf{M})$. In fact, for any discrete group $\mathscr{G} = (S, + \mod 1)$ with respect to addition modulo 1, there exists a matrix $\mathbf{M}$, whose pattern $\mathscr{P}(\mathbf{M})$ coincides with the set S [2, Theorem 1.8]. Figure 1 gives an example of a pattern $\mathscr{P}(\mathbf{M})$ of a matrix $\mathbf{M}$ and a generating set $\mathscr{G}(\mathbf{M}^{\mathrm{T}})$, where the matrix is an upper triangular matrix. By scaling and shearing, the points of the pattern lie dense along a certain direction.

The discrete Fourier transform is defined by applying the Fourier matrix

$$\mathscr{F}(\mathbf{M}) := \frac{1}{\sqrt{m}} \left(\mathrm{e}^{-2\pi \mathrm{i} \mathbf{h}^{\mathrm{T}} \mathbf{y}}\right)_{\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}), \mathbf{y} \in \mathscr{P}(\mathbf{M})}$$

to a vector $\mathbf{a} = \left(a_{\mathbf{y}}\right)_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} \in \mathbb{C}^m$ having the same order of elements as the columns of $\mathscr{F}(\mathbf{M})$. We write for the Fourier transform $\hat{\mathbf{a}} := \left(\hat{a}_{\mathbf{h}}\right)_{\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})} = \sqrt{m}\mathscr{F}(\mathbf{M})\mathbf{a}$, where the vector $\hat{\mathbf{a}}$ is ordered in the same way as the rows of the Fourier matrix $\mathscr{F}(\mathbf{M})$. Further investigations of patterns and generating sets, especially concerning subpatterns and shift invariant spaces, can be found in [14], which is extended in [1] with respect to bases and certain orderings of the elements of both sets to obtain a fast Fourier transform.

Finally, we denote by $\lambda_1(\mathbf{M}), \ldots, \lambda_d(\mathbf{M})$ the eigenvalues of $\mathbf{M}$ including their multiplicities in increasing order, i.e., for $i < j$ we get $|\lambda_i(\mathbf{M})| \leq |\lambda_j(\mathbf{M})|$.

For the rest of this paper, let $|\lambda_d(\mathbf{M})| \geq 2$. To emphasize this fact, we call a matrix $\mathbf{M}$ that fulfills $|\lambda_d(\mathbf{M})| \geq 2$ an expanding matrix.

2.2 Function Spaces

For functions $f : \mathbb{T}^d \to \mathbb{C}$ on the torus $\mathbb{T}^d := \mathbb{R}^d/2\pi\mathbb{Z}^d$ consider the Banach spaces $\mathrm{L}_p(\mathbb{T}^d)$, $1 \le p \le \infty$, with norm

$$\|f \mid \mathrm{L}_p(\mathbb{T}^d)\|^p := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} |f(\mathbf{x})|^p \,\mathrm{d}\mathbf{x}$$

and the usual modification for $p = \infty$, that $\|f \mid \mathrm{L}_\infty(\mathbb{T}^d)\| = \operatorname{ess\,sup}_{\mathbf{x} \in \mathbb{T}^d} |f(\mathbf{x})|$. Analogously, for sequences $\mathbf{c} := \{c_\mathbf{z}\}_{\mathbf{z} \in \mathscr{X}}$, $\mathscr{X} \subseteq \mathbb{Z}^d$, the Banach spaces $\ell_q(\mathscr{X})$, $1 \le q \le \infty$, are defined with norm

$$\|\mathbf{c} \mid \ell_q(\mathscr{X})\|^q := \sum_{\mathbf{k} \in \mathscr{X}} \left|c_\mathbf{k}\right|^q,$$

again with the usual modification for $q = \infty$. For $f \in \mathrm{L}_1(\mathbb{T}^d)$ the Fourier coefficients are given by

$$c_\mathbf{k}(f) := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} f(\mathbf{x}) \mathrm{e}^{-\mathrm{i}\mathbf{k}^\mathrm{T}\mathbf{x}} \,\mathrm{d}\mathbf{x}, \quad \mathbf{k} \in \mathbb{Z}^d.$$

For $\beta \ge 0$, we define the ellipsoidal weight function $\sigma_\beta^\mathbf{M}$, which was similarly introduced in [2, Sect. 1.2],

$$\sigma_\beta^\mathbf{M}(\mathbf{k}) := \left(1 + \|\mathbf{M}\|_2^2 \|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2\right)^{\beta/2}, \quad \mathbf{k} \in \mathbb{Z}^d,$$

to define for $q \ge 1$ the spaces

$$A^\beta_{\mathbf{M},q}(\mathbb{T}^d) := \left\{ f \in \mathrm{L}_1(\mathbb{T}^d);\ \left\| f \mid A^\beta_{\mathbf{M},q} \right\| < \infty \right\},$$

where

$$\left\| f \mid A^\beta_{\mathbf{M},q} \right\| := \left\| \{\sigma_\beta^\mathbf{M}(\mathbf{k}) c_\mathbf{k}(f)\}_{\mathbf{k} \in \mathbb{Z}^d} \mid \ell_q(\mathbb{Z}^d) \right\|.$$

A special case is given by $A(\mathbb{T}^d) := A^0_{\mathbf{M},1}(\mathbb{T}^d)$, which is the Wiener algebra of all functions with absolutely convergent Fourier series. We see that $\|\mathbf{M}^\mathrm{T}\|_2^2 = \lambda_d(\mathbf{M}^\mathrm{T}\mathbf{M}) > 1$. For any diagonal matrix $\mathbf{M} = \operatorname{diag}(N, \ldots, N)$, $N \in \mathbb{N}$, the weight function simplifies to $(1 + \|\mathbf{k}\|_2^2)^{\beta/2}$ and these spaces resemble the spaces used in [23] to derive error bounds for interpolation by translates. Even more, if we fix $\alpha \in \mathbb{R}$ and $q \ge 1$, due to the inequalities

$$\left(1 + \|\mathbf{M}\|_2^2 \|\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2\right)^{\alpha/2} \le \left(\frac{\lambda_d(\mathbf{M}^\mathrm{T}\mathbf{M})}{\lambda_1(\mathbf{M}^\mathrm{T}\mathbf{M})}\right)^{\alpha/2} \left(1 + \|\mathbf{z}\|_2^2\right)^{\alpha/2}$$

and

$$(1 + \|\mathbf{z}\|_2^2)^{\alpha/2} = (1 + \|\mathbf{M}^{\mathrm{T}}\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2)^{\alpha/2} \leq (1 + \|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2)^{\alpha/2}$$

we have, that all spaces $A^{\alpha}_{\mathbf{M},q}$ of regular integer matrices $\mathbf{M}$ are equal to $A^{\alpha}_{\mathbf{E}_d,q}$, which is the same as A^{α}_{q} in [23]. However, each of the different norms provides a different quantification of the functions $f \in A^{\alpha}_{q}$. We keep the matrix $\mathbf{M}$ in the notation of the space in order to distinguish the specific norm that we will use.

For the weight $\sigma^{\mathbf{M}}_{\beta}$ we finally have the following lemma.

Lemma 1 *For a regular expanding matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$, i.e., $|\lambda_d(\mathbf{M})| \geq 2$, and an ellipsoidal weight function $\sigma^{\mathbf{M}}_{\beta}$, where $\beta > 0$ we have*

$$\sigma^{\mathbf{M}}_{\beta}(\mathbf{k} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) \leq \|\mathbf{M}\|_2^{\beta}\sigma^{\mathbf{M}}_{\beta}(\mathbf{k})\sigma^{\mathbf{M}}_{\beta}(\mathbf{z}) \quad \textit{for } \mathbf{k},\ \mathbf{z} \in \mathbb{Z}^d. \tag{1}$$

Proof We have $2 \leq \|\mathbf{M}\|_2 = \sqrt{\lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M})}$. For $\mathbf{z} = \mathbf{0}$ or $\mathbf{k} = \mathbf{0}$ the assertion holds. For $\mathbf{k}, \mathbf{z} \neq \mathbf{0}$ we apply the triangle inequality and the submultiplicativity of the spectral norm to obtain

$$\begin{aligned}\sigma^{\mathbf{M}}_{\beta}(\mathbf{k} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) &= (1 + \|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}(\mathbf{k} + \mathbf{M}^{\mathrm{T}}\mathbf{z})\|_2^2)^{\beta/2} \\ &\leq \|\mathbf{M}\|_2^{\beta}(1 + \|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2 + 2\|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2\|\mathbf{z}\|_2 + \|\mathbf{z}\|_2^2)^{\beta/2}.\end{aligned}$$

Using $\|\mathbf{M}\|_2\|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2 \geq 1$, $\|\mathbf{z}\|_2 \geq 1$ and $\|\mathbf{M}\|_2 \geq 2$, we further get

$$\begin{aligned}&\sigma^{\mathbf{M}}_{\beta}(\mathbf{k} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) \\ &\quad\leq \|\mathbf{M}\|_2^{\beta}(1 + \|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2 + \|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2\|\mathbf{z}\|_2^2 + \|\mathbf{M}^{\mathrm{T}}\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2)^{\beta/2} \\ &\quad\leq \|\mathbf{M}\|_2^{\beta}(1 + \|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2)^{\beta/2}(1 + \|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2)^{\beta/2} \\ &\quad= \|\mathbf{M}\|_2^{\beta}\sigma^{\mathbf{M}}_{\beta}(\mathbf{k})\sigma^{\mathbf{M}}_{\beta}(\mathbf{z}).\end{aligned}$$

□

Remark 1 In the same way one would obtain $\sigma^{\mathbf{M}}_{\beta}(\mathbf{k} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) \leq 2^{\beta}\|\mathbf{M}\|_2^{\beta}\sigma^{\mathbf{M}}_{\beta}(\mathbf{k})\sigma^{\mathbf{M}}_{\beta}(\mathbf{z})$ for all regular integral matrices with $\|\mathbf{M}\|_2 \geq 1$ with slightly bigger constant 2^{β}. For the matrices of interest, this slight difference is not very important and we focus on the former one for simplicity.

3 Interpolation and the Strang-Fix Condition

This section is devoted to interpolation on a pattern $\mathscr{P}(\mathbf{M})$ and its corresponding periodic Strang-Fix conditions. The periodic Strang-Fix conditions were introduced in [5, 16] for tensor product grids as a counterpart to the strong Strang-Fix

conditions on the Euclidean space $\mathbb{R}^d$ and generalized in [21, 23]. We generalize them to arbitrary patterns on the torus.

A space of functions V is called $\mathbf{M}$-invariant, if for all $\mathbf{y} \in \mathscr{P}(\mathbf{M})$ and all functions $\varphi \in V$ the translates $\mathrm{T}_{\mathbf{y}}\varphi := \varphi(\circ - 2\pi\mathbf{y}) \in V$. Especially the space

$$V_{\mathbf{M}}^{\varphi} := \operatorname{span}\left\{ \mathrm{T}_{\mathbf{y}}\varphi \,;\, \mathbf{y} \in \mathscr{P}(\mathbf{M}) \right\}$$

of translates of φ is $\mathbf{M}$-invariant. A function $\xi \in V_{\mathbf{M}}^{\varphi}$ is of the form $\xi = \sum_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} a_{\mathbf{y}} \mathrm{T}_{\mathbf{y}}\varphi$. For $\varphi \in \mathrm{L}_1(\mathbb{T}^d)$ an easy calculation on the Fourier coefficients using the unique decomposition of $\mathbf{k} \in \mathbb{Z}^d$ into $\mathbf{k} = \mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}, \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}), \mathbf{z} \in \mathbb{Z}^d$, yields, that $\xi \in V_{\mathbf{M}}^{\varphi}$ holds if and only if

$$c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\xi) = \hat{a}_{\mathbf{h}} c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\varphi) \quad \text{for all } \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}), \mathbf{z} \in \mathbb{Z}^d, \tag{2}$$

is fulfilled, where $\hat{\mathbf{a}} = \left(\hat{a}_{\mathbf{h}}\right)_{\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})} = \sqrt{m}\mathscr{F}(\mathbf{M})\mathbf{a}$ denotes the discrete Fourier transform of $\mathbf{a} \in \mathbb{C}^m$.

Further, the space of trigonometric polynomials on the generating set $\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$ is denoted by

$$\mathscr{T}_{\mathbf{M}} := \left\{ \varphi \,;\, \varphi = \sum_{\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} a_{\mathbf{h}} \mathrm{e}^{\mathrm{i}\mathbf{h}^{\mathrm{T}}\circ},\ a_{\mathbf{h}} \in \mathbb{C} \right\}.$$

For any function $\varphi \in \mathrm{L}_1(\mathbb{T}^d)$ the Fourier partial sum $\mathrm{S}_{\mathbf{M}}\varphi \in \mathscr{T}_{\mathbf{M}}$ given by

$$\mathrm{S}_{\mathbf{M}}\varphi := \sum_{\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} c_{\mathbf{h}}(\varphi) \mathrm{e}^{\mathrm{i}\mathbf{h}^{\mathrm{T}}\circ}$$

is such a trigonometric polynomial.

The discrete Fourier coefficients of a pointwise given φ are defined by

$$c_{\mathbf{h}}^{\mathbf{M}}(\varphi) := \frac{1}{m} \sum_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} \varphi(2\pi\mathbf{y}) \mathrm{e}^{-2\pi\mathrm{i}\mathbf{h}^{\mathrm{T}}\mathbf{y}}, \quad \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}), \tag{3}$$

which are related to the Fourier coefficients for $\varphi \in A(\mathbb{T}^d)$ by the following aliasing formula.

Lemma 2 *Let $\varphi \in A(\mathbb{T}^d)$ and the regular matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$ be given. Then the discrete Fourier coefficients $c_{\mathbf{h}}^{\mathbf{M}}(\varphi)$ are given by*

$$c_{\mathbf{k}}^{\mathbf{M}}(\varphi) = \sum_{\mathbf{z} \in \mathbb{Z}^d} c_{\mathbf{k}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\varphi), \quad \mathbf{k} \in \mathbb{Z}^d. \tag{4}$$

Proof Writing each point evaluation of φ in (3) as its Fourier series, we obtain due to the absolute convergence of the series

$$
\begin{aligned}
c_{\mathbf{k}}^{\mathbf{M}}(\varphi) &= \frac{1}{m} \sum_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} \left(\sum_{\mathbf{h} \in \mathbb{Z}^d} c_{\mathbf{h}}(\varphi) \mathrm{e}^{2\pi \mathrm{i} \mathbf{h}^{\mathrm{T}} \mathbf{y}} \right) \mathrm{e}^{-2\pi \mathrm{i} \mathbf{k}^{\mathrm{T}} \mathbf{y}} \\
&= \frac{1}{m} \sum_{\mathbf{h} \in \mathbb{Z}^d} c_{\mathbf{h}}(\varphi) \sum_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} \mathrm{e}^{-2\pi \mathrm{i} (\mathbf{k}-\mathbf{h})^{\mathrm{T}} \mathbf{y}} \\
&= \sum_{\mathbf{z} \in \mathbb{Z}^d} c_{\mathbf{k}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(f).
\end{aligned}
$$

The last equality is valid because the sum over $\mathbf{y}$ simplifies to m if $\mathbf{k} \equiv \mathbf{h} \bmod \mathbf{M}^{\mathrm{T}}$, and vanishes otherwise, cf. [20, Lemma 2.7]. □

Definition 1 Let $\mathbf{M} \in \mathbb{Z}^{d \times d}$ be a regular matrix. A function $\mathrm{I}_{\mathbf{M}} \in V_{\mathbf{M}}^{\varphi}$ is called fundamental interpolant or Lagrange function of $V_{\mathbf{M}}^{\varphi}$ if

$$
\mathrm{I}_{\mathbf{M}}(2\pi \mathbf{y}) := \delta_{\mathbf{0},\mathbf{y}}^{\mathbf{E}_d}, \quad \mathbf{y} \in \mathscr{P}(\mathbf{M}), \quad \text{where } \delta_{\mathbf{x},\mathbf{y}}^{\mathbf{M}} := \begin{cases} 1 & \text{if } \mathbf{y} \equiv \mathbf{x} \bmod \mathbf{M}, \\ 0 & \text{else.} \end{cases}
$$

The following lemma characterizes the existence of such a fundamental interpolant.

Lemma 3 *Given a regular matrix* $\mathbf{M} \in \mathbb{Z}^{d \times d}$ *and a function* $\varphi \in A(\mathbb{T}^d)$*, the fundamental interpolant* $\mathrm{I}_{\mathbf{M}} \in V_{\mathbf{M}}^{\varphi}$ *exists if and only if*

$$
\sum_{\mathbf{z} \in \mathbb{Z}^d} c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\varphi) \neq 0, \quad \textit{for all } \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}). \tag{5}
$$

If the fundamental interpolant $\mathrm{I}_{\mathbf{M}} \in V_{\mathbf{M}}^{\varphi}$ *exists, it is uniquely determined.*

Proof Assume the fundamental interpolant $\mathrm{I}_{\mathbf{M}} \in V_{\mathbf{M}}^{\varphi}$ exists. Hence, there exists a vector $\mathbf{a} = (a_{\mathbf{h}})_{\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})}$ such that for its Fourier transform $\hat{\mathbf{a}} = \sqrt{m} \mathscr{F}(\mathbf{M}) \mathbf{a}$ it holds due to (2) that

$$
c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}}) = \hat{a}_{\mathbf{h}} c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\varphi), \quad \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}), \ \mathbf{z} \in \mathbb{Z}^d.
$$

Applying this equality to the discrete Fourier coefficients of $\mathrm{I}_{\mathbf{M}}$ yields

$$
c_{\mathbf{h}}^{\mathbf{M}}(\mathrm{I}_{\mathbf{M}}) = \sum_{\mathbf{z} \in \mathbb{Z}^d} c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}}) = \hat{a}_{\mathbf{h}} \sum_{\mathbf{z} \in \mathbb{Z}^d} c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\varphi) = \hat{a}_{\mathbf{h}} c_{\mathbf{h}}^{\mathbf{M}}(\varphi). \tag{6}
$$

The discrete Fourier coefficients are known by Definition 1 and [20, Lemma 2.7] as $c_{\mathbf{h}}^{\mathbf{M}}(\mathrm{I}_{\mathbf{M}}) = \frac{1}{m}$, $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$, which is nonzero for all $\mathbf{h}$ and hence (5) follows.

On the other hand, if (5) is fulfilled, then the function ξ, which is defined by

$$
c_{\mathbf{k}}(\xi) = \frac{c_{\mathbf{k}}(\varphi)}{m c_{\mathbf{k}}^{\mathbf{M}}(\varphi)}, \quad \mathbf{k} \in \mathbb{Z}^d, \tag{7}
$$

is in the space $V_{\mathbf{M}}^{\varphi}$ having the coefficients $\hat{a}_{\mathbf{h}} = (mc_{\mathbf{h}}^{\mathbf{M}}(\varphi))^{-1}$, $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$. The discrete Fourier coefficients also fulfill $c_{\mathbf{h}}^{\mathbf{M}}(\xi) = \frac{1}{m}$. Hence, again by Definition 1 and [20, Lemma 2.7], ξ is a fundamental interpolant with respect to the pattern $\mathscr{P}(\mathbf{M})$. If the fundamental interpolant $\mathrm{I}_{\mathbf{M}}$ exists, (7) also provides uniqueness. □

The associated interpolation operator $\mathrm{L}_{\mathbf{M}} f$ is given by

$$\mathrm{L}_{\mathbf{M}} f := \sum_{\mathbf{y} \in \mathscr{P}(\mathbf{M})} f(2\pi \mathbf{y})\, \mathrm{T}_{\mathbf{y}}\, \mathrm{I}_{\mathbf{M}} = m \sum_{\mathbf{k} \in \mathbb{Z}^d} c_{\mathbf{k}}^{\mathbf{M}}(f) c_{\mathbf{k}}(\mathrm{I}_{\mathbf{M}}) \mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}} \circ}. \tag{8}$$

The following definition introduces the periodic Strang-Fix conditions, which require the Fourier coefficients $c_{\mathbf{k}}(\mathrm{I}_{\mathbf{M}})$ of the fundamental interpolant to decay in a certain ellipsoidal way. The condition number $\kappa_{\mathbf{M}}$ of $\mathbf{M}$ is given by

$$\kappa_{\mathbf{M}} := \sqrt{\frac{\lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M})}{\lambda_1(\mathbf{M}^{\mathrm{T}}\mathbf{M})}} = \|\mathbf{M}\|_2 \|\mathbf{M}^{-1}\|_2.$$

Definition 2 Given a regular expanding matrix $\mathbf{M} \in \mathbb{Z}^{d \times d}$, a fundamental interpolant $\mathrm{I}_{\mathbf{M}} \in \mathrm{L}_1(\mathbb{T}^d)$ fulfills the ellipsoidal (periodic) Strang-Fix conditions of order $s > 0$ for $q \geq 1$ and an $\alpha \in \mathbb{R}^+$, if there exists a nonnegative sequence $\mathbf{b} = \{b_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{Z}^d} \subset \mathbb{R}_0^+$, such that for all $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$, $\mathbf{z} \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ we have

1. $|1 - mc_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})| \leq b_{\mathbf{0}} \kappa_{\mathbf{M}}^{-s} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^s$,
2. $|mc_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}})| \leq b_{\mathbf{z}} \kappa_{\mathbf{M}}^{-s} \|\mathbf{M}\|_2^{-\alpha} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^s$

with

$$\gamma_{\mathrm{SF}} := \left\| \{\sigma_{\alpha}^{\mathbf{M}}(\mathbf{z}) b_{\mathbf{z}}\}_{\mathbf{z} \in \mathbb{Z}^d} \middle| \ell_q(\mathbb{Z}^d) \right\| < \infty.$$

For both properties we enforce a stronger decay than by the ellipse defined by the level curves of $\|\mathbf{M}^{-\mathrm{T}} \circ\|_2$, i.e., we have an upper bound by

$$(\kappa_{\mathbf{M}}^{-s} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^s \leq \kappa_{\mathbf{M}}^{-s} \|\mathbf{M}^{-\mathrm{T}}\|_2^s \|\mathbf{h}\|_2^s = \left(\lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M})\right)^{-s/2} \|\mathbf{h}\|^s).$$

The second one enforces a further stronger decay with respect to α, i.e.,

$$(\kappa_{\mathbf{M}}^{-s} \|\mathbf{M}\|_2^{-\alpha} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^s \leq (\lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M}))^{-(\alpha+s)/2} \|\mathbf{h}\|_2^s).$$

For the one-dimensional case or the tensor product case, i.e., $\mathbf{M} = \mathrm{diag}(N, \ldots, N)$ we have $\kappa_{\mathbf{M}} = 1$, $\lambda_1(\mathbf{M}^{\mathrm{T}}\mathbf{M}) = \lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M}) = N$, and this simplifies to the already known case $N^{-\alpha-s}\|\mathbf{h}\|_2$. Looking at the level curves of the map $\|\mathbf{M}^{-\mathrm{T}} \circ \|$, we see they produce ellipsoids, where $|\lambda_d(\mathbf{M})|$ is the length of the longest axis. Hence the decay is normalized with respect to the longest axis of the ellipsoid.

4 Error Bounds for Interpolation

In order to investigate the error of interpolation $\|f - \mathrm{L}_{\mathbf{M}} f\|$, where $\mathrm{L}_{\mathbf{M}}$ is the interpolation operator into $V_{\mathbf{M}}^{\varphi}$ for certain $\varphi \in A(\mathbb{T}^d)$, we use the triangle inequality with respect to any norm

$$\|f - \mathrm{L}_{\mathbf{M}} f\| \le \|\mathrm{S}_{\mathbf{M}} f - \mathrm{L}_{\mathbf{M}} \mathrm{S}_{\mathbf{M}} f\| + \|f - \mathrm{S}_{\mathbf{M}} f\| + \|\mathrm{L}_{\mathbf{M}}(f - \mathrm{S}_{\mathbf{M}} f)\|$$

and look at these three terms separately.

Theorem 1 *For an expanding regular matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$, a trigonometric polynomial $f \in \mathscr{T}_{\mathbf{M}}$ and a fundamental interpolant $\mathrm{I}_{\mathbf{M}} \in A(\mathbb{T}^d)$ fulfilling the ellipsoidal Strang-Fix conditions for fixed values $s \ge 0$, $\alpha > 0$, and $q \ge 1$ we have*

$$\left\| f - \mathrm{L}_{\mathbf{M}} f \middle| A^{\alpha}_{\mathbf{M},q} \right\| \le \left(\frac{1}{\|\mathbf{M}\|_2}\right)^{s} \gamma_{\mathrm{SF}} \left\| f \middle| A^{\alpha+s}_{\mathbf{M},q} \right\|. \tag{9}$$

Proof The proof is given for $q < \infty$. For $q = \infty$ the same arguments apply with the usual modifications with respect to the norm. Looking at the Fourier coefficients of $\mathrm{L}_{\mathbf{M}} f$ in (8) for $f \in \mathscr{T}_{\mathbf{M}}$ yields

$$c_{\mathbf{h}}(\mathrm{L}_{\mathbf{M}} f) = m c^{\mathbf{M}}_{\mathbf{h}}(f) c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}}) = m c_{\mathbf{h}}(f) c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}}), \quad \mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}}),$$

and hence we have

$$f - \mathrm{L}_{\mathbf{M}} f = \sum_{\mathbf{k}\in\mathbb{Z}^d} \left(c_{\mathbf{k}}(f) - m c^{\mathbf{M}}_{\mathbf{k}}(f) c_{\mathbf{k}}(\mathrm{I}_{\mathbf{M}})\right) \mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}}\circ}.$$

Using the unique decomposition of $\mathbf{k} \in \mathbb{Z}^d$ into $\mathbf{k} = \mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}$, $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$, $\mathbf{z} \in \mathbb{Z}^d$, yields

$$f - \mathrm{L}_{\mathbf{M}} f = \sum_{\mathbf{h}\in\mathscr{G}(\mathbf{M}^{\mathrm{T}})} c_{\mathbf{h}}(f) e^{\mathrm{i}\mathbf{h}^{\mathrm{T}}\circ} \Big(\big(1 - m c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})\big) - \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} m c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}}) \mathrm{e}^{\mathrm{i}\mathbf{M}^{\mathrm{T}}\mathbf{z}\circ} \Big).$$

Applying the definition of the norm in $A^{\alpha}_{\mathbf{M},q}(\mathbb{T}^d)$, we obtain

$$\begin{aligned} &\left\| f - \mathrm{L}_{\mathbf{M}} f \middle| A^{\alpha}_{\mathbf{M},q} \right\|^q \\ &= \sum_{\mathbf{h}\in\mathscr{G}(\mathbf{M}^{\mathrm{T}})} |c_{\mathbf{h}}(f)|^q \left(\left| (1 - m c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})) \sigma^{\mathbf{M}}_{\alpha}(\mathbf{h}) \right|^q + \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \left| m c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}}) \sigma^{\mathbf{M}}_{\alpha}(\mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) \right|^q \right). \end{aligned}$$

Using the Strang-Fix conditions of the fundamental interpolant $\mathrm{I}_{\mathbf{M}}$ and Lemma 1 we get the following upper bound

$$\begin{aligned}
\left\| f - \mathrm{L}_{\mathbf{M}} f \middle| A^{\alpha}_{\mathbf{M},q} \right\|^q &\leq \sum_{\mathbf{h}\in\mathscr{G}(\mathbf{M}^{\mathrm{T}})} |c_{\mathbf{h}}(f)|^q \Bigg(b_{\mathbf{0}}^q \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{sq} \sigma^{\mathbf{M}}_{\alpha q}(\mathbf{h}) \kappa_{\mathbf{M}}^{-sq} \\
&\quad + \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} b_{\mathbf{z}}^q \kappa_{\mathbf{M}}^{-sq} \|\mathbf{M}\|_2^{-\alpha q} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{sq} \sigma^{\mathbf{M}}_{\alpha q}(\mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}) \Bigg) \\
&\leq \Bigg(\sum_{\mathbf{h}\in\mathscr{G}(\mathbf{M}^{\mathrm{T}})} |c_{\mathbf{h}}(f)|^q \|\mathbf{M}^{-\mathrm{T}}\|_2^{sq} \kappa_{\mathbf{M}}^{-sq} \|\mathbf{h}\|_2^{sq} \sigma^{\mathbf{M}}_{\alpha q}(\mathbf{h}) \Bigg) \\
&\quad \Bigg(\sum_{\mathbf{z}\in\mathbb{Z}^d} (\sigma^{\mathbf{M}}_{\alpha}(\mathbf{z}) b_{\mathbf{z}})^q \Bigg) \\
&\leq \gamma_{\mathrm{SF}}^q \|\mathbf{M}\|_2^{-sq} \left\| f \middle| A^{\alpha+s}_{\mathbf{M},q} \right\|^q .
\end{aligned}$$

□

Theorem 2 *Let* $\mathbf{M} \in \mathbb{Z}^{d\times d}$ *be regular. If* $f \in A^{\mu}_{\mathbf{M},q}(\mathbb{T}^d)$, $q \geq 1$, $\mu \geq \alpha \geq 0$, *then*

$$\left\| f - \mathrm{S}_{\mathbf{M}} f \middle| A^{\alpha}_{\mathbf{M},q} \right\| \leq \left(\frac{2}{\|\mathbf{M}\|_2} \right)^{\mu-\alpha} \left\| f \middle| A^{\mu}_{\mathbf{M},q} \right\| .$$

Proof This proof is given for $q < \infty$. For $q = \infty$ the same arguments apply with the usual modifications with respect to the norm. We examine the left-hand side of the inequality, apply $\sigma^{\mathbf{M}}_{\alpha}(\mathbf{k}) = \sigma^{\mathbf{M}}_{\alpha-\mu}(\mathbf{k})\sigma^{\mathbf{M}}_{\mu}(\mathbf{k})$, and obtain

$$\begin{aligned}
\left\| f - \mathrm{S}_{\mathbf{M}} f \mid A^{\alpha}_{\mathbf{M},q} \right\| &= \left\| \{\sigma^{\mathbf{M}}_{\alpha}(\mathbf{k}) c_{\mathbf{k}}(f)\}_{\mathbf{k}\in\mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \mid \ell_q\left(\mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})\right) \right\| \\
&\leq \max_{\mathbf{k}\in\mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \sigma^{\mathbf{M}}_{\alpha-\mu}(\mathbf{k}) \left\| f \mid A^{\mu}_{\mathbf{M},q} \right\| .
\end{aligned}$$

The decomposition of $\mathbf{k} \in \mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$ into $\mathbf{k} = \mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}$, $\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$, yields that $\mathbf{0} \neq \mathbf{z} \in \mathbb{Z}^d$ and hence none of these integral points lies inside the parallelotope $\mathbf{M}^{\mathrm{T}}\left[-\frac{1}{2}, \frac{1}{2}\right)^d$. Hence, $\mathbf{M}^{-\mathrm{T}}\mathbf{k}$ lies outside $\left[-\frac{1}{2}, \frac{1}{2}\right)^d$ and we have

$$\begin{aligned}
\max_{\mathbf{k}\in\mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \sigma^{\mathbf{M}}_{\alpha-\mu}(\mathbf{k}) &= \max_{\mathbf{k}\in\mathbb{Z}^d\setminus\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \left(1 + \|\mathbf{M}\|_2^2 \|\mathbf{M}^{-\mathrm{T}}\mathbf{k}\|_2^2\right)^{\frac{\alpha-\mu}{2}} \\
&\leq \max_{j\in\{1,\ldots,d\}} \left(1 + \frac{\|\mathbf{M}\|_2^2}{4}\right)^{\frac{\alpha-\mu}{2}} \\
&\leq \left(\frac{\|\mathbf{M}\|_2^2}{4}\right)^{\frac{\alpha-\mu}{2}} .
\end{aligned}$$

□

Indeed, Theorem 2 does hold for any regular matrix $\mathbf{M}$. It is not required that the matrix has to be expanding. For the following theorem, let $|\mathbf{z}| := \big(|z_1|, \ldots, |z_d|\big)^{\mathrm{T}}$ denote the vector of the absolute values of the elements of the vector $\mathbf{z} \in \mathbb{Z}^d$.

Theorem 3 *For an expanding regular matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$ let $\mathrm{I}_{\mathbf{M}}$ be a fundamental interpolant such that*

$$\gamma_{\mathrm{IP}} := m \begin{cases} \max\limits_{\mathbf{h}\in\mathscr{G}_S(\mathbf{M}^{\mathrm{T}})} \left(|c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})|^q + \|\mathbf{M}\|_2^{\alpha q} \sum\limits_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} |\sigma_\alpha^{\mathbf{M}}(\mathbf{z}) c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}})|^q \right)^{1/q} & \\ & \text{for } q<\infty, \\ \max\limits_{\mathbf{h}\in\mathscr{G}_S(\mathbf{M}^{\mathrm{T}})} \sup\Big\{ |c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})|,\ \|\mathbf{M}\|_2^{\alpha} |\sigma_\alpha^{\mathbf{M}}(\mathbf{z}) c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}})|;\ \mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\} \Big\} & \\ & \text{for } q=\infty \end{cases}$$

is finite. Then we get for $f \in A^{\mu}_{\mathbf{M},q}(\mathbb{T}^d)$, $q \geq 1$, $\mu \geq \alpha \geq 0$, and $\mu > d(1-1/q)$

$$\Big\| \mathrm{L}_{\mathbf{M}}(f - \mathrm{S}_{\mathbf{M}} f) \Big| A^{\alpha}_{\mathbf{M},q} \Big\| \leq \gamma_{\mathrm{IP}}\gamma_{\mathrm{Sm}} \left(\frac{1}{\|\mathbf{M}\|_2} \right)^{\mu-\alpha} \big\| f \big| A^{\mu}_{\mathbf{M},q} \big\|,$$

where

$$\gamma_{\mathrm{Sm}} := (1+d)^{\alpha/2} 2^{-\mu} \begin{cases} \left(\sum\limits_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \|2|\mathbf{z}| - \mathbf{1}\|_2^{-p\mu} \right)^{1/p} & \text{for } \mathrm{q} > 1,\ \frac{1}{\mathrm{p}} + \frac{1}{\mathrm{q}} = 1, \\ \sup\limits_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \|2|\mathbf{z}| - \mathbf{1}\|_2^{-\mu} & \text{for } \mathrm{q} = 1. \end{cases}$$

Proof This proof is given for $q < \infty$. For $q = \infty$ the same arguments apply with the usual modifications with respect to the norm. We write the norm on the left-hand side of the inequality as

$$\begin{aligned} \Big\| \mathrm{L}_{\mathbf{M}}(f - \mathrm{S}_{\mathbf{M}} f) \Big| A^{\alpha}_{\mathbf{M},q} \Big\|^q &= \Big\| \sum_{\mathbf{k}\in\mathbb{Z}^d} \sigma_\alpha^{\mathbf{M}}(\mathbf{k}) c_{\mathbf{k}} \left(\mathrm{L}_{\mathbf{M}}(f - \mathrm{S}_{\mathbf{M}} f) \right) \mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}}\circ} \Big| A^{\alpha}_{\mathbf{M},q} \Big\|^q \\ &= \sum_{\mathbf{k}\in\mathbb{Z}^d} \Big| \sigma_\alpha^{\mathbf{M}}(\mathbf{k}) m c_{\mathbf{k}}^{\mathbf{M}}(f - \mathrm{S}_{\mathbf{M}} f) c_{\mathbf{k}}(\mathrm{I}_{\mathbf{M}}) \Big|^q . \end{aligned}$$

By decomposing $\mathbf{k} = \mathbf{h} + \mathbf{M}^{\mathrm{T}}\mathbf{z}$, $\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$, $\mathbf{z} \in \mathbb{Z}^d$, and using Lemma 1 we obtain

$$\begin{aligned}\left\|\mathrm{L}_{\mathbf{M}}(f-\mathrm{S}_{\mathbf{M}}f)\middle|A^{\alpha}_{\mathbf{M},q}\right\|^q &\le \sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}\left|\sigma^{\mathbf{M}}_{\alpha}(\mathbf{h})mc^{\mathbf{M}}_{\mathbf{h}}(f-\mathrm{S}_{\mathbf{M}}f)\right|^q\\ &\quad\times\left(|c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})|^q+\|\mathbf{M}\|_2^{\alpha q}\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}|\sigma^{\mathbf{M}}_{\alpha}(\mathbf{z})c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}})|^q\right)\\ &\le \gamma^q_{\mathrm{IP}}\sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}|\sigma^{\mathbf{M}}_{\alpha}(\mathbf{h})c^{\mathbf{M}}_{\mathbf{h}}(f-\mathrm{S}_{\mathbf{M}}f)|^q.\end{aligned}$$

In the remaining sum we first apply the aliasing formula (4). Then, the Hölder inequality yields

$$\begin{aligned}\sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}|\sigma^{\mathbf{M}}_{\alpha}(\mathbf{h})c^{\mathbf{M}}_{\mathbf{h}}(f-\mathrm{S}_{\mathbf{M}}f)|^q &= \sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}\sigma^{\mathbf{M}}_{\alpha q}(\mathbf{h})\Big(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}|c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(f)|\Big)^q\\ &\le \sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}\sigma^{\mathbf{M}}_{\alpha q}(\mathbf{h})\Big(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\sigma^{\mathbf{M}}_{-\mu p}(\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z})\Big)^{q/p}\\ &\quad\times\Big(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}|\sigma^{\mathbf{M}}_{\mu}(\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z})c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(f)|^q\Big).\end{aligned}$$

The first sum over $\mathbf{z}$ converges due to $p\mu > d$, i.e., analogously to the proof of Theorem 2, we get for $\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$

$$\begin{aligned}\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\sigma^{\mathbf{M}}_{-\mu p}(\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}) &= \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\big(1+\|\mathbf{M}\|_2^2\|\mathbf{M}^{-\mathrm{T}}\mathbf{h}+\mathbf{z}\|_2^2\big)^{-p\mu/2}\\ &\le \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\big(1+\|\mathbf{M}\|_2^2\big\|\,|\mathbf{z}|-\tfrac{1}{2}\mathbf{1}\big\|_2^2\big)^{-p\mu/2}\\ &\le \sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\left(\frac{\|\mathbf{M}\|_2^2}{4}\|2|\mathbf{z}|-\mathbf{1}\|_2^2\right)^{-p\mu/2}\\ &= \|\mathbf{M}\|_2^{-p\mu}2^{-p\mu}\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\|2|\mathbf{z}|-\mathbf{1}\|_2^{-p\mu}.\end{aligned}$$

Using for $\alpha\ge 0$

$$\begin{aligned}\max_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}\sigma^{\mathbf{M}}_{\alpha}(\mathbf{h}) &\le \Big(1+\|\mathbf{M}\|_2^2\big\|\mathbf{M}^{-\mathrm{T}}\mathbf{M}^{\mathrm{T}}\tfrac{1}{2}\mathbf{1}\big\|_2^2\Big)^{\alpha/2}\\ &= \big(1+\tfrac{d}{4}\|\mathbf{M}\|_2^2\big)^{\alpha/2}\\ &\le (1+d)^{\alpha/2}\|\mathbf{M}\|_2^{\alpha},\end{aligned}\tag{10}$$

the upper bound for the last factor can be given as

$$\begin{aligned}
&\sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} |\sigma_{\alpha}^{\mathbf{M}}(\mathbf{h})c_{\mathbf{h}}^{\mathbf{M}}(f-\mathrm{S}_{\mathbf{M}}f)|^q \\
&\quad\le \sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \sigma_{\alpha q}^{\mathbf{M}}(\mathbf{h}) 2^{-\mu q}\|\mathbf{M}\|_2^{-\mu q} \left(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \||2\mathbf{z}|-\mathbf{1}\|^{-p\mu}\right)^{q/p} \\
&\qquad\times \left(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} |\sigma_{\mu}^{\mathbf{M}}(\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z})c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(f)|^q\right) \\
&\quad\le 2^{-\mu q}\|\mathbf{M}\|_2^{-\mu q} \left(\max_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \sigma_{q\alpha}^{\mathbf{M}}(\mathbf{h})\right) \left(\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \||2\mathbf{z}|-\mathbf{1}\|^{-p\mu}\right)^{q/p} \\
&\qquad\times \left(\sum_{\mathbf{h}\in\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})}\sum_{\mathbf{z}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} |\sigma_{\mu}^{\mathbf{M}}(\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z})c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(f)|^q\right) \\
&\quad\le \|\mathbf{M}\|_2^{(\alpha-\mu)q}\gamma_{\mathrm{Sm}}^q \|f|A_{\mathbf{M},q}^{\mu}\|^q. \qquad \square
\end{aligned}$$

Remark 2 It is easy to see that for a fundamental interpolant $\mathrm{I}_{\mathbf{M}}$ satisfying the ellipsoidal periodic Strang-Fix conditions of order s for q and α we have

$$\gamma_{\mathrm{IP}} \le C\cdot\gamma_{\mathrm{SF}}$$

where the constant C depends on $\mathbf{M}$, α, s and q but is especially independent of f.

We summarize our treatment of the interpolation error in the following theorem.

Theorem 4 *Let an expanding regular matrix $\mathbf{M}\in\mathbb{Z}^{d\times d}$ and a fundamental interpolant $\mathrm{I}_{\mathbf{M}}$ fulfilling the periodic ellipsoidal Strang-Fix conditions of order s for $q\ge 1$, and $\alpha\ge 0$ be given. Then for $f\in A_{\mathbf{M},q}^{\mu}(\mathbb{T}^d)$, $\mu\ge\alpha\ge 0$ and $\mu > d(1-1/q)$, we have*

$$\|f-\mathrm{L}_{\mathbf{M}}f|A_{\mathbf{M},q}^{\alpha}\| \le C_{\rho}\left(\frac{1}{\|\mathbf{M}\|_2}\right)^{\rho}\|f|A_{\mathbf{M},q}^{\mu}\|,$$

where $\rho := \min\{s,\mu-\alpha\}$ and

$$C_{\rho} := \begin{cases} \gamma_{\mathrm{SF}} + 2^{\mu-\alpha} + \gamma_{\mathrm{IP}}\gamma_{\mathrm{Sm}} & \text{for } \rho = \mathrm{s}, \\ (1+d)^{s+\alpha-\mu}\gamma_{\mathrm{SF}} + 2^{\mu-\alpha} + \gamma_{\mathrm{IP}}\gamma_{\mathrm{Sm}} & \text{for } \rho = \mu-\alpha. \end{cases}$$

Proof For $\rho = s$ Theorems 1–3 can be applied directly due to $\|f|A_{\mathbf{M},q}^{\alpha+s}\| \le \|f|A_{\mathbf{M},q}^{\mu}\|$. If $\rho = \mu-\alpha$, we have to replace Theorem 1 by an upper bound with respect to μ. Using this theorem and the inequality in (10), we get

$$
\begin{aligned}
&\| \mathrm{S}_{\mathbf{M}} f - \mathrm{L}_{\mathbf{M}} \mathrm{S}_{\mathbf{M}} f | A^{\alpha}_{\mathbf{M},q} \| \\
&\quad \le \gamma_{\mathrm{SF}} \|\mathbf{M}\|_2^{-s} \left\| \{\sigma^{\mathbf{M}}_{\alpha+s}(\mathbf{h}) c_{\mathbf{h}}(f)\}_{\in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \middle| \ell_q(\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})) \right\| \\
&\quad \le \gamma_{\mathrm{SF}} \max_{\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \sigma^{\mathbf{M}}_{\alpha+s-\mu}(\mathbf{h}) \|\mathbf{M}\|_2^{-s} \left\| \{\sigma^{\mathbf{M}}_{\mu}(\mathbf{h}) c_{\mathbf{h}}(f)\}_{\in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})} \middle| \ell_q(\mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})) \right\| \\
&\quad \le \gamma_{\mathrm{SF}} (1+d)^{s+\alpha-\mu} \|\mathbf{M}\|_2^{s+\alpha-\mu} \ \|\mathbf{M}\|_2^{-s} \| f | A^{\mu}_{\mathbf{M},q} \|. \qquad \square
\end{aligned}
$$

Remark 3 The factor $\kappa_{\mathbf{M}}^{-s}$ in both constraints of the Strang-Fix conditions, cf. Definition 2, enforces a strong decay on the Fourier coefficients of the fundamental interpolant $\mathrm{I}_{\mathbf{M}}$. Omitting this factor in both constraints, i.e., leaving just $\|\mathbf{M}\|_2^{-\alpha}$ in the second one, weakens to a less restrictive constraint on the fundamental interpolant $\mathrm{I}_{\mathbf{M}}$. This changes the decay rate from

$$
\left(\frac{1}{\|\mathbf{M}\|_2} \right)^s = \left(\frac{1}{\sqrt{\lambda_d(\mathbf{M}^{\mathrm{T}}\mathbf{M})}} \right)^s
$$

to

$$
\left(\frac{\kappa_{\mathbf{M}}}{\|\mathbf{M}\|_2} \right)^s = \left(\|\mathbf{M}^{-\mathrm{T}}\|_2 \right)^s = \left(\frac{1}{\sqrt{\lambda_1(\mathbf{M}^{\mathrm{T}}\mathbf{M})}} \right)^s,
$$

which is then also the rate of decay in Theorem 4. Hence, while this formulation eases the constraints with respect to the decay by restricting it just to the shortest axis of the ellipsoid given by $\|\mathbf{M}^{-\mathrm{T}} \circ \|_2 = 1$, the rate of convergence is also relaxed. On the other hand, the stronger formulation in Theorems 1–4 requires the fundamental interpolant to fulfill stronger constraints.

When increasing the number of sampling points, i.e., the determinant $|\det \mathbf{M}|$, for both variations there are cases where the bound is not decreased. Namely, in the first one if the value $\|\mathbf{M}\|_2$ is not increased, in the second version if the value $\|\mathbf{M}^{-\mathrm{T}}\|_2$ is not decreased.

Again, for the tensor product case $\mathbf{M} = \mathrm{diag}(N, \ldots, N)$ and the one-dimensional setting, both formulations of the Strang-Fix conditions and the resulting error bounds are equal.

5 The Three-Directional Box Splines

For a function $g \in \mathrm{L}_1(\mathbb{R}^d)$ on the Euclidean space $\mathbb{R}^d$ the Fourier transform is given by

$$
\hat{g}(\xi) := \int_{\mathbb{R}^d} g(\mathbf{x}) \mathrm{e}^{\mathrm{i}\xi^{\mathrm{T}}\mathbf{x}} \, \mathrm{d}\mathbf{x}, \quad \xi \in \mathbb{R}^d,
$$

and we introduce the periodization with respect to a regular matrix $\mathbf{M} \in \mathbb{Z}^{d\times d}$ for a function $g : \mathbb{R}^d \to \mathbb{C}$ having compact support

$$g^{\mathbf{M}} : \mathbb{T}^d \to \mathbb{C}, \quad g^{\mathbf{M}} := \sum_{\mathbf{z}\in\mathbb{Z}^d} g\big(\tfrac{1}{2\pi}\mathbf{M}(\circ - 2\pi\mathbf{z})\big).$$

Its Fourier coefficients $c_{\mathbf{k}}(g^{\mathbf{M}})$, $\mathbf{k} \in \mathbb{Z}^d$, can be obtained from $\hat{g}$ by using the substitution $\mathbf{y} = \frac{1}{2\pi}\mathbf{M}\mathbf{x}$, i.e., $\mathrm{d}\mathbf{y} = \frac{1}{(2\pi)^d} m\,\mathrm{d}\mathbf{x}$. Hence

$$\begin{aligned} c_{\mathbf{k}}(g^{\mathbf{M}}) &= \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} \sum_{\mathbf{z}\in\mathbb{Z}^d} g\big(\tfrac{1}{2\pi}\mathbf{M}(\mathbf{x}-2\pi\mathbf{z})\big)\mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}}\mathbf{x}}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^d} \sum_{\mathbf{z}\in\mathbb{Z}^d} \int_{\mathbb{T}^d} g\big(\tfrac{1}{2\pi}\mathbf{M}(\mathbf{x}-2\pi\mathbf{z})\big)\mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}}\mathbf{x}}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{m} \int_{\mathbb{R}^d} g(\mathbf{y})\mathrm{e}^{\mathrm{i}\mathbf{k}^{\mathrm{T}}\left(2\pi\mathbf{M}^{-1}\mathbf{y}\right)}\,\mathrm{d}\mathbf{y} \\ &= \frac{1}{m}\hat{g}\big(2\pi\mathbf{M}^{-\mathrm{T}}\mathbf{k}\big). \end{aligned}$$

The same applied to the Lagrange interpolation symbol $\tilde{g}(\xi) := \sum_{\mathbf{z}\in\mathbb{Z}^d} g(\mathbf{z})\mathrm{e}^{\mathrm{i}\xi^{\mathrm{T}}\mathbf{z}}$ yields $c_{\mathbf{h}}^{\mathbf{M}}(\tilde{g}^{\mathbf{M}}) = \frac{1}{m}\tilde{g}(2\pi\mathbf{M}^{-\mathrm{T}}\mathbf{h})$, $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$.

We look at an example for the case $d = 2$. The three-directional box splines $B_{\mathbf{p}}$, $\mathbf{p} = (p_1, p_2, p_3) \in \mathbb{N}^3$, $p_j \geq 1$, $j = 1, 2, 3$, are given by their Fourier transform

$$\hat{B}_{\mathbf{p}}(\xi) := \big(\operatorname{sinc}\tfrac{1}{2}\xi_1\big)^{p_1}\big(\operatorname{sinc}\tfrac{1}{2}\xi_2\big)^{p_2}\big(\operatorname{sinc}\tfrac{1}{2}(\xi_1+\xi_2)\big)^{p_3}.$$

Applying the periodization, we obtain the function $B_{\mathbf{p}}^{\mathbf{M}} : \mathbb{T}^2 \to \mathbb{C}$ by its Fourier coefficients

$$c_{\mathbf{k}}(B_{\mathbf{p}}^{\mathbf{M}}) = \tfrac{1}{m}\big(\operatorname{sinc}\pi\mathbf{k}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{e}_1)^{p_1}\big(\operatorname{sinc}\pi\mathbf{k}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{e}_2)^{p_2}\big(\operatorname{sinc}\pi\mathbf{k}^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{e}_1+\mathbf{e}_2)\big)^{p_3}.$$

Due to positivity of $\hat{B}_{\mathbf{p}}(\xi), \xi \in [-\pi,\pi]^2$, cf. [3, Sect. 4], we know that $c_{\mathbf{h}}(B_{\mathbf{p}}^{\mathbf{M}}) \neq 0$ for $\mathbf{h} \in \mathscr{G}(\mathbf{M}^{\mathrm{T}})$. Hence by [14, Corollary 3.5] the translates $\mathrm{T}_{\mathbf{y}}B_{\mathbf{p}}^{\mathbf{M}}$, $\mathbf{y} \in \mathscr{P}(\mathbf{M})$, form a basis of $V_{\mathbf{M}}^{B_{\mathbf{p}}^{\mathbf{M}}}$.

Theorem 5 *Let $\mathbf{M} \in \mathbb{Z}^{2\times 2}$ be a regular matrix, $\mathbf{p} \in \mathbb{N}^3$, $p_j \geq 1$, $j = 1, 2, 3$ a vector, $s := \min\{p_1+p_2, p_1+p_3, p_2+p_3\}$ and $\alpha \geq 0$, $q \geq 1$, such that $s - \alpha > 2$.*

The fundamental interpolant $\mathrm{I}_{\mathbf{M}} \in V_{\mathbf{M}}^{B_{\mathbf{p}}^{\mathbf{M}}}$ of the periodized 3-directional box spline $B_{\mathbf{p}}^{\mathbf{M}}$ fulfills the periodic ellipsoidal Strang-Fix conditions of order $s-\alpha$ for α and q, which depends on $\mathbf{p}$.

Proof We first examine the case $\alpha = 0$. Taking a look at the second Strang-Fix condition, we obtain for $\mathbf{h} \in \mathcal{G}_S(\mathbf{M}^T)$ and $\mathbf{z} \in \mathbb{Z}^d \setminus \{\mathbf{0}\}$, following the same steps as in the proof of Theorem 1.10 in [16], the inequality

$$\begin{aligned}
|mc_{\mathbf{h}+\mathbf{M}^T\mathbf{z}}(\mathrm{I}_\mathbf{M})| &= \left| \frac{c_{\mathbf{h}+\mathbf{M}^T\mathbf{z}}(B_\mathbf{p}^\mathbf{M})}{c_\mathbf{h}^\mathbf{M}(B_\mathbf{p}^\mathbf{M})} \right| \\
&\leq \frac{1}{c_\mathbf{h}^\mathbf{M}(B_\mathbf{p}^\mathbf{M})} \frac{|\sin \pi \mathbf{h}^T\mathbf{M}^{-1}(\mathbf{e}_1+\mathbf{e}_2)|^{p_3}}{|\pi(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T(\mathbf{e}_1+\mathbf{e}_2)|^{p_3}} \prod_{j=1}^{2} \frac{|\sin \pi \mathbf{h}^T\mathbf{M}^{-1}\mathbf{e}_j|^{p_j}}{|\pi(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T\mathbf{e}_j|^{p_j}} \\
&\leq \frac{1}{c_\mathbf{h}^\mathbf{M}(B_\mathbf{p}^\mathbf{M})} \frac{|\mathbf{h}^T\mathbf{M}^{-1}(\mathbf{e}_1+\mathbf{e}_2)|^{p_3}}{|(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T(\mathbf{e}_1+\mathbf{e}_2)|^{p_3}} \prod_{j=1}^{2} \frac{|\mathbf{h}^T\mathbf{M}^{-1}\mathbf{e}_j|^{p_j}}{|(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T\mathbf{e}_j|^{p_j}}.
\end{aligned}$$

For $z_1 \neq 0, z_2 \neq 0, z_1+z_2 \neq 0$, and $|z_1+z_2| \neq 1$ it holds using $|(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T\mathbf{e}_j| \geq (|z_j| - \frac{1}{2})$, $j = 1, 2$, and $|(\mathbf{M}^{-T}\mathbf{h}+\mathbf{z})^T(\mathbf{e}_1+\mathbf{e}_2|) \geq (|z_1+z_2|-1)$ that

$$|mc_{\mathbf{h}+\mathbf{M}^T\mathbf{z}}(B_\mathbf{p}^\mathbf{M})| \leq \left(|\mathbf{h}^T\mathbf{M}^{-1}\mathbf{e}_1| + |\mathbf{h}^T\mathbf{M}^{-1}\mathbf{e}_2|\right)^s \frac{1}{(|z_1+z_2|-1)^{p_3}} \prod_{j=1}^{2} \frac{1}{(|z_j|-\frac{1}{2})^{p_j}},$$

where applying the Cauchy-Schwarz inequality $|h_1| + |h_2| \leq \sqrt{2}\|\mathbf{h}\|_2$ yields

$$|mc_{\mathbf{h}+\mathbf{M}^T\mathbf{z}}(B_\mathbf{p}^\mathbf{M})| \leq \|\mathbf{M}^{-T}\mathbf{h}\|_2^s \frac{2^{s/2}}{(|z_1|-\frac{1}{2})^{p_1}(|z_2|-\frac{1}{2})^{p_2}(|z_1+z_2|-1)^{p_3}}.$$

Defining

$$A := \left(\min_{\mathbf{h} \in \mathcal{G}_S(\mathbf{M}^T)} mc_\mathbf{h}^\mathbf{M}(B_\mathbf{p}^\mathbf{M}) \right)^{-1}$$

we can use the last inequality to obtain that the fundamental interpolant $\mathrm{I}_\mathbf{M}$ corresponding to $B_\mathbf{p}^\mathbf{M}$ fulfills the Strang-Fix conditions of order s with $\alpha = 0$, where the series for γ_{SF} is given by

$$b_\mathbf{z} = b_\mathbf{z}^0 = \frac{2^{s/2}A}{(|z_1|-\frac{1}{2})^{p_1}(|z_2|-\frac{1}{2})^{p_2}(|z_1+z_2|-1)^{p_3}},$$

at least for $\mathbf{z} = (z_1, z_2)^T$ with $z_1 \neq 0$, $z_2 \neq 0$, $z_1+z_2 \neq 0$ and $|z_1+z_2| \neq 1$. An upper bound for the remaining indices $\mathbf{z}$ can be established using similar arguments as for this case. These estimates can be directly transcribed from the already mentioned proof, cf. [16, pp. 51–57], including the bound for the first of the Strang-Fix conditions, i.e., $b_\mathbf{0}^0$. This concludes the proof for the case $\alpha = 0$.

For $\alpha \geq 0$ we define the series $b_\mathbf{z} := 2^{-\alpha/2}\|\mathbf{M}^{-T}\|_2^{-\alpha} b_\mathbf{z}^0$, $\mathbf{z} \in \mathbb{Z}^2$, and obtain for the first Strang-Fix condition with $\mathbf{h} \in \mathcal{G}_S(\mathbf{M}^T)$ and using $\|\mathbf{M}\|_2^\alpha \geq 1$ that

$$\begin{aligned}|1 - m c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}})| &\le b_{\mathbf{0}}^{0} \kappa_{\mathbf{M}}^{-s} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^s \\ &\le \kappa_{\mathbf{M}}^{-\alpha} 2^{-\alpha/2} b_{\mathbf{0}}^{0} \kappa_{\mathbf{M}}^{-(s-\alpha)} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{s-\alpha} \\ &\le b_{\mathbf{0}} \kappa_{\mathbf{M}}^{-(s-\alpha)} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{s-\alpha}.\end{aligned}$$

For the second condition we get

$$\begin{aligned}|m c_{\mathbf{h}+\mathbf{M}^{\mathrm{T}}\mathbf{z}}(\mathrm{I}_{\mathbf{M}})| &\le b_{\mathbf{z}}^{0} \kappa_{\mathbf{M}}^{-\alpha} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{\alpha} \kappa_{\mathbf{M}}^{-(s-\alpha)} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{s-\alpha} \\ &\le \|\mathbf{M}^{-\mathrm{T}}\|_2^{-\alpha} 2^{-\alpha/2} b_{\mathbf{z}}^{0} \|\mathbf{M}\|_2^{-\alpha} \kappa_{\mathbf{M}}^{-(s-\alpha)} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{s-\alpha} \\ &\le b_{\mathbf{z}} \|\mathbf{M}\|_2^{-\alpha} \kappa_{\mathbf{M}}^{-(s-\alpha)} \|\mathbf{M}^{-\mathrm{T}}\mathbf{h}\|_2^{s-\alpha},\end{aligned}$$

where the first inequality in both cases is mentioned for completeness. The series that is used to define γ_{SF} is given by

$$\gamma_{\mathrm{SF}}^{q} = \sum_{\mathbf{z} \in \mathbb{Z}^2} |(1 + \|\mathbf{M}\|_2^2 \|\mathbf{M}^{-\mathrm{T}}\mathbf{z}\|_2^2)^{\alpha/2} b_{\mathbf{z}}|^q,$$

which converges for $s - \alpha > 2$ by applying again the same inequalities that were used for the case of a diagonal matrix $\mathbf{M} = \mathrm{diag}(N, \ldots, N), q = 2$, and $\alpha = 0$ in Theorem 1.10 of [16]. □

This can also be applied to the d-variate case, $d > 2$, using the $\frac{d(d+1)}{2}$-directional box spline $B_{\mathbf{p}}$, $\mathbf{p} \in \mathbb{N}^{\frac{d(d+1)}{2}}$, consisting of the directions $\mathbf{e}_j$, $j = 1, \ldots, d$, and $\mathbf{e}_j + \mathbf{e}_i$, $i, j = 1, \ldots, d, i \neq j$, the corresponding four-directional box spline [16, Theorem. 1.11], and its multivariate version, the d^2-directional box spline, which can be generated analogously to the $\frac{d(d+1)}{2}$-directional box spline, i.e., using the directions $\mathbf{e}_j$, $\mathbf{e}_i + \mathbf{e}_j$ and $\mathbf{e}_i - \mathbf{e}_j$. Nevertheless, for the periodized d^2-directional box spline $B_{\mathbf{q}}^{\mathbf{M}}$, $\mathbf{q} \in \mathbb{N}^{d^2}$, the fundamental interpolant $\mathrm{I}_{\mathbf{M}}$ does not exist. This can be seen by looking at $B_{\mathbf{q}}^{\mathbf{M}}$ in the Fourier domain, where it does contain at least one two-dimensional four-directional box spline as a factor. Hence the non-normal interpolation of the four-directional box spline, which was investigated in [12] carries over to the higher dimensional case. In order to apply the above-mentioned theorems, we have to use the so-called incorrect interpolation, i.e., we set $c_{\mathbf{h}}(\mathrm{I}_{\mathbf{M}}) = m^{-1}$ for $\mathbf{h} \in \mathscr{G}_{\mathrm{S}}(\mathbf{M}^{\mathrm{T}})$, where $c_{\mathbf{h}}^{\mathbf{M}}(B_{\mathbf{q}}^{\mathbf{M}}) = 0$.

Acknowledgments We thank both the anonymous reviewers for their valuable remarks which improved the presentation of this paper.

References

1. Bergmann, R.: The fast Fourier transform and fast wavelet transform for patterns on the torus. Appl. Comput. Harmon. Anal. **35**(1), 39–51 (2013)
2. Bergmann, R.: Translationsinvariante Räume multivariater anisotroper Funktionen auf dem Torus. Dissertation, Universität zu Lübeck (2013)

3. de Boor, C., Höllig, K., Riemenschneider, S.D.: Bivariate cardinal interpolation by splines on a three-direction mesh. Illinois J. Math. **29**(4), 533–566 (1985)
4. de Boor, C., Höllig, K., Riemenschneider, S.D.: Box splines. Springer, New York (1993)
5. Brumme, G.: Error estimates for periodic interpolation by translates. In: Laurent, P.J., Le Méhauté, A., Schumaker, L.L. (eds.) Wavelets, Images and Surface Fitting, pp. 75–82. AK Peters Ltd., Boston (1994)
6. Delvos, F.J.: Periodic interpolation on uniform meshes. J. Approx. Theory **51**, 71–80 (1987)
7. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005)
8. Fadili, M.J., Starck, J.L.: Curvelets and ridgelets. Encyclopedia of Complexity and Systems Science, vol. 3, pp. 1718–1738. Springer, New York (2007)
9. Goh, S.S., Lee, S.L., Teo, K.M.: Multidimensional periodic multiwavelets. J. Approx. Theory **98**(1), 72–103 (1999)
10. Guo, K., Labate, D.: Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. Appl. Comput. Harmon. Anal. **30**(2), 231–242 (2010)
11. Guo, K., Labate, D., Lim, W.Q., Weiss, G., Wilson, E.: Wavelets with composite dilations and their MRA properties. Appl. Comput. Harmon. Anal. **20**(2), 202–236 (2006)
12. Jetter, K., Stöckler, J.: Algorithms for cardinal interpolation using box splines and radial basis functions. Numer. Math. **60**(1), 97–114 (1991)
13. Krishtal, I.A., Blanchard, J.D.: Matrical filters and crystallographic composite dilation wavelets. Math. Comp. **81**(278), 905–922 (2012)
14. Langemann, D., Prestin, J.: Multivariate periodic wavelet analysis. Appl. Comput. Harmon. Anal. **28**(1), 46–66 (2010)
15. Locher, F.: Interpolation on uniform meshes by the translates of one function and related attenuation factors. Math. Comp. **37**(156), 403–416 (1981)
16. Pöplau, G.: Multivariate periodische Interpolation durch Translate und deren Anwendung. Dissertation, Universität Rostock (1995)
17. Pöplau, G., Sprengel, F.: Some error estimates for periodic interpolation on full and sparse grids. In: le Méhauté, A., Rabut, C., Schumaker, L.L. (eds.) Curves and Surfaces with Applications in CAGD, pp. 355–362. Vanderbilt University Press (1997)
18. Skopina, M.: Multiresolution analysis of periodic functions. East J. Approx. **3**(2), 203–224 (1997)
19. Skopina, M.: Wavelet approximation of periodic functions. J. Approx. Theory **104**(2), 302–329 (2000)
20. Sloan, I.H., Joe, S.: Lattice methods for multiple integration. Oxford University Press, USA (1994)
21. Sprengel, F.: Interpolation und Waveletzerlegung multivariater periodischer Funktionen. Dissertation, Universität Rostock (1997)
22. Sprengel, F.: Periodic interpolation and wavelets on sparse grids. Numer. Algorithms **17**(1), 147–169 (1998)
23. Sprengel, F.: A class of periodic function spaces and interpolation on sparse grids. Numer. Funct. Anal. Optim. **21**(1), 273–293 (2000)

A Generalized Class of Hard Thresholding Algorithms for Sparse Signal Recovery

Jean-Luc Bouchot

Abstract We introduce a whole family of hard thresholding algorithms for the recovery of sparse signals $\mathbf{x} \in \mathbb{C}^N$ from a limited number of linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, with $m \ll N$. Our results generalize previous ones on hard thresholding pursuit algorithms. We show that uniform recovery of all s-sparse vectors $\mathbf{x}$ can be achieved under a certain restricted isometry condition. While these conditions might be unrealistic in some cases, it is shown that with high probability, our algorithms select a correct set of indices at each iteration, as long as the active support is smaller than the actual support of the vector to be recovered, with a proviso on the shape of the vector. Our theoretical findings are illustrated by numerical examples.

Keywords Compressive sensing · Sparse recovery · Hard thresholding · Sparse approximation

1 Compressive Sensing and Sparse Signal Recovery

This paper is concerned with the standard compressive sensing problem, i.e., we analyze the reconstruction of sparse signals $\mathbf{x} \in \mathbb{C}^N$ based only on a few number of (linear) measurements $\mathbf{y} \in \mathbb{C}^m$ where $m \ll N$. It is known from the compressive sensing literature that recovery of s-sparse signals $\mathbf{x}$ is ensured when the sensing (or measurement) matrix is random (Gaussian or sub-Gaussian for instance) and when the number of measurements scales linearly with the sparsity of the signal up to a log factor. Known methods arise mainly from optimization theory such as the ℓ_1 minimization [3, 14], reweighted norm minimizations [5, 13], primal-dual optimization [6], or from iterative solvers (see for instance [1, 10–12]).

J.-L. Bouchot (✉)
Department of Mathematics, Drexel University, 3141 Chestnut Street, Philadelphia, PA, USA
e-mail: jean-luc.bouchot@drexel.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_4,

We investigate in particular some variations of the Hard Thresholding Pursuit (HTP) algorithm [7], an iterative thresholding-based method, and its graded approach, a recent variation that does not require prior knowledge of the sparsity [2]. We analyze the reconstruction abilities of these algorithms in both idealized and realistic settings. In particular we introduce a generalization that improve the speed performance of (GHTP).

The idealized setting is characterized by the fact that the signal to be recovered $\mathbf{x}$ is exactly s-sparse and that the measurements occur in an error-free manner. In this case exact recovery is ensured by all such algorithms provided that a certain *restricted isometry condition (RIC)* is met by the sensing matrix [4, 8]. In comparison, we may consider a more realistic setting in which the vector $\mathbf{x}$ may suffer a sparsity defect and the measurements through the matrix $\mathbf{A}$ may be inaccurate. In this case, we have $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{C}^m$ represents the error induced by the measurement process. The sparsity defect can be integrated into this error term by considering $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_{\overline{S}}$ where $\mathbf{x}_S$ corresponds to the s most important components (i.e., the largest absolute entries) of $\mathbf{x}$. Thus, we may incorporate the remaining components into the noise as $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{A}\mathbf{x}_S + (\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}) = \mathbf{A}\mathbf{x}_S + \mathbf{e}'$ where $\mathbf{e}' = \mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e} \in \mathbb{C}^m$ contains both the sparsity defect and the measurement noise.

The remainder of this article is organized as follows. We start in Sect. 2 by reviewing some previous work regarding the (HTP) and (GHTP) algorithms (Sect. 2.1). This leads to introduce a family of algorithms that generalizes the two previous ones in Sect. 2.2. These algorithms are studied theoretically in the following sections in both uniform (see Sect. 3) and nonuniform settings (Sect. 4). Finally, Sect. 5 compares and validates numerically our findings. Throughout this paper we use the following notations:

- $\mathbf{x}^*$ represents the nonincreasing rearrangement of a vector $\mathbf{x}$:

$$x_1^* \geq x_2^* \geq \cdots \geq x_N^* \geq 0$$

  and there exists a permutation π of $\{1, \ldots, N\}$ such that $x_j^* = |x_{\pi(j)}|$.
- S is the support of an s-sparse vector $\mathbf{x}$ or the set of indices of its s largest absolute entries.
- $\mathbf{x}_T$ corresponds to the vector $\mathbf{x}$ either restricted to the set T or such that $x_{Ti} = x_i$ for $i \in T$ and 0 elsewhere, depending on the context.
- The complement of a set T in $\{1, \ldots, N\}$ is denoted by $\overline{T}$
- $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote respectively the ceil and floor functions.
- δ_s corresponds to the restricted isometry constant of order s of a given matrix $\mathbf{A}$ and is defined as the smallest δ such that

$$(1 - \delta) \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta) \|\mathbf{x}\|_2^2$$

  holds for any s-sparse vector $\mathbf{x}$.

2 (HTP), (GHTP), and their Generalizations

2.1 Previous Results

The Hard Thresholding Pursuit [7] and its graded variant [2] can be summarized by the following steps:

$$S^n := \left\{\text{indices of } k \text{ largest entries of } \left|\mathbf{x}^{n-1} + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n-1})\right|\right\}, \quad (\text{GHTP}_1)$$
$$\mathbf{x}^n := \operatorname{argmin}\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subset S^n\}, \quad (\text{GHTP}_2)$$

with $k = s$ for (HTP) and $k = n$ for (GHTP).

It was shown that robust and stable recovery is achieved under some RIC:

Theorem 1 *If the restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ obeys*

$$\delta_{3s} \leq \frac{1}{3} \text{ for (HTP)}, \quad \text{and} \quad \delta_{9s} \leq \frac{1}{3} \text{ for (GHTP)},$$

then the sequences $(\mathbf{x}^n)$ produced by (HTP) or (GHTP) with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \in \mathbb{C}^m$ for some s-sparse $\mathbf{x} \in \mathbb{C}^N$ and some $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_2 \leq \gamma\, x_s^*$ satisfy

$$\|\mathbf{x} - \mathbf{x}^{\overline{n}}\|_2 \leq d\|\mathbf{e}\|_2, \qquad \overline{n} \leq c\, s.$$

The constants $c \leq 3$ for (HTP) and $c \leq 4$ for (GHTP), $d \leq 2.45$, and $\gamma \geq 0.079$ depend only on δ_{3s} or δ_{9s}.

It is worth mentioning that reshuffling the index set in the (GHTP) algorithm adds robustness to (GHTP) (as seen in the numerical experiments in Sect. 5) over *Orthogonal Matching Pursuit (OMP)* at the cost that its implementation cannot be done using QR updates.

2.2 Generalizations

We investigate here some generalizations of the Graded Hard Thresholding Pursuit that improve the speed of the algorithm while only slightly deteriorating its reconstruction capability. In order to speed up the convergence we need to lower the number of iterations. Following an index selection process similar to the (HTP) and (GHTP) algorithms, we introduce (f-HTP) that relies on a different number of indices selected per iteration (note that Generalized HTP would be a confusing name with regard to the (GHTP) algorithm).

Let $f : \mathbb{N} \to \mathbb{N}$ be a nondecreasing function such that there exists $n_0 \geq 0$ with $f(n) \geq s$ for any $n \geq n_0$. The (f-HTP) algorithm is defined by the following sequence of operations:

$$S^n := \left\{\text{indices of } f(n) \text{largest entries of } \left|\mathbf{x}^{n-1} + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n-1})\right|\right\}, \quad (f\text{-HTP}_1)$$
$$\mathbf{x}^n := \operatorname{argmin}\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subset S^n\}. \quad (f\text{-HTP}_2)$$

Observe that the constant function $f(n) = s$ yields the original (HTP) algorithm while $f(n) = n$ corresponds to (GHTP). Particularly interesting in terms of speed and number of iterations is the case $f(n) = 2^{n-1}$ which shall be refer to as (GHTP^2).

2.3 First Results

We first provide some preliminary results as we did for the original graded algorithm (GHTP) in [2]. We show that a similar geometric decay of the error at each iteration $\|\mathbf{x} - \mathbf{x}^n\|_2$ holds for the generalization (f-HTP), see (3). It also ensures that a certain number of indices of largest entries may be included in the support after a given number of iterations (see Lemma 1). These results will allow us to prove the main result for uniform recovery (as stated in Theorem 2) by induction.

2.3.1 Geometric Decay

In the remainder of this article, we will define n_0 as the smallest integer such that $f(n) \geq s$, for all $n \geq n_0$. In particular, $n_0 = 0$ for (HTP), s for (GHTP) and $\lceil \log_2(s) \rceil$ for (GHTP^2). Using the results from [2, 7] we have the following estimates, for $n \geq n_0$:

$$\left\|\mathbf{x}^{n+1} - \mathbf{x}\right\|_2 \leq \sqrt{\frac{1}{1 - \delta^2_{f(n+1)+s}}} \left\|\left(\mathbf{x}^{n+1} - \mathbf{x}\right)_{\overline{S^{n+1}}}\right\|_2 + \frac{1}{1 - \delta_{f(n+1)+s}} \left\|\left(\mathbf{A}^*\mathbf{e}\right)_{S^{n+1}}\right\|_2, \quad (1)$$

$$\left\|\left(\mathbf{x}^{n+1} - \mathbf{x}\right)_{\overline{S^{n+1}}}\right\|_2 \leq \sqrt{2}\delta_{f(n)+f(n+1)+s} \left\|\mathbf{x}^n - \mathbf{x}\right\| + \sqrt{2} \left\|\left(\mathbf{A}^*\mathbf{e}\right)_{S \Delta S^{n+1}}\right\|_2. \quad (2)$$

Combining these two estimates yields the geometric decay

$$\left\|\mathbf{x}^{n+1} - \mathbf{x}\right\|_2 \leq \sqrt{\frac{2\delta^2_{f(n)+f(n+1)+s}}{1 - \delta^2_{f(n+1)+s}}} \left\|\mathbf{x}^n - \mathbf{x}\right\|_2 + \tau_{f(n+1)+s} \|\mathbf{e}\|_2 \quad (3)$$

with $\tau_{f(n+1)+s} = \sqrt{\frac{2}{\left(1-\delta_{f(n+1)+s}\right)^2}} + \frac{\sqrt{1+\delta_{f(n+1)+s}}}{1-\delta_{f(n+1)+s}}$. These results are the same as in our previous paper up to the RIC that needs to be adapted. In a more concise way

we can write, where the multiplicative coefficient can be written depending only on $\delta_{2f(n+1)+s}$:

$$\left\|\mathbf{x}^{n+1}-\mathbf{x}\right\|_2 \leq \rho_{2f(n+1)+s}\left\|\mathbf{x}^{n}-\mathbf{x}\right\|_2+\tau_{f(n+1)+s}\|\mathbf{e}\|_2$$

with $\rho_{2f(n+1)+s}=\sqrt{\frac{2\delta^2_{2f(n+1)+s}}{1-\delta^2_{2f(n+1)+s}}}$.

2.3.2 Preparatory Lemma

As for the original (GHTP) algorithm [2] we can show that, if the p largest absolute entries are contained in the support at iteration n, then k further iterations of (f-HTP) are sufficient to recover the q following largest entries, as stated in the following lemma.

Lemma 1 *Let $\mathbf{x}\in\mathbb{C}^N$ be s-sparse and let (S^n) be the sequence of index sets produced by (f-HTP) with $\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{e}$ for some $\mathbf{e}\in\mathbb{C}^m$. For integers $p\geq 0$ and $n\geq n_0$, suppose that S^n contains the indices of p largest absolute entries of $\mathbf{x}$. Then, for integers $k,q\geq 1$, S^{n+k} contains the indices of $p+q$ largest absolute entries of $\mathbf{x}$, provided*

$$x^*_{p+q}>\rho^k_{s+2f(n+k)}\|\mathbf{x}^*_{\{p+1,\ldots,s\}}\|_2+\kappa_{n+k-1}\|\mathbf{e}\|_2, \tag{4}$$

with the constants ρ_{n+k-1} as defined above and $\kappa_{n+k-1}=\frac{\sqrt{2}\delta_{s+f(n+k-1)}\sqrt{1-\delta_{s+f(n-1)}}}{1-\delta_{s+f(n-2)}}$ $+\frac{\sqrt{2}}{1-\delta_{s+f(n+1)}}\frac{\delta_{s+f(n+k-1)}}{1-\rho_{n+k-1}}+\sqrt{2}\sqrt{1+\delta_2}$ depending only on the restricted isometry constant $\delta_{s+f(n+k)}$.

Remark 1 The proof of Lemma 1 is not provided here. It follows directly from the proof of Lemma 3 and 4 from [2] with changes imposed on the current iteration number and the number of indices selected, which is replaced by $f(n)$ everywhere.

Remark 2 Lemma 1 is not ideal in the sense that the number of iterations needed for the recovery of the next q largest entries, does not depend on the actual index selection method and whether we select exponentially many new indices at each iteration or just a linear number of new candidates. This leads to overestimate the number of iterations needed. As we see in the following section, it creates RIC that are not always realistic, as they yield RIP of order up to s^s (in the worst, but fastest, scenario). It shows however, that there exist conditions under which the convergence of the algorithm is guaranteed. We also see that in particular cases (namely when dealing with power or flat vectors) the conditions from Theorem 2 can be drastically improved. Moreover, as suggested by the numerical experiments in Sect. 5, these results are only a rough over estimation of the actual number of iterations needed.

Table 1 Examples of RIC that a sensing matrix should fulfill for uniform recovery, according to Theorem 2

f	s	n	$2n$	$n(n+1)/2$	2^{n-1}
RIC	$3s$[a]	$9s$	$s/2+16s$	$\sqrt{2s}+8s^2$	$s+s^s$
Name	(HTP)	(GHTP)			(GHTP^2)

[a] This result actually coincides with the one from the original paper about (HTP) [7]

3 Uniform Recovery via (f-HTP)

3.1 General Results

This section is dedicated to the problem of uniform recovery of all s-sparse vectors $\mathbf{x}$ given a certain sensing matrix. While this gives some ideas of why the (f-HTP) algorithms may converge, our proof yields, for certain choices of f, unrealistic and unapplicable conditions. Such considerations are detailed in Table 1.

Theorem 2 *If the restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ obeys*

$$\delta_{2f(\overline{n})+n_0} \leq \frac{1}{3},$$

then the sequence $(\mathbf{x}^n)$ produced by (f-HTP) with $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{e} \in \mathbb{C}^m$ for some s-sparse $\mathbf{x} \in \mathbb{C}^N$ and some $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_2 \leq \gamma\, x_s^*$ satisfies

$$\|\mathbf{x}-\mathbf{x}^{\overline{n}}\|_2 \leq d\|\mathbf{e}\|_2, \qquad \overline{n} \leq c\,s.$$

The constants $c \leq 4$, $d \leq 2.45$, and $\gamma \geq 0.079$ depend only on $\delta_{s+2f(\overline{n})}$.

This theorem generalizes the one obtained first for (HTP) and (GHTP) to more general index selection schemes. It is purely an adaptation of our previous results and does not depend on the index selection function f. Therefore, as stated above it generates unrealistic restricted isometry conditions. For instance, when considering the case of $f(n) = 2^{n-1}$ we would need to ensure an RIC of order $\Omega(s^s)$. Table 1 gives some examples of RIC for different choices of f.

While the two last conditions are unrealistic, the first cases still yield reasonable RIC. For instance the case $f = 2n$ yields an RIC at the order $16s$ which is still in a comparable range as for (OMP) (see [15] for a stable recovery under RIC of order $13s$).

Fortunately these strong conditions can be improved in the particular cases of power vector or almost flat vectors. The recovery of power vectors is analyzed in the following section where we show in the case of (GHTP^2) that the matrix only needs to obey an RIC of order $\Omega(s^{\text{polylog}(s)})$. (Other examples are given in Table 2.)

Table 2 Examples of RIC-orders that the measurement matrix needs to obey for different (f-HTP) algorithms (these are just order of magnitude)

f	s	n	$2n$	$n(n+1)/2$	2^{n-1}
RIC	$3s$	$3s + 4C\text{polylog(s)}$	$3s + 8C\text{polylog(s)}$	$3s + 4C\text{polylog}^2(s)$	$s + 2s^{\text{polylog}(s)}$
Name	(HTP)	(GHTP)			(GHTP^2)

3.2 The Case of Power Vectors

We investigate the convergence of the family of generalized algorithms when facing particular power vectors. Our results rely on the following lemma used for decomposing the support:

Lemma 2 *Any set $S \subseteq \{1, \ldots, N\}$ of size $s \leq N$ can be decomposed in r subsets $S_1, \ldots, S_r$ such that*

1. $r = \lfloor \log_2(s) \rfloor + 1$
2. $S = \bigcup_{i=1}^{r} S_i$
3. $S_i \cap S_j = \varnothing$, *for* $j \neq i$
4. $|S_i| \leq \lceil s/2^i \rceil$.

Proof We show this result by induction on the set size s. For $s = 1$, $S_1 = S$ fulfills all the criteria. Assume now that Lemma 2 holds for all $1 \leq n \leq s - 1$. Without loss of generality, we can consider the set $S = \{1, \ldots, s\}$. Writing $S = S_1 \cup T$ with $S_1 = \{1, \ldots, \lceil s/2 \rceil\}$ and $T = S \backslash S_1$, we have $|T| = s - \lceil s/2 \rceil < s$ and therefore, applying the induction hypothesis yields $T = \bigcup_{j=1}^{r_T} T_j$ with $r_T = \lfloor \log_2(|T|) \rfloor + 1$ and $|T_j| \leq \lceil |T|/2^j \rceil$. We now define $S_i := T_{i-1}$ for $i > 1$ and therefore the partition $S_1, \ldots, S_r$ fulfills the three first criteria of the lemma. To verify the last statement of the lemma we consider two separated cases:

If s **is even**, then there exists a $k \in \mathbb{N}$ such that $s = 2k$ and $|S_1| = |T| = k$. The induction hypothesis implies, for $j \geq 1$

$$|T_j| \leq \left\lceil k/2^j \right\rceil \quad \text{and} \quad |S_{j+1}| \leq \left\lceil s/2^{j+1} \right\rceil$$

which proves the last point of the lemma.

If s **is odd**, then there exists a $k \in \mathbb{N}$ such that $s = 2k + 1$ and $|S_1| = k + 1$ and $|T| = k$. The induction hypothesis implies, for $j \geq 1$

$$|T_j| \leq \left\lceil k/2^j \right\rceil \quad \text{and} \quad |S_{j+1}| \leq \left\lceil s/2^{j+1} - 1/2^{j+1} \right\rceil \leq \left\lceil s/2^{j+1} \right\rceil$$

which finishes the proof of the lemma. □

Consider vectors $\mathbf{x}$ such that for all $1 \leq j \leq s, x_j^* = 1/j^\alpha$ **for some** $\alpha > 1/2$ (other cases will be considered later). We have

$$\|\mathbf{x}^*_{\{p+1,\dots,s\}}\|_2^2 = \frac{1}{(p+1)^{2\alpha}} + \dots + \frac{1}{s^{2\alpha}} \leq \int_p^s \frac{1}{x^{2\alpha}}\mathrm{d}x = \frac{1}{2\alpha-1}\left(\frac{1}{p^{2\alpha-1}} - \frac{1}{s^{2\alpha-1}}\right) \leq \frac{1}{2\alpha-1}\frac{1}{p^{2\alpha-1}}.$$

With this, it is sufficient to find k and q such that

$$\frac{1}{(p+q)^{2\alpha}} > \rho^{2k}\frac{1}{2\alpha-1}\frac{1}{p^{2\alpha-1}},$$

for condition (4) from Lemma 1 to be valid.

This condition is equivalent to

$$k > \frac{1}{\log(1/\rho^2)}\log\left(\frac{p}{2\alpha-1}\left(1+\frac{q}{p}\right)^{2\alpha}\right).$$

In conclusion, $\{1,\dots,p\} \subset S^n \Rightarrow \{1,\dots,p+q\} \subset S^{n+k}$ holds provided that

$$k > \frac{2\alpha\log\left(p\left(1+\frac{q}{p}\right)\right)}{\log(1/\rho^2)} - \frac{\log(2\alpha-1)}{\log(1/\rho^2)}.$$

If we now consider r subsets $S_1,\dots,S_r$, $r = \lfloor \log_2(s) \rfloor + 1$ as suggested by Lemma 2, then we can successively apply Lemma 1 to each r subsets S_i. Defining $S_0 = \varnothing$, $q_i = |S_i|$, for $i \geq 0$, k_i the number of iterations needed to add subset S_i, using $k_0 = n_0$, and $p_i = \sum_{j=1}^{i-1} q_j$, we finally get that the number of iterations for uniform recovery is bounded by

$$\begin{aligned}
\overline{n} &\leq \sum_{i=0}^{r} k_i \leq \frac{2\alpha}{\log(1/\rho^2)}\sum_{i=1}^{r}\log\left(p_i\left(1+\frac{q_i}{p_i}\right)\right) - r\frac{\log(2\alpha-1)}{\log(1/\rho^2)} + n_0 \\
&\leq \frac{2\alpha}{\log(1/\rho^2)}\sum_{i=1}^{r}\log\left(\sum_{j=1}^{i} q_j\right) - r\frac{\log(2\alpha-1)}{\log(1/\rho^2)} + n_0 \\
&\leq \frac{2\alpha}{\log(1/\rho^2)}\sum_{i=1}^{r}\log\left(\sum_{j=1}^{i}\left(s/2^j+1\right)\right) - r\frac{\log(2\alpha-1)}{\log(1/\rho^2)} + n_0 \qquad (5) \\
&\leq \frac{2\alpha}{\log(1/\rho^2)}\sum_{i=1}^{r}\log\left(s\sum_{j=1}^{i}1/2^j + i\right) - r\frac{\log(2\alpha-1)}{\log(1/\rho^2)} + n_0 \\
&\leq \frac{2\alpha}{\log(1/\rho^2)}\sum_{i=1}^{r}\log(2s) - r\frac{\log(2\alpha-1)}{\log(1/\rho^2)} + n_0,
\end{aligned}$$

where we have used the fact that $q_j \leq \lceil s/2^j \rceil \leq s/2^j + 1$ in inequality (5). With such a partition we have that $r = \lfloor \log_2(s) \rfloor + 1 \leq \log_2(s) + 1$ and hence $\overline{n}$ can be bounded by

$$\overline{n} \leq \left(\log_2(s) + 1\right) \left(\log(2s) \frac{2\alpha}{\log(1/\rho^2)} - \frac{\log(2\alpha - 1)}{\log(1/\rho^2)} \right) + n_0.$$

Using this, we only need to ensure the RIC to the order $2^{\overline{n}}$ which is $\Omega(s \cdot s^{\log(s)})$ when using the (GHTP^2). This is not yet acceptable for real-world applications but much less critical then what Theorem 2 suggests. Moreover, it corresponds also to the worst case scenario for (f-HTP) algorithms.

If we now consider **the case** $\alpha = 1/2$, a similar analysis yields

$$\|\mathbf{x}^*_{\{p+1,\ldots,s\}}\|_2^2 \leq \int_p^s \frac{1}{x} \mathrm{d}x = \log(s - p),$$

and condition (4) reads $\frac{1}{p+q} > \rho^{2k} \log(s - p)$. Therefore, Lemma 1 holds for

$$k > \frac{\log(\log(s)) + \log(p + q)}{\log(1/\rho^2)}.$$

Using a partition as given in Lemma 2 gives a sufficient number of iterations

$$\begin{aligned} \overline{n} = \sum_{i=0}^{r} k_i &= \sum_{i=1}^{r} \frac{\log(\log(s)) + \log(\sum_{j=1}^{i} q_j)}{\log(1/\rho^2)} + n_0 \\ &\leq \left(\log_2(s) + 1\right) \frac{\log(\log(s)) + \log(2s)}{\log(1/\rho^2)} + n_0. \end{aligned}$$

Again, in this case, the RIC has to be valid at the order $\Omega(s \cdot s^{\mathrm{polylog(s)}})$ for (GHTP^2).

The case $0 < \alpha < 1/2$ can be treated in the exact same way by approximating the 2-norm with an integral. This yields, using the same support decomposition,

$$\overline{n} \leq \left(\log_2(s) + 1\right) \left(\frac{(1 - 2\alpha)\log(s)}{\log(1/\rho^2)} + \frac{2\alpha \log(2s)}{\log(1/\rho^2)} - \frac{\log(1 - 2\alpha)}{\log(1/\rho^2)} \right).$$

Consider **an almost flat s-sparse vector $\mathbf{x}$** such that there exists an $\varepsilon \geq 0$ with $1 - \varepsilon \leq x_j^* \leq 1$, for $j = 1, \ldots, s$ (this corresponds to $\alpha = 0$). In this case, we have that

$$(1 - \varepsilon)^2 (s - p) \leq \|\mathbf{x}_{\{p+1,\ldots,s\}^*}\|_2^2 \leq s - p$$

Hence condition (4) now reads $1 > \rho^{2k} \frac{s-p}{1-\varepsilon}$ and is fulfilled whenever

$$k > \frac{\log\left(\frac{s-p}{1-\varepsilon}\right)}{\log(1/\rho^2)}.$$

Using the decomposition given in Lemma 2, we get that

$$\overline{n} \leq \frac{\log\left(s/\left(1-\varepsilon\right)\right)^2}{\log(1/\rho^2)} + n_0$$

iterations are sufficient to recover the signal $\mathbf{x}$. This gives a RIC in the order of $\Omega(s^{\log(s)})$ when considering (GHTP^2) for the power vector case. All of the previous results can be summarized in the following corollary:

Corollary 1 *Let* $\mathbf{x}$ *be an* s*-sparse vector such that its nondecreasing rearrangement can be written as* $x_j^* = 1/j^\alpha$*, for all* $1 \leq j \leq s$*, for some* $\alpha \geq 0$ *or* $1-\varepsilon \leq x_j^* \leq 1$*, for some* $\varepsilon \geq 0$*. Then for any matrix* $\mathbf{A} \in \mathbb{C}^{m\times N}$*,* $\mathbf{x}$ *can be recovered from* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *in at most* $\overline{n} = C\frac{\text{polylog}(s)}{\log(1/\rho^2)} + n_0$ *iterations of (*f*-HTP) provided that the RIP conditions are satisfied at the order* $\Omega(s + 2f(\text{polylog(s)}))$*. The constant* C *and the polynomial involved depend only on* α *and* ρ

As a consequence, (GHTP^2) requires a RIC in the order of $\Omega(s + 2s^{\log(s)})$. Similarly, considering $f(n) = 2n$ yields a RIC in the order of $\Omega(3s + \text{polylog}(s))$ which is tractable and still provides a strong speed improvement over the original (GHTP) (even if the complexity remains in the same order, the constant in front is much lower). Some examples are summarized in Table 2.

4 Nonuniform Recovery via (f-HTP)

We consider here the problem of recovering a particular fixed vector $\mathbf{x}$ instead of recovering any vector for a given matrix $\mathbf{A}$.

4.1 Useful Inequalities

We recall here some results regarding the tail distribution of some random variables and the probability distribution of the smallest singular value of a subgaussian matrix [9]. These results play an important role in proving the nonuniform recovery of vectors via (f-HTP).

Lemma 3 ([9]) *Let* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *be a subgaussian matrix, the following inequalities hold*

$$\mathbb{P}(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{I}\|_{2\to 2} > \delta) \leq 2\exp(-c'\delta^2 m) \tag{6}$$

$$\mathbb{P}\left(|\langle a_\ell, \mathbf{v}\rangle| > t\|\mathbf{v}\|_2\right) \leq 4\exp\left(-c''t^2 m\right), \tag{7}$$

where c'' *depends only on the distribution.*

4.2 Recovery

Following [2, Prop. 9], we can see that with high probability the algorithms make no mistakes when selecting the indices. This statement is true while the size of the index set selected at a given iteration is strictly smaller than the actual sparsity of the signal under a condition on the shape of the vector to be recovered. This result is summarized in the following proposition:

Proposition 1 *Let $\lambda \geq 1$ and let $\mathbf{x} \in \mathbb{C}^N$ be an s-sparse vector such that $x_1^* \leq \lambda x_s^*$. If $A \in \mathbb{R}^{m \times N}$ is a sub-Gaussian matrix with*

$$m \geq Cs \ln(N),$$

then with high probability ($\geq 1 - 2N^{-c}$) and for any error vector $\mathbf{e} \in \mathbb{C}^m$ such that $\|\mathbf{e}\|_2 \leq \gamma x_s^*$ the sequences S^n and $\mathbf{x}^n$ produced by (f-HTP) with $\mathbf{y} = A\mathbf{x} + \mathbf{e}$ satisfy, at iteration $n_0 - 1$ (where n_0 denotes the smallest integer such that $f(n) \geq s$):

$$S^{n_0 - 1} \subset S \tag{8}$$

where the constant γ depends only on λ and the constant C on λ and c.

Remark 3 It is worth mentioning that the proof of this Proposition does not apply to the (HTP) algorithm. Indeed, the result holds only while the number of indices is strictly smaller than the actual sparsity. This condition is never met with $f(n) = s$.

Proof The proof follows from our previous results. We show that, with high probability, $S^n \subseteq S$ for all $1 \leq n \leq n_0 - 1$. For this we need to show that $\chi_n > \zeta_n$ where we define

$$\chi_n := \left[\left(\mathbf{x}^{n-1} + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n-1}) \right)_S \right]^*_{f(n)},$$
$$\zeta_n := \left[\left(\mathbf{x}^{n-1} + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n-1}) \right)_{\overline{S}} \right]^*_1.$$

Literally, with $\mathbf{z}^n := \mathbf{x}^n - \mathbf{A}^*(y - \mathbf{A}\mathbf{x}^n)$ χ_n is the $f(n)^{th}$ largest absolute entry of $\mathbf{z}^n$ on the support S of $\mathbf{x}$ while ζ_n is the largest absolute entry of $\mathbf{z}^n$ on its complement. Now $\chi_n > \zeta_n$ for all $1 \leq n \leq n_0 - 1$ is true with failure probability P, and we have

$$P := \mathbb{P}(\exists n \in \{1, \ldots, n_0 - 1\} : \zeta_n \geq \chi_n \text{ and } (\chi_{n-1} > \zeta_{n-1}, \ldots, \chi_1 > \zeta_1)) \tag{9}$$

$$P \leq \mathbb{P}\left(\|\mathbf{A}^*_{S \cup \{\ell\}} \mathbf{A}_{S \cup \{\ell\}} - \mathbf{I}\|_{2 \to 2} > \delta \text{ for some } \ell \in \overline{S} \right) \tag{10}$$

$$+ \sum_{n=1}^{n_0 - 1} \mathbb{P}\big(\zeta_n \geq \chi_n, (\chi_{n-1} > \zeta_{n-1}, \ldots, \chi_1 > \zeta_1), \ (\|\mathbf{A}^*_{S \cup \{\ell\}} \mathbf{A}_{S \cup \{\ell\}} - \mathbf{I}\|_{2 \to 2} \leq \delta \text{ for all } \ell \in \overline{S}) \big), \tag{11}$$

Defining $T^{s-f(n-1)}$ as the set of indices corresponding to the $s-f(n-1)$ smallest absolute entries of $\mathbf{z}^n$ on S we can easily verify that

$$\chi_n \geq \frac{1}{\sqrt{s-f(n-1)}}\left(\|\mathbf{x}_{T^{s-f(n-1)}}\|_2 - \delta\|\mathbf{x}-\mathbf{x}^{n-1}\|_2 - \sqrt{1+\delta}\|\mathbf{e}\|_2\right).$$

Similarly, we have

$$\zeta_n \leq \max_{\ell\in\overline{S}} |\langle a_\ell, \mathbf{A}(\mathbf{x}-\mathbf{x}^{n-1})\rangle| + \sqrt{1+\delta}\|\mathbf{e}\|_2.$$

Finally, with $\mathbb{P}'(E)$ denoting the probability of an event E intersected with the event $\left((\chi_{n-1}>\zeta_{n-1},\dots,\chi_1>\zeta_1), (\|\mathbf{A}^*_{S\cup\{\ell\}}\mathbf{A}_{S\cup\{\ell\}}-\mathbf{I}\|_{2\to 2}\leq\delta \text{ for all } \ell\in\overline{S})\right)$ inequality (11) reads

$$\begin{aligned}\mathbb{P}'\Big(\zeta_n \geq \chi_n\Big) &\leq \mathbb{P}'\Big(\max_{\ell\in\overline{S}} |\langle a_\ell, \mathbf{A}(\mathbf{x}-\mathbf{x}^{n-1})\rangle| > \frac{1}{\sqrt{s-f(n-1)}}\\ &\qquad\times\left(\|\mathbf{x}_{T^{s-f(n-1)}}\|_2 - \delta\|\mathbf{x}-\mathbf{x}^{n-1}\|_2\right) - 2\sqrt{1+\delta}\|\mathbf{e}\|_2\Big)\\ &\leq \mathbb{P}'\Big(\max_{\ell\in\overline{S}} \left|\langle a_\ell, \mathbf{A}\left(\mathbf{x}-\mathbf{x}^{n-1}\right)\rangle\right| > \frac{\delta}{\sqrt{s-f(n-1)}}\|\mathbf{x}-\mathbf{x}^{n-1}\|_2\Big)\end{aligned} \tag{12}$$

where the last inequality follows from the fact that

$$\begin{aligned}&\frac{1}{\sqrt{s-f(n-1)}}(\|\mathbf{x}_{T^{s-f(n-1)}}\|_2 - \delta\|\mathbf{x}-\mathbf{x}^{n-1}\|_2) - 2\sqrt{1+\delta}\|\mathbf{e}\|_2\\ &\qquad\geq \frac{\delta}{\sqrt{s-f(n-1)}}\|\mathbf{x}-\mathbf{x}^{n-1}\|_2\end{aligned} \tag{13}$$

whenever

$$1-2\sqrt{1+\delta}\gamma \geq 2\delta\left(\frac{\lambda}{\sqrt{1-\delta^2}} + \frac{\sqrt{1+\delta}}{1-\delta}\gamma\right). \tag{14}$$

Indeed, inequality (13) is equivalent to

$$\frac{1}{\sqrt{s-f(n-1)}}\|\mathbf{x}_{T^{s-f(n-1)}}\|_2 - 2\sqrt{1+\delta}\|\mathbf{e}\|_2 \geq \frac{2\delta}{\sqrt{s-f(n-1)}}\|\mathbf{x}-\mathbf{x}^{n-1}\|_2. \tag{15}$$

The left-hand side can be estimated by

$$x_s^* - 2\sqrt{1+\delta}\gamma x_s^* = x_s^*\left(1 - 2\sqrt{1+\delta}\right),$$

and the right-hand side is estimated by

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}^{n-1}\|_2 &\leq \frac{1}{\sqrt{1-\delta^2}}\|\mathbf{x}_{\overline{S^{n-1}}}\|_2 + \frac{1}{1-\delta}\|\left(\mathbf{A}^*\mathbf{e}\right)_{S^{n-1}}\|_2, \text{ using estimate (1)},\\
&\leq \frac{\sqrt{s - f(n-1)}}{\sqrt{1-\delta^2}}x_1^* + \frac{\sqrt{1+\delta}}{1-\delta}\|\mathbf{e}\|_2,\\
&\leq \frac{\sqrt{s - f(n-1)}}{\sqrt{1-\delta^2}}\lambda x_s^* + \frac{\sqrt{1+\delta}}{1-\delta}\gamma x_s^*,\\
&\leq \sqrt{s - f(n-1)}\left(\frac{\lambda}{\sqrt{1-\delta^2}} + \frac{\sqrt{1+\delta}}{1-\delta}\gamma\right)x_s^*.
\end{aligned}$$

Hence, condition (14) is verified by choosing δ then γ (depending on λ) small enough.

Finally, using the fact that $\|\mathbf{A}\left(\mathbf{x} - \mathbf{x}^{n-1}\right)\|_2 \leq \sqrt{1+\delta}\|\mathbf{x} - \mathbf{x}^{n-1}\|_2$, the failure probability from (12), can be further approximated by

$$\mathbb{P}'\left(\zeta_n > \chi_n\right) \leq \mathbb{P}'\left(\max_{\ell \in \overline{S}} \left|\langle a_\ell, \mathbf{A}\left(\mathbf{x} - \mathbf{x}^{n-1}\right)\rangle\right| > \frac{\delta}{\sqrt{1+\delta}}\|\mathbf{A}\left(\mathbf{x} - \mathbf{x}^{n-1}\right)\|_2\right) \tag{16}$$

Combining these results with (7) and (6), we finally get that

$$P \leq 2\left(N - s\right)\exp\left(-c'\delta^2 m\right) + 4\left(N - s\right)\left(n_0 - 1\right)\exp\left(-\frac{c''\delta^2 m}{(1+\delta)s}\right).$$

This leads to

$$P \leq 2N^2\exp\left(-\frac{c_\delta''' m}{s}\right)$$

with an appropriate choice of c_δ'''. □

4.3 Hybrid Algorithms

We may ask ourselves whether Proposition 1 is of interest or not as it does not lead to the complete recovery of $\mathbf{x}$. However, Proposition 1 ensures us that we can create hybrid algorithm with the (f-HTP) framework where we can make large steps first until a certain criterion is met, and then adaptively reduce the increase of the index set's size until it gets to the sparsity of the signal.

The following gives an example of such a hybrid algorithm.

Algorithm 1. Example of an hybrid algorithm for sparse signal recovery.

Data: A matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$, a measurement vector $\mathbf{y} \in \mathbb{C}^m$, a switching step $n \in \mathbb{N}, n \leq n_0$
Result: an s sparse signal $\mathbf{x}$
Set $S^0 = \varnothing$, $\mathbf{x}^0 = \mathbf{0}$, $nIter = 0$;
while $nIter \leq n$ **do**
  Do an iteration of (GHTP^2);
  $nIter = nIter + 1$;
end
while *Convergence is not done* **do**
  Do an iteration of (GHTP);
  $nIter = nIter + 1$;
end

The only important thing to be careful of is that we stay below the sparsity when we start reducing the number of indices added at each iteration. Moreover, even if Proposition 1 does not ensure convergence of the algorithm until the very last important index, it was shown in [2] that (GHTP) does converge in s iterations. This ensures us that such an hybrid algorithm can be used for nonuniform recovery and that it converges in a number of iterations $\overline{n} \leq s$.

5 Numerical Results

This section validates our theoretical findings with some numerical experiments. Note that all the necessary Matlab files can be found on the author's webpage.[1] Validation is being made with some (f-HTP) examples compared to the (HTP), (GHTP), and (OMP) algorithms. The following particular index selection functions f are used:

- $f(n) = s$: (HTP),
- $f(n) = n$: (GHTP),
- $f(n) = 2n$: (GHTP2n),
- $f(n) = n(n+1)/2$: (GHTPn^2),
- $f(n) = 2^{n-1}$: (GHTP^2).

Moreover we will denote by (Hyb1) and (Hyb2) the two algorithms such that the functions f are defined, respectively by

$$f(n) = \begin{cases} 2^{n-1}, & \text{if } n < n_0, \\ 2^{n_0-2} + n - n_0 + 1, & \text{otherwise,} \end{cases} \qquad \text{(Hyb1)}$$

$$f(n) = \begin{cases} n(n-1)/2, & \text{if } n < n_0, \\ (n_0-1)n_0/2 + n - n_0 + 1, & \text{otherwise.} \end{cases} \qquad \text{(Hyb2)}$$

[1] http://www.math.drexel.edu/~jb3455/publi.html

The algorithms were tested using 100 randomly generated Gaussian matrices $\mathbf{A} \in \mathbb{R}^{200\times 1000}$ each of which were used to recover 10 vectors with randomly generated supports (which represents a total of 1,000 random tests for each vector kind and sparsity level). The tests were carried out on three different kinds of vectors to assess the dependence of the algorithms on the decay of the vector $\mathbf{x}$; "flat" vectors with $x_j^* = 1$ for $j \in \{1, \ldots, s\}$, "linear" vectors with $x_j^* = (s+1-j)/s$ for $j \in \{1, \ldots, s\}$, and Gaussian vectors whose s nonzero entries are independent standard normal random variables.

5.1 Successful Recovery and Area of Convergence

We first want to assess the recovery ability of our algorithms by recording the frequency of success as a function of the sparsity. As stopping criterion here we have used the natural one for (HTP) ($S^n = S^{n-1}$) and [$S \subseteq S^n$ or $\|\mathbf{x}-\mathbf{x}^n\|_2/\|\mathbf{x}\|_2 < 10^{-4}$] for ($f$-HTP) and (OMP).[2] A recovered $\mathbf{x}$ is recorded as a success whenever the relative error is smaller than 10^{-4}.

As expected, the steeper the index selection function the harder it is for the algorithm to converge. As a consequence (see Fig. 1) (GHTP^2) performs the worst. However, for reasonable functions f (up to quadratic functions) the range of convergence of the algorithm is similar to the original one. Moreover, due to the reshuffling of the index set, our family of functions tend to perform better than a classical (OMP).

5.2 Number of Iterations for Successful Recovery

One important reason for introducing this generalized family of functions is to lower the number of iterations needed for convergence. Indeed, while the reshuffling of the active set can be seen as an advantage in terms of recovery capability of our algorithms, it takes away any chance of faster implementation, using for instance QR updates in the inner loop. The following set of graphs (depicted in Fig. 2) analyzes the maximum number of iterations needed for recovery.

Three things are worth mentioning. First, as already stated in Remark 2, the maximum number of iterations suggested by Theorem 2 is a very rough overestimation of the actual number of iterations. This is mainly due to the fact that the proof of Theorem 2 relies on the geometric decay of $\|\mathbf{x}^n - \mathbf{x}\|_2$ that can only be proven for $n \geq n_0$. However, as we describe in the next Section, the algorithm picks correct indices much earlier than the n_0th iteration. This also shows that the proof Theorem 2 is not optimal as clearly, for most of these algorithms, the RIP suggested is not respected.

[2] Compared to real applications, we have access here to the true sparsity and the true support of the signal $\mathbf{x}$. This stopping criterion needs to be adapted for real-world examples.

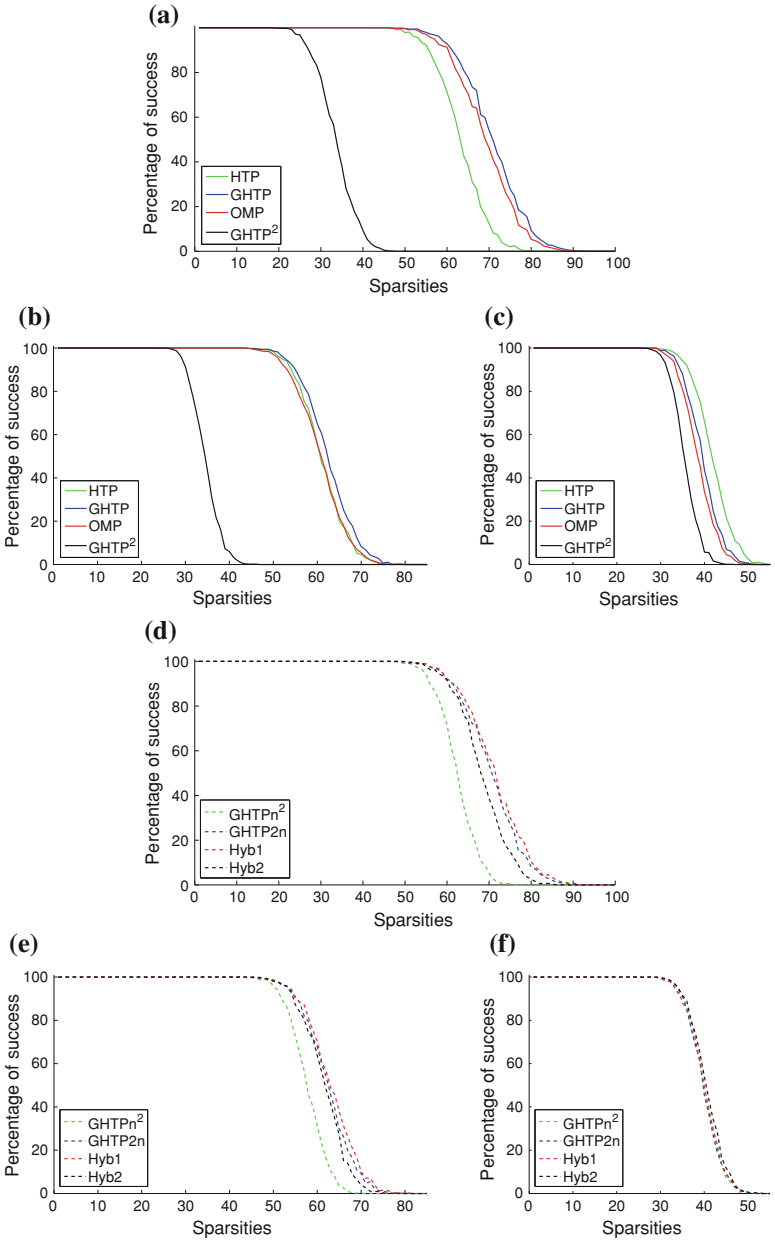


Fig. 1 Frequency of success for the original algorithms [(HTP), (GHTP^2), (GHTP), and (OMP), *two firsts row*] and the new generalized approach [(GHTP2n), (GHTPn^2), (Hyb1) and (Hyb2), *bottom rows*] when the original vector is Gaussian (*first* and *third rows*), linear (*left column*) or flat (*right column*). **a** Gaussian vectors—original algorithms, **b** Linear vectors—original algorithms, **c** Flat vectors—original algorithms, **d** Gaussian vectors—generalized algorithms, **e** Linear vectors—generalized algorithms, **f** Flat vectors—generalized algorithms

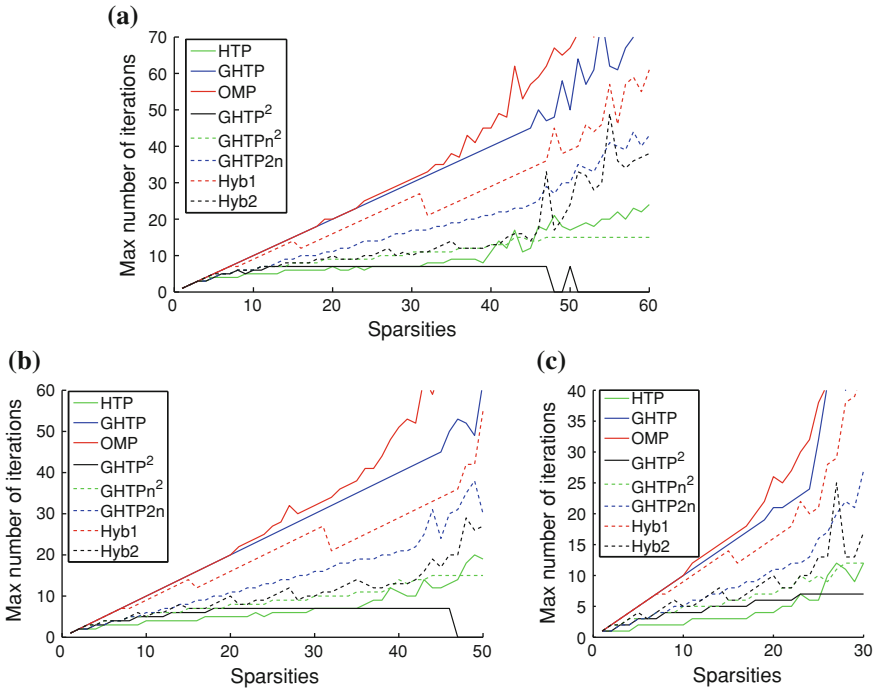


Fig. 2 Maximum number of iterations for exact recovery for the different algorithms when considering Gaussian (*top plot*), linear (*bottom left*), or flat (*bottom right*) vectors. **a** Gaussian vectors, **b** Linear vectors, **c** Flat vectors

Second, when the algorithms converge, their number of iterations scale according to the underlying function f. The number of iterations behaves like a logarithm for (GHTP^2), like a square root for (GHTPn^2) and linearly for both (GHTP2n) and (GHTP). Again, (OMP) needs a few more iterations, mainly to compensate the wrong indices that have been picked at an earlier stage of the algorithm.

Finally, it is reasonable to think that the analysis carried out in Corollary 1 can be extended to more general vector shapes. However, to improve the estimation of the number of iterations we would need to adapt the proof to earlier iterations, instead of starting counting at n_0.

5.3 Indices Correctly Captured

We investigate now the ability of our family of algorithms to pick correct indices at each iteration. Figure 3 shows these quantities for the three kinds of vectors (Gaussian to the left, linear in the middle and flat on the right) when dealing with different sparsities and index selection functions (see legend for more details).

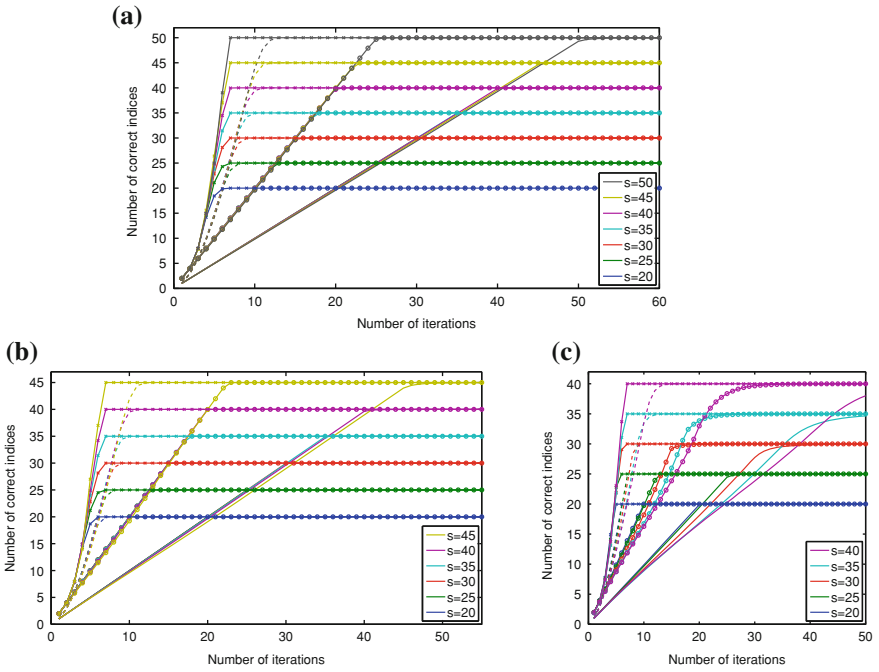


Fig. 3 Minimum number of correct indices picked at each iteration for different sparsity levels. *Continuous lines* correspond to (OMP), *circles* to (GHTP2n), *dashed lines* to (GHTPn^2), and *crosses* to (GHTP^2). **a** Gaussian vectors, **b** Linear vectors, **c** Flat vectors

As expected most of the algorithms made no mistakes when picking a current active set. This suggests that Proposition 1 can be improved to more general vector shapes.

6 Conclusion

This article introduced a class of algorithms that generalizes the Hard Thresholding Pursuit. It allows to overcome both the lack of a priori knowledge regarding the sparsity of the signal to recover and the convergence issue noticed in an earlier extension. We have shown that uniform and nonuniform convergence is possible for all algorithms of this type, but sometimes under unrealistic restricted isometrty conditions.

Fortunately, our numerical results tend to show that the number of iterations implied by our results may be a really rough overestimates. This will drive our future research which would also imply some improved restricted isometry conditions. Moreover, by using a combination of index selecting functions, we are able to produce hybrid algorithms that are both reliable and fast, at least in a nonuniform setting. For

such algorithms, a selection of an adequate turning point is needed which is also left for further study.

Acknowledgments The author wants to thank Simon Foucart and Michael Minner for their fruitful comments and suggested literature. The author is also thankful to the NSF for funding his work under the grant number (DMS-1120622).

References

1. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**, 265–274 (2009)
2. Bouchot, J.L., Foucart, S., Hitczenko, P.: Hard thresholding pursuit algorithms: number of iterations. (2013, submitted)
3. Candès, E.J., Romberg, J.: ℓ_1-magic: recovery of sparse signals via convex programming. http://www.acm.caltech.edu/l1magic/downloads/l1magic.pdf4 (2005)
4. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 4203–4215 (2005)
5. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. Appl. **14**, 877–905 (2008)
6. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imag. Vis. **40**, 120–145 (2011)
7. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. SIAM J. Numer. Anal. **49**, 2543–2563 (2011)
8. Foucart, S.: Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In: Neamtu, M., Schumaker, L. (eds.) Approximation Theory XIII: San Antonio 2010. Springer Proceedings in Mathematics, vol. 13, pp. 65–77. Springer, New York (2012)
9. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhaüser, Basel (2013)
10. Needell, D., Tropp, J.A.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**, 301–321 (2009)
11. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**, 2231–2242 (2004)
12. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory **53**, 4655–4666 (2007)
13. Wipf, D., Nagarajan, S.: Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. IEEE J. Sel. Top. Sig. Process. **4**, 317–329 (2010)
14. Yang, J., Zhang, Y.: Alternating direction algorithms for *ell*_1-problems in compressive sensing. SIAM J. Sci. Comput. **33**, 250–278 (2011)
15. Zhang, T.: Sparse recovery with orthogonal matching pursuit under RIP. IEEE Trans. Inf. Theory **57**, 6215–6221 (2011)

On a New Proximity Condition for Manifold-Valued Subdivision Schemes

Tom Duchamp, Gang Xie and Thomas Yu

Abstract An open theoretical problem in the study of subdivision algorithms for approximation of manifold-valued data has been to give necessary and sufficient conditions for a manifold-valued subdivision scheme, based on a linear subdivision scheme, to share the same regularity as the linear scheme. This is called the *smoothness equivalence problem*. In a companion paper, the authors introduced a *differential proximity condition* that solves the smoothness equivalence problem. In this paper, we review this condition, comment on a few of its unanticipated features, and as an application, show that the single basepoint log-exp scheme suffers from an intricate breakdown of smoothness equivalence. We also show that the differential proximity condition is coordinate independent, even when the linear scheme is not assumed to possess the relevant smoothness.

Keywords Differential proximity condition · Nonlinear subdivision · Curvature · Resonance · Super-convergence · Nonlinear dynamical system

T. Duchamp
Department of Mathematics, University of Washington, PO Box 354350,
Seattle, WA 98195-4350, USA
e-mail: duchamp@math.washington.edu

G. Xie (✉)
Department of Mathematics, East China University of Science and Technology,
Shanghai 200237, China
e-mail: rpi1004@ecust.edu.cn

T. Yu
Department of Mathematics, Drexel University, 3141 Chestnut Street, 206 Korman Center,
Philadelphia, PA 19104, USA
e-mail: yut@drexel.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_5,

1 Introduction

In recent years, manifold-valued data has become ubiquitous; the configuration spaces of robots and space of anisotropic diffusion tensors are but two examples. Although manifolds, by definition, can be locally parameterized by points in Euclidean space, such a local parametric representation is insufficient when the topology of the underlying space is nontrivial (e.g., configuration space), and even in the case of trivial topology (anisotropic diffusion) it is desirable to respect the natural symmetry and metric structure of the underlying manifold. For these reasons, genuinely nonlinear, differential geometric, and approximation methods have come to play an important role.

Recently, several research groups [3–8, 11–14, 16–21] have studied subdivision methods for manifold-valued data. Roughly speaking, a subdivision method takes as input coarse scale data and recursively generates data at successively finer scales with the hope that in the limit a function with desired regularity properties is obtained. Such algorithms have attracted the interest of applied analysts not only because of their intrinsic beauty, but also because of their connection with wavelet-like representations. In this context, various approximation-theoretic questions come to mind, such as: How much regularity does the limit function possess? At what rate does it approximate the underlying function from which the coarse data originates?

A number of different subdivision schemes for manifold-valued data were introduced in the above references: some exploit the exponential map, others a retraction map, and some the Karcher mean, yet others are based on an embedding of the manifold into Euclidean space. But in all cases, the subdivision method is modeled on an underlying linear subdivision scheme. It is therefore natural to seek conditions under which the limit function of a manifold-valued subdivision method enjoys the same limit properties as the limit function of underlying linear subdivision scheme. This is called the *smoothness equivalence problem*. We and others [3–6, 12, 13, 16–18, 21], have introduced various *proximity conditions* that are sufficient for a manifold-valued scheme to have the smoothness equivalence property. Although numerical evidence for the necessity of these proximity conditions were given in [2, 17, 21], necessity has remained an open problem.

In a companion paper [2], we present a complete solution of the smoothness equivalence problem in terms of a new proximity condition, which we call the *differential proximity condition*. Here, we review this condition, comment on a few of its unanticipated features, and as an application, we show why the single basepoint log-exp scheme suffers from an intricate breakdown of smoothness equivalence. We also prove that the differential proximity condition is coordinate independent. The coordinate independence result established in Sect. 4 is stronger than what would follow immediately from the main result in [2].

2 Smooth Compatibility and the Differential Proximity Condition

Let M be a differentiable manifold of dimension n. A map $S : \ell(\mathbb{Z} \to M) \to \ell(\mathbb{Z} \to M)$ is called a *subdivision scheme on* M if it is of the form

$$(S\mathbf{x})_{2i+\sigma} = q_\sigma(x_{i-m_\sigma}, \cdots, x_{i-m_\sigma+L_\sigma}), \quad \sigma = 0, 1, \ i \in \mathbb{Z}, \tag{1}$$

where $L_\sigma, m_\sigma \in \mathbb{Z}$, $L_\sigma > 1$, and q_σ are continuous maps

$$q_\sigma : \underbrace{M \times \cdots \times M}_{L_\sigma + 1 \text{copies}} \to M, \quad \sigma = 0, 1, \tag{2}$$

defined in a neighborhood of the hyper-diagonal of $M \times \cdots \times M$ and satisfying the condition

$$q_\sigma(x, \ldots, x) = x\,. \tag{3}$$

The maps q_0, q_1 are usually referred to as the *even and odd rules* of the subdivision scheme S. In general, q_σ are only defined in a neighborhood of the hyper-diagonal, and therefore S is only defined for locally sufficiently dense sequences. We call L_σ the *locality factors* and m_σ the *phase factors* of the subdivision scheme S. The above definition was used, for example, in [15, 20].

We now impose additional conditions on S.

Definition 1 Let S be a subdivision scheme on M. Let S_{lin} be a linear subdivision scheme with the same phase and locality factors as S and let $q_{\text{lin},\sigma}$, $\sigma = 0, 1$, be the (linear) maps associated with S_{lin}, as in (1). We say that S is *smoothly compatible*[1] *with* S_{lin} if

(a) q_0 and q_1 are (C^∞) smooth maps, and
(b) for any $x \in M$, $dq_\sigma|_{(x,\ldots,x)} : T_xM \times \cdots \times T_xM \to T_xM$ satisfies the condition

$$dq_\sigma|_{(x,\ldots,x)}(X_0, \ldots, X_{L_\sigma}) = q_{\text{lin},\sigma}(X_0, \ldots, X_{L_\sigma}), \quad \sigma = 0, 1.$$

Remark 1 The maps $q_{\text{lin},\sigma}, \sigma = 0, 1$, are the even and odd rules of S_{lin}. The compatibility condition in Definition 1 is satisfied by all the manifold-valued data subdivision schemes seen in the literature [4–6, 9, 12, 13, 16–18, 21].

Assume that S satisfies the compatibility condition in Definition 1. Our differential proximity condition is defined in terms of a finite-dimensional map Q. From Eqs. (1) and (2), it follows that there is a unique integer K such that any $K+1$ consecutive entries in any (dense enough) sequence $\mathbf{x}$ determines *exactly* $K+1$, and no more, consecutive entries in $S\mathbf{x}$. We may call $K+1$ the size of a *minimal invariant neighborhood* of S. For any linear C^k subdivision scheme,

[1] In [7, Definition 3.5], Grohs gives a similar compatibility condition.

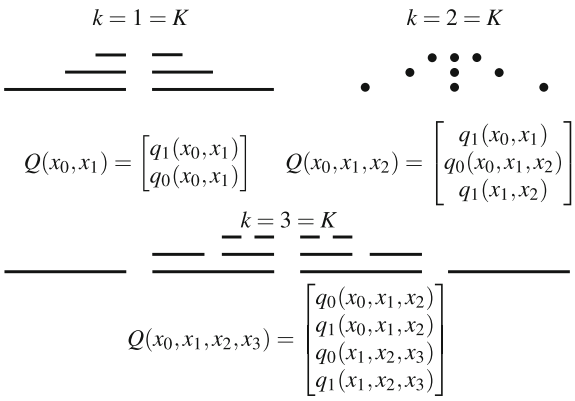


Fig. 1 If S is the symmetric C^k (degree $k+1$) B-spline subdivision scheme, the corresponding map Q has a minimal invariant neighborhood of size $K+1 = k+1$. The figure shows two subdivision steps starting from $k+1$ entries of the initial sequence (*Dots* and *intervals* are used only because of the primal and dual symmetries in the B-spline subdivision schemes for odd and even k. The symmetry properties, however, play no role here)

$$K \geq k,$$

with equality attained by the C^k, degree $k+1$, B-spline subdivision scheme (see Fig. 1). It follows that there is a map

$$Q : U \longrightarrow U \subset \underbrace{M \times \cdots \times M}_{K+1 \text{copies}}, \tag{4}$$

for U a sufficiently small open neighborhood of the hyper-diagonal, such that if $\mathbf{y} = S\mathbf{x}$, then

$$Q([\mathbf{x}_i, \ldots, \mathbf{x}_{i+K}]) = [\mathbf{y}_{2i+s}, \ldots, \mathbf{y}_{2i+s+K}], \tag{5}$$

for all i. The integer s, called a *shift factor*, is a constant independent of i but dependent on the phase factors of S. A basic property of S is that when the input sequence $\mathbf{x}$ is shifted by one entry, then the subdivided sequence $\mathbf{y}$ is shifted by two entries. This property is also reflected in Eq. (5).

The compatibility condition implies that

$$dQ|_{(x,\ldots,x)} = Q_{\text{lin}}, \quad \forall\, x \in M, \tag{6}$$

where $Q_{\text{lin}} : T_xM \times \cdots \times T_xM \to T_xM \times \cdots \times T_xM$ is the corresponding linear self-map defined by the maps $q_{\text{lin},\sigma}$ in the compatibility condition.

We shall define our new order k proximity condition based on the higher order behavior of the map Q. At this point, we work in local coordinates on M. Let Q be the map $Q(x_0, x_1, \ldots, x_K)$ expressed in local coordinates around $x_0 \in M$, and define $\Psi\colon \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ ($K+1$ copies) $\to \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ by

$$\Psi := \nabla \circ Q \circ \Sigma, \tag{7}$$

where ∇, $\Sigma = \nabla^{-1} : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ are the linear maps defined by the correspondence

$$(x_0, x_1, \ldots, x_K) \underset{\Sigma}{\overset{\nabla}{\rightleftarrows}} (\delta_0 = x_0, \delta_1, \ldots, \delta_K), \tag{8}$$

where $\delta_k :=$ k-th order difference of $x_0, x_1, \ldots, x_k$, so

$$\delta_k = \sum_{\ell=0}^{k} (-1)^{k-\ell} \binom{k}{\ell} x_\ell, \text{ and } x_k = \sum_{\ell=0}^{k} \binom{k}{\ell} \delta_\ell. \tag{9}$$

Note that Ψ is only defined in a neighborhood of $(x_0, 0, \ldots, 0)$. (Here, by abuse of notation, we identify points in M with the corresponding points in $\mathbb{R}^n$ under the given coordinate chart.) We write

$$\Psi = (\Psi_0, \Psi_1, \ldots, \Psi_K), \quad \Psi_\ell : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^n,$$

when referring to the different components of Ψ.

We remark that Eq. (6), together with linearity of Σ and ∇, implies the identity

$$d\Psi|_{(x,0,\ldots,0)} = \Psi_{\text{lin}} := \nabla \circ Q_{\text{lin}} \circ \Sigma, \quad \forall\, x. \tag{10}$$

Definition 2 Let S be a subdivision scheme on M smoothly compatible with S_{lin}. Let $k \geq 1$. We say that S and S_{lin} satisfy an *order k differential proximity condition* if for any $x_0 \in M$,

$$D^\nu \Psi_\ell|_{(x_0,0,\ldots,0)} = 0, \quad \text{when } |\nu| \geq 2, \quad \text{weight}(\nu) := \sum j\nu_j \leq \ell, \quad \forall\, \ell = 1, \ldots, k, \tag{11}$$

where $D^\nu \Psi_\ell$ denotes the derivative of Ψ_ℓ with respect to the multi-index $\nu = (\nu_1, \ldots, \nu_K)$.

Remark 2 Above, $\nu = (\nu_1, \ldots, \nu_K)$ does not have a 0-th component, so D^ν does not differentiate with respect to the 0-th argument. But since (11) has to hold for arbitrary x_0, then under the smooth compatibility assumption, condition (11) would be unaltered if we interpret ν as $(\nu_0, \nu_1, \ldots, \nu_K)$.

Remark 3 In fact, the above condition is equivalent to the following seemingly stronger condition:

$$D^\nu \Psi_\ell|_{(x_0,0,\ldots,0)} = 0, \quad \text{when } |\nu| \geq 2, \quad \text{weight}(\nu) \leq \begin{cases} \ell, & 1 \leq \ell \leq k \\ k, & \ell > k \end{cases}. \tag{12}$$

The proof, however, is rather technical as it relies on a major algebraic structure found in the proof of the sufficiency part of the following main result. See the sufficiency section of [2].

In Sect. 4, we need the following property of linear subdivision schemes:

Lemma 1 *If S_{lin} reproduces Π_k (= the space of polynomials of degree not exceeding k), then Ψ_{lin} has the block* upper triangular *form:*

$$\Psi_{\text{lin},\ell}(\delta_0, \delta_1, \ldots, \delta_K) = \begin{cases} \frac{1}{2^\ell}\delta_\ell + \sum_{\ell'=\ell+1}^{K} U_{\ell,\ell'}\delta_{\ell'}, & \ell = 0, \ldots, k \\ \sum_{\ell'=k+1}^{K} U_{\ell,\ell'}\delta_{\ell'}, & \ell = k+1, \ldots, K \end{cases}, \tag{13}$$

where $U_{\ell,\ell'}$ are scalars*-dependent only on the mask of S_{lin}. Moreover, if S_{lin} is C^k smooth, the spectral radius of the lower right block $[U_{\ell,\ell'}]_{\ell,\ell'=k+1,\ldots,K}$ is strictly smaller than $1/2^k$.*

Remark 4 We may combine Lemma 1 with (11) to restate the differential proximity condition as:

$$D^\nu \Psi_\ell|_{(x_0,0,\ldots,0)} = \begin{cases} \frac{1}{2^\ell}\text{id}, & |\nu| = 1 \text{ and } \text{weight}(\nu) = \ell, \\ 0, & |\nu| = 1 \text{ and } \text{weight}(\nu) < \ell, \text{ or} \\ & |\nu| \geq 2 \text{ and } \text{weight}(\nu) \leq \ell, \end{cases} \tag{14}$$

for $\ell = 1, \ldots, k$.

In [2], we establish the following:

Theorem 1 *Let S be a subdivision scheme on a manifold smoothly compatible with a stable C^k smooth linear scheme S_{lin}. Then S is C^k smooth* if and only if *it satisfies the order k differential proximity condition.*

Unlike the compatibility condition, the differential proximity condition is expressed in local coordinates. A natural question is whether the latter condition is invariant under change of coordinates. For the original proximity conditions, the invariance question was answered in the affirmative in [20]. Armed with Theorem 1, we know that the order k differential proximity condition, being equivalent to the C^k smoothness of S, cannot be satisfied in one chart but not another, as the notion of smoothness is coordinate independent. In summary, we have:

Corollary 1 *If S is smoothly compatible with a stable C^k linear subdivision scheme S_{lin}, then the order k differential proximity condition is invariant under change of coordinates.*

3 What's New?

The original proximity condition, used in our previous work, reads as

$$\|\Delta^{j-1} S\mathbf{x} - \Delta^{j-1} S_{\text{lin}}\mathbf{x}\|_\infty \leq C\, \Omega_j(\mathbf{x}), \quad j = 1, \ldots, k, \tag{15}$$

where

$$\Omega_j(\mathbf{x}) := \sum_{\gamma \in \Gamma_j} \prod_{i=1}^{j} \|\Delta^i \mathbf{x}\|_\infty^{\gamma_i}, \quad \Gamma_j := \left\{ \gamma = (\gamma_1, \cdots, \gamma_j) \;\middle|\; \gamma_i \in \mathbb{Z}^+, \sum_{i=1}^{j} i\,\gamma_i = j+1 \right\}.$$

It is well known that this condition is a sufficient condition for the C^k-equivalence property ([17, Theorem 2.4].) Moreover, years of usage of this condition (15) and numerical evidence suggests that it is also necessary.

This original proximity condition does not explicitly assume a compatibility condition between S and S_{lin}, making it difficult to formulate a precise necessary condition. In our new formulation, we explicitly impose the smooth compatibility condition in Definition 1, which enables us to address the problem of necessity.

Our new formulation also addresses a perplexing aspect of condition (15). A careful inspection of the proof of [17, Theorem 2.4], shows that only the following proximity condition is needed[2]:

$$\|\Delta^j S\mathbf{x} - \Delta^j S_{\text{lin}}\mathbf{x}\|_\infty \le C\,\Omega_j(\mathbf{x}), \quad j = 1, \ldots, k, \tag{16}$$

provided that we have already established C^0 regularity of S. We are thus faced with a dilemma: Despite the strong empirical evidence for the necessity of the proximity condition (15), it appears that it is unnecessarily strong!

A moment's thought suggests that the new proximity condition (11) is merely a differential version of the weaker condition (16). In fact, in all previous work a proximity condition is always established by a local Taylor expansion of $S\mathbf{x} - S_{\text{lin}}\mathbf{x}$ (recall that subdivision schemes act locally). Consequently, the differential aspect of (11) is hardly anything new. But once the differential proximity condition is written in the form (11) (or in the equivalent form (14)), we see a natural interpretation: Viewing the map

$$Q : U \to U$$

as a discrete dynamical system, the proximity conditions can be interpreted in terms of the rate of approach of points in U to the hyper-diagonal, which is the fixed-point set of Q:

- Condition (14) then suggests that the linear term $2^{-\ell}\text{id}$ is the dominant term, so, generically, the k-th order differences of the subdivision data *within any invariant neighborhood* (Fig. 1) decays like $O(2^{-jk})$.
  If k is the first order at which the differential proximity condition fails, then there is a weight k term in the Taylor expansion of Ψ_k, such a nonlinear term is called a *resonance term* in the dynamical system literature, and the dynamical system interpretation would suggest that the k-th order differences of the subdivision data decays slower than $O(2^{-jk})$. More precisely, the presence of resonance slows down the decay to $O(j2^{-jk})$.

[2] Use $\|\Delta\mathbf{x}\|_\infty \le 2\|\mathbf{x}\|_\infty$ to see that (15) implies (16).

- This in turn suggests the necessity result. However, proving the lower bound result needed and tackling the lack of stability condition in the nonlinear subdivision theory are technically difficult. The former requires us to come up with a delicate argument to show that the initial data exists so that the effect of resonance terms would not dissipate away in the course of iteration. The latter requires us to exploit a subtle superconvergence property.
- The same dynamical system interpretation suggests that our differential proximity condition may be too weak. For, unlike (15) or (16), it involves a fixed, although arbitrary, invariant neighborhood, and therefore does not appear to capture the expanding nature of a subdivision scheme.[3] Worse, the sufficiency part of the theorem concerns *establishing* the C^k-smoothness of the limiting function, and that would require one to analyze the decay rate of order $k+1$, not k, differences. If one examines Fig. 1, one sees that a minimal invariant neighborhood may very well be too small to allow for the computation of any $k+1$ order difference. Fortunately, an unexpected algebraic structure we discovered in [2] not only makes the seemingly impossible mission of proving sufficiency possible, but also explains simultaneously why the apparent stronger than necessary proximity condition (15) always holds true whenever C^k equivalence holds.

4 Coordinate Independence

Corollary 1 suggests that there is an intrinsic, coordinate-free, reformulation of the differential proximity condition waiting to be discovered. With this in mind, we establish here the following coordinate independence result.

Theorem 2 *If S is smoothly compatible with a Π_k reproducing linear subdivision scheme S_{lin}, then the order k differential proximity condition is invariant under change of coordinates.*

Note that this result is stronger than Corollary 1, because a stable C^k linear subdivision scheme must reproduce Π_k, but the converse is far from being true.

Let $\chi(x) = \overline{x}$ be the change of coordinate map on M, and let $Q(x_0, x_1, \ldots, x_K)$ and $\overline{Q}(\overline{x}_0, \overline{x}_1, \ldots, \overline{x}_K)$ denote the expressions for map Q in these two coordinate systems. Writing

$$\chi_{\text{vec}}(x_0, x_1, \ldots, x_K) := (\chi(x_0), \chi(x_1), \ldots, \chi(x_K)),$$

shows that $Q(x_0, x_1, \ldots, x_K)$ and $\overline{Q}(\overline{x}_0, \overline{x}_1, \ldots, \overline{x}_K)$ are related by the formula

$$\overline{Q} = \chi_{\text{vec}} \circ Q \circ \chi_{\text{vec}}^{-1}. \tag{17}$$

[3] For instance, it is well known from the linear theory that the spectral property of Ψ_{lin} alone is insufficient for characterizing the regularity property S_{lin}.

The map Ψ, by the definition (7), then takes the following forms in the two coordinate systems:

$$\Psi(\delta_0, \delta_1, \ldots, \delta_K) = \nabla \circ Q \circ \Sigma(\delta_0, \delta_1, \ldots, \delta_K),$$
$$\overline{\Psi}(\overline{\delta}_0, \overline{\delta}_1, \ldots, \overline{\delta}_K) = \nabla \circ \overline{Q} \circ \Sigma(\overline{\delta}_0, \overline{\delta}_1, \ldots, \overline{\delta}_K).$$

It then follows that a change of coordinates induces the following transformation rule for Ψ:

$$\overline{\Psi} = \nabla \circ \overline{Q} \circ \Sigma = \nabla \circ \chi_{\text{vec}} \circ Q \circ \chi_{\text{vec}}^{-1} \circ \Sigma = \underbrace{\nabla \circ \chi_{\text{vec}} \circ \Sigma}_{=:\Xi} \circ \underbrace{\nabla \circ Q \circ \Sigma}_{=\Psi} \circ \underbrace{\nabla \circ \chi_{\text{vec}}^{-1} \circ \Sigma}_{=\Xi^{-1}}. \tag{18}$$

Proof The proof proceeds in two steps:

Step 1. Note the following structure of the Taylor expansion of $\Xi_\ell(\delta_0, \delta_1, \ldots, \delta_K)$ around a point $(\delta_0 = x_0, 0, \ldots, 0)$. Note also that $\Xi_0(\delta_0, \delta_1, \ldots, \delta_K) = \chi(x_0)$. For $\ell \geq 1$, compute as follows:

$$\begin{aligned}
\Xi_\ell(\delta) &= \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \chi(x_i) \\
&= \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \left[\chi(x_0) + \chi'(x_0)(x_i - x_0) + \sum_{k\geq 2} \frac{1}{k!} \chi^{(k)}(x_0)(x_i - x_0)^k \right] \\
&= \chi'(x_0) \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} x_i + \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \sum_{k\geq 2} \frac{1}{k!} \chi^{(k)}(x_0)(x_i - x_0)^k \\
&= \chi'(x_0)\delta_\ell + \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \sum_{k\geq 2} \frac{1}{k!} \chi^{(k)}(x_0) \Big[\sum_{j=1}^{i} \tbinom{i}{j} \delta_j \Big]^k \\
&= \chi'(x_0)\delta_\ell + \sum_{k\geq 2} \frac{1}{k!} \sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \sum_{j_1,\ldots,j_k \geq 1} \tbinom{i}{j_1} \cdots \tbinom{i}{j_k} \chi^{(k)}(x_0)(\delta_{j_1}, \ldots, \delta_{j_k}) \\
&= \chi'(x_0)\delta_\ell + \sum_{k\geq 2} \frac{1}{k!} \sum_{j_1,\ldots,j_k \geq 1} \left[\sum_{i=0}^{\ell} (-1)^{\ell-i} \tbinom{\ell}{i} \tbinom{i}{j_1} \cdots \tbinom{i}{j_k} \right] \chi^{(k)}(x_0)(\delta_{j_1}, \ldots, \delta_{j_k}),
\end{aligned}$$

where we have repeatedly used the multilinearity of $\chi^{(k)}(x_0)$.

Note that, for fixed $j_1, \ldots, j_k$, $\binom{i}{j_1} \cdots \binom{i}{j_k}$ is a polynomial in i of degree $j_1 + \cdots + j_k$. Then, by (9), $\sum_{i=0}^{\ell} (-1)^{\ell-i} \binom{\ell}{i} \binom{i}{j_1} \cdots \binom{i}{j_k}$ is an ℓ-th order difference of uniform samples of a degree $j_1 + \cdots + j_k$ polynomial, which vanishes when $j_1 + \cdots + j_k < \ell$.

Consequently, the ℓ-th component of Ξ has no (linear or nonlinear) terms of weight less than ℓ, thus

$$D^\nu \Xi_\ell|_{(x_0,0,\ldots,0)} = 0, \quad \text{weight}(\nu) < \ell.$$

The same conditions hold with Ξ replaced by Ξ^{-1}—simply replace χ with χ^{-1} in the derivation above.

We now know that the ℓ-th component of both Ξ and Ξ^{-1} do not have terms of weight *strictly* less than ℓ, the proximity condition of Ψ says that its ℓ-th component Ψ_ℓ does not have terms of weight ℓ and lower. These facts alone only guarantee that $\overline{\Psi}_\ell$ does not have terms of weight $\ell - 1$ and lower. (The second part of the proof explains this along the way.)

Step 2. To complete the proof, we now show that all weight ℓ terms in $\overline{\Psi}_\ell$ vanish. Assume that Ψ satisfies the order k differential proximity condition. By Remark 4 and Step 1, we have for $\ell = 2, \ldots, k$,

$$\begin{aligned}\overline{\Psi}_\ell(\overline{\delta}) &= \Xi_\ell(\Psi \circ \Xi^{-1}(\overline{\delta})) \\ &= \sum_{\text{weight}(\nu)=\ell} \frac{1}{\nu!} D^\nu \Xi_\ell|_{(x_0,0,\ldots,0)} \left(\Psi_1(\Xi^{-1}(\overline{\delta}))^{\nu_1}, \ldots, \Psi_\ell(\Xi^{-1}(\overline{\delta}))^{\nu_\ell}\right) \qquad (19) \\ &\qquad + (\text{weight} > \ell \text{ terms}).\end{aligned}$$

For each $i = 1, \ldots, \ell$, again by Remark 4 and Step 1,

$$\Psi_i(\Xi^{-1}(\overline{\delta})) = \frac{1}{2^i} \sum_{\text{weight}(\eta)=i} \frac{1}{\eta!} D^\eta \Xi_i^{-1}|_{(\overline{x}_0,0,\ldots,0)} \overline{\delta}^\eta + (\text{weight} > i \text{ terms}). \quad (20)$$

An inspection then reveals that the only weight ℓ terms in $\overline{\Psi}_\ell(\overline{\delta})$ are

$$\begin{aligned}&\sum_{\text{weight}(\nu)=\ell} \frac{1}{\nu!} D^\nu \Xi_\ell \left(\left(2^{-1} \sum_{\text{weight}(\eta)=1} \frac{1}{\eta!} D^\eta \Xi_i^{-1} \overline{\delta}^\eta \right)^{\nu_1}, \ldots, \right. \\ &\qquad \left. \left(2^{-\ell} \sum_{\text{weight}(\eta)=\ell} \frac{1}{\eta!} D^\eta \Xi_i^{-1} \overline{\delta}^\eta \right)^{\nu_\ell} \right) \\ &= 2^{-\ell} \sum_{\text{weight}(\nu)=\ell} \frac{1}{\nu!} D^\nu \Xi_\ell \left(\left(\sum_{\text{weight}(\eta)=1} \frac{1}{\eta!} D^\eta \Xi_i^{-1} \overline{\delta}^\eta \right)^{\nu_1}, \ldots, \right. \\ &\qquad \left. \left(\sum_{\text{weight}(\eta)=\ell} \frac{1}{\eta!} D^\eta \Xi_i^{-1} \overline{\delta}^\eta \right)^{\nu_\ell} \right). \qquad (21)\end{aligned}$$

By yet another inspection, we see that by virtue of the chain rule the weight ℓ terms in the Taylor expansion of $(\Xi \circ \Xi^{-1})_\ell$ are given by the summation after the $2^{-\ell}$ factor in (21). But $\Xi \circ \Xi^{-1} =$ identity, so any nonlinear term in its Taylor expansion must vanish. In other words, all the nonlinear (i.e., degree > 1) terms in (21) vanish. This implies that the Taylor expansion of $\overline{\Psi}_\ell(\overline{\delta})$ has the linear term $2^{-\ell}\overline{\delta}_\ell$ as its only

weight ℓ term, and all other terms, linear or nonlinear, are of weight strictly greater than ℓ. In other words, $\overline{\Psi}$ satisfies the same differential proximity condition as Ψ. □

It is worth stressing the role of the Π_k reproduction property of S_{lin} in the coordinate independence proof above: it induces a kind of "upper-triangular" structure in Ψ_{lin} (Lemma 1) and enters the proof in Step 2 above. In particular, the dyadic eigenvalues in Ψ_{lin} are the key to the derivation of (21). Indeed, (21) implies that as far as the lowest weight terms (i.e., weight ℓ) in the ℓ-th component are concerned, the map $\Xi \circ \Psi \circ \Xi^{-1}$ is the same as $2^{-\ell} \Xi \circ \Xi^{-1}$.

5 The Log-exp Scheme on Surfaces

As an application of Theorem 1, we show that the single basepoint log-exp scheme introduced in [9] does not satisfy the differential proximity condition. Consequently, thanks to Theorem 1, we can conclude a breakdown of smoothness equivalence in the single basepoint scheme.

The paper [3] studies the proximity condition for the single basepoint schemes defined by general retraction maps. Since [3] predates the development of Theorem 1, the results therein were derived from the original proximity condition. As the discussion in Sect. 3 hinted, the breakdown results based on the original proximity condition from [3] easily imply corresponding breakdown results based on our differential proximity condition. Therefore, the anticipated breakdown of smoothness equivalence in the more general setting once again follows from Theorem 1.

As the computations in [3] are rather involved, we present here the special case of the single basepoint log-exp scheme based on the C^5, symmetric B-spline, whose subdivision mask is

$$(a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2, a_3, a_4) = \frac{1}{64}(1, 7, 21, 35, 35, 21, 7, 1)\,, \tag{22}$$

and in the simple case where M is a two-dimensional Riemannian manifold. In this case,

$$q_0(x_0, x_1, x_2, x_3) = \exp_{x_2}\left(a_4 \log_{x_2}(x_0) + a_2 \log_{x_2}(x_1) + a_{-2} \log_{x_2}(x_3)\right), \tag{23a}$$

$$q_1(x_0, x_1, x_2, x_3) = \exp_{x_1}\left(a_3 \log_{x_1}(x_0) + a_{-1} \log_{x_1}(x_2) + a_{-3} \log_{x_1}(x_3)\right), \tag{23b}$$

$$Q(x_0, x_1, x_2, x_3, x_4, x_5) = \begin{pmatrix} q_1(x_0, x_1, x_2, x_3) \\ q_0(x_0, x_1, x_2, x_3) \\ q_1(x_1, x_2, x_3, x_4) \\ q_0(x_1, x_2, x_3, x_4) \\ q_1(x_2, x_3, x_4, x_5) \\ q_0(x_2, x_3, x_4, x_5) \end{pmatrix}, \tag{24}$$

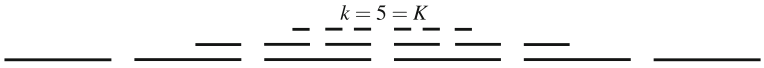


Fig. 2 Minimal invariant neighborhood of the C^5 B-spline subdivision scheme

and $\Psi = (\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5)$ is defined according to (7) (Fig. 2).

Theorem 3 *The nonlinear scheme S defined by (23) satisfies the C^5-equivalence property if and only if the manifold M has vanishing curvature.*

One direction is clear, suppose M has vanishing curvature, we may then choose local coordinates about any point in M in which the Riemannian metric is the Euclidean metric. But in these coordinates, S coincides with the (linear) C^5 B-spline scheme.

Now assume that M has nonzero curvature at the point x_0. Then by Theorems 1 and 2, it suffices to choose coordinates centered at x_0 in which the derivative

$$D^{\nu}\Psi_5|_{(x_0,0,\ldots,0)}\,, \text{ for } \nu = (1,2,0,0,0) \tag{25}$$

does not vanish.

Notice that, although ν has weight 5, it has degree 3. Consequently, to compute this derivative, we need to only compute the Taylor expansion of Ψ_5 up to order 3 and weight 5 in some coordinate system.

The computations are vastly simplified if we perform them in *Riemann normal coordinates* centered at x_0. We merely summarize the results from Riemannian geometry we need. (A detailed treatment of normal coordinates is given in Chap. 4 of [10] as well as in [1], particularly pages 41–42).

Let $x = (u, v)$ denote normal coordinates on $\mathbb{R}^2$ centered at the origin. Let (u, v, U, V) denote the corresponding coordinates on the tangent bundle TM, where (U, V) are the components of the tangent vector based at (u, v). *Riemann's Theorem* then states that in these coordinates the coefficients of the Riemannian metric are given by

$$\begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{K_0}{3} \begin{pmatrix} v^2 & uv \\ uv & u^2 \end{pmatrix},$$

where K_0 denotes the Gauss curvature at $(0, 0)$. A standard computation using this formula and the differential equations for geodesics yields the following Taylor expansion for the exponential map about $(0, 0, 0, 0)$ up to degree 3 in (u, v, U, V):

$$\exp_{(u,v)}(U, V) \approx (u, v) + (U, V) + \frac{2}{3} K_0 \det \begin{pmatrix} U & V \\ u & v \end{pmatrix} \cdot (V, -U)\,.$$

From this, one finds that up to degree 3 in (u_0, v_0, u, v), the Taylor expansion of log is given by

$$\log_{(u_0,v_0)}(u, v) \approx (u, v) - (u_0, v_0) - \frac{2}{3} K_0 \det \begin{pmatrix} u_0 & v_0 \\ u & v \end{pmatrix} \cdot (-(v - v_0), (u - u_0))\,.$$

Setting $\delta_\ell = (\delta_{\ell,u}, \delta_{\ell,v})$, substituting the expansions for exp and log into the definition of q_σ yields the Taylor expansion of q_σ up to degree 3. Substituting these Taylor expansions into Ψ_5, and dropping all terms in δ_ℓ of degree larger than 3 and weight larger than 5 yields (after a straightforward, but lengthy computation) the formula

$$\Psi_5(\delta) \approx \frac{1}{2^5}(\delta_{5,u}, \delta_{5,v}) + \frac{7}{16}K_0\Big\{ \det \begin{pmatrix} \delta_{1,u} & \delta_{1,v} \\ \delta_{2,u} & \delta_{2,v} \end{pmatrix} (-\delta_{2,v}, \delta_{2,u}) + \det \begin{pmatrix} \delta_{1,u} & \delta_{1,v} \\ \delta_{3,u} & \delta_{3,v} \end{pmatrix} (-\delta_{1,v}, \delta_{1,u}) \Big\}.$$

The weight 5 terms are nonzero exactly when $K_0 \neq 0$, so Theorem 3 is proved. This formally disproves the smoothness equivalence conjecture first posted in [9].

While Theorem 3 says that nonvanishing curvature is the root cause of the C^5-breakdown in the nonlinear scheme defined by (22)–(23), one can show by a similar computation that the same scheme satisfies C^4-equivalence *regardless of the curvature of* M. In [3, 21], such a C^4-equivalence property was found to be attributable to *both* a special property of the exponential map and the dual time-symmetry property of the scheme (22)–(23). More precisely,

- If one replaces the exponential map by an arbitrary retraction map, then the resulting scheme will satisfy the C^2-equivalence property but suffer a C^3-breakdown on a general manifold.
- If one replaces the underlying linear scheme by a stable C^4 linear subdivision scheme *without* a dual time-symmetry, then the resulting scheme will satisfy a C^3-equivalence property but suffer a C^4-breakdown on a general manifold.

To illustrate the latter point, we next consider the single basepoint log-exp scheme based on the C^6 B-spline, whose subdivision mask is

$$(a_{-4}, a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2, a_3, a_4) = \frac{1}{128}(1, 8, 28, 56, 70, 56, 28, 8, 1). \quad (26)$$

In this case,

$$q_0(x_0, x_1, x_2, x_3, x_4) = \exp_{x_2}\big(a_4 \log_{x_2}(x_0) + a_2 \log_{x_2}(x_1) + a_{-2} \log_{x_2}(x_4) + a_{-4} \log_{x_2}(x_5)\big), \quad (27a)$$

$$q_1(x_0, x_1, x_2, x_3) = \exp_{x_1}\big(a_3 \log_{x_1}(x_0) + a_{-1} \log_{x_1}(x_2) + a_{-3} \log_{x_1}(x_3)\big), \quad (27b)$$

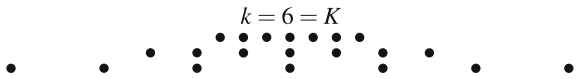


Fig. 3 Minimal invariant neighborhood of the C^6 B-spline subdivision scheme

$$Q(x_0, x_1, x_2, x_3, x_4, x_5, x_6) = \begin{pmatrix} q_1(x_0, x_1, x_2, x_3) \\ q_0(x_0, x_1, x_2, x_3, x_4) \\ q_1(x_1, x_2, x_3, x_4) \\ q_0(x_1, x_2, x_3, x_4, x_5) \\ q_1(x_2, x_3, x_4, x_5) \\ q_0(x_2, x_3, x_4, x_5, x_5) \\ q_1(x_3, x_5, x_6, x_6) \end{pmatrix}, \tag{28}$$

and $\Psi = (\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, \Psi_6)$ is defined according to (7) (Fig. 3).

Note that the underlying scheme in this case is even smoother than before (C^6 instead of C^5), and it has a primal symmetry. However, the resulting nonlinear scheme, based on the single basepoint strategy, fails to inherit such a primal symmetry.

Theorem 4 *The nonlinear scheme S defined by (27) satisfies the C^4-equivalence property if and only if the manifold M has vanishing curvature.*

One direction is clear, for suppose M has vanishing curvature, we may then choose local coordinates about any point in M in which the Riemannian metric is the Euclidean metric. But in these coordinates, S coincides with the (linear) C^6 B-spline scheme.

Now assume that M has nonzero curvature at the point x_0. Then by Theorems 1 and 2 it suffices to choose coordinates centered at x_0 in which the derivative

$$D^\nu \Psi_4|_{(x_0,0,\ldots,0)}, \quad \text{for} \quad \nu = (2, 1, 0, 0, 0) \tag{29}$$

does not vanish.

Notice that, although ν has weight 4, it has degree 3. Consequently, to compute this derivative, we need to only compute the Taylor expansion of Ψ_4 up to order 3 and weight 4. We proceed as before, we substitute the Taylor expansions for q_0 and q_1 into Ψ_4, and drop all terms in δ_ℓ of degree larger than 3 and weight larger than 4 to arrive at the expansion

$$\Psi_4(\delta) \approx \frac{1}{2^4}(\delta_{4,u}, \delta_{4,v}) + \frac{1}{3} K_0 \det \begin{pmatrix} \delta_{1,u} & \delta_{1,v} \\ \delta_{2,u} & \delta_{2,v} \end{pmatrix} (-\delta_{1,v}, \delta_{1,u}).$$

The weight 4 terms are nonzero exactly when $K_0 \neq 0$, so Theorem 4 is proved.

Acknowledgments Tom Duchamp gratefully acknowledges the support and hospitality provided by the IMA during his visit from April to June 2011, when part of the work in this article was completed, as well as travel support through the PIMS CRG on Applied and Computational Harmonic Analysis. Gang Xie's research was supported by the Fundamental Research Funds for the Central

Universities and the National, Natural Science Foundation of China (No.11101146). Thomas Yu's research was partially supported by the National Science Foundation grants DMS 0915068 and DMS 1115915, as well as a fellowship offered by the Louis and Bessie Stein family.

References

1. Bishop, R.L.: Riemannian manifolds. arXiv:1303.5390v2 [math.DG] (2013)
2. Duchamp, T., Xie, G., Yu, T.P.-Y.: A new proximity condition for manifold-valued subdivision schemes. (2014)
3. Duchamp, T., Xie, G., Yu, T.P.-Y.: Single basepoint subdivision schemes for manifold-valued data: time-symmetry without space-symmetry. Found. Comput. Math. **13**(5), 693–728 (2013). doi:10.1007/s10208-013-9144-1
4. Grohs, P.: Smoothness equivalence properties of univariate subdivision schemes and their projection analogues. Numer. Math. **113**(2), 163–180 (2009)
5. Grohs, P.: Smoothness of interpolatory multivariate subdivision in Lie groups. IMA J. Numer. Anal. **29**(3), 760–772 (2009)
6. Grohs, P.: A general proximity analysis of nonlinear subdivision schemes. SIAM J. Math. Anal. **42**(2), 729–750 (2010)
7. Grohs, P.: Stability of manifold-valued subdivision schemes and multiscale transformations. Constr. Approx. **32**(3), 569–596 (2010)
8. Grohs, P.: Finite elements of arbitrary order and quasiinterpolation for Riemannian data. IMA J. Numer. Anal. **33**(3), 849–874 (2013)
9. Ur Rahman, I., Drori, I., Stodden, V.C., Donoho, D.L., Schröder, P.: Multiscale representations for manifold-valued data. Multiscale Model. Simul. **4**(4):1201–1232 (2005)
10. Spivak M.: A comprehensive introduction to differential geometry, vol. 2. 3rd edn. Publish or Perish, Wilmington (2005)
11. Wallner, J.: Smoothness analysis of subdivision schemes by proximity. Constr. Approx. **24**(3), 289–318 (2006)
12. Wallner, J., Dyn, N.: Convergence and C^1 analysis of subdivision schemes on manifolds by proximity. Comput. Aided Geom. Des. **22**(7), 593–622 (2005)
13. Wallner, J., Nava Yazdani, E., Grohs, P.: Smoothness properties of Lie group subdivision schemes. Multiscale Model. Simul. **6**(2), 493–505 (2007)
14. Wallner, J., Nava Yazdani, E., Weinmann, A.: Convergence and smoothness analysis of subdivision rules in Riemannian and symmetric spaces. Adv. Comput. Math. **34**(2), 201–218 (2011)
15. Xie, G., Yu, T.P.-Y.: Smoothness analysis of nonlinear subdivision schemes of homogeneous and affine invariant type. Constr. Approx. **22**(2), 219–254 (2005)
16. Xie, G., Yu, T.P.-Y.: Smoothness equivalence properties of manifold-valued data subdivision schemes based on the projection approach. SIAM J. Numer. Anal. **45**(3), 1200–1225 (2007)
17. Xie, G., Yu, T.P.-Y.: Smoothness equivalence properties of general manifold-valued data subdivision schemes. Multiscale Model. Simul. **7**(3), 1073–1100 (2008)
18. Xie, G., Yu, T.P.-Y.: Smoothness equivalence properties of interpolatory Lie group subdivision schemes. IMA J. Numer. Anal. **30**(3), 731–750 (2009)
19. Xie, G., Yu, T.P.-Y.: Approximation order equivalence properties of manifold-valued data subdivision schemes. IMA J. Numer. Anal. **32**(2), 687–700 (2011)
20. Xie, G., Yu, T.P.-Y.: Invariance property of the proximity condition in nonlinear subdivision. J. Approx. Theory **164**(8), 1097–1110 (2012)
21. Nava Yazdani, E., Yu, T.P.-Y.: On Donoho's log-exp subdivision scheme: choice of retraction and time-symmetry. Multiscale Model. Simul. **9**(4), 1801–1828 (2011)

Wachspress and Mean Value Coordinates

Michael S. Floater

Abstract This paper gives a brief survey of two kinds of generalized barycentric coordinates, Wachspress and mean value coordinates, and their applications. Applications include surface parameterization in geometric modeling, curve and surface deformation in computer graphics, and their use as nodal shape functions for polygonal and polyhedral finite element methods.

Keywords Barycentric coordinates · Wachspress coordinates · Mean value coordinates

1 Introduction

There is no unique way to generalize barycentric coordinates to polygons and polyhedra. However, two specific choices have turned out to be useful in several applications: Wachspress and mean value coordinates, and the purpose of this paper is to survey their main properties, applications, and generalizations.

For convex polygons, the coordinates of Wachspress and their generalizations due to Warren and others [15, 22, 30–33] are arguably the simplest since they are rational functions (quotients of bivariate polynomials), and it is relatively simple to evaluate them and their derivatives. Some simple bounds on their gradients have been found recently in [6], justifying their use as shape functions for polygonal finite elements.

For star-shaped polygons, and arbitrary polygons, Wachspress coordinates are not well-defined, and mean value coordinates are perhaps the most popular choice, due to their generality and surprising robustness over complex geometric shapes [1, 2, 4, 8, 13, 16], even though they are no longer positive if the polygon is not star-shaped.

M. S. Floater (✉)
Department of Mathematics, University of Oslo, Blindern, Moltke Moes vei 35, PO Box 1053, 0316 Oslo, Norway
e-mail: michaelf@ifi.uio.no

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_6,

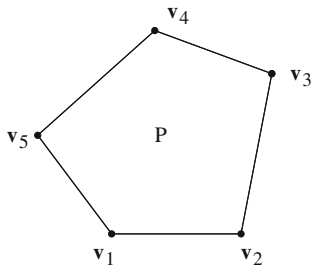


Fig. 1 Vertex ordering for a polygon

They have been employed in various tasks in geometric modeling, such as surface parameterization and plane and space deformation, as well as shading and animation in computer graphics.

While most of this paper surveys previous results, we add two new ones. The first is a new formula for the gradients of mean value coordinates, which could be used in finite element methods. The second is an alternative formula for the mean value coordinates themselves, which is valid on the boundary of the polygon. Though it may not be of practical value, it offers an alternative way of showing that these coordinates extend continuously to the polygon boundary.

2 Barycentric Coordinates on Polygons

Let $P \subset \mathbb{R}^2$ be a convex polygon, viewed as an open set, with vertices $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, $n \geq 3$, in some anticlockwise ordering. Figure 1 shows an example with $n = 5$. We call any functions $\phi_i : P \to \mathbb{R}$, $i = 1, \ldots, n$, (generalized) barycentric coordinates if, for $\mathbf{x} \in P$, $\phi_i(\mathbf{x}) \geq 0$, $i = 1, \ldots, n$, and

$$\sum_{i=1}^{n} \phi_i(\mathbf{x}) = 1, \qquad \sum_{i=1}^{n} \phi_i(\mathbf{x})\mathbf{v}_i = \mathbf{x}. \tag{1}$$

For $n = 3$, the functions ϕ_1, ϕ_2, ϕ_3 are uniquely determined and are the usual triangular barycentric coordinates w.r.t. the triangle with vertices $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. For $n \geq 4$, the choice of $\phi_1, \ldots, \phi_n$ is no longer unique. However, they share some basic properties, derived in [7]:

- The functions ϕ_i have a unique continuous extension to ∂P, the boundary of P.
- Lagrange property: $\phi_i(\mathbf{v}_j) = \delta_{ij}$.
- Piecewise linearity on ∂P:

$$\phi_i((1-\mu)\mathbf{v}_j + \mu\mathbf{v}_{j+1}) = (1-\mu)\phi_i(\mathbf{v}_j) + \mu\phi_i(\mathbf{v}_{j+1}), \qquad \mu \in [0,1]. \tag{2}$$

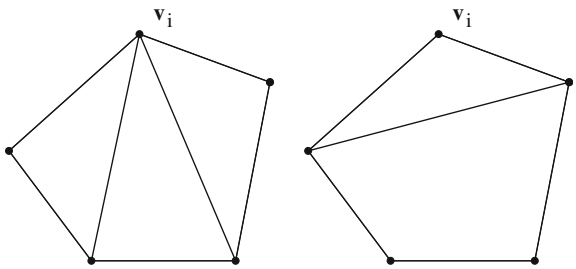


Fig. 2 Partitions for L_i and ℓ_i

(Here and throughout, vertices are indexed cyclically, i.e., $\mathbf{v}_{n+1} := \mathbf{v}_1$ etc.)

- Interpolation: if

$$g(\mathbf{x}) = \sum_{i=1}^{n} \phi_i(\mathbf{x}) f(\mathbf{v}_i), \qquad \mathbf{x} \in P, \tag{3}$$

then $g(\mathbf{v}_i) = f(\mathbf{v}_i)$. We call g a barycentric interpolant to f.

- Linear precision: if f is linear then $g = f$.
- $\ell_i \le \phi_i \le L_i$ where $L_i, \ell_i : P \to \mathbb{R}$ are the continuous, piecewise linear functions over the partitions of P shown in Fig. 2 satisfying $L_i(\mathbf{v}_j) = \ell_i(\mathbf{v}_j) = \delta_{ij}$.

3 Wachspress Coordinates

Wachspress coordinates were developed by Wachspress [30], and Warren [32]. They can be defined by the formula

$$\phi_i(\mathbf{x}) = \frac{w_i(\mathbf{x})}{\sum_{j=1}^{n} w_j(\mathbf{x})}, \tag{4}$$

where

$$w_i(\mathbf{x}) = \frac{A(\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1})}{A(\mathbf{x}, \mathbf{v}_{i-1}, \mathbf{v}_i) A(\mathbf{x}, \mathbf{v}_i, \mathbf{v}_{i+1})},$$

and $A(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ denotes the signed area of the triangle with vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$,

$$A(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) := \frac{1}{2} \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix},$$

where $\mathbf{x}_k = (x_k, y_k)$; see Fig. 3. The original proof that these coordinates are barycentric was based on the so-called adjoint of P; see Wachspress [30], and Warren [32].

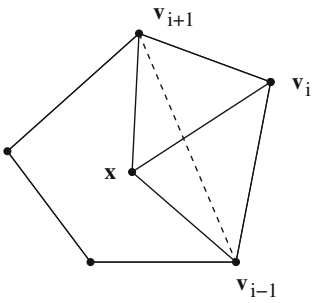


Fig. 3 Triangles defining Wachspress coordinates

The following proof is due to Meyer et al. [22]. Due to (4), it is sufficient to show that

$$\sum_{i=1}^{n} w_i(\mathbf{x})(\mathbf{v}_i - \mathbf{x}) = 0. \tag{5}$$

Fix $\mathbf{x} \in P$ and let

$$A_i = A_i(\mathbf{x}) = A(\mathbf{x}, \mathbf{v}_i, \mathbf{v}_{i+1}) \qquad \text{and} \qquad B_i = A(\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}).$$

Then we can express $\mathbf{x}$ as a barycentric combination of $\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}$:

$$\mathbf{x} = \frac{A_i}{B_i}\mathbf{v}_{i-1} + \frac{(B_i - A_{i-1} - A_i)}{B_i}\mathbf{v}_i + \frac{A_{i-1}}{B_i}\mathbf{v}_{i+1},$$

regardless of whether $\mathbf{x}$ lies inside or outside the triangle formed by $\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}$. This equation can be rearranged in the form

$$\frac{B_i}{A_{i-1}A_i}(\mathbf{v}_i - \mathbf{x}) = \frac{1}{A_{i-1}}(\mathbf{v}_i - \mathbf{v}_{i-1}) - \frac{1}{A_i}(\mathbf{v}_{i+1} - \mathbf{v}_i).$$

Summing both sides of this over i, and observing that the right hand side then cancels to zero, gives

$$\sum_{i=1}^{n} \frac{B_i}{A_{i-1}A_i}(\mathbf{v}_i - \mathbf{x}) = 0,$$

which proves (5).

3.1 Rational Functions

Another way of expressing these coordinates is in the form

$$\phi_i(\mathbf{x}) = \frac{\hat{w}_i(\mathbf{x})}{\sum_{j=1}^{n} \hat{w}_j(\mathbf{x})}, \qquad \hat{w}_i(\mathbf{x}) = B_i \prod_{j \neq i-1,i} A_j(\mathbf{x}), \tag{6}$$

and since each area $A_j(\mathbf{x})$ is linear in $\mathbf{x}$, we see from this that ϕ_i is a rational (bivariate) function, with total degree $\leq n-2$ in the numerator and denominator. In fact, the denominator, $W = \sum_{j=1}^{n} \hat{w}_j$, has total degree $\leq n-3$ due to linear precision: since (5) holds with w_i replaced by $\hat{w}_i$, it implies that

$$\sum_{i=1}^{n} \hat{w}_i(\mathbf{x})\mathbf{v}_i = W(\mathbf{x})\mathbf{x}.$$

The left hand side is a (vector-valued) polynomial of degree $\leq n-2$ in $\mathbf{x}$ and since $\mathbf{x}$ has degree 1, the degree of W must be at most $n-3$.

The degrees, $n-2$ and $n-3$, of the numerator and denominator of ϕ_i agree with the triangular case where $n=3$ and the coordinates are linear functions.

We note that the 'global' form of $\phi_i(\mathbf{x})$ in (6) is also valid for $\mathbf{x} \in \partial P$, unlike the 'local' form (4), though it requires more computation for large n.

3.2 Perpendicular Distances to Edges

An alternative way of expressing Wachspress coordinates is in terms of the perpendicular distances of $\mathbf{x}$ to the edges of P. This is the form used by Warren et al. [33], and it generalizes in a natural way to higher dimension.

For each i, let $\mathbf{n}_i \in \mathbb{R}^2$ be the outward unit normal to the edge $e_i = [\mathbf{v}_i, \mathbf{v}_{i+1}]$, and for any $\mathbf{x} \in P$ let $h_i(\mathbf{x})$ be the perpendicular distance of $\mathbf{x}$ to the edge e_i, so that

$$h_i(\mathbf{x}) = (\mathbf{v}_i - \mathbf{x}) \cdot \mathbf{n}_i = (\mathbf{v}_{i+1} - \mathbf{x}) \cdot \mathbf{n}_i,$$

see Fig. 4. Then the coordinates in (4) can be expressed as

$$\phi_i(\mathbf{x}) = \frac{\tilde{w}_i(\mathbf{x})}{\sum_{j=1}^{n} \tilde{w}_j(\mathbf{x})}, \tag{7}$$

where

$$\tilde{w}_i(\mathbf{x}) := \frac{\mathbf{n}_{i-1} \times \mathbf{n}_i}{h_{i-1}(\mathbf{x})h_i(\mathbf{x})}, \tag{8}$$

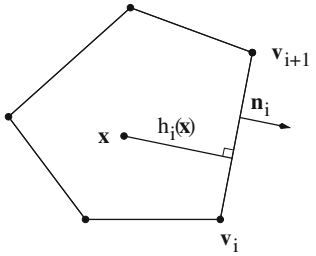


Fig. 4 Perpendicular distances

and

$$\mathbf{x}_1 \times \mathbf{x}_2 := \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}.$$

for $\mathbf{x}_k = (x_k, y_k)$. To see this, observe that with $L_j = |\mathbf{v}_{j+1} - \mathbf{v}_j|$ (and $|\cdot|$ the Euclidean norm) and β_i the interior angle of the polygon at $\mathbf{v}_i$,

$$A(\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}) = \frac{1}{2} \sin \beta_i L_{i-1} L_i,$$

and

$$A(\mathbf{x}, \mathbf{v}_{i-1}, \mathbf{v}_i) = \frac{1}{2} h_{i-1}(\mathbf{x}) L_{i-1}, \qquad A(\mathbf{x}, \mathbf{v}_i, \mathbf{v}_{i+1}) = \frac{1}{2} h_i(\mathbf{x}) L_i,$$

so that

$$w_i(\mathbf{x}) = 2\tilde{w}_i(\mathbf{x}).$$

3.3 Gradients

The gradient of a Wachspress coordinate can be found quite easily from the perpendicular form (7 and 8). Since $\nabla h_i(\mathbf{x}) = -\mathbf{n}_i$, the gradient of $\tilde{w}_i$ is [6]

$$\nabla \tilde{w}_i(\mathbf{x}) = \tilde{w}_i(\mathbf{x}) \left(\frac{\mathbf{n}_{i-1}}{h_{i-1}(\mathbf{x})} + \frac{\mathbf{n}_i}{h_i(\mathbf{x})} \right). \tag{9}$$

Thus the (vector-valued) ratio $\mathbf{R}_i := \nabla \tilde{w}_i / \tilde{w}_i$ is simply

$$\mathbf{R}_i(\mathbf{x}) = \frac{\mathbf{n}_{i-1}}{h_{i-1}(\mathbf{x})} + \frac{\mathbf{n}_i}{h_i(\mathbf{x})}.$$

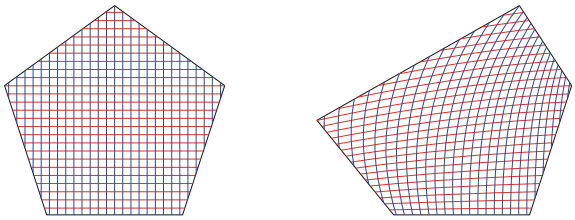


Fig. 5 Barycentric mapping

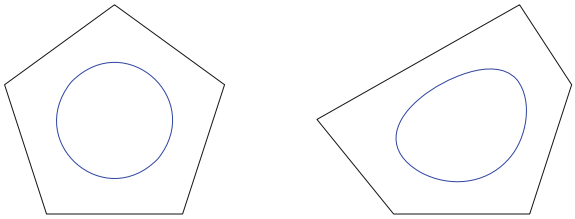


Fig. 6 Curve deformation

Using the formula [6]

$$\nabla\phi_i = \phi_i(\mathbf{R}_i - \sum_{j=1}^{n} \phi_j \mathbf{R}_j) \tag{10}$$

for any function ϕ_i of the form (7), we thus obtain $\nabla\phi_i(\mathbf{x})$ for $\mathbf{x} \in P$.

3.4 Curve Deformation

While Wachspress's motivation for these coordinates was finite element methods over polygonal partitions, Warren suggested their use in deforming curves. The coordinates can be used to define a barycentric mapping of one polygon to another, and such a mapping will then map, or deform, a curve embedded in the first polygon into a new one, with the vertices of the polygon acting as control points, with an effect similar to those of Bézier and spline curves and surfaces.

Assuming the second polygon is P' with vertices $\mathbf{v}'_1, \ldots, \mathbf{v}'_n$, the barycentric mapping $\mathbf{g} : P \to P'$ is defined as follows. Given $\mathbf{x} \in P$,

1. express $\mathbf{x}$ in Wachspress coordinates, $\mathbf{x} = \sum_{i=1}^{n} \phi_i(\mathbf{x})\mathbf{v}_i$,
2. set $\mathbf{g}(\mathbf{x}) = \sum_{i=1}^{n} \phi_i(\mathbf{x})\mathbf{v}'_i$.

Figure 5 shows such a mapping. Figure 6 shows the effect of using the mapping to deform a curve (a circle in this case).

It is now known that Wachspress mappings between convex polygons are always injective; as shown in [9]. The basic idea of the proof is to show that $\mathbf{g}$ has a positive

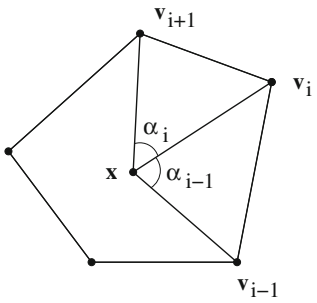


Fig. 7 Notation for mean value coordinates

Jacobian determinant $J(\mathbf{g})$. To do this one first shows that $J(\mathbf{g})$ can be expressed as

$$J(\mathbf{g}) = 2 \sum_{1 \le i < j < k \le n} \begin{vmatrix} \phi_i & \phi_j & \phi_k \\ \partial_1 \phi_i & \partial_1 \phi_j & \partial_1 \phi_k \\ \partial_2 \phi_i & \partial_2 \phi_j & \partial_2 \phi_k \end{vmatrix} A(\mathbf{v}'_i, \mathbf{v}'_j, \mathbf{v}'_k).$$

By the convexity of P', the signed areas $A(\mathbf{v}'_i, \mathbf{v}'_j, \mathbf{v}'_k)$ in the sum are all positive, and so $J(\mathbf{g}) > 0$ if all the 3×3 determinants in the sum are positive, and this turns out to be the case for Wachspress coordinates ϕ_i.

4 Mean Value Coordinates

As we have seen, Wachspress coordinates are relatively simple functions, and lead to well-behaved barycentric mappings. They are, however, limited to convex polygons. For a nonconvex polygon they are not well-defined, since the denominator in the rational expression becomes zero at certain points in the polygon. An alternative set of coordinates for convex polygons is the mean value coordinates [4], which have a simple generalization to nonconvex polygons, though positivity is in general lost. Suppose initially that P is convex as before, then the mean value (MV) coordinates are defined by (4) and

$$w_i(\mathbf{x}) = \frac{\tan(\alpha_{i-1}/2) + \tan(\alpha_i/2)}{|\mathbf{v}_i - \mathbf{x}|}, \tag{11}$$

with the angles $\alpha_j = \alpha_j(\mathbf{x})$, with $0 < \alpha_j < \pi$, as shown in Fig. 7. To show that these coordinates are barycentric, it is sufficient, as in the Wachspress case, to show that the w_i in (11) satisfy (5). This can be done in four steps:

1. Express the unit vectors $\mathbf{e}_i := (\mathbf{v}_i - \mathbf{x})/|\mathbf{v}_i - \mathbf{x}|$ in polar coordinates:

$$\mathbf{e}_i = (\cos\theta_i, \sin\theta_i),$$

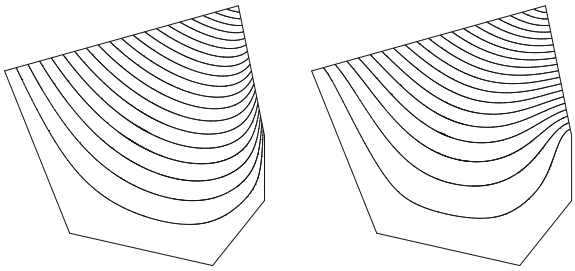


Fig. 8 Wachspress (*left*). Mean value (*right*)

and note that $\alpha_i = \theta_{i+1} - \theta_i$.

2. Use the fact that the integral of the unit normals $\mathbf{n}(\theta) = (\cos\theta, \sin\theta)$ on a circle is zero:

$$\int_0^{2\pi} \mathbf{n}(\theta)\, d\theta = 0.$$

3. Split this integral according to the θ_i:

$$\int_0^{2\pi} \mathbf{n}(\theta)\, d\theta = \sum_{i=1}^{n} \int_{\theta_i}^{\theta_{i+1}} \mathbf{n}(\theta)\, d\theta. \tag{12}$$

4. Show by trigonometry that

$$\int_{\theta_i}^{\theta_{i+1}} \mathbf{n}(\theta)\, d\theta = \frac{1-\cos\alpha_i}{\sin\alpha_i}(\mathbf{e}_i + \mathbf{e}_{i+1}) = \tan(\alpha_i/2)(\mathbf{e}_i + \mathbf{e}_{i+1}).$$

Substituting this into the sum in (12) and rearranging gives (5).

We can compute $\tan(\alpha_i/2)$ from the formulas

$$\cos\alpha_i = \mathbf{e}_i \cdot \mathbf{e}_{i+1}, \qquad \sin\alpha_i = \mathbf{e}_i \times \mathbf{e}_{i+1}. \tag{13}$$

Figure 8 compares the contour lines of a Wachspress coordinate, on the left, with the corresponding MV coordinate, on the right.

4.1 Gradients

Similar to the Wachspress case, the gradient $\nabla\phi_i$ of the MV coordinate ϕ_i can be computed from the formula (10) if we can find the ratio $\mathbf{R}_i := \nabla w_i / w_i$, with w_i in

(11). Let $r_i = |\mathbf{v}_i - \mathbf{x}|$ and $t_i = \tan(\alpha_i/2)$ so that

$$w_i = \frac{t_{i-1} + t_i}{r_i}.$$

Further, define

$$\mathbf{c}_i = \frac{\mathbf{e}_i}{r_i} - \frac{\mathbf{e}_{i+1}}{r_{i+1}},$$

and for a vector $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2$, let $\mathbf{a}^\perp := (-a_2, a_1)$.

Theorem 1 *For the MV coordinates,*

$$\mathbf{R}_i = \left(\frac{t_{i-1}}{t_{i-1} + t_i}\right) \frac{\mathbf{c}_{i-1}^\perp}{\sin\alpha_{i-1}} + \left(\frac{t_i}{t_{i-1} + t_i}\right) \frac{\mathbf{c}_i^\perp}{\sin\alpha_i} + \frac{\mathbf{e}_i}{r_i}.$$

We will show this using two lemmas.

Lemma 1 *For* $\mathbf{u} \in \mathbb{R}^2$, *let* $\mathbf{e} = (e_1, e_2) = (\mathbf{u} - \mathbf{x})/|\mathbf{u} - \mathbf{x}|$ *and* $r = |\mathbf{u} - \mathbf{x}|$. *Then*

$$\nabla e_1 = \frac{e_2 \mathbf{e}^\perp}{r}, \qquad \nabla e_2 = -\frac{e_1 \mathbf{e}^\perp}{r}.$$

Proof If $\mathbf{d} = (d_1, d_2) = \mathbf{u} - \mathbf{x}$, then using the fact that

$$\nabla d_1 = (-1, 0), \quad \nabla d_2 = (0, -1), \quad \text{and} \quad \nabla r = -\mathbf{d}/r,$$

the result follows from the quotient rule:

$$\nabla e_k = \nabla\left(\frac{d_k}{r}\right) = \frac{r\nabla d_k - d_k \nabla r}{r^2}, \qquad k = 1, 2. \qquad \square$$

Lemma 2 *Suppose* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, *and let*

$$\mathbf{e} = (\mathbf{u} - \mathbf{x})/|\mathbf{u} - \mathbf{x}|, \qquad r = |\mathbf{u} - \mathbf{x}|,$$
$$\mathbf{f} = (\mathbf{v} - \mathbf{x})/|\mathbf{v} - \mathbf{x}|, \qquad s = |\mathbf{v} - \mathbf{x}|.$$

Then

$$\nabla(\mathbf{e} \cdot \mathbf{f}) = -(\mathbf{e} \times \mathbf{f})\mathbf{c}^\perp \quad \textit{and} \quad \nabla(\mathbf{e} \times \mathbf{f}) = (\mathbf{e} \cdot \mathbf{f})\mathbf{c}^\perp,$$

where

$$\mathbf{c} = \frac{\mathbf{e}}{r} - \frac{\mathbf{f}}{s}.$$

Proof With $\mathbf{e} = (e_1, e_2)$ and $\mathbf{f} = (f_1, f_2)$,

$$\nabla(\mathbf{e} \cdot \mathbf{f}) = f_1 \nabla e_1 + e_1 \nabla f_1 + f_2 \nabla e_2 + e_2 \nabla f_2,$$
$$\nabla(\mathbf{e} \times \mathbf{f}) = f_2 \nabla e_1 + e_1 \nabla f_2 - f_1 \nabla e_2 - e_2 \nabla f_1,$$

and applying Lemma 1 to ∇e_k and ∇f_k, $k = 1, 2$, gives the result. □

We now prove Theorem 1. Recalling (13), Lemma 2 shows that

$$\nabla(\cos\alpha_i) = -(\sin\alpha_i)\mathbf{c}_i^{\perp}, \qquad \nabla(\sin\alpha_i) = (\cos\alpha_i)\mathbf{c}_i^{\perp}. \tag{14}$$

From this it follows that

$$\nabla t_i = \frac{t_i}{\sin\alpha_i}\mathbf{c}_i^{\perp}.$$

Since, $\nabla r_i = -\mathbf{e}_i$, this means that

$$\nabla\left(\frac{t_j}{r_i}\right) = \frac{t_j}{r_i}\left(\frac{\mathbf{c}_j^{\perp}}{\sin\alpha_i} + \frac{\mathbf{e}_i}{r_i}\right), \qquad j = i-1, i.$$

Therefore,

$$\nabla w_i = \frac{t_{i-1}}{r_i}\left(\frac{\mathbf{c}_{i-1}^{\perp}}{\sin\alpha_{i-1}}\right) + \frac{t_i}{r_i}\left(\frac{\mathbf{c}_i^{\perp}}{\sin\alpha_i}\right) + w_i\frac{\mathbf{e}_i}{r_i},$$

which, after dividing by w_i, proves Theorem 1.

Incidentally, though we did not use it, we note that both equations in (14) imply that

$$\nabla\alpha_i = \mathbf{c}_i^{\perp}.$$

Another derivative formula for MV coordinates can be found in [28].

4.2 Alternative Formula

We saw that Wachspress coordinates can be expressed in the "global form" (6) in which $\phi_i(\mathbf{x})$ is well-defined for $\mathbf{x} \in \partial P$ as well as for $\mathbf{x} \in P$. It turns out that MV coordinates also have a global form with the same property, though for large n, the resulting expression requires more computation, and involves more square roots, than the local form based on (11). Let $\mathbf{d}_i = \mathbf{v}_i - \mathbf{x}$, $i = 1, \ldots, n$.

Theorem 2 *The MV coordinates in (4) can be expressed as*

$$\phi_i(\mathbf{x}) = \frac{\hat{w}_i(\mathbf{x})}{\sum_{j=1}^{n}\hat{w}_j(\mathbf{x})}, \tag{15}$$

where

$$\hat{w}_i = (r_{i-1}r_{i+1} - \mathbf{d}_{i-1} \cdot \mathbf{d}_{i+1})^{1/2} \prod_{j \neq i-1,i} (r_j r_{j+1} + \mathbf{d}_j \cdot \mathbf{d}_{j+1})^{1/2}. \qquad (16)$$

Proof From the addition formula for sines, we have

$$w_i = \frac{1}{r_i}\left(\frac{\sin(\alpha_{i-1}/2)}{\cos(\alpha_{i-1}/2)} + \frac{\sin(\alpha_i/2)}{\cos(\alpha_i/2)}\right) = \frac{\sin((\alpha_{i-1}+\alpha_i)/2)}{r_i \cos(\alpha_{i-1}/2)\cos(\alpha_i/2)}.$$

Then, to get rid of the half-angles we use the identities

$$\sin(A/2) = \sqrt{(1-\cos A)/2},$$
$$\cos(A/2) = \sqrt{(1+\cos A)/2},$$

to obtain

$$w_i = \frac{1}{r_i}\left(\frac{2(1-\cos(\alpha_{i-1}+\alpha_i))}{(1+\cos\alpha_{i-1})(1+\cos\alpha_i)}\right)^{1/2}.$$

Now we substitute in the scalar product formula,

$$\cos(\alpha_{i-1}+\alpha_i) = \frac{\mathbf{d}_{i-1}\cdot\mathbf{d}_{i+1}}{r_{i-1}r_{i+1}},$$

and similarly for $\cos\alpha_{i-1}$ and $\cos\alpha_i$, and the $1/r_i$ term cancels out:

$$w_i = \left(\frac{2(r_{i-1}r_{i+1} - \mathbf{d}_{i-1}\cdot\mathbf{d}_{i+1})}{(r_{i-1}r_i + \mathbf{d}_{i-1}\cdot\mathbf{d}_i)(r_i r_{i+1} + \mathbf{d}_i\cdot\mathbf{d}_{i+1})}\right)^{1/2},$$

which gives 15 and 16. □

One can easily check that this formula gives the correct values (2) for $\mathbf{x} \in \partial P$.

4.3 Star-Shaped Polygons

The original motivation for these coordinates was for parameterizing triangular meshes [3, 5, 29]. In this application, the point $\mathbf{x}$ is a vertex in a planar triangulation, with $\mathbf{v}_1, \ldots, \mathbf{v}_n$ its neighbouring vertices. Thus, in this case, the polygon P (with vertices $\mathbf{v}_1, \ldots, \mathbf{v}_n$) is not necessarily convex, but always star-shaped, with $\mathbf{x}$ a point in its kernel, i.e., every vertex $\mathbf{v}_i$ is "visible" from $\mathbf{x}$; see Fig. 9. In this case the angles α_i in (11) are again positive, and the weight $w_i(\mathbf{x})$ is again positive. Thus the MV coordinates of $\mathbf{x}$ remain positive in this star-shaped case. The advantage ofthis is

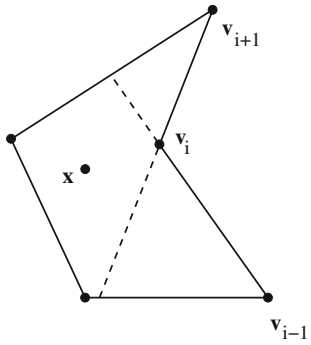


Fig. 9 A star-shaped polygon and its kernel

that when these coordinates are applied to the parameterization of triangular meshes, the piecewise linear mapping is guaranteed to be injective, i.e., none of the triangles "fold over," when the boundary of the mesh is mapped to a convex polygon.

4.4 Arbitrary Polygons

It was later observed, in [13], that the coordinates are still well-defined, though not necessarily positive, when P is an arbitrary polygon, provided that the angles α_i are treated as signed angles: i.e., we take α_i in (11) to have the same sign as $\mathbf{e}_i \times \mathbf{e}_{i+1}$, which will be the case if we use the formulas (13). The reason for this is that even though $w_i(\mathbf{x})$ in (11) may be negative for some i, when P is arbitrary, the sum $\sum_{i=1}^{n} w_i(\mathbf{x})$ is nevertheless positive for any $\mathbf{x}$ in P. This was shown in [13], where it was also shown that these more general MV coordinates have the Lagrange and piecewise linearity properties on ∂P.

This generalization of MV coordinates allows the curve deformation method to be extended to arbitrary polygons. It was further observed in [13] that MV coordinates even have a natural generalization to any set of polygons, as long as the polygons do not intersect one another. The polygons may or may not be nested. These generalized MV coordinates were applied to image warping in [13].

5 Polygonal Finite Elements

There has been steadily growing interest in using generalized barycentric coordinates for finite element methods on polygonal (and polyhedral) meshes [6, 11, 23, 26, 27, 34]. In order to establish the convergence of the finite element method, one would need to derive a bound on the gradients of the coordinates in terms of the geometry

of the polygon P. Various bounds on

$$\sup_{\mathbf{x}\in P} |\nabla \phi_i(\mathbf{x})|$$

were derived in [11] for Wachspress (and other) coordinates, and in [23] for MV coordinates. For the Wachspress coordinates, a simpler bound was derived in [6]. If we define, for $\mathbf{x} \in P$,

$$\lambda(\mathbf{x}) := \sum_{i=1}^{n} |\nabla \phi_i(\mathbf{x})|, \tag{17}$$

then λ plays a role similar to the Lebesgue function in the theory of polynomial interpolation because for g in (3),

$$|\nabla g(\mathbf{x})| \le \sum_{i=1}^{n} |\nabla \phi_i(\mathbf{x})||f(\mathbf{v}_i)| \le \lambda(\mathbf{x}) \max_{i=1,\dots,n} |f(\mathbf{v}_i)|.$$

It was shown in [6] that with

$$\Lambda := \sup_{\mathbf{x}\in P} \lambda(\mathbf{x}) \tag{18}$$

the corresponding 'Lebesgue constant', and with ϕ_i the Wachspress coordinates,

$$\Lambda \le \frac{4}{h_*},$$

where

$$h_* = \min_{i=1,\dots,n} \min_{j\neq i,i+1} h_i(\mathbf{v}_j).$$

6 Curved Domains

Consider again the barycentric interpolant g in (3). Since g is piecewise linear on the boundary ∂P, it interpolates f on ∂P if f itself is piecewise linear on ∂P. Warren et al. [33] proposed a method of interpolating any continuous function f defined on the boundary of any convex domain, by, roughly speaking, taking a continuous "limit" of the polygonal interpolants g in (3). Specifically, suppose that the boundary of some convex domain $P \subset \mathbb{R}^2$ is represented as a closed, parametric curve $\mathbf{c} : [a, b] \to \mathbb{R}^2$, with $\mathbf{c}(b) = \mathbf{c}(a)$. Then any sequence of parameter values, $t_1, \dots, t_n$, with $a \le t_1 < t_2 < \cdots < t_n < b$, with mesh size $h = \max_i (t_{i+1} - t_i)$, defines a convex polygon P_h with vertices $\mathbf{v}_i = \mathbf{c}(t_i)$; see Fig. 10. The barycentric interpolant g in (3) with respect to this polygon is then

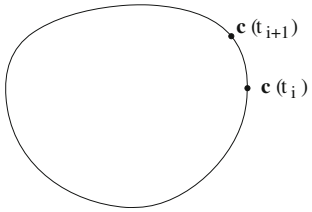


Fig. 10 From polygons to curved domains

$$g_h(\mathbf{x}) = \sum_{i=1}^{n} \phi_i(\mathbf{x}) f(\mathbf{c}(t_i)). \tag{19}$$

Taking the limit $g = \lim_{h\to 0} g_h$ over a sequence of such polygons, and letting the ϕ_i be the Wachspress coordinates, gives

$$g(\mathbf{x}) = \int_a^b w(\mathbf{x}, t) f(\mathbf{c}(t))\, dt \Big/ \int_a^b w(\mathbf{x}, t)\, dt, \qquad \mathbf{x} \in P, \tag{20}$$

where

$$w(\mathbf{x}, t) = \frac{(\mathbf{c}'(t) \times \mathbf{c}''(t))}{((\mathbf{c}(t) - \mathbf{x}) \times \mathbf{c}'(t))^2}.$$

It was shown in [33] that the barycentric property also holds for this g: if $f : \mathbb{R}^2 \to \mathbb{R}$ is linear, i.e., $f(\mathbf{x}) = ax + by + c$, then $g = f$. However, it also follows from the fact that if f is linear, $g_h = f$ for all h.

There is an analogous continuous MV interpolant, with g also given by (20), but with the weight function $w(\mathbf{x}, t)$ replaced by

$$w(\mathbf{x}, t) = \frac{(\mathbf{c}(t) - \mathbf{x}) \times \mathbf{c}'(t)}{|\mathbf{c}(t) - \mathbf{x}|^3}. \tag{21}$$

One can also derive the barycentric property of this continuous interpolant by applying the unit circle construction of Sect. 4 directly to the curved domain P. Figure 11 shows the MV interpolant to the function $\cos(2\theta)$, $0 \le \theta < 2\pi$, on the boundary of the unit circle.

Similar to the generalization of MV coordinates to nonconvex polygons, the continuous MV interpolant also extends to arbitrarily shaped curve domains: one simply applies the same formula (21). Even though the cross product,

$$(\mathbf{c}(t) - \mathbf{x}) \times \mathbf{c}'(t)$$

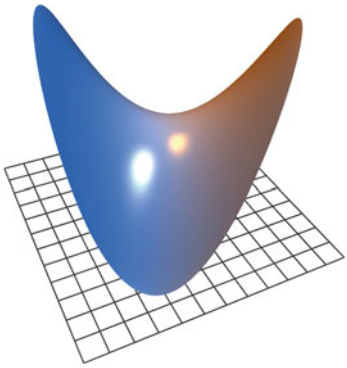


Fig. 11 An MV interpolant on a *circle*

may be negative for some values of t, the integral $\int_a^b w(\mathbf{x}, t)\, dt$ of w in (21) remains positive [2].

6.1 Hermite Interpolation

If the normal derivative of f is also known on the boundary of the domain, we could consider matching both the values and normal derivatives of f. In [2, 10] two distinct approaches were used to construct such a Hermite interpolant, both based on the construction of MV interpolants. To motivate this, let π_n denote the linear space of polynomials of degree $\leq n$ in one real variable. Suppose that $f : [0, 1] \to \mathbb{R}$ has a first derivative at $x = 0$ and $x = 1$. Then there is a unique cubic polynomial, $p \in \pi_3$, such that

$$p^{(k)}(i) = f^{(k)}(i), \qquad i = 0, 1, \quad k = 0, 1.$$

There are various ways of expressing p. One is as

$$p = l_0(x) + \omega(x) l_1(x),$$

where

$$l_0(x) = (1 - x) f(0) + x f(1), \quad \omega(x) = x(1 - x), \quad l_1(x) = (1 - x) m_0 + x m_1,$$

and

$$m_0 = f'(0) - (f(1) - f(0)), \qquad m_1 = (f(1) - f(0)) - f'(1).$$

The basic idea of the Hermite interpolant in [2] is to generalize this construction to a general planar domain, replacing the linear interpolants l_0 and l_1 by MV interpolants, and replacing the weight function ω by an MV "weight" function. This gives a Hermite interpolant in 2D, but it does not in general have cubic precision. Another way of expressing p above is as the minimizer of a functional. For a fixed $x \in (0, 1)$, $p(x)$ is the value $s(x)$ of the spline s that minimizes the functional

$$E(s) = \int_0^1 (s''(y))^2 \, dy,$$

in the spline space

$$S = \{s \in C^1[0, 1] : s|_{[0,x]}, s|_{[x,1]} \in \pi_3\},$$

subject to the boundary conditions

$$s^{(k)}(i) = f^{(k)}(i), \qquad i = 0, 1, \quad k = 0, 1.$$

A generalization of this minimization was used in [10] to generate a function on a curved domain that appears, numerically, to interpolate the boundary data, but a mathematical proof of this is still missing. The cubic construction in [10] was recently derived independently through certain mean value properties of biharmonic functions by Li et al. [19]. They also give a closed-form expression for the coordinates on a polygonal domain when a suitable definition of the boundary data is used along the edges.

7 Coordinates in Higher Dimensions

So far we have only considered coordinates for points in $\mathbb{R}^2$, but there are applications of barycentric coordinates for points in a polyhedron in $\mathbb{R}^3$, such as in Fig. 12, or more generally for points in a polytope in $\mathbb{R}^d$. Both Wachspress and MV coordinates have been generalized to higher dimensions.

7.1 Wachspress Coordinates in 3D

Warren [32] generalized the coordinates of Wachspress to simple convex polyhedra: convex polyhedra in which all vertices have three incident faces. In [33], Warren et al. derived the same coordinates in a different way (avoiding the so-called "adjoint"), generalizing (7) as follows. Let $P \subset \mathbb{R}^3$ be a simple convex polyhedron, with faces

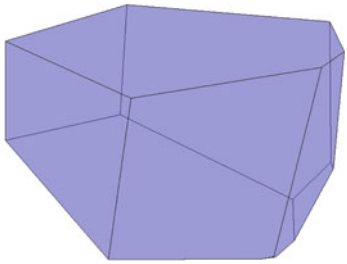


Fig. 12 Simple, convex polyhedron

F and vertices V. For each face $f \in F$, let $\mathbf{n}_f \in \mathbb{R}^3$ denote its unit outward normal, and for any $\mathbf{x} \in P$, let $h_f(\mathbf{x})$ denote the perpendicular distance of $\mathbf{x}$ to f, which can be expressed as the scalar product

$$h_f(\mathbf{x}) = (\mathbf{v} - \mathbf{x}) \cdot \mathbf{n}_f,$$

for any vertex $\mathbf{v} \in V$ belonging to f. For each vertex $\mathbf{v} \in V$, let f_1, f_2, f_3 be the three faces incident to $\mathbf{v}$, and for $\mathbf{x} \in P$, let

$$w_{\mathbf{v}}(\mathbf{x}) = \frac{\det(\mathbf{n}_{f_1}, \mathbf{n}_{f_2}, \mathbf{n}_{f_3})}{h_{f_1}(\mathbf{x}) h_{f_2}(\mathbf{x}) h_{f_3}(\mathbf{x})}, \tag{22}$$

where it is understood that f_1, f_2, f_3 are ordered such that the determinant in the numerator is positive. Here, for vectors $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$,

$$\det(\mathbf{a}, \mathbf{b}, \mathbf{c}) := \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}.$$

Thus the ordering of f_1, f_2, f_3 must be anticlockwise around $\mathbf{v}$, seen from outside P. In this way, $w_{\mathbf{v}}(\mathbf{x}) > 0$, and it was shown in [33] that the functions

$$\phi_{\mathbf{v}}(\mathbf{x}) := \frac{w_{\mathbf{v}}(\mathbf{x})}{\sum_{\mathbf{u} \in V} w_{\mathbf{u}}(\mathbf{x})} \tag{23}$$

are barycentric coordinates for $\mathbf{x} \in P$ in the sense that

$$\sum_{\mathbf{v} \in V} \phi_{\mathbf{v}}(\mathbf{x}) = 1, \qquad \sum_{\mathbf{v} \in V} \phi_{\mathbf{v}}(\mathbf{x}) \mathbf{v} = \mathbf{x}. \tag{24}$$

To deal with nonsimple polyhedra, it was suggested in [33] that one might decompose a nonsimple vertex into simple ones by perturbing its adjacent facets. Later, Ju et al. [15] found a cleaner solution, using properties of the so-called *polar dual*. With

respect to each $\mathbf{x}$ in a general convex polyhedron $P \subset \mathbb{R}^3$, there is a dual polyhedron,

$$\tilde{P}_{\mathbf{x}} := \{\mathbf{y} \in \mathbb{R}^3 : \mathbf{y} \cdot (\mathbf{z} - \mathbf{x}) \leq 1, \mathbf{z} \in P\}.$$

It contains the origin $\mathbf{y} = 0$, and its vertices are the endpoints of the vectors

$$\mathbf{p}_f(\mathbf{x}) := \frac{\mathbf{n}_f}{h_f(\mathbf{x})}, \qquad f \in F,$$

when placed at the origin. Suppose that a vertex $\mathbf{v} \in V$ has k incident faces, $f_1, \ldots, f_k$, for some $k \geq 3$, where we again assume they are ordered in some anticlockwise fashion around $\mathbf{v}$, as seen from outside P. The endpoints of the k vectors $\mathbf{p}_{f_1}(\mathbf{x}), \ldots \mathbf{p}_{f_k}(\mathbf{x})$ form a k-sided polygon. This polygon is the face of $\tilde{P}_{\mathbf{x}}$, dual to the vertex $\mathbf{v}$ of P. This face and the origin in $\mathbb{R}^3$ form a polygonal pyramid, $Q_{\mathbf{v}} \subset \tilde{P}_{\mathbf{x}}$. It was shown in [15] that if we define

$$w_{\mathbf{v}}(\mathbf{x}) = \mathrm{vol}(Q_{\mathbf{v}}),$$

then the functions $\phi_{\mathbf{v}}$ in (23) are again barycentric coordinates. In practice, we could triangulate the face dual to $\mathbf{v}$ by connecting the endpoint of $\mathbf{p}_{f_1}(\mathbf{x})$ to the endpoints of all the other $\mathbf{p}_{f_i}(\mathbf{x})$, and so compute $\mathrm{vol}(Q_{\mathbf{v}})$ as a sum of volumes of tetrahedra. Thus, we could let

$$w_{\mathbf{v}}(\mathbf{x}) = \sum_{i=2}^{k-1} \det(\mathbf{p}_{f_1}(\mathbf{x}), \mathbf{p}_{f_i}(\mathbf{x}), \mathbf{p}_{f_{i+1}}(\mathbf{x})). \tag{25}$$

Some matlab code for evaluating these coordinates and their gradients can be found in [6].

7.2 MV Coordinates in 3D

MV coordinates were generalized to three dimensions in [8, 16], the basic idea being to replace integration over the unit circle, as in Sect. 4, by integration over the unit sphere.

Consider first the case that $P \subset \mathbb{R}^3$ is a convex polyhedron with triangular faces (though it does not need to be simple). Fix $\mathbf{x} \in P$ and consider the radial projection of the boundary of P onto the unit sphere centered at $\mathbf{x}$. A vertex $\mathbf{v} \in V$ is projected to the point (unit vector) $\mathbf{e}_{\mathbf{v}} := (\mathbf{v} - \mathbf{x})/|\mathbf{v} - \mathbf{x}|$. A face $f \in F$ is projected to a spherical triangle $f_{\mathbf{x}}$ whose vertices are $\mathbf{e}_{\mathbf{v}}$, $\mathbf{v} \in V_f$, where $V_f \subset V$ denotes the set of (three) vertices of f. Let $\mathbf{I}_f$ denote the (vector-valued) integral of its unit normals,

$$\mathbf{I}_f := \int_{f_{\mathbf{x}}} \mathbf{n}(\mathbf{y})\, d\mathbf{y}.$$

Since the three vectors $\mathbf{e}_{\mathbf{v}}$, $\mathbf{v} \in V_f$, are linearly independent, there are three unique weights $w_{\mathbf{v},f} > 0$ such that

$$\mathbf{I}_f = \sum_{\mathbf{v}\in V_f} w_{\mathbf{v},f}\mathbf{e}_{\mathbf{v}}. \tag{26}$$

The weights can be found as ratios of 3×3 determinants from Cramer's rule. Since the integral of all unit normals of the unit sphere is zero, and letting $F_{\mathbf{v}} \subset F$ denote the set of faces that are incident on the vertex $\mathbf{v}$, we find, by switching summations, that

$$0 = \sum_{f\in F} \mathbf{I}_f = \sum_{f\in F} \sum_{\mathbf{v}\in V_f} w_{\mathbf{v},f}\mathbf{e}_{\mathbf{v}} = \sum_{\mathbf{v}\in V} \sum_{f\in F_{\mathbf{v}}} w_{\mathbf{v},f}\mathbf{e}_{\mathbf{v}},$$

and so the functions

$$w_{\mathbf{v}} := \sum_{f\in F_{\mathbf{v}}} \frac{w_{\mathbf{v},f}}{|\mathbf{v}-\mathbf{x}|}, \tag{27}$$

satisfy

$$\sum_{\mathbf{v}\in V} w_{\mathbf{v}}(\mathbf{x})(\mathbf{v}-\mathbf{x}) = 0.$$

It follows that the functions $\phi_{\mathbf{v}}$ given by (23) with $w_{\mathbf{v}}$ given by (27) are barycentric coordinates, i.e., they are positive in P and satisfy (24).

It remains to find the integral $\mathbf{I}_f$ in terms of the points $\mathbf{v} \in V_f$ and $\mathbf{x}$. We follow the observation made in [8]. The spherical triangle $f_{\mathbf{x}}$ and the point $\mathbf{x}$ form a wedge of the solid unit sphere centered at $\mathbf{x}$. Since the integral of all unit normals over this wedge is zero, the integral $\mathbf{I}_f$ is minus the sum of the integrals over the three planar faces of the wedge. Suppose $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are the vertices of f in anticlockwise order, and let $\mathbf{e}_i = \mathbf{e}_{\mathbf{v}_i}$. For $i = 1, 2, 3$, the ith side of the wedge is the sector of the unit circle formed by the two unit vectors $\mathbf{e}_i$ and $\mathbf{e}_{i+1}$, with the cyclic notation $\mathbf{v}_{i+3} := \mathbf{v}_i$. If $\beta_i \in (0, \pi)$ is the angle between $\mathbf{e}_i$ and $\mathbf{e}_{i+1}$ then the area of the sector is $\beta_i/2$, and hence

$$\mathbf{I}_f = \frac{1}{2}\sum_{i=1}^{3} \beta_i \mathbf{m}_i, \tag{28}$$

where

$$\mathbf{m}_i = \frac{\mathbf{e}_i \times \mathbf{e}_{i+1}}{|\mathbf{e}_i \times \mathbf{e}_{i+1}|}.$$

Equating this with (26) gives

$$w_{\mathbf{v}_i,f} = \frac{1}{2}\sum_{j=1}^{3} \beta_j \frac{\mathbf{m}_j \cdot \mathbf{m}_{i+1}}{\mathbf{e}_i \cdot \mathbf{m}_{i+1}}.$$

These 3D MV coordinates were used for surface deformation in [16] when the surface is represented as a dense triangular mesh. Some contour plots of the coordinate functions can be found in [8].

For a polyhedron with faces having arbitrary numbers of vertices, the same approach can be applied, but there is no longer uniqueness. Suppose $f \in F$ is a face with $k \geq 3$ vertices. The integral $\mathbf{I}_f$ is again well-defined, and can be computed as the sum of k terms, generalizing (28). However, there is no unique choice of the local weights $w_{\mathbf{v},f}$ in (26) for $k > 3$, since there are k of these. Langer et al. [17] proposed using a certain type of spherical polygonal MV coordinates to determine the $w_{\mathbf{v},f}$, but other choices are possible.

8 Final Remarks

We have not covered here other kinds of generalized barycentric coordinates, and related coordinates, which include Sibson's natural neighbor coordinates [24], Sukumar's maximum entropy coordinates [25], Gordon and Wixom coordinates [12], spherical barycentric coordinates [17], harmonic coordinates [14], Green coordinates [21], Poisson coordinates [18], Positive MV coordinates [20] and others. A more general survey paper is being planned in which some of these other coordinates will be included.

References

1. Bruvoll, S., Floater, M.S.: Transfinite mean value interpolation in general dimension. J. Comp. Appl. Math. **233**, 1631–1639 (2010)
2. Dyken, C., Floater, M.S.: Transfinite mean value interpolation. Comp. Aided Geom. Des. **26**, 117–134 (2009)
3. Floater, M.S.: Parametrization and smooth approximation of surface triangulations. Comp. Aided Geom. Des. **14**, 231–250 (1997)
4. Floater, M.S.: Mean value coordinates. Comp. Aided Geom. Des. **20**, 19–27 (2003)
5. Floater, M.S.: One-to-one piecewise linear mappings over triangulations. Math. Comp. **72**, 685–696 (2003)
6. Floater, M.S., Gillette, A., Sukumar, N.: Gradient bounds for Wachspress coordinates on polytopes. SIAM J. Numer. Anal. **52**, 515–532 (2014)
7. Floater, M.S., Hormann, K., Kós, G.: A general construction of barycentric coordinates over convex polygons. Adv. Comp. Math. **24**, 311–331 (2006)
8. Floater, M.S., Kos, G., Reimers, M.: Mean value coordinates in 3D. Comp. Aided Geom. Des. **22**, 623–631 (2005)

9. Floater, M., Kosinka, J.: On the injectivity of Wachspress and mean value mappings between convex polygons. Adv. Comp. Math. **32**, 163–174 (2010)
10. Floater, M., Schulz, C.: Pointwise radial minimization: Hermite interpolation on arbitrary domains. Comp. Graphics Forum (Proc. Symp. Geom. Process. 2008) **27**, 1505–1512 (2008)
11. Gillette, A., Rand, A., Bajaj, C.: Error estimates for generalized barycentric interpolation. Adv. Comp. Math. **37**, 417–439 (2012)
12. Gordon, W.J., Wixom, J.A.: Pseudo-harmonic interpolation on convex domains. SIAM J. Numer. Anal. **11**, 909–933 (1974)
13. Hormann, K., Floater, M.S.: Mean value coordinates for arbitrary planar polygons. ACM Trans. Graph. **25**, 1424–1441 (2006)
14. Joshi, P., Meyer, M., DeRose, T.., Green, B., Sanocki, T.: Harmonic coordinates for character articulation. ACM Trans. Graph. **26**, 71 (2007)
15. Ju, T., Schaefer, S., Warren, J., Desbrun, M.: A geometric construction of coordinates for convex polyhedra using polar duals. In: Desbrun, M., Pottman H. (eds.) Geometry Processing 2005, Eurographics Association 2005, pp. 181–186 (2005)
16. Ju, T., Schaefer, S., Warren, J.: Mean value coordinates for closed triangular meshes. ACM TOG **24**, 561–566 (2005)
17. Langer, T., Belyaev, A., Seidel, H.-P.: Spherical barycentric coordinates. In: Polthier, K., Sheffer A. (eds.) Eurographics Symposium on Geometry Processing, pp. 81–88 (2006)
18. Li, X.-Y., Hu, S.-M.: Poisson coordinates. IEEE Trans. Visual. Comput. Graphics **19**, 344–352 (2013)
19. Li, X.-Y., Ju, T., Hu, S.-M.: Cubic mean value coordinates. ACM Trans. Graph. **32**, 1–10 (2013)
20. Lipman, Y., Kopf, J., Cohen-Or, D., Levin, D.: GPU-assisted positive mean value coordinates for mesh deformation. In: Symposium on Geometry Processing, pp. 117–123 (2007)
21. Lipman, Y., Levin, D., Cohen-Or, D.: Green coordinates. ACM Trans. Graph. **27**, 1–10 (2008)
22. Meyer, M., Barr, A., Lee, H., Desbrun, M.: Generalized barycentric coordinates on irregular polygons. J. Graph. Tools **7**, 13–22 (2002)
23. Rand, A., Gillette, A., Bajaj, C.: Interpolation error estimates for mean value coordinates over convex polygons. Adv. Comp. Math. **39**, 327–347 (2013)
24. Sibson, R.: A brief description of natural neighbour interpolation. In: Barnett, V. (ed.) Interpreting Multivariate Data, pp. 21–36. John Wiley, Chichester (1981)
25. Sukumar, N.: Construction of polygonal interpolants: a maximum entropy approach. Int. J. Num. Meth. Eng. **61**, 2159–2181 (2004)
26. Sukumar, N., Tabarraei, A.: Conforming polygonal finite elements. Int. J. Num. Meth. Eng. **61**, 2045–2066 (2004)
27. Talischi, C., Paulino, G.H., Le, C.H.: Honeycomb Wachspress finite elements for structural topology optimization. Struct. Multidisc. Optim. **37**, 569–583 (2009)
28. Thiery, J.-M., Tierny, J., Boubekeur, T.: Jacobians and Hessians of mean value coordinates for closed triangular meshes. Vis. Comput. **29**, 217–229 (2013)
29. Tutte, W.T.: How to draw a graph. Proc. London Math. Soc. **13**, 743–768 (1963)
30. Wachspress, E.: A rational finite element basis. Academic Press, New York (1975)
31. Wachspress, E.L.: Barycentric coordinates for polytopes. Comput. Math. Appl. **61**, 3319–3321 (2011)
32. Warren, J.: Barycentric coordinates for convex polytopes. Adv. Comp. Math. **6**, 97–108 (1996)
33. Warren, J., Schaefer, S., Hirani, A., Desbrun, M.: Barycentric coordinates for convex sets. Adv. Comp. Math. **27**, 319–338 (2007)
34. Wicke, M., Botsch, M., Gross, M.: A finite element method on convex polyhedra. Proc. Eurograph. **07**, 355–364 (2007)

Hermite and Bernstein Style Basis Functions for Cubic Serendipity Spaces on Squares and Cubes

Andrew Gillette

Abstract We introduce new Hermite style and Bernstein style geometric decompositions of the cubic serendipity finite element spaces $\mathcal{S}_3(I^2)$ and $\mathcal{S}_3(I^3)$, as defined in the recent work of Arnold and Awanou [*Found. Comput. Math.* **11** (2011), 337–344]. The serendipity spaces are substantially smaller in dimension than the more commonly used bicubic and tricubic Hermite tensor product spaces—12 instead of 16 for the square and 32 instead of 64 for the cube—yet are still guaranteed to obtain cubic order a priori error estimates in H^1 norm when used in finite element methods. The basis functions we define have a canonical relationship both to the finite element degrees of freedom as well as to the geometry of their graphs; this means the bases may be suitable for applications employing *isogeometric analysis* where domain geometry and functions supported on the domain are described by the same basis functions. Moreover, the basis functions are linear combinations of the commonly used bicubic and tricubic polynomial Bernstein or Hermite basis functions, allowing their rapid incorporation into existing finite element codes.

Keywords Finite elements · Serendipity elements · Multivariate polynomial interpolation · Tensor product interpolation · Hermite interpolation

1 Introduction

Serendipity spaces offer a rigorous means to reduce the degrees of freedom associated to a finite element method while still ensuring optimal order convergence. The "serendipity" moniker came from the observation of this phenomenon among finite element practitioners before its mathematical justification was fully understood; see e.g., [6, 11, 12, 15]. Recent work by Arnold and Awanou [1, 2] classifies

A. Gillette(✉)
Department of Mathematics, University of Arizona, 617 N. Santa Rita Ave., PO Box 210089, Tucson, AZ 85721, USA
e-mail: agillette@math.arizona.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_7,

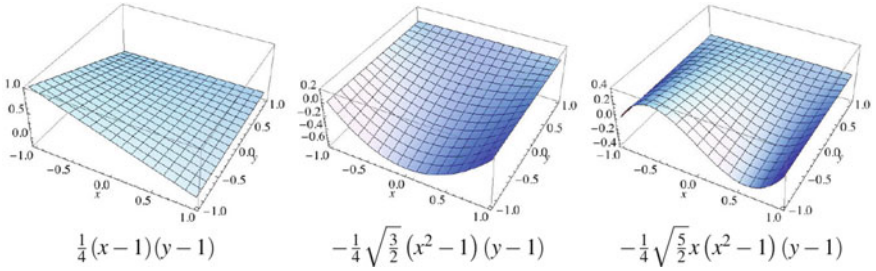


Fig. 1 Cubic serendipity functions on I^2 from [16]. The *left* function is associated to the vertex below the peak. The *middle* and *right* functions are associated to the edge $y = -1$ but do not correspond to the domain points $(\pm\frac{1}{3}, -1)$ in any canonical or symmetric fashion, making them less useful for geometric modeling or isogeometric analysis

serendipity spaces on cubical meshes in $n \geq 2$ dimensions by giving a simple and precise definition of a space of polynomials $\mathcal{S}_r(I^n)$ that must be spanned, as well as a unisolvent set of degrees of freedom for them. Crucially, the space $\mathcal{S}_r(I^n)$ contains all polynomials in n variables of total degree at most r, a property shared by the space of polynomials $\mathcal{Q}_r(I^n)$ spanned by the standard order r tensor product method. This property allows the derivation of an a priori error estimate for serendipity methods of the same order (with respect to the width of a mesh element) as their standard tensor product counterparts.

In this paper, we provide two coordinate-independent geometric decompositions for both $\mathcal{S}_3(I^2)$ and $\mathcal{S}_3(I^3)$, the cubic serendipity spaces in two and three dimensions, respectively. More precisely, we present sets of polynomial basis functions, prove that they provide a basis for the corresponding cubic serendipity space, and relate them canonically to the domain geometry. Each basis is designated as either Bernstein or Hermite style, as each function restricts to one of these common basis function types on each edge of the square or cube. The standard pictures for $\mathcal{S}_3(I^2)$ and $\mathcal{S}_3(I^3)$ serendipity elements, shown on the right of Figs. 2 and 4, have one dot for each vertex and two dots for each edge of the square or cube. We refer to these as **domain points** and will present a canonical relationship between the defined bases and the domain points.

To the author's knowledge, the only basis functions previously available for cubic serendipity finite element purposes employ Legendre polynomials, which lack a clear relationship to the domain points. Definitions of these basis functions can be found in Szabó and Babuška [16, Sect. 6.1 and 13.3]; the two functions from [16] associated to the edge $y = -1$ of I^2, are shown in Fig. 1 (middle and right). The restriction of these functions to the edge gives an even polynomial in one case and an odd polynomial in the other, forcing an *ad hoc* choice of how to associate the functions to the corresponding domain points $(\pm\frac{1}{3}, -1)$. The functions presented in this paper do have a natural correspondence to the domain points of the geometry.

Maintaining a concrete and canonical relationship between domain points and basis functions is an essential component of the growing field of *isogeometric*

analysis (IGA). One of the main goals of IGA is to employ basis functions that can be used both for geometry modeling and finite element analysis, exactly as we provide here for cubic serendipity spaces. Each function is a linear combination of bicubic or tricubic Bernstein or Hermite polynomials; the specific coefficients of the combination are given in the proofs of the theorems. This makes the incorporation of the functions into a variety of existing application contexts relatively easy. Note that tensor product bases in two and three dimensions are commonly available in finite element software packages (e.g., deal.II [4]) and cubic tensor products in particular are commonly used both in modern theory (e.g., isogeometric analysis [9]) and applications (e.g., cardiac electrophysiology models [17]). Hence, a variety of areas of computational science could directly employ the new cubic serendipity basis functions presented here.

The benefit of serendipity finite element methods is a significant reduction in the computational effort required for optimal order (in this case, cubic) convergence. Cubic serendipity methods on meshes of squares requires 12 functions per element, an improvement over the 16 functions per element required for bicubic tensor product methods. On meshes of cubes, the cubic serendipity method requires 32 functions per element instead of the 64 functions per element required for tricubic tensor product methods. Using fewer basis functions per element reduces the size of the overall linear system that must be solved, thereby saving computational time and effort. An additional computational advantage occurs when the functions presented here are used in an isogeometric fashion. The process of converting between computational geometry bases and finite element bases is a well-known computational bottleneck in engineering applications [8] but is easily avoided when basis functions suited to both purposes are employed.

The outline of the paper is as follows: In Sect. 2, we fix notation and summarize relevant background on Bernstein and Hermite basis functions as well as serendipity spaces. In Sect. 3, we present polynomial Bernstein and Hermite style basis functions for $\mathcal{S}_3(I^2)$ that agree with the standard bicubics on edges of I^2 and provide a novel geometric decomposition of the space. In Sect. 4, we present polynomial Bernstein and Hermite style basis functions for $\mathcal{S}_3(I^3)$ that agree with the standard tricubics on edges of I^3, reduce to our bases for I^2 on faces of I^3, and provide a novel geometric decomposition of the space. Finally, we state our conclusions and discuss future directions in Sect. 5.

2 Background and Notation

2.1 Serendipity Elements

We first review the definition of serendipity spaces and their accompanying notation from the work of Arnold and Awanou [1, 2].

Definition 1 The **superlinear degree** of a monomial in n variables, denoted $\mathrm{sldeg}(\cdot)$, is given by

$$\mathrm{sldeg}(x_1^{e_1}x_2^{e_2}\cdots x_n^{e_n}) := \left(\sum_{i=1}^{n} e_i\right) - \#\{e_i \ : \ e_i = 1\}. \tag{1}$$

In words, $\mathrm{sldeg}(q)$ is the ordinary degree of q, ignoring variables that enter linearly. For instance, the superlinear degree of xy^2z^3 is 5.

Definition 2 Define the following spaces of polynomials, each of which is restricted to the domain $I^n = [-1, 1]^n \subset \mathbb{R}^n$:

$$\begin{aligned}
\mathcal{P}_r(I^n) &:= \mathrm{span}_{\mathbb{R}}\{\text{monomials in } n \text{ variables with total degree at most } r\}\\
\mathcal{S}_r(I^n) &:= \mathrm{span}_{\mathbb{R}}\{\text{monomials in } n \text{ variables with superlinear degree at most } r\}\\
\mathcal{Q}_r(I^n) &:= \mathrm{span}_{\mathbb{R}}\{\text{monomials in } n \text{ variables of degree at most } r \text{ in each variable}\}.
\end{aligned}$$

Note that $\mathcal{P}_r(I^n) \subset \mathcal{S}_r(I^n) \subset \mathcal{Q}_r(I^n)$, with proper containments when $r, n > 1$. The space $\mathcal{S}_r(I^n)$ is called the degree r **serendipity space** on the n-dimensional cube I^n. In the notation of the recent paper by Arnold and Awanou [2], the serendipity spaces discussed in this work would be denoted $\mathcal{S}_r\Lambda^0(I^n)$, indicating that they are differential 0-form spaces. The space $\mathcal{Q}_r(I^n)$ is associated with standard tensor product finite element methods; the fact that $\mathcal{S}_r(I^n)$ satisfies the containments above is one of the key features allowing it to retain an $O(h^r)$ a priori error estimate in H^1 norm, where h denotes the width of a mesh element [5]. The spaces have dimension given by the following formulas (cf. [1]).

$$\begin{aligned}
\dim \mathcal{P}_r(I^n) &= \binom{n+r}{n},\\
\dim \mathcal{S}_r(I^n) &= \sum_{d=0}^{\min(n,\lfloor r/2\rfloor)} 2^{n-d}\binom{n}{d}\binom{r-d}{d},\\
\dim \mathcal{Q}_r(I^n) &= (r+1)^n.
\end{aligned}$$

We write out standard bases for these spaces more precisely in the cubic cases of concern here.

$$\mathcal{P}_3(I^2) = \mathrm{span}\{1, \underbrace{x, y}_{\text{linear}}, \underbrace{x^2, y^2, xy}_{\text{quadratic}}, \underbrace{x^3, y^3, x^2y, xy^2}_{\text{cubic}}\}, \tag{2}$$

$$\mathcal{S}_3(I^2) = \mathcal{P}_3(I^2) \cup \mathrm{span}\{\underbrace{x^3y, xy^3}_{\text{superlinear cubic}}\}, \tag{3}$$

$$\mathcal{Q}_3(I^2) = \mathcal{S}_3(I^2) \cup \mathrm{span}\{x^2y^2, x^3y^2, x^2y^3, x^3y^3\}. \tag{4}$$

Observe that the dimensions of the three spaces are 10, 12, and 16, respectively.

$$\mathcal{P}_3(I^3) = \text{span}\{1, \underbrace{x, y, z}_{\text{linear}}, \underbrace{x^2, y^2, z^2, xy, xz, yz}_{\text{quadratic}}, \underbrace{x^3, y^3, z^3, x^2y, x^2z, xy^2, y^2z, xz^2, yz^2, xyz}_{\text{cubic}}\} \tag{5}$$

$$\mathcal{S}_3(I^3) = \mathcal{P}_3(I^3) \cup \text{span}\{\underbrace{x^3y, x^3z, y^3z, xy^3, xz^3, yz^3, x^2yz, xy^2z, xyz^2, x^3yz, xy^3z, xyz^3}_{\text{superlinear cubic}}\} \tag{6}$$

$$\mathcal{Q}_3(I^3) = \mathcal{S}_3(I^3) \cup \text{span}\{x^3y^2, \ldots, x^3y^3z^3\}. \tag{7}$$

Observe that the dimensions of the three spaces are 20, 32, and 64, respectively.

The serendipity spaces are associated to specific **degrees of freedom** in the classical finite element sense. For a face f of I^n of dimension $d \geq 0$, the degrees of freedom associated to f for $\mathcal{S}_r(I^n)$ are (cf. [1])

$$u \longmapsto \int_f uq, \qquad q \in \mathcal{P}_{r-2d}(f).$$

For the cases considered in this work, $n = 2$ or 3 and $r = 3$, so the only nonzero degrees of freedom are when f is a vertex ($d = 0$) or an edge ($d = 1$). Thus, the degrees of freedom for our cases are the values

$$u(v), \quad \int_e u\, dt, \quad \text{and} \quad \int_e ut\, dt, \tag{8}$$

for each vertex v and each edge e of the square or cube.

2.2 Cubic Bernstein and Hermite Bases

For cubic order approximation on square or cubical grids, tensor product bases are typically built from one of two alternative bases for $\mathcal{P}_3([0, 1])$:

$$[\beta] = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} := \begin{bmatrix} (1-x)^3 \\ (1-x)^2x \\ (1-x)x^2 \\ x^3 \end{bmatrix} \qquad [\psi] = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{bmatrix} := \begin{bmatrix} 1 - 3x^2 + 2x^3 \\ x - 2x^2 + x^3 \\ x^2 - x^3 \\ 3x^2 - 2x^3 \end{bmatrix}$$

The set $\{\beta_1, 3\beta_2, 3\beta_3, \beta_4\}$ is the **cubic Bernstein** basis and the set $[\psi]$ is the **cubic Hermite** basis. Bernstein functions have been used recently to provide a geometric decomposition of finite element spaces over simplices [3]. Hermite functions, while more common in geometric modeling contexts [13] have also been studied in finite element contexts for some time [7]. The Hermite functions have the following important property relating them to the geometry of the graph of their associated interpolant:

$$u = u(0)\psi_1 + u'(0)\psi_2 - u'(1)\psi_3 + u(1)\psi_4, \qquad \forall u \in \mathcal{P}_3([0,1]). \tag{9}$$

We have chosen these sign and basis ordering conventions so that both bases have the same symmetry property:

$$\beta_k(1-x) = \beta_{5-k}(x), \qquad \psi_k(1-x) = \psi_{5-k}(x). \tag{10}$$

The bases $[\beta]$ and $[\psi]$ are related by $[\beta] = \mathbb{V}[\psi]$ and $[\psi] = \mathbb{V}^{-1}[\beta]$ where

$$\mathbb{V} = \begin{bmatrix} 1 & -3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -3 & 1 \end{bmatrix}, \qquad \mathbb{V}^{-1} = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 1 \end{bmatrix}. \tag{11}$$

Let $[\beta^n]$ denote the tensor product of n copies of $[\beta]$. Denote $\beta_i(x)\beta_j(y) \in [\beta^2]$ by β_{ij} and $\beta_i(x)\beta_j(y)\beta_k(z) \in [\beta^3]$ by β_{ijk}. In general, $[\beta^n]$ is a basis for $\mathcal{Q}_3([0,1]^n)$, but we will make use of the specific linear combination used to prove this, as stated in the following proposition.

Proposition 1 *For $0 \le r, s, t \le 3$, the reproduction properties of $[\beta]$, $[\beta^2]$, and $[\beta^3]$ take on the respective forms*

$$x^r = \sum_{i=1}^{4} \binom{3-r}{4-i} \beta_i, \tag{12}$$

$$x^r y^s = \sum_{i=1}^{4}\sum_{j=1}^{4} \binom{3-r}{4-i}\binom{3-s}{4-j} \beta_{ij}, \tag{13}$$

$$x^r y^s z^t = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{k=1}^{4} \binom{3-r}{4-i}\binom{3-s}{4-j}\binom{3-t}{4-k} \beta_{ijk}. \tag{14}$$

The proof is elementary. We have a similar property for tensor products of the Hermite basis $[\psi]$, using analogous notation. The proof is a simple matter of swapping the order of summation.

Proposition 2 *Let*

$$\varepsilon_{r,i} := \sum_{a=1}^{4} \binom{3-r}{4-a} v_{ai} \tag{15}$$

where v_{ai} denotes the (a, i) entry (row, column) of $\mathbb{V}$ from (11). For $0 \le r, s, t \le 3$, the reproduction properties of $[\psi]$, $[\psi^2]$, and $[\psi^3]$ take on the respective forms

$$x^r = \sum_{i=1}^{4} \varepsilon_{r,i} \psi_i, \tag{16}$$

$$x^r y^s = \sum_{i=1}^{4} \sum_{j=1}^{4} \varepsilon_{r,i} \varepsilon_{s,j} \psi_{ij}, \tag{17}$$

$$x^r y^s z^t = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \varepsilon_{r,i} \varepsilon_{s,j} \varepsilon_{t,k} \psi_{ijk}. \tag{18}$$

Transforming the bases $[\beta]$ and $[\psi]$ to domains other than $[0, 1]$ is straightforward. If $T : [a, b] \to [0, 1]$ is linear, then replacing x with $T(x)$ in each basis function expression for $[\beta]$ and $[\psi]$ gives bases for $\mathcal{P}_3([a, b])$. Note, however, that the derivative interpolation property for $[\psi]$ must be adjusted to account for the scaling:

$$\begin{aligned} u(x) = {} & u(a)\psi_1(T(x)) + (b-a)u'(a)\psi_2(T(x)) \\ & - (b-a)u'(b)\psi_3(T(x)) + u(b)\psi_4(T(x)), \quad \forall u \in \mathcal{P}_3([a, b]). \end{aligned} \tag{19}$$

In geometric modeling applications, the coefficient $(b - a)$ is sometimes left as an adjustable parameter, usually denoted s for scale factor [10], however, $(b - a)$ is the only choice of scale factor that allows the representation of u given in (19). For all the Hermite and Hermite style functions, we will use **derivative-preserving scaling** which will include scale factors on those functions related to derivatives; this will be made explicit in the various contexts where it is relevant.

Remark 1 Both $[\beta]$ and $[\psi]$ are Lagrange-like at the endpoints of $[0, 1]$, i.e., at an endpoint, the only basis function with nonzero value is the function associated to that endpoint (β_1 or ψ_1 for 0, β_4 or ψ_4 for 1). This means the two remaining basis functions of each type (β_2, β_3 or ψ_2, ψ_3) are naturally associated to the two edge degrees of freedom (8). We will refer to these associations between basis functions and geometrical objects as the standard **geometrical decompositions** of $[\beta]$ and $[\psi]$.

3 Local Bases for $\mathcal{S}_3(I^2)$

Before defining local bases on the square, we fix notation for the domain points to which they are associated. For $[0, 1]^2$, define the set of ordered pairs

$$X := \{\{i, j\} \mid i, j \in \{1, \ldots, 4\}\}.$$

Then X is the disjoint union $V \cup E \cup D$ where

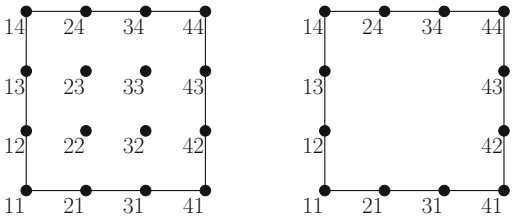


Fig. 2 On the *left*, ordered pairs from X are shown next to the domain point of $[0,1]^2$ to which they correspond. On the *right*, only those ordered pairs used for the serendipity basis are shown. The correspondences $V \leftrightarrow$ vertices, $E \leftrightarrow$ edge points, and $D \leftrightarrow$ domain interior points are evident

$$V := \{\{i,j\} \in X \mid i,j \in \{1,4\}\}; \tag{20}$$

$$E := \{\{i,j\} \in X \mid \text{exactly one of } i,j \text{ is an element of } \{1,4\}\}; \tag{21}$$

$$D := \{\{i,j\} \in X \mid i,j \in \{2,3\}\}. \tag{22}$$

The V indices are associated with vertices of $[0,1]^2$, the E indices to edges of $[0,1]^2$, and the D vertices to the domain interior to $[0,1]^2$. The relation between indices and domain points of the square is shown in Fig. 2. We will frequently denote an index set $\{i,j\}$ as ij to reduce notational clutter.

3.1 A Local Bernstein Style Basis for $\mathcal{S}_3(I^2)$

We now establish a local Bernstein style basis for $\mathcal{S}_3(I^3)$ where $I := [-1,1]$. Define the following set of 12 functions, indexed by $V \cup E$; note the scaling by $1/16$.

$$[\xi^2] = \begin{bmatrix} \xi_{11} \\ \xi_{14} \\ \xi_{41} \\ \xi_{44} \\ \xi_{12} \\ \xi_{13} \\ \xi_{42} \\ \xi_{43} \\ \xi_{21} \\ \xi_{31} \\ \xi_{24} \\ \xi_{34} \end{bmatrix} = \begin{bmatrix} (1-x)(1-y)(-2-2x+x^2-2y+y^2) \\ (1-x)(y+1)(-2-2x+x^2+2y+y^2) \\ (x+1)(1-y)(-2+2x+x^2-2y+y^2) \\ (x+1)(y+1)(-2+2x+x^2+2y+y^2) \\ (1-x)(1-y)^2(y+1) \\ (1-x)(1-y)(y+1)^2 \\ (x+1)(1-y)^2(y+1) \\ (x+1)(1-y)(y+1)^2 \\ (1-x)^2(x+1)(1-y) \\ (1-x)(x+1)^2(1-y) \\ (1-x)^2(x+1)(y+1) \\ (1-x)(x+1)^2(y+1) \end{bmatrix} \cdot \frac{1}{16}. \tag{23}$$

Fix the basis orderings

$$[\xi^2] := [\ \underbrace{\xi_{11}, \xi_{14}, \xi_{41}, \xi_{44}}_{\text{indices in } V},\ \underbrace{\xi_{12}, \xi_{13}, \xi_{42}, \xi_{43}, \xi_{21}, \xi_{31}, \xi_{24}, \xi_{34}}_{\text{indices in } E}\ ], \tag{24}$$

$$[\beta^2] := [\ \underbrace{\beta_{11}, \beta_{14}, \beta_{41}, \beta_{44}}_{\text{indices in } V},\ \underbrace{\beta_{12}, \beta_{13}, \beta_{42}, \beta_{43}, \beta_{21}, \beta_{31}, \beta_{24}, \beta_{34}}_{\text{indices in } E},\ \underbrace{\beta_{22}, \beta_{23}, \beta_{32}, \beta_{33}}_{\text{indices in } D}\ ] \tag{25}$$

The following theorem will show that $[\xi^2]$ is a geometric decomposition of $\mathcal{S}_3(I^2)$, by which we mean that each function in $[\xi^2]$ has a natural association to a specific degree of freedom, i.e., to a specific domain point of the element.

Theorem 1 *Let $\beta^I_{\ell m}$ denote the scaling of $\beta_{\ell m}$ to I^2, i.e.*

$$\beta^I_{\ell m} := \beta_\ell((x+1)/2)\beta_m((y+1)/2).$$

The set $[\xi^2]$ has the following properties:

(i) $[\xi^2]$ *is a basis for* $\mathcal{S}_3(I^2)$.
(ii) *For any* $\ell m \in V \cup E$, $\xi_{\ell m}$ *is identical to* $\beta^I_{\ell m}$ *on the edges of* I^2.
(iii) $[\xi^2]$ *is a geometric decomposition of* $\mathcal{S}_3(I^2)$.

Proof For (i), we scale $[\xi^2]$ to $[0,1]^2$ to take advantage of a simple characterization of the reproduction properties. Let $[\xi^2]^{[0,1]}$ denote the set of scaled basis functions $\xi^{[0,1]}_{\ell m}(x,y) := \xi_{\ell m}(2x-1, 2y-1)$. Given the basis orderings in (24) and (25), it can be confirmed directly that $[\xi^2]^{[0,1]}$ is related to $[\beta^2]$ by

$$[\xi^2]^{[0,1]} = \mathbb{B}[\beta^2] \tag{26}$$

where $\mathbb{B}$ is the 12×16 matrix with the structure

$$\mathbb{B} := \left[\ \mathbb{I}\ \middle|\ \mathbb{B}'\ \right], \tag{27}$$

where $\mathbb{I}$ is the 12×12 identity matrix and $\mathbb{B}'$ is the 12×4 matrix

$$\mathbb{B}' = \begin{bmatrix} -4 & -2 & -2 & -1 \\ -2 & -4 & -1 & -2 \\ -2 & -1 & -4 & -2 \\ -1 & -2 & -2 & -4 \\ 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}. \tag{28}$$

Using $ij \in X$ to denote an index for β_{ij} and $\ell m \in V \cup E$ to denote an index for $\xi^{[0,1]}_{\ell m}$, the entries of $\mathbb{B}$ can be denoted by $b^{\ell m}_{ij}$ so that

$$\mathbb{B} := \begin{bmatrix} b^{11}_{11} & \cdots & b^{11}_{ij} & \cdots & b^{11}_{33} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b^{\ell m}_{11} & \cdots & b^{\ell m}_{ij} & \cdots & b^{\ell m}_{33} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b^{34}_{11} & \cdots & b^{34}_{ij} & \cdots & b^{34}_{33} \end{bmatrix}. \tag{29}$$

We now observe that for each $ij \in X$,

$$\binom{3-r}{4-i}\binom{3-s}{4-j} = \sum_{\ell m \in V \cup E} \binom{3-r}{4-\ell}\binom{3-s}{4-m} b^{\ell m}_{ij}, \tag{30}$$

for all (r, s) pairs such that $\text{sldeg}(x^r y^s) \leq 3$ (recall Definition 1). Note that this claim holds trivially for the first 12 columns of $\mathbb{B}$, i.e., for those $ij \in V \cup E \subset X$. For $ij \in D \subset X$, (30) defines an invertible linear system of 12 equations with 12 unknowns whose solution is the ij column of $\mathbb{B}'$; the 12 (r, s) pairs correspond to the exponents of x and y in the basis ordering of $\mathcal{S}_3(I^2)$ given in (2) and (3). Substituting (30) into (13) yields:

$$x^r y^s = \sum_{ij \in X} \left(\sum_{\ell m \in V \cup E} \binom{3-r}{4-\ell}\binom{3-s}{4-m} b^{\ell m}_{ij} \right) \beta_{ij}.$$

Swapping the order of summation and regrouping yields

$$x^r y^s = \sum_{\ell m \in V \cup E} \binom{3-r}{4-\ell}\binom{3-s}{4-m} \left(\sum_{ij \in X} b^{\ell m}_{ij} \beta_{ij} \right).$$

The inner summation is exactly $\xi^{[0,1]}_{\ell m}$ by (26), implying that

$$x^r y^s = \sum_{\ell m \in V \cup E} \binom{3-r}{4-\ell}\binom{3-s}{4-m} \xi^{[0,1]}_{\ell m}, \tag{31}$$

for all (r, s) pairs with $\text{sldeg}(x^r y^s) \leq 3$. Since $[\xi^2]^{[0,1]}$ has 12 elements which span the 12-dimensional space $\mathcal{S}_3([0, 1]^2)$, it is a basis for $\mathcal{S}_3([0, 1]^2)$. By scaling, $[\xi^2]$ is a basis for $\mathcal{S}_3(I^2)$.

For (ii), note that an edge of $[0, 1]^2$ is described by an equation of the form $\{x \text{ or } y\} = \{0 \text{ or } 1\}$. Since $\beta_2(t)$ and $\beta_3(t)$ are equal to 0 at $t = 0$ and $t = 1$, $\beta_{ij} \equiv 0$

on the edges of $[0,1]^2$ for any $ij \in D$. By the structure of $\mathbb{B}$ from (27), we see that for any $\ell m \in V \cup E$,

$$\xi_{\ell m}^{[0,1]} = \beta_{\ell m} + \sum_{ij \in D} b_{ij}^{\ell m} \beta_{ij}. \tag{32}$$

Thus, on the edges of $[0,1]^2$, $\xi_{\ell m}^{[0,1]}$ and $\beta_{\ell m}$ are identical. After scaling back, we have $\xi_{\ell m}$ and $\beta_{\ell m}^I$ identical on the edges of I^2, as desired.

For (iii), the geometric decomposition is given by the indices of the basis functions, i.e., the function $\xi_{\ell m}$ is associated to the domain point for $\ell m \in V \cup E$. This follows immediately from (ii), the fact that $[\beta^2]$ is a tensor product basis, and Remark 1 □

Remark 2 It is worth noting that the basis $[\xi^2]$ was derived by essentially the reverse order of the proof of part (i) of the theorem. More precisely, the 12 coefficients in each column of $\mathbb{B}$ define an invertible linear system given by (30). After solving for the coefficients, we can immediately derive the basis functions via (26). By the nature of this approach, the edge agreement property (ii) is guaranteed by the symmetry properties of the basis $[\beta]$. This technique was inspired by a previous work for Lagrange-like quadratic serendipity elements on convex polygons [14].

3.2 A Local Hermite Style Basis for $\mathcal{S}_3(I^2)$

We now establish a local Hermite style basis $[\vartheta^2]$ for $\mathcal{S}_3(I^2)$ using the bicubic Hermite basis $[\psi^2]$ for $\mathcal{Q}_3([0,1]^2)$. Define the following set of 12 functions, indexed by $V \cup E$; note the scaling by $1/8$.

$$[\vartheta^2] = \begin{bmatrix} \vartheta_{11} \\ \vartheta_{14} \\ \vartheta_{41} \\ \vartheta_{44} \\ \vartheta_{12} \\ \vartheta_{13} \\ \vartheta_{42} \\ \vartheta_{43} \\ \vartheta_{21} \\ \vartheta_{31} \\ \vartheta_{24} \\ \vartheta_{34} \end{bmatrix} = \begin{bmatrix} -(1-x)(1-y)(-2+x+x^2+y+y^2) \\ -(1-x)(y+1)(-2+x+x^2-y+y^2) \\ -(x+1)(1-y)(-2-x+x^2+y+y^2) \\ -(x+1)(y+1)(-2-x+x^2-y+y^2) \\ (1-x)(1-y)^2(y+1) \\ (1-x)(1-y)(y+1)^2 \\ (x+1)(1-y)^2(y+1) \\ (x+1)(1-y)(y+1)^2 \\ (1-x)^2(x+1)(1-y) \\ (1-x)(x+1)^2(1-y) \\ (1-x)^2(x+1)(y+1) \\ (1-x)(x+1)^2(y+1) \end{bmatrix} \cdot \frac{1}{8}. \tag{33}$$

Fix the basis orderings

$$[\vartheta^2] := [\ \underbrace{\vartheta_{11}, \vartheta_{14}, \vartheta_{41}, \vartheta_{44}}_{\text{indices in } V},\ \underbrace{\vartheta_{12}, \vartheta_{13}, \vartheta_{42}, \vartheta_{43}, \vartheta_{21}, \vartheta_{31}, \vartheta_{24}, \vartheta_{34}}_{\text{indices in } E}\ ], \tag{34}$$

$$[\psi^2] := [\ \underbrace{\psi_{11}, \psi_{14}, \psi_{41}, \psi_{44}}_{\text{indices in } V},\ \underbrace{\psi_{12}, \psi_{13}, \psi_{42}, \psi_{43}, \psi_{21}, \psi_{31}, \psi_{24}, \psi_{34}}_{\text{indices in } E},\ \underbrace{\psi_{22}, \psi_{23}, \psi_{32}, \psi_{33}}_{\text{indices in } D}\ ] \tag{35}$$

Theorem 2 *Let $\psi^I_{\ell m}$ denote the derivative-preserving scaling of $\psi_{\ell m}$ to I^2, i.e.*

$$\psi^I_{\ell m} := \psi_\ell((x+1)/2)\psi_m((y+1)/2), \qquad \ell m \in V,$$
$$\psi^I_{\ell m} := 2\psi_\ell((x+1)/2)\psi_m((y+1)/2), \qquad \ell m \in E.$$

The set $[\vartheta^2]$ has the following properties:

(i) *$[\vartheta^2]$ is a basis for $\mathcal{S}_3(I^2)$.*
(ii) *For any $\ell m \in V \cup E$, $\xi_{\ell m}$ is identical to $\psi^I_{\ell m}$ on the edges of I^2.*
(iii) *$[\vartheta^2]$ is a geometric decomposition of $\mathcal{S}_3(I^2)$.*

Proof The proof follows that of Theorem 1 so we abbreviate proof details that are similar. For (i), let $[\vartheta^2]^{[0,1]}$ denote the derivative-preserving scaling of $[\vartheta^2]$ to $[0, 1]^2$; the scale factor is $1/2$ for functions with indices in E. Given the basis orderings in (34) and (35), we have

$$[\vartheta^2]^{[0,1]} = \mathbb{H}[\psi^2] \tag{36}$$

where $\mathbb{H}$ is the 12×16 matrix with the structure

$$\mathbb{H} := \left[\ \mathbb{I}\ \middle|\ \mathbb{H}'\ \right], \tag{37}$$

where $\mathbb{I}$ is the 12×12 identity matrix and $\mathbb{H}'$ is the 12×4 matrix with

$$\mathbb{H}' = \begin{bmatrix} -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \tag{38}$$

Denote the entries of $\mathbb{H}$ by $h_{ij}^{\ell m}$ (cf. (29)). Recalling (15), observe that for each $ij \in X$,

$$\varepsilon_{r,i}\varepsilon_{s,j} = \sum_{\ell m \in V \cup E} \varepsilon_{r,\ell}\varepsilon_{s,m} h_{ij}^{\ell m}, \tag{39}$$

for all (r, s) pairs such that $\operatorname{sldeg}(x^r y^s) \leq 3$. Similar to the Bernstein case, we substitute (39) into (17), swap the order of summation and regroup, yielding

$$x^r y^s = \sum_{\ell m \in V \cup E} \varepsilon_{r,\ell}\varepsilon_{s,m} \left(\sum_{ij \in X} h_{ij}^{\ell m} \psi_{ij} \right).$$

The inner summation is exactly $\vartheta_{\ell m}^{[0,1]}$ by (36), implying that

$$x^r y^s = \sum_{\ell m \in V \cup E} \varepsilon_{r,\ell}\varepsilon_{s,m} \vartheta_{\ell m}^{[0,1]}, \tag{40}$$

for all (r, s) pairs with $\operatorname{sldeg}(x^r y^s) \leq 3$, proving that $[\vartheta^2]^{[0,1]}$ is a basis for $\mathcal{S}_3([0, 1]^2)$. By derivative-preserving scaling, $[\vartheta^2]$ is a basis for $\mathcal{S}_3(I^2)$.

For (ii), observe that for any $ij \in D$, $\psi_{ij} \equiv 0$ on the edges of $[0, 1]^2$ by virtue of the bicubic Hermite basis functions' definition. By the structure of $\mathbb{H}$ from (37), we see that for any $\ell m \in V \cup E$,

$$\vartheta_{\ell m}^{[0,1]} = \psi_{\ell m} + \sum_{ij \in D} h_{ij}^{\ell m} \psi_{ij}. \tag{41}$$

Thus, on the edges of $[0, 1]^2$, $\vartheta_{\ell m}^{[0,1]}$ and $\psi_{\ell m}$ are identical. After scaling back, we have $\vartheta_{\ell m}$ and $\psi_{\ell m}^I$ identical on the edges of I^2, as desired.

For (iii), the geometric decomposition is given by the indices of the basis functions, i.e., the function $\vartheta_{\ell m}$ is associated to the domain point for $\ell m \in V \cup E$. This follows immediately from (ii), the fact that $[\psi^2]$ is a tensor product basis, and Remark 1 at the end of Sect. 2. See also Fig. 3. □

4 Local Bases for $\mathcal{S}_3(I^3)$

Before defining local bases on the cube, we fix notation for the domain points to which they are associated. For $[0, 1]^3$, define the set of ordered triplets

$$Y := \{\{i, j, k\} \mid i, j, k \in \{1, \ldots, 4\}\}.$$

Then Y is the disjoint union $V \cup E \cup F \cup M$ where

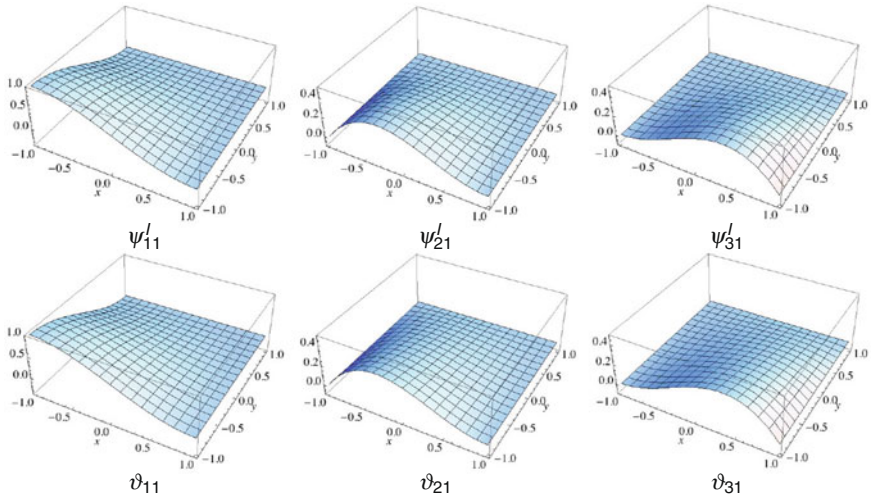


Fig. 3 The *top row* shows 3 of the 16 bicubic Hermite functions on I^2 while the *bottom row* shows 3 of the 12 cubic Hermite style serendipity functions. The visual differences are subtle, although some changes in concavity can be observed. Note that functions in the same column have the same values on the edges of I^2

$$V := \{\{i, j, k\} \in Y \mid i, j, k \in \{1, 4\}\}; \tag{42}$$
$$E := \{\{i, j, k\} \in Y \mid \text{exactly two of } i, j, k \text{ are elements of } \{1, 4\}\}; \tag{43}$$
$$F := \{\{i, j, k\} \in Y \mid \text{exactly one of } i, j, k \text{ is an element of } \{1, 4\}\}; \tag{44}$$
$$M := \{\{i, j, k\} \in Y \mid i, j, k \in \{2, 3\}\}. \tag{45}$$

The V indices are associated with vertices of $[0, 1]^3$, the E indices to edges of $[0, 1]^3$, the F indices to face interior points of $[0, 1]^3$, and the M vertices to the domain interior of $[0, 1]^3$. The relation between indices and domain points of the cube is shown in Fig. 4.

4.1 A Local Bernstein Style Basis for $\mathcal{S}_3(I^3)$

Under the notation and conventions established in Sect. 2, we are ready to establish a local Bernstein style basis for $\mathcal{S}_3(I^3)$ where $I := [-1, 1]$. In Fig. 5, we define a set of 32 functions, indexed by $V \cup E \subset Y$; note the scaling by $1/32$. We fix the following basis orderings, with omitted basis functions ordered lexicographically by index.

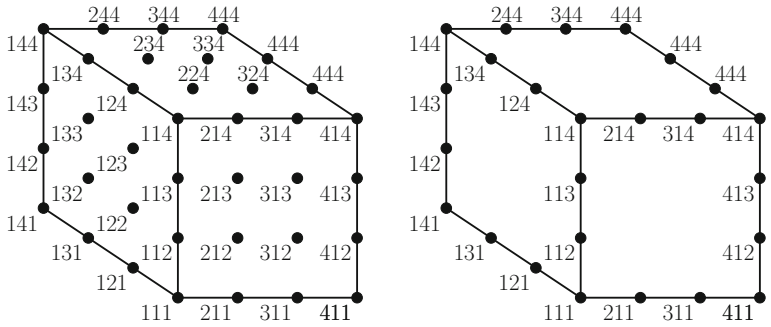


Fig. 4 On the *left*, ordered triplets from Y are shown next to the domain point of $[0, 1]^3$ to which they correspond. Points hidden by the perspective are not shown. The origin is at the point labeled 111; the positive x, y, and z axes go right, back, and up, respectively. On the *right*, only those indices used for the serendipity basis are shown. The correspondences $V \leftrightarrow$ vertices, $E \leftrightarrow$ edge points, $F \leftrightarrow$ face interior points, and $M \leftrightarrow$ domain interior points are evident

$$[\xi^3] := [\ \underbrace{\xi_{111}, \ldots, \xi_{444}}_{\text{indices in } V},\ \underbrace{\xi_{112}, \ldots, \xi_{443}}_{\text{indices in } E}\ ], \tag{46}$$

$$[\beta] := [\ \underbrace{\beta_{111}, \ldots, \beta_{444}}_{\text{indices in } V},\ \underbrace{\beta_{112}, \ldots, \beta_{443}}_{\text{indices in } E},\ \underbrace{\beta_{122}, \ldots, \beta_{433}}_{\text{indices in } F},\ \underbrace{\beta_{222}, \ldots, \beta_{333}}_{\text{indices in } M},\ ] \tag{47}$$

Theorem 3 *Let $\beta^I_{\ell mn}$ denote the scaling of $\beta_{\ell mn}$ to I^3, i.e.*

$$\beta^I_{\ell mn} := \beta_\ell((x+1)/2)\beta_m((y+1)/2)\beta_n((z+1)/2).$$

The set $[\xi^3]$ has the following properties:

(i) $[\xi^3]$ *is a basis for* $\mathcal{S}_3(I^3)$.
(ii) $[\xi^3]$ *reduces to* $[\xi^2]$ *on faces of* I^3.
(iii) *For any* $\ell mn \in V \cup E$, $\xi_{\ell mn}$ *is identical to* $\beta^I_{\ell mn}$ *on edges of* I^3.
(iv) $[\xi^3]$ *is a geometric decomposition of* $\mathcal{S}_3(I^3)$.

The proof is similar to that of Theorem 1. Note that for (ii), the claim can be confirmed directly by calculation, for instance, $\xi_{111}(x, y, -1) = \xi_{11}(x, y)$ or $\xi_{142}(x, 1, z) = \xi_{12}(x, z)$. A complete proof can be found in a longer version of this paper appearing online at arXiv:1208.5973 [math.NA].

4.2 A Local Hermite Style Basis for $\mathcal{S}_3(I^3)$

We now establish a local Hermite style basis $[\vartheta^3]$ for $\mathcal{S}_3(I^3)$ using the tricubic Hermite basis $[\psi^3]$ for $\mathcal{Q}_3([0, 1]^3)$. In Fig. 6, we define a set of 32 functions, indexed

$$[\xi^3] = \begin{bmatrix} \xi_{111} \\ \xi_{114} \\ \xi_{141} \\ \xi_{144} \\ \xi_{411} \\ \xi_{414} \\ \xi_{441} \\ \xi_{444} \\ \xi_{112} \\ \xi_{113} \\ \xi_{121} \\ \xi_{124} \\ \xi_{131} \\ \xi_{134} \\ \xi_{142} \\ \xi_{143} \\ \xi_{211} \\ \xi_{214} \\ \xi_{241} \\ \xi_{244} \\ \xi_{311} \\ \xi_{314} \\ \xi_{341} \\ \xi_{344} \\ \xi_{412} \\ \xi_{413} \\ \xi_{421} \\ \xi_{424} \\ \xi_{431} \\ \xi_{434} \\ \xi_{442} \\ \xi_{443} \end{bmatrix} = \begin{bmatrix} (1-x)(1-y)(1-z)(-5-2x+x^2-2y+y^2-2z+z^2) \\ (1-x)(1-y)(z+1)(-5-2x+x^2-2y+y^2+2z+z^2) \\ (1-x)(y+1)(1-z)(-5-2x+x^2+2y+y^2-2z+z^2) \\ (1-x)(y+1)(z+1)(-5-2x+x^2+2y+y^2+2z+z^2) \\ (x+1)(1-y)(1-z)(-5+2x+x^2-2y+y^2-2z+z^2) \\ (x+1)(1-y)(z+1)(-5+2x+x^2-2y+y^2+2z+z^2) \\ (x+1)(y+1)(1-z)(-5+2x+x^2+2y+y^2-2z+z^2) \\ (x+1)(y+1)(z+1)(-5+2x+x^2+2y+y^2+2z+z^2) \\ (1-x)(1-y)(1-z)^2(z+1) \\ (1-x)(1-y)(1-z)(z+1)^2 \\ (1-x)(1-y)^2(y+1)(1-z) \\ (1-x)(1-y)^2(y+1)(z+1) \\ (1-x)(1-y)(y+1)^2(1-z) \\ (1-x)(1-y)(y+1)^2(z+1) \\ (1-x)(y+1)(1-z)^2(z+1) \\ (1-x)(y+1)(1-z)(z+1)^2 \\ (1-x)^2(x+1)(1-y)(1-z) \\ (1-x)^2(x+1)(1-y)(z+1) \\ (1-x)^2(x+1)(y+1)(1-z) \\ (1-x)^2(x+1)(y+1)(z+1) \\ (1-x)(x+1)^2(1-y)(1-z) \\ (1-x)(x+1)^2(1-y)(z+1) \\ (1-x)(x+1)^2(y+1)(1-z) \\ (1-x)(x+1)^2(y+1)(z+1) \\ (x+1)(1-y)(1-z)^2(z+1) \\ (x+1)(1-y)(1-z)(z+1)^2 \\ (x+1)(1-y)^2(y+1)(1-z) \\ (x+1)(1-y)^2(y+1)(z+1) \\ (x+1)(1-y)(y+1)^2(1-z) \\ (x+1)(1-y)(y+1)^2(z+1) \\ (x+1)(y+1)(1-z)^2(z+1) \\ (x+1)(y+1)(1-z)(z+1)^2 \end{bmatrix} \cdot \frac{1}{32}$$

Fig. 5 Bernstein style basis functions for $\mathcal{S}_3(I^3)$ with properties given by Theorem 3

by $V \cup E \subset Y$; note the scaling by $1/16$. We fix the following basis orderings, with omitted basis functions ordered lexicographically by index.

$$[\vartheta^3] := [\ \underbrace{\vartheta_{111}, \ldots, \vartheta_{444}}_{\text{indices in } V},\ \underbrace{\vartheta_{112}, \ldots, \vartheta_{443}}_{\text{indices in } E}\ ], \tag{48}$$

$$[\beta] := [\ \underbrace{\psi_{111}, \ldots, \psi_{444}}_{\text{indices in } V},\ \underbrace{\psi_{112}, \ldots, \psi_{443}}_{\text{indices in } E},\ \underbrace{\psi_{122}, \ldots, \psi_{433}}_{\text{indices in } F},\ \underbrace{\psi_{222}, \ldots, \psi_{333}}_{\text{indices in } M},\ ] \tag{49}$$

Theorem 4 *Let $\psi^I_{\ell mn}$ denote the derivative-preserving scaling of $\psi_{\ell mn}$ to I^3, i.e.,*

$$\psi^I_{\ell m} := \psi_\ell((x+1)/2)\psi_m((y+1)/2)\psi_n((z+1)/2), \qquad \ell mn \in V,$$
$$\psi^I_{\ell mn} := 2\psi_\ell((x+1)/2)\psi_m((y+1)/2)\psi_n((z+1)/2), \qquad \ell mn \in E.$$

$$[\vartheta^3] = \begin{bmatrix} \vartheta_{111} \\ \vartheta_{114} \\ \vartheta_{141} \\ \vartheta_{144} \\ \vartheta_{411} \\ \vartheta_{414} \\ \vartheta_{441} \\ \vartheta_{444} \\ \vartheta_{112} \\ \vartheta_{113} \\ \vartheta_{121} \\ \vartheta_{124} \\ \vartheta_{131} \\ \vartheta_{134} \\ \vartheta_{142} \\ \vartheta_{143} \\ \vartheta_{211} \\ \vartheta_{214} \\ \vartheta_{241} \\ \vartheta_{244} \\ \vartheta_{311} \\ \vartheta_{314} \\ \vartheta_{341} \\ \vartheta_{344} \\ \vartheta_{412} \\ \vartheta_{413} \\ \vartheta_{421} \\ \vartheta_{424} \\ \vartheta_{431} \\ \vartheta_{434} \\ \vartheta_{442} \\ \vartheta_{443} \end{bmatrix} = \begin{bmatrix} -(1-x)(1-y)(1-z)(-2+x+x^2+y+y^2+z+z^2) \\ -(1-x)(1-y)(z+1)(-2+x+x^2+y+y^2-z+z^2) \\ -(1-x)(y+1)(1-z)(-2+x+x^2-y+y^2+z+z^2) \\ -(1-x)(y+1)(z+1)(-2+x+x^2-y+y^2-z+z^2) \\ -(x+1)(1-y)(1-z)(-2-x+x^2+y+y^2+z+z^2) \\ -(x+1)(1-y)(z+1)(-2-x+x^2+y+y^2-z+z^2) \\ -(x+1)(y+1)(1-z)(-2-x+x^2-y+y^2+z+z^2) \\ -(x+1)(y+1)(z+1)(-2-x+x^2-y+y^2-z+z^2) \\ (1-x)(1-y)(1-z)^2(z+1) \\ (1-x)(1-y)(1-z)(z+1)^2 \\ (1-x)(1-y)^2(y+1)(1-z) \\ (1-x)(1-y)^2(y+1)(z+1) \\ (1-x)(1-y)(y+1)^2(1-z) \\ (1-x)(1-y)(y+1)^2(z+1) \\ (1-x)(y+1)(1-z)^2(z+1) \\ (1-x)(y+1)(1-z)(z+1)^2 \\ (1-x)^2(x+1)(1-y)(1-z) \\ (1-x)^2(x+1)(1-y)(z+1) \\ (1-x)^2(x+1)(y+1)(1-z) \\ (1-x)^2(x+1)(y+1)(z+1) \\ (1-x)(x+1)^2(1-y)(1-z) \\ (1-x)(x+1)^2(1-y)(z+1) \\ (1-x)(x+1)^2(y+1)(1-z) \\ (1-x)(x+1)^2(y+1)(z+1) \\ (x+1)(1-y)(1-z)^2(z+1) \\ (x+1)(1-y)(1-z)(z+1)^2 \\ (x+1)(1-y)^2(y+1)(1-z) \\ (x+1)(1-y)^2(y+1)(z+1) \\ (x+1)(1-y)(y+1)^2(1-z) \\ (x+1)(1-y)(y+1)^2(z+1) \\ (x+1)(y+1)(1-z)^2(z+1) \\ (x+1)(y+1)(1-z)(z+1)^2 \end{bmatrix} \cdot \frac{1}{16}$$

Fig. 6 Hermite style basis functions for $\mathcal{S}_3(I^3)$ with properties given by Theorem 4

The set $[\vartheta^3]$ *has the following properties:*

(i) $[\vartheta^3]$ *is a basis for* $\mathcal{S}_3(I^3)$.
(ii) $[\vartheta^3]$ *reduces to* $[\vartheta^2]$ *on faces of* I^3.
(iii) *For any* $\ell mn \in V \cup E$, $\vartheta_{\ell mn}$ *is identical to* $\psi^I_{\ell mn}$ *on edges of* I^3.
(iv) $[\vartheta^3]$ *is a geometric decomposition of* $\mathcal{S}_3(I^3)$.

The proof is similar to that of Theorem 1. A complete proof can be found in a longer version of this paper appearing online at arXiv:1208.5973 [math.NA].

5 Conclusions and Future Directions

The basic functions presented in this work are well-suited for use in finite element applications, as discussed in the introduction. For geometric modeling purposes, some adaptation of traditional techniques will be required as the bases do not have the classical properties of positivity and do not form a partition of unity. Nevertheless, we are already witnessing the successful implementation of the basis $[\vartheta^3]$ in the geometric modeling and finite element analysis package Continuity developed by the Cardiac Mechanics Research Group at UC San Diego. In that context, the close similarities of $[\vartheta^3]$ and $[\psi^3]$ has allowed a straightforward implementation procedure with only minor adjustments to the geometric modeling subroutines.

Additionally, the proof techniques used for the theorems suggest a number of promising extensions. Similar techniques should be able to produce Bernstein style bases for higher polynomial order serendipity spaces, although the introduction of interior degrees of freedom that occurs when $r > 3$ requires some additional care to resolve. Some higher order Hermite style bases may also be available, although the association of directional derivative values to vertices is somewhat unique to the $r = 3$ case. Preconditioners for finite element methods employing our bases are still needed, as is a thorough analysis of the tradeoffs between the approach outlined here and alternative approaches to basis reduction, such as static condensation. The fact that all the functions defined here are fixed linear combinations of standard bicubic or tricubic basis functions suggests that appropriate preconditioners will have a straightforward and computationally advantageous construction.

Acknowledgments Support for this work was provided in part by NSF Award 0715146 and the National Biomedical Computation Resource while the author was at the University of California, San Diego.

References

1. Arnold, D., Awanou, G.: The serendipity family of finite elements. Found. Comput. Math. **11**(3), 337–344 (2011)
2. Arnold, D.N., Awanou, G.: Finite element differential forms on cubical meshes. Math. Comput. **83**, 1551–1570 (2014)
3. Arnold, D., Falk, R., Winther, R.: Geometric decompositions and local bases for spaces of finite element differential forms. Comput. Methods Appl. Mech. Eng. **198**(21–26), 1660–1672 (2009)
4. Bangerth, W., Hartmann, R., Kanschat, G.: Deal. ii—A general-purpose object-oriented finite element library. ACM Trans. Math. Softw. (TOMS) **33**(4), 24–es (2007)
5. Brenner, S., Scott, L.: The Mathematical Theory of Finite Element Methods. Springer, New York (2002)
6. Ciarlet, P.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, vol. 40, 2nd edn. SIAM, Philadelphia (2002)
7. Ciarlet, P., Raviart, P.: General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods. Arch. Ration. Mechan. Anal. **46**(3), 177–199 (1972)

8. Cottrell, J., Hughes, T., Bazilevs, Y.: Isogeometric Analysis: Toward Integration of CAD and FEA. Wiley, Chichester (2009)
9. Evans, J.A., Hughes, T.J.R.: Explicit trace inequalities for isogeometric analysis and parametric hexahedral finite elements. Numer. Math. **123**(2), 259–290 (2013)
10. Hoschek, J., Lasser, D.: Fundamentals of Computer Aided Geometric Design. AK Peters, Wellesley (1993)
11. Hughes, T.: The Finite Element Method. Prentice Hall, Englewood Cliffs (1987)
12. Mandel, J.: Iterative solvers by substructuring for the *p*-version finite element method. Comput. Methods Appl. Mech. Eng. **80**(1–3), 117–128 (1990)
13. Mortenson, M.: Geometric Modeling, 3rd edn. Wiley, New York (2006)
14. Rand, A., Gillette, A., Bajaj, C.: Quadratic serendipity finite elements on polygons using generalized barycentric coordinates. Math. Comput. http://www.ams.org/journals/mcom/0000-000-00/S0025-5718-2014-02807-X/home.html (2014)
15. Strang, G., Fix, G.J.: An Analysis of the Finite Element Method. Prentice-Hall, Englewood Cliffs (1973)
16. Szabó, B., Babuška, I.: Finite Element Analysis. Wiley-Interscience, New York (1991)
17. Zhang, Y., Liang, X., Ma, J., Jing, Y., Gonzales, M.J., Villongco, C., Krishnamurthy, A., Frank, L.R., Nigam, V., Stark, P., Narayan, S.M., McCulloch, A.D.: An atlas-based geometry pipeline for cardiac hermite model construction and diffusion tensor reorientation. Med. Image Anal. **16**(6), 1130–1141 (2012)

Suitability of Parametric Shepard Interpolation for Nonrigid Image Registration

A. Ardeshir Goshtasby

Abstract Shepard interpolation is known to produce flat horizontal spots at and around data points. The phenomenon is caused by Shepard's use of rational inverse-distance weight functions, producing similar values around data points. In this paper, a parametric version of Shepard interpolation is introduced that avoids flat horizontal spots. Because Shepard interpolation or its parametric version does not require the solution of a system of equations, the interpolation is stable under varying density and organization of data points as well as under highly varying data values. The suitability of parametric Shepard interpolation in nonrigid image registration is investigated and its speed and accuracy are compared with those of multiquadric, thin-plate spline, and moving least-squares.

Keywords Shepard interpolation · Parametric Shepard interpolation · Image registration

1 Introduction

Image registration is the process of finding correspondence between all points in two images of a scene. This correspondence is needed to fuse information in the images, to detect changes occurring in the scene between the times the images are obtained, and to recover the scene's geometry [1].

Image registration is generally achieved in three steps: (1) a set of points is detected in one image, (2) the corresponding points are located in the second image, and (3) from the coordinates of corresponding points in the images, a transformation function is determined to warp the geometry of the second image to resemble that of the first.

A. A. Goshtasby (✉)
Department of Computer Science and Engineering, 303 Russ Engineering Center, Wright State University, 3640 Colonel Glenn Highway, Dayton, OH 45435, USA
e-mail: ardy@wright.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_8,

This transformation function makes it possible to spatially align the images and establish correspondence between all scene points appearing in both images. When one image is simply a translated and/or rotated version of the other, the process is known as rigid registration. Otherwise, the process is called nonrigid registration.

The problem of finding a transformation function for registration of two images can be described as follows: Given the coordinates of n corresponding points in the images:

$$\{(x_i, y_i), (X_i, Y_i) : i = 1, \ldots, n\}, \tag{1}$$

we would like to determine two functions f and g that satisfy

$$\begin{array}{l} X_i = f(x_i, y_i), \\ Y_i = g(x_i, y_i), \end{array} \quad i = 1, \ldots, n. \tag{2}$$

f can be considered a single-valued surface interpolating 3-D points

$$\{(x_i, y_i, X_i) : i = 1, \ldots, n\}, \tag{3}$$

and g can be considered another single-valued surface interpolating 3-D points

$$\{(x_i, y_i, Y_i) : i = 1, \ldots, n\}. \tag{4}$$

Coordinates (x_i, y_i) represent the column and row numbers of the ith point in the first image and coordinates (X_i, Y_i) represent the column and row numbers of the ith point in the second image. Coordinate x increases from left to right and varies between 0 and $n_c - 1$ and coordinate y increases from top to bottom and varies between 0 and $n_r - 1$. The integers n_c and n_r are, respectively, the number of columns and the number of rows in the first image. Similarly, coordinate X increases from left to right and varies between 0 and $N_c - 1$ and coordinate Y increases from top to bottom and varies between $N_r - 1$. N_c and N_r are, respectively, the number of columns and the number of rows in the second image.

Throughout this paper, the first image will be referred to as the *reference image* and the second image will be referred to as the *sensed image*. Also, the points given in the images will be referred to as the *control points*. Therefore, (x_i, y_i) and (X_i, Y_i) represent the coordinates of the ith corresponding control points in the images.

Functions f and g are the components of the transformation, relating coordinates of points in the sensed image to the coordinates of the corresponding points in the reference image. By knowing the coordinates of n corresponding control points in two images of a scene, we would like to find a transformation with components f and g that will map the sensed image point-by-point to the reference image. This mapping, in effect, transforms the geometry of the sensed image to resemble that of the reference image.

A component of a transformation function is a single-valued function that takes a point (x, y) in the reference image and determines the X- or the Y-coordinate of the corresponding point in the sensed image. Many interpolation functions exist in

the literature that can be used for this purpose. We are interested in a function that is locally sensitive and stable.

Local sensitivity is required to keep an error in the location of a control point local. Due to noise and other factors, some point correspondences may be inaccurate. Such inaccuracies should not be propagated over the entire interpolation (registration) domain. Rather, the influence of an inaccurate control point location should be kept local to the point. For this reason, it is necessary to define a component of a transformation function in terms of monotonically decreasing rather than monotonically increasing basis functions.

Stability in solution is required to ensure that two images can always be registered independent of the geometric difference between them. Methods that require the solution of systems of equations are generally not desired because the equations to be solved may become ill-conditioned and impossible to solve.

In this paper, a component of a transformation function is defined by a parametric version of the Shepard interpolation [16]. Registration speed and accuracy of parametric Shepard interpolation (PSI) are measured and compared with those of multiquadric (MQ) [4, 5], thin-plate spline (TPS) [2, 3, 12], and moving least-squares (MLS) [8, 9].

2 Parametric Shepard Interpolation

One of the earliest methods for interpolation of scattered data was proposed by Shepard [16]. Shepard interpolation is a weighted mean method that uses rational inverse-distance weights. Given data points $\{(x_i, y_i) : i = 1, \ldots, n\}$ with associating data values $\{F_i : i = 1, \ldots, n\}$, Shepard interpolation estimates the value at point (x, y) in the interpolation domain from

$$f(x, y) = \sum_{i=1}^{n} W_i(x, y) F_i, \tag{5}$$

where

$$W_i(x, y) = \frac{R_i(x, y)}{\sum_{j=1}^{n} R_j(x, y)}, \tag{6}$$

and

$$R_i(x, y) = \{(x - x_i)^2 + (y - y_i)^2\}^{-\frac{1}{2}}. \tag{7}$$

Function $f(x, y)$ interpolates the data without solving a system of equations. The interpolation value at (x, y) is obtained by simply evaluating the right side of Eq. (5).

Function $f(x, y)$ can be considered a single-valued surface interpolating 3-D points

$$\{(x_i, y_i, F_i) : i = 1, \ldots, n\}. \tag{8}$$

Table 1 Coordinates of 9 uniformly spaced points in the xy domain with associating height (data) values sampled from the plane in Fig. 1a

i	1	2	3	4	5	6	7	8	9
x_i	0	1	2	0	1	2	0	1	2
y_i	0	0	0	1	1	1	2	2	2
F_i	0	1	2	0	1	2	0	1	2

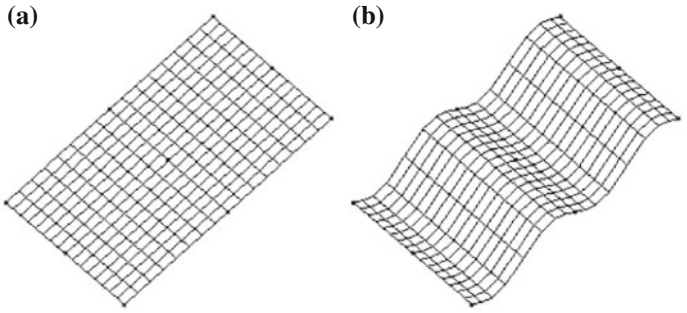


Fig. 1 **a** The planar surface from which the points in Table 1 are sampled, and **b** the surface interpolating the points in Table 1 by Shepard interpolation

The 3-D points listed in Table 1 represent uniformly spaced samples from the plane depicted in Fig. 1a. The surface interpolating the points as computed by the Shepard's method is depicted in Fig. 1b.

The reason for the flat horizontal spots at and around the data points is the nonlinear relation between xy and f. Flat horizontal spots result because similar interpolation values are obtained at and in the vicinity of each data point. This artifact can be removed by subjecting x and y to the same nonlinearity that f is subjected to. Representing x, y, and f as functions of new parameters u and v using Shepard's Eq. (5), we have

$$x(u,v) = \sum_{i=1}^{n} w_i(u,v)x_i, \tag{9}$$

$$y(u,v) = \sum_{i=1}^{n} w_i(u,v)y_i, \tag{10}$$

$$h(u,v) = \sum_{i=1}^{n} w_i(u,v)F_i, \tag{11}$$

where

$$w_i(u,v) = \frac{r_i(u,v)}{\sum_{j=1}^{n} r_j(u,v)}, \tag{12}$$

$$r_i(u,v) = \{(u-u_i)^2 + (v-v_i)^2\}^{-\frac{1}{2}}. \tag{13}$$

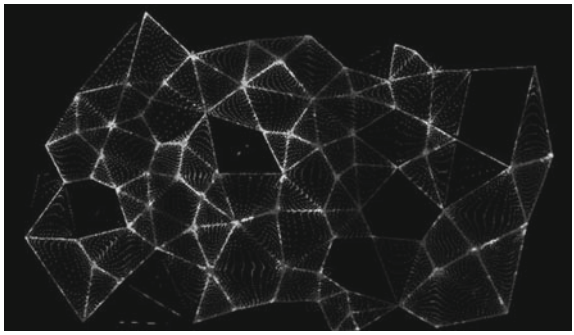


Fig. 2 Density of (x, y, f) points in PSI obtained at uniformly spaced parameters u and v

The ith data point has parameter coordinates (u_i, v_i), where $u_i = x_i/(n_c - 1)$ and $v_i = y_i/(n_r - 1)$. The integers n_c and n_r are, respectively, the number of columns and the number of rows in the reference image. As x is varied between 0 and $n_c - 1$, u varies between 0 and 1, and as y is varied between 0 and $n_r - 1$, v varies between 0 and 1.

An example of a parametric Shepard surface representing a component of a transformation function is given in Fig. 2. Parameters u and v are varied from 0 to 1 with increments 0.02 in both u and v to obtain the surface points shown in the figure. Uniformly spaced (u, v) coordinates produce surface points that have higher densities near the interpolation points and also near edges connecting the points. Although the highly varying density of surface points in Fig. 2 suggests a surface with a nearly polyhedral shape, the surface is actually very smooth and represents the height values shown in Fig. 11c. We would like to determine parameter coordinates (u, v) that correspond to (x, y) pixel coordinates in the reference image, and by using those parameter coordinates estimate $h(u, v)$ and use it as $f(x, y)$.

To find the interpolation value at (x, y), PSI requires the solution of two nonlinear equations to find the corresponding parameter coordinates (u, v). The obtained parameter coordinates are then used to find $h(u, v)$, which is considered the same as the value for $f(x, y)$. For image registration purposes, however, solution of nonlinear equations is not necessary. Surface coordinates that are within half a pixel of the actual coordinates are sufficient to resample the sensed image to the geometry of the reference image by the nearest-neighbor resampling method. Therefore, gaps between estimated surface points can be filled with required accuracy by bilinear interpolation of points obtained at uniformly spaced u and v if increments in u and v are sufficiently small.

The algorithm for calculating a component of a transformation function by PSI is described below. By notation "if $a \notin [b \pm 0.5]$," it is implied "if $a < b - 0.5$ or $a > b + 0.5$".

Algorithm PSI: Given 3-D points $\{(x_i, y_i, F_i) : i = 1, \ldots, n\}$, calculate entries of array $F[x, y]$ the size of the reference image with entry $[x, y]$ showing the X- or the

Y-component of the point in the sensed image corresponding to point (x, y) in the reference image, depending on whether F_i represents X_i or Y_i.

1. Let $u_i = x_i/(n_c - 1)$ and $v_i = y_i/(n_r - 1)$. This will map control points in the reference image to parameter coordinates in the range 0 to 1.
2. Initially, let increments in u and v be $\Delta u = 0.5$ and $\Delta v = 0.5$.
3. For $u = 0$ to 1 with increment Δu and for $v = 0$ to 1 with increment Δv:

   If $[x(u, v) + x(u + \Delta u, v)]/2 \notin [x(u + \Delta u/2, v) \pm 0.5]$ or if $[y(u, v) + y(u + \Delta u, v)]/2 \notin [y(u + \Delta u/2, v) \pm 0.5]$ or if $[h(u, v) + h(u + \Delta u, v)]/2 \notin [h(u + \Delta u/2, v) \pm 0.5]$ or,

   if $[x(u, v) + x(u, v + \Delta v)]/2 \notin [x(u, v + \Delta v/2) \pm 0.5]$ or if $[y(u, v) + y(u, v + \Delta v)]/2 \notin [y(u, v + \Delta v/2) \pm 0.5]$ or if $[h(u, v) + h(u, v + \Delta v)]/2 \notin [h(u, v + \Delta v/2) \pm 0.5]$ or,

   if $[x(u, v) + x(u + \Delta u, v) + x(u, v + \Delta v) + x(u + \Delta u, v + \Delta v)]/4 \notin [x(u + \Delta u/2, v + \Delta v/2) \pm 0.5]$ or if $[y(u, v) + y(u + \Delta u, v) + y(u, v + \Delta v) + y(u + \Delta u, v + \Delta v)]/4 \notin [y(u + \Delta u/2, v + \Delta v/2) \pm 0.5]$ or if $[h(u, v) + h(u + \Delta u, v) + h(u, v + \Delta v) + h(u + \Delta u, v + \Delta v)]/4 \notin [h(u + \Delta u/2, v + \Delta v/2) \pm 0.5]$,

   reduce Δu and Δv by a factor of 2 and go back to Step 3. Otherwise, continue. (This step determines the largest increment in u and v that can produce an accuracy of half a pixel or better in image resampling.)
4. For $u = 0$ to 1 with increment Δu and for $v = 0$ to 1 with increment Δv:
   Calculate $[x(u, v), y(u, v), h(u, v)]$, $[x(u+\Delta u, v), y(u+\Delta u, v), h(u+\Delta u, v)]$, $[x(u + \Delta u, v + \Delta v), y(u + \Delta u, v + \Delta v), h(u + \Delta u, v + \Delta v)]$, $[x(u, v + \Delta v), y(u, v + \Delta v), h(u, v + \Delta v)]$. Then estimate values within each local patch defined by parameters $[u, u + \Delta u] \times [v, v + \Delta v]$ by bilinear interpolation of values at the four corners of the patch, saving the estimated surface value for $h(u, v)$ at $F[x(u, v), y(u, v)]$.
5. For $i = 1, \ldots, n$:
   If $F_i \notin [h(u_i, v_i) \pm 0.5]$, reduce Δu and Δv by a factor of 2 and go back to Step 4.
6. Return array F.

Analysis of Algorithm PSI: Step 3 of the algorithm recursively subdivides each surface patch into 4 smaller patches until the distance between the center of each patch to the center of its base and the distances between midpoints of its bounding curves to the corresponding base edges fall below half a pixel. When the approximation error for all patches falls below half a pixel, each patch is replaced with its base, which is the bilinear interpolation of the four corners of the patch. This will be the speed-up achieved by not calculating the surface value at all pixels in the reference image directly and instead estimating some of the values by bilinear interpolation.

Computation of Step 3 of the algorithm is depicted in Fig. 3a. For a patch with corner points at (u, v), $(u + \Delta u, v)$, $(u, v + \Delta v)$, and $(u + \Delta u, v + \Delta v)$, distances of the midpoints of the bounding curves of the patch to the midpoint of the line connecting the corresponding bounding curve endpoints are computed. When all such distances fall below half a pixel, and also when the distance of the center of the patch at $(u + \Delta/2u, v + \Delta v/2)$ to the center of the base of the patch obtained from the average of its four corners becomes smaller than half a pixel in x, y, and F directions, subdivision is stopped.

Note that since two adjacent patches share a bounding curve, for each patch there is a need to carry out the computations at only two of the bounding curves. The bounding curves at opposing sides of the patch are considered when processing the opposing patches. By computing the errors at midpoints of two bounding curves for each patch, the bounding curves of patches with parameter $u = 1$ or parameter $v = 1$ will remain. After computing errors at all patches, errors are computed at patches with boundary parameters $u = 1$ or $v = 1$, and if all such errors become smaller than half a pixel, subdivision of Step 3 is stopped. If the smaller side of an image contains M pixels and m is the largest number such that $2^m < M$, Step 3 needs to be repeated m or fewer times. For instance, when the reference image is of size 1200×1600 pixels, Step 3 will be repeated a maximum of 10 times.

After estimating the largest increment in u and v to obtain desired accuracy at the center of each patch as well as at midpoints of its bounding curves, the patch is approximated in Step 4 by bilinear interpolation of its four corners.

To ensure that the employed bilinear interpolation creates an overall surface that is within half a pixel of the points it is supposed to interpolate, in Step 5, the surface values estimated in Step 4 are compared with the given F_i values, which are, in effect, X_i or Y_i. If the difference between any such values is larger than half a pixel, Δu and Δv are reduced by a factor or 2 and Steps 4 and 5 are repeated until the approximating surface falls within half a pixel of the required values. Note that in Step 3 the patches are not generated; only values at bounding curve midpoints and patch centers are calculated. In most situations, this finds the required increment in u and v to produce the required accuracy in interpolation. In rare cases, the process may not produce a surface sufficiently close to the interpolation points. Step 5 is included to ensure that the obtained surface does, in fact, pass within half a pixel of the points it is supposed to interpolate (Fig. 3b).

It should be mentioned that due to the nonlinear relation between (x, y) and (u, v), by varying u and v from 0 to 1, the computed x and y values may not completely cover the image domain. To cover the image domain in its entirety, it may be necessary to start u and v from values slightly below 0 and continue to values slightly past 1 to cover all pixels in the image domain. This requires decrementing u from 0 by Δu and v from 0 by Δv in Step 4 of the algorithm until values for pixels along the lower and left sides of the reference image are obtained. Also, it is required to increment u from 1 by Δu and v from 1 by Δv in Step 4 of the algorithm until values for pixels along the right and upper sides of the reference image are obtained.

Algorithm PSI uses the same Δu and Δv everywhere when calculating the surface points. Since density of points vary across the interpolation domain, Δu and Δv can

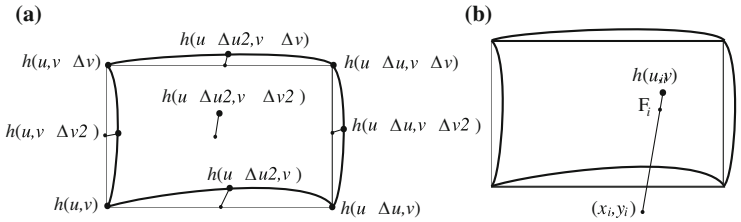


Fig. 3 **a** The subdivision method used in Step 3 of Algorithm PSI. **b** Ensuring the approximating surface passes within half a pixel of the point it is supposed to interpolate in Step 5 of the algorithm

be made local so that after each subdivision those patches that are within half a pixel of their base are not subdivided, and only patches that are farther from their bases by more than half a pixel are subdivided. This requires keeping track of each patch individually. The bookkeeping time involved in doing so is usually higher than the time saved by not subdividing some patches.

If the reference image contains N pixels and n corresponding points are available, the worst case requires using Eqs. (9)–(11) to calculate the X- and the Y-component of the transformation at all pixels. In that case, the computational complexity of the algorithm will be of order Nn. If Δu and Δv are relatively large so that interpolation values at many pixels are calculated from bilinear interpolation of known values, computation time reduces and, at best, the computational complexity of the algorithm will be of order N. Therefore, the computational complexity of the algorithm when the reference image contains N pixels and n correspondences are available will be between N and Nn depending on whether the geometry to be estimated is simple, such as a plane, or very complex, such as the geometric difference between images showing different views of an urban scene.

3 Evaluation

Various interpolation functions may be used as the components of a transformation function in image registration. Properties desired of a transformation function and generally of an interpolation function are:

1. *Monotonicity, convexity, and nonnegativity preserving*: These properties ensure that a chosen transformation function is well behaved and does not produce high fluctuations and overshoots in estimated values. Such properties are generally achieved by implicitly or explicitly imposing geometric gradients at the control points, making the interpolating surface take desired shapes at the control points. Lu and Schumaker [11] and Li [10] derive monotonicity-preserving conditions, Lai [7], Renka [13], and Schumaker and Speleers [15] derive convexity-preserving conditions, and Schumaker and Speleers [14] and

Hussain and Hussain [6] derive nonnegativity preserving conditions for piecewise smooth interpolation of data at irregularly spaced points.

2. *Linearity preserving*: If data values in the image domain vary linearly, the function interpolating the data should also vary linearly. This property ensures that a transformation function does not introduce nonlinearity into the resampled image when reference and sensed images are related linearly.
3. *Adaptive to irregular spacing of the control points*: Since control points in an image are rarely uniformly spaced, a transformation function should have the ability to adapt to the local density and organization of the control points. Spacing between the control points across the image domain can vary greatly. If the transformation function is defined by rational basis functions, the shapes of the functions adapt well to the spacing between the points.

From Eqs. (9)–(11), we see that at the vicinity of the ith data site,

$$a \equiv \frac{\Delta h(u, v)}{\Delta x(u, v)} = \frac{F_i}{x_i} \tag{14}$$

and

$$b \equiv \frac{\Delta h(u, v)}{\Delta y(u, v)} = \frac{F_i}{y_i}. \tag{15}$$

Therefore, the surface at the vicinity of the ith data point takes slopes $(F_i/x_i, F_i/y_i)$. Note that in the traditional Shepard interpolation the slopes of the interpolating surface at each data site are 0, resulting in a flat horizontal spot. In parametric Shepard, the slopes of the interpolating surface at an interpolation point are no longer 0 and represent slopes of plane

$$a(x - x_i) + b(y - y_i) + (F - F_i) = 0. \tag{16}$$

Since x monotonically increases with u and y monotonically increases with v, and x and y are single-valued functions of u and v, for any unique (u, v) a unique (x, y) is obtained. Therefore, the obtained interpolating surface does not contain folds and for any unique (x, y) a single interpolation value is obtained. This property is required of a component of a transformation function in an image registration.

The interpolating surface is defined by a convex combination of tangent planes of the form given by Eq. (16). Since the tangent planes extend beyond the convex-hull of the interpolation points, the interpolating surface covers the entire image domain. This makes it possible to find for each point in the reference image, the corresponding point in the sensed image, making it possible to establish correspondence between all points in the images. Note that a convex combination of the tangent planes passing through the interpolation points is not the same as the convex combination of the interpolating points. Therefore, like TPS, PSI may produce overshoots away from the interpolation points depending on the arrangement of the data points (the control points in the reference image) and the highly varying nature of the data values.

Fig. 4 The *face image* set, showing a facial stone carving captured from different distances and views of the scene. Corresponding control points in the image are also shown. The control points marked with a *black* '+' are used to determine the transformation function and the control points marked with a *white* '+' are used to determine the registration accuracy

To determine the suitability of PSI in image registration, experiments were carried out using the image sets shown in Figs. 4, 5, 6, 7, 8, and 9. These image sets exhibit varying degrees of geometric differences. The control points used to obtain the transformation function in each case are also shown. Methods to find corresponding control points in images of a scene can be found in [1]. The control points marked with a black '+' are used to determine a transformation function and the control points marked with a white '+' are used to determine the registration accuracy using the obtained transformation function.

The images in Fig. 4 are captured from different views of a facial stone carving. The geometric difference between the images varies locally. We will refer to these images as *Face* images. There are 80 control points in each image. Forty of the control points are used to estimate the components of the transformation function, and the remaining 40 are used to evaluate the registration accuracy with the obtained transformation function.

The images in Fig. 5 represent aerial images taken from different views and distances of a few buildings. The images contain small local and global geometric differences. We will refer to these as the *Aerial* images. There are 31 corresponding control points in the images, of which 16 are used to estimate the transformation function and the remaining 15 are used to evaluate the registration accuracy with the obtained transformation.

The images in Fig. 6 represent two views of a terrain scene. There is depth discontinuity near the center of the images. We will call this the *Terrain* image set.

Fig. 5 The *aerial image* set, representing two aerial images captured from different views and distances of a few buildings

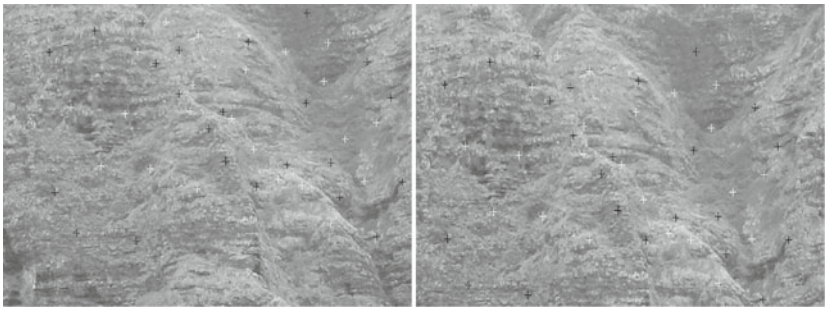


Fig. 6 The *terrain image* set, representing two views of a terrain scene

There are 46 corresponding control points in the images, of which half are used to determine the transformation function and the remaining half are used to evaluate the registration accuracy.

The images in Fig. 7 show close-up views of a small area in the terrain scene. The images in this set will be referred to as the *Rock* images. The geometric difference between the images varies greatly across the image domain. There are 58 corresponding control points in the images, of which half are used to estimate the transformation function and the remaining half are used to evaluate the registration accuracy.

The images in Fig. 8 show two views of a partially snow-covered rocky mountain. We will refer to these as *Mountain* images. The geometric difference between the images varies considerably across the image domain. There are 165 corresponding control points in the images, of which 83 are used to determine the transformation function and the remaining 82 are used to evaluate the registration accuracy.

The images in Fig. 9 show a parking lot taken from the same viewpoint but with different views. These images are related by a homography. We will refer to these images as the *Parking* images. The images contain 32 corresponding control points, of which half are used to find the components of the transformation function and the remaining half are used to determine the registration accuracy.

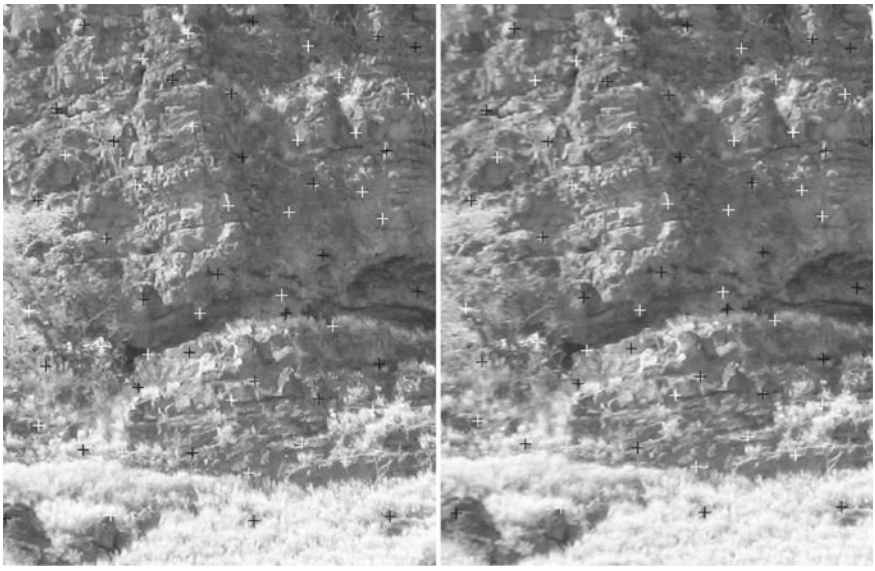


Fig. 7 The *rock image* set, showing close-up views of a small area in the terrain scene

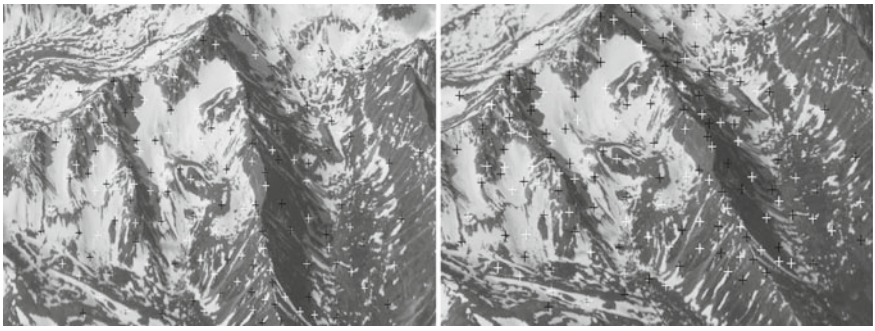


Fig. 8 The *mountain image* set, representing two views of a snowy mountain with sharp peaks and valleys

We will compare the speed and accuracy of PSI with those obtained by MQ, TPS, and MLS in the registration of these six sets of images. For each method, the time to determine the transformation function plus the time to resample the sensed image to the geometry of the reference image is determined. Since the true geometric relation between the images is not known, half of the control points are used to determine the transformation parameters and the remaining half are used to measure the accuracy of the transformation in mapping the remaining control points in the sensed image to the corresponding control points in the reference image. Points marked in black '+' in Figs. 4, 5, 6, 7, 8 and 9 are used to determine a transformation function and

Fig. 9 The *parking image* set, captured from the same viewpoint of a parking lot but with slightly different view angles

points marked in white '+' are used to determine the registration accuracy with the obtained transformation function.

The components of a transformation function are calculated by (1) MQ, (2) TPS, (3) MLS, and (4) PSI. Then, root-mean-squared

$$\mathrm{RMS} = \sqrt{\frac{1}{n'}\sum_{j=1}^{n'}\big(X_j - f(x_j, y_j)\big)^2 + \big(Y_j - g(x_j, y_j)\big)^2} \tag{17}$$

and maximum

$$\mathrm{MAX} = \max_{j=1}^{n'}\left\{\sqrt{\big(X_j - f(x_j, y_j)\big)^2 + \big(Y_j - g(x_j, y_j)\big)^2}\right\} \tag{18}$$

errors in finding known corresponding points are calculated and used to evaluate the registration accuracy. Here, n' represents the number of control-point correspondences not used to estimate the transformation parameters but are only used to determine the registration accuracy. Errors obtained by the four methods on the six image sets are shown in Table 2. These results show that thin-plate spline has the highest speed in spite of the fact that it solves a system of equations to find each component of a transformation function. This happens to be the case when there are up to a few hundred corresponding control points in the images.

A single method could not produce the best RMS or MAX error for all images and methods vary in accuracy depending on the organization of the points and the severity of the geometric difference between the images. Most frequently, best accuracy is

Table 2 Performance measures of multiquadric (MQ), thin-plate spline (TPS), moving least-squares (MLS), and parametric Shepard interpolation (PSI) in registration of the Face, Aerial, Terrain, Rock, Mountain, and Parking image sets

Method	Measure	Face	Aerial	Terrain	Rock	Mountain	Parking
	TIME	1.34	0.19	0.73	0.70	2.48	0.39
MQ	RMS	4.05	6.80	10.28	4.08	4.62	5.89
	MAX	**9.00**	13.89	26.38	9.10	30.62	14.33
	TIME	**1.09**	**0.14**	**0.58**	**0.61**	**1.93**	**0.31**
TPS	RMS	**3.85**	1.34	2.16	**1.51**	**4.47**	**0.98**
	MAX	10.68	2.43	4.26	3.34	32.18	1.79
	TIME	1.98	0.41	1.15	1.06	3.35	0.67
MLS	RMS	3.96	**1.16**	**1.62**	1.52	5.46	0.95
	MAX	9.32	**2.13**	**3.40**	3.69	33.17	**1.45**
	TIME	1.93	0.27	1.05	1.05	2.98	0.69
PSI	RMS	4.32	1.38	1.79	1.59	4.91	1.13
	MAX	11.64	2.35	5.10	**3.04**	33.97	1.70

Performance measures are: computation time in seconds (TIME), root-mean-squared (RMS) error in pixels, and maximum (MAX) error, also in pixels. The transformation functions are determined using half of the corresponding control points in each image set, and registration errors are computed using the remaining half. The best and the worst performances obtained for each image set are shown in *bold* and *underlined*, respectively

achieved by MLS while worst accuracy is achieved by MQ. TPS is the fastest in all cases while MLS is the slowest in most cases.

To view the quality of registration achieved by the four methods, registration of the Mountain image set by the four methods is shown in Fig. 10. MQ is accurate within the convex-hull of the control points. But errors are very large outside the convex-hull of the control points, contributing to high MAX errors in all except one of the datasets.

Considering both speed and accuracy, we see that the overall best results are obtained by TPS on the image sets tested. PSI has been a stable method with a performance that has been close to the best method in each case. It does not require the solution of a system of equations, giving it a great advantage over TPS and MQ, which require the solution of large systems of equations when a large set of correspondences is given. When thousands of control points are available, solving large systems of equations not only slows the computation, depending on the organization of the points, the systems of equations to be solved may become ill-conditioned and unsolvable.

A transformation function is required to be locally sensitive. This is needed so that an error in the location of a control point does not influence registration of the entire image. Among the methods tested, PSI and MLS are locally sensitive and are suitable for the registration of images containing some inaccurate correspondences. An example to demonstrate this is given below. The components of the transformation obtained from the 82 corresponding points in the Mountain image set (Fig. 8) by TPS and PSI are shown in Fig. 11. Intensity at (x, y) in a component of a trans-

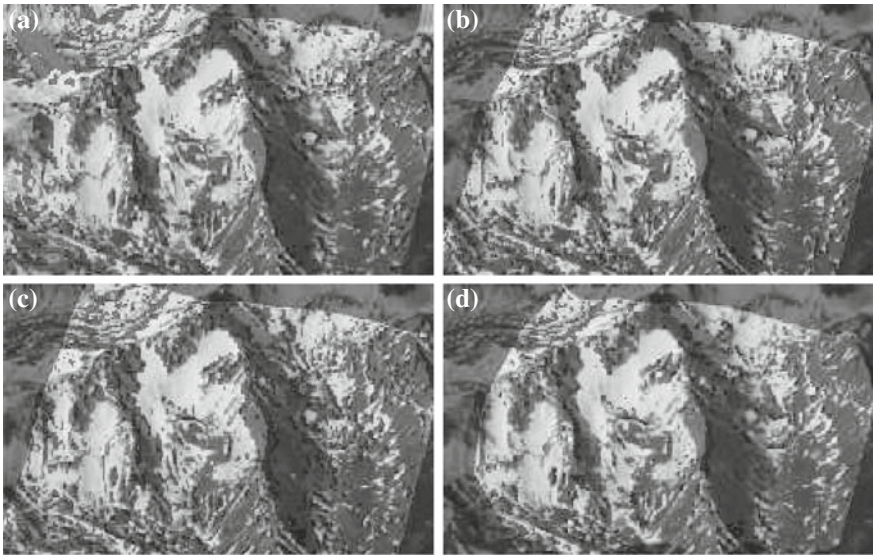


Fig. 10 Resampling of the sensed image and overlaying with the reference image in the Mountain image set using **a** MQ, **b** TPS, **c** MLS, and **d** PSI. The *dark spots* appearing in these images are areas where registration is poor. Areas near the image borders in the reference image that do not appear in the sensed image are shown darker

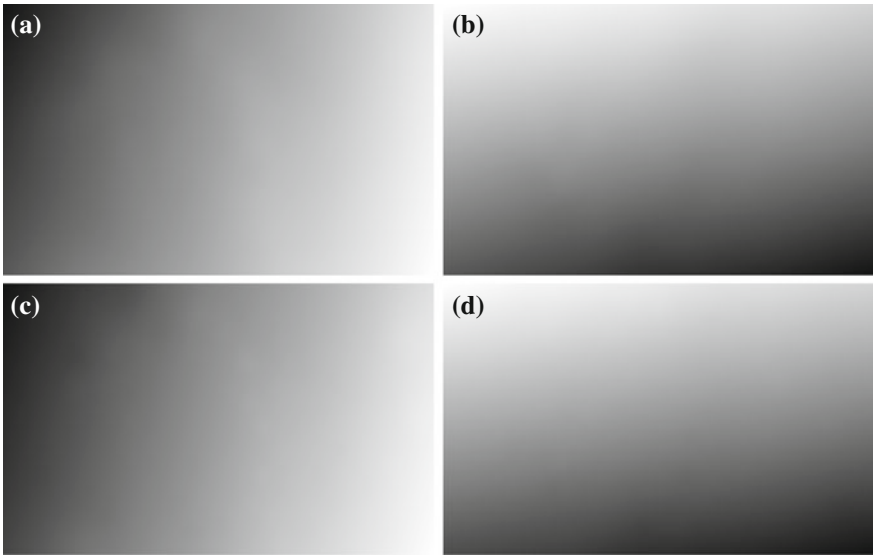


Fig. 11 **a** X-component and **b** Y-component of the transformation function obtained by TPS using the control-point correspondences depicted in Fig. 8. **c** X-component and **d** Y-component of the transformation obtained by PSI. The intensity at (x, y) in these images is set proportional to the X-coordinate and the Y-coordinate of the point in the sensed image corresponding to point (x, y) in the reference image

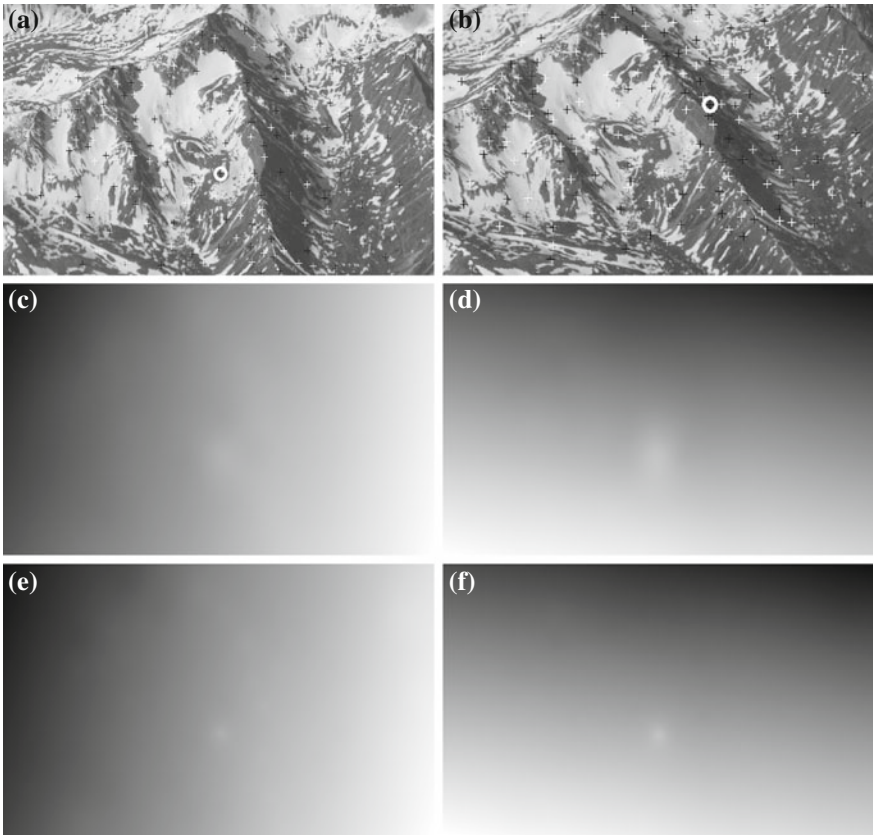


Fig. 12 **a**, **b** These images are the same as those shown in Fig. 8 except for moving one control point in the sensed image, introducing an incorrect correspondence. The incorrect corresponding points are encircled in these images. **c** X-component and **d** Y-component of the transformation function obtained by TPS using the control-point correspondences shown in (**a**) and (**b**). **e** X-component and **f** Y-component of the transformation function obtained by PSI

formation is proportional to the X- or the Y-component of the point in the sensed image corresponding to point (x, y) in the reference image. When correspondences are accurate, the components of transformation obtained by the two methods are very similar.

By moving one of the control points in the sensed image, we create a pair of points that do not correspond to each other. The incorrect correspondence pair are encircled in Fig. 12a, b. The components of the transformation obtained by TPS are shown in Fig. 12c, d and those obtained by PSI are shown in Fig. 12e, f. The bright spot in a component of a transformation shows the location of the error and can be used as a guide to identify the incorrect correspondence. While PSI keeps such errors local, TPS spreads the errors to a wider area, affecting the registration of more pixels.

Sharper details are obtained in the components of the transformation obtained by PSI when compared to those obtained by TPS. This shows that PSI can accommodate sharper geometric differences between the images than TPS.

4 Concluding Remarks

To register two images, not only is a set of corresponding control points from the images required, a transformation function is required that can use information about the correspondences to estimate correspondence between the remaining points in the images. A transformation function fills in the gaps between corresponding control points, establishing correspondence between all points in the images.

Comparing the performances of MQ, TPS, MLS, and PSI using six sets of images with varying degrees of local and global geometric differences, it is found that although none of the transformation functions can outperform all others on all image sets, some transformation functions generally perform better than others. Among the transformation functions tested, MLS and PSI are found to be the most stable, producing consistent accuracies on all image sets. With the six image sets tested, TPS is found to produce the lowest RMS error for most cases while requiring the least computation time.

Both MQ and TPS are global methods, and so an inaccurate correspondence can influence registration accuracy of a large area in the image domain. MLS is a locally sensitive method in the sense that an inaccurate correspondence affects the registration of pixels mostly in the neighborhood of the inaccurate correspondence. Although PSI is defined globally, but since its weight functions are monotonically decreasing, an inaccurate correspondence affects registration accuracy of points mostly in its neighborhood. The influence of an inaccurate correspondence on registration of distant points becomes negligible and vanishes beyond a certain point due to the quantization step involved in image resampling.

The main contribution of this work is considered to be parametric formulation of the Shepard interpolation, removing its weakness of creating flat horizontal spots at the data points while maintaining its strength of not requiring the solution of a system of equations to find the coefficients of the function. Experimental results show that although the speed and accuracy of PSI do not surpass those of top performing interpolation methods in the literature, its speed and accuracy are close to those of top performing methods.

Overall, if up to a few hundred correspondences is available and the correspondences are known to be accurate, TPS is the method of choice. If more than a few hundred correspondences are available or if some correspondences are known to be inaccurate, MLS is the method of choice. The proposed PSI is faster than MLS, but its accuracy falls short of MLS. PSI is the only method that does not require the solution of a system of equations; therefore, it is the most stable method among those tested, always producing a result independent of the severity of the geometric difference between the images.

Acknowledgments The author would like to thank the reviewers for their insightful comments, the State of Ohio for support of this work, and Image Registration and Fusion Systems for the images used in this study. The editorial assistance of Libby Stephens in preparation of this manuscript is also greatly appreciated.

References

1. A. Goshtasby, Image Registration: Principles, Tools, and Methods, Springer, 2012.
2. Goshtasby, A.: Registration of images with geometric distortions. IEEE Trans. Geosci. Remote Sens. **26**(1), 60–64 (1988)
3. Harder, R.L., Desmarais, R.N.: Interpolation using surface splines. J. Aircraft **9**(2), 189–191 (1972)
4. Hardy, R.L.: Multiquadric equations of topography and other irregular surfaces. J. Geophys. Res. **76**(8), 1905–1915 (1971)
5. Hardy, R.L.: Theory and applications of the multiquadric-biharmonic method—20 years of discovery—1969–1988. Comput. Math. Appl. **19**(8/9), 163–208 (1990)
6. Hussain, M.Z., Hussain, M.: C^1 positive scattered data interpolation. Comput. Math. Appl. **59**, 457–567 (2010)
7. Lai, M.-J.: Convex preserving scattered data interpolation using bivariate C^1 cubic spline. J. Comput. Appl. Math. **119**, 249–258 (2000)
8. Lancaster, P., Šalkauskas, K.: Surfaces generated by moving least squares methods. Math. Comput. **37**(155), 141–158 (1981)
9. Lancaster, P. Šalkauskas, K.: Curve and Surface Fitting: An Introduction, pp. 55–62, 225–244. Academic Press, New York (1986)
10. Li, A.: Convexity preserving interpolation. Comput. Aided Geom. Des. **16**, 127–147 (1999)
11. Lu, H., Schumaker, L.L.: Monotone surfaces to scattered data using C^1 piecewise cubics. SIAM J. Numer. Anal. **34**(2), 596–585 (1997)
12. Meinguet, J.: An intrinsic approach to multivariate spline interpolation at arbitrary points. In: Sahney, B.N. (ed.) Polynomial and Spline Approximation, pp. 163–190. D. Reidel Publishing Company, Dordrecht (1979)
13. Renka, R.J.: Algorithm 833: CSRFPAXK-interpolation of scattered data with a C^1 convexity-preserving surface. ACM Trans. Math. Softw. **30**(2), 200–211 (2004)
14. Schumaker, L.L., Speleers, H.: Nonnegativity preserving macro-element interpolation of scattered data. Comput. Aided Geom. Des. **27**, 245–261 (2010)
15. Schumaker, L.L., Speleers, H.: Convexity preserving splines over triangulations. Comput. Aided Geom. Des. **28**, 270–284 (2011)
16. Shepard, D.: A two-dimensional interpolation function for irregularly spaced data, In: Proceedings of 23rd National Conference ACM, pp. 517–524 (1968)

Parabolic Molecules: Curvelets, Shearlets, and Beyond

Philipp Grohs, Sandra Keiper, Gitta Kutyniok and Martin Schäfer

Abstract Anisotropic representation systems such as curvelets and shearlets have had a significant impact on applied mathematics in the last decade. The main reason for their success is their superior ability to optimally resolve anisotropic structures such as singularities concentrated on lower dimensional embedded manifolds, for instance, edges in images or shock fronts in solutions of transport dominated equations. By now, a large variety of such anisotropic systems have been introduced, for instance, second-generation curvelets, bandlimited shearlets, and compactly supported shearlets, all based on a parabolic dilation operation. These systems share similar approximation properties, which are usually proven on a case-by-case basis for each different construction. The novel concept of parabolic molecules, which was recently introduced by two of the authors, allows for a unified framework encompassing all known anisotropic frame constructions based on parabolic scaling. The main result essentially states that all such systems share similar approximation properties. One main consequence is that at once all the desirable approximation properties of one system within this framework can be deduced virtually for any other system based on parabolic scaling. This paper motivates and surveys recent results in this direction.

Keywords Curvelets · Nonlinear approximation · Parabolic scaling · Shearlets

P. Grohs (✉)
Seminar for Applied Mathematics, ETH Zürich, 8092 Zürich, Switzerland
e-mail: philipp.grohs@sam.math.ethz.ch

S. Keiper · G. Kutyniok · M. Schäfer
Department of Mathematics, Technische Universität Berlin, 10623 Berlin, Germany
e-mail: keiper@math.tu-berlin.de

G. Kutyniok
e-mail: kutyniok@math.tu-berlin.de

M. Schäfer
e-mail: schaefer@math.tu-berlin.de

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_9,

1 Introduction

Wavelets have had a tremendous impact on applications requiring an efficient representation system such as image compression or PDE solvers. However, multivariate data does typically exhibit the distinct property of being governed by anisotropic features, whose wavelets—as an isotropic system—are not capable of resolving optimally in the sense of optimal approximation rates. In imaging sciences, this fact is even backed up by neurophysiology, since it is generally accepted today that neurons are highly directional-based, thereby reacting most strongly to curvelike structures.

This observation has led to the introduction of various novel representation systems, which are designed to accommodate the anisotropic nature of most multivariate data. The considered model situation are functions with singularities along lower dimensional embedded manifolds such as edges or rays in imaging applications, with the goal to provide optimally sparse approximations of these objects. Some of the most well-known termed *directional representation systems* nowadays are ridgelets [4], curvelets [5], and shearlets [19, 28]. With the introduction of such a variety of systems, the appeal has grown to extract the underlying principles of these new constructions and build an abstract common framework that can unite many of these systems "under one roof." The framework should be general enough to include as many constructions as possible, while on the other hand, it should also be specific enough to still capture their main features and properties. Such a framework would help to gain deeper insights into the properties of such systems. Moreover, it bears an obvious economical advantage. Up to now the properties of each new system, e.g., their approximation rates of anisotropic features, have been proven more or less from scratch, although the proofs often resemble one another in many ways. From the higher level viewpoint provided by such a framework, it becomes possible to provide proofs which build upon abstract properties, and are therefore independent of the specific constructions. Thus, results can be established for many systems simultaneously.

The introduction of *parabolic molecules* in 2011 by two of the authors [17] was the first step in this direction. A system of parabolic molecules can be regarded as being generated from a set of functions via parabolic dilations, rotations, and translations. Each element in a system of parabolic molecules is therefore naturally associated with a certain scale, orientation, and spatial location. The central conceptual idea is now to allow the generators to vary, as long as they obey a prescribed time-frequency localization, which also explains the terminology “molecules.”

At the heart of this is the fundamental observation that it is the foremost time-frequency localizations of the functions in a system that determines its properties and performance. This concept of *variable generators*, where in the extreme case every element is allowed to have its own individual generator, is a key feature of the framework and gives it a great amount of flexibility. Additional flexibility is achieved by *parameterizations* to allow generic indexing of the elements. Another fruitful idea is the relaxation of the rigid vanishing moment conditions imposed

on the generators of most classical constructions by requiring the moments to only *vanish asymptotically* at high scales without changing the asymptotic behavior of the approximation.

It was shown in [17] that the concept of parabolic molecules can unify shear-based and rotation-based constructions under one roof. In particular, it enables to treat the classical shearlets and curvelets simultaneously, although these specific constructions are based on different construction principles: For curvelets the scaling is done by a dilation with respect to polar coordinates and the orientation is enforced by rotations. Shearlets, on the other hand, are based on affine scaling of a single generator and the directionality is generated by the action of shear matrices. As an example application, in [17] parabolic molecules were used to show that these systems feature a similar approximation behavior, thereby not only unifying the approximation results for curvelets [5] and shearlets [20, 27], but proving optimal sparse approximations for a much larger class of systems belonging to the class of parabolic molecules.

Our exposition is organized as follows: We begin with a general introduction to the problem of sparsely representing multivariate data in Sect. 2. The main issue with such data is the possible occurrence of anisotropic phenomena, which otherwise impairs the good performance of classical wavelet systems. This motivates the need for so-called directional representation systems, some classical constructions of which we present in Sect. 3, namely classical curvelets and shearlets. Here we emphasize their similar approximation performance, which is almost optimal for cartoon-like images.

After this exposition we turn to parabolic molecules as a unifying framework. We first establish the basic concepts in Sect. 4 and state one main result, namely the cross-Gramian of two systems of parabolic molecules exhibits a strong off-diagonal decay. This property will become essential in Sect. 6, where we discuss the approximation behavior of parabolic molecules. Before moving there, however, we pause for a while in Sect. 5 to illustrate the versatility of the framework by giving some examples. After we have convinced the reader of their applicability, we then turn to the section on approximation, where we essentially prove that any two systems of parabolic molecules, which are consistent and have sufficiently high order, exhibit the same approximation behavior.

2 Representation of Multivariate Data

Most applications require efficient encoding of multivariate data in the sense of optimal (sparse) approximation rates by a suitable representation system. This is typically phrased as a problem of best N-term approximation (see Sect. 2.1). The performance of an approximation scheme is then usually analyzed with respect to certain subclasses of the Hilbert space $L^2(\mathbb{R}^d)$, which is the standard continuum domain model for d-dimensional data, in particular, in imaging science. As elaborated upon before, the key feature of most multivariate data is the appearance of anisotropic

phenomena. Hence, such a subclass of $L^2(\mathbb{R}^d)$ is required to provide a suitable model for this fact, which, for $d = 2$, is fulfilled by the subclass of so-called cartoon-like images as introduced in Sect. 2.2. It can then be easily seen that wavelets do not deliver optimal approximation rates (Sect. 2.3), which then naturally leads to the theory of directional representation systems.

In the sequel, we will use the "analyst's brackets" $\langle x \rangle := \sqrt{1 + x^2}$, for $x \in \mathbb{R}$. Also, for two quantities $A, B \in \mathbb{R}$, which may depend on several parameters we shall write $A \lesssim B$, if there exists a constant $C > 0$ such that $A \leq CB$, uniformly in the parameters. If the converse inequality holds true, we write $A \gtrsim B$ and if both inequalities hold, we shall write $A \asymp B$.

2.1 Sparse Approximation

We start by briefly discussing some aspects of approximation theory. From a practical standpoint, a function $f \in L^2(\mathbb{R}^2)$ is a rather intractable object. In order to analyze f, the most common approach is to represent it with respect to some representation system $(m_\lambda)_{\lambda \in \Lambda} \subseteq L^2(\mathbb{R}^2)$, i.e., to expand f as

$$f = \sum_{\lambda \in \Lambda} c_\lambda m_\lambda, \tag{1}$$

and then consider the coefficients $c_\lambda \in \mathbb{R}$. In practice we have to account for noise, hence it is necessary to ensure the robustness of such a representation. This leads to the notion of a frame (cf. [8, 9]).

A frame is a generalization of the notion of an orthonormal basis to include redundant systems while still ensuring stability. More precisely, a system $(m_\lambda)_{\lambda \in \Lambda} \subseteq L^2(\mathbb{R}^2)$ forms a *frame* for $L^2(\mathbb{R}^2)$, if there exist constants $0 < A \leq B < \infty$ such that

$$A \|f\|_2^2 \leq \sum_{\lambda \in \Lambda} |\langle f, m_\lambda \rangle|^2 \leq B \|f\|_2^2 \quad \text{for all} \quad f \in L^2(\mathbb{R}^2).$$

A frame is called *tight*, if $A = B$ is possible, and *Parseval*, if $A = B = 1$. Since the *frame operator* $S : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ defined by $Sf = \sum_{\lambda \in \Lambda} \langle f, m_\lambda \rangle m_\lambda$ is invertible, it follows that one sequence of coefficients in (1)—note that for a redundant system this sequence is not unique anymore—can be computed as

$$c_\lambda = \langle f, S^{-1} m_\lambda \rangle, \quad \lambda \in \Lambda,$$

where $(S^{-1} m_\lambda)_\lambda$ is usually referred to as the *canonical dual frame*. This particular coefficient sequence has the distinct property that it minimizes the ℓ_2-norm.

When representing f with respect to a frame $(m_\lambda)_\lambda \subseteq L^2(\mathbb{R}^2)$, we are confronted with yet another problem. Since in real-world applications infinitely many coefficients are infeasible, the function f has to be approximated by a finite subset of

this system. Letting N be the number of elements allowed in this approximation, we obtain what is called an *N-term approximation* for f with respect to $(m_\lambda)_\lambda$. The *best N-term approximation*, typically denoted by f_N, is optimal among those in terms of a minimal approximation error and is defined by

$$f_N = \operatorname*{argmin}_{(c_\lambda)_{\lambda\in\Lambda_N}} \|f - \sum_{\lambda\in\Lambda_N} c_\lambda m_\lambda\|_2^2 \quad \text{subject to} \quad \#\Lambda_N \leq N.$$

An appropriate measure for the approximation behavior of a system $(m_\lambda)_\lambda$ for a subclass $\mathscr{C}$, say, of $L^2(\mathbb{R}^2)$ is the decay of the L^2-error of the best N-term approximation $\|f - f_N\|_2$ as $N \to \infty$, thus the *asymptotic approximation rate*. As discussed before, the representation system might not form an orthonormal basis in which case the computation of the best N-term approximation is far from being understood. The delicacy of this problem can, for instance, be seen in [13]. A typical approach to circumvent this problem is to consider instead the N-term approximation by the N largest coefficients $(c_\lambda)_{\lambda\in\Lambda}$. It is evident that this error also provides a bound for the error of best N-term approximation.

There indeed exists a close relation between the N-term approximation rate achieved by a frame and the decay rate of the corresponding frame coefficients. By measuring this decay rate in terms of the ℓ_p-(quasi)-norms for $p > 0$, the following lemma shows that membership of the coefficient sequence to an ℓ_p-space for small p implies "good" N-term approximation rates. For the proof, we refer to [10, 27].

Lemma 1 *Let $f = \sum c_\lambda m_\lambda$ be an expansion of $f \in L^2(\mathbb{R}^2)$ with respect to a frame $(m_\lambda)_{\lambda\in\Lambda}$. Further, assume that the coefficients satisfy $(c_\lambda)_\lambda \in \ell^{2/(2k+1)}$ for some $k > 0$. Then the best N-term approximation rate is at least of order N^{-k}, i.e.*

$$\|f - f_N\|_2 \lesssim N^{-k}.$$

2.2 Image Data and Anisotropic Phenomena

To model the fact that multivariate data appearing in applications is typically governed by anisotropic features—in the two-dimensional case curvilinear structures—the so-called *cartoon-like functions* were introduced in [11]. This class is by now widely used as a standard model, in particular, for natural images. It mimics the fact that natural images often consist of nearly smooth parts separated by discontinuities as illustrated in Fig. 2.

The first rigorous mathematical definition was given in [11] and extensively employed starting from the work in [5]. It postulates that images consist of $C^2(\mathbb{R}^2)$-regions separated by smooth $C^2(\mathbb{R})$-curves. This leads to the next definition (see also Fig. 2).

Fig. 1 *1* Illustration of the appearance of "cartoon-like parts" in natural images. *2* Illustration of the fact that the human brain is able to deduce the image (*2a*) just from its "cartoon-like" ingredients (*2b*)

Fig. 2 Example of a cartoon-like function

Definition 1 The *class* $\mathscr{E}^2(\mathbb{R}^2)$ *of cartoon-like functions* is the set of functions $f:\mathbb{R}^2 \to \mathbb{C}$ of the form

$$f = f_0 + f_1\chi_B,$$

where $B \subset [0,1]^2$ is a set with ∂B being a continuous and piecewise C^2-curve with bounded curvature and $f_i \in C^2(\mathbb{R}^2)$ are functions with supp $f_0 \subset [0,1]^2$ and $\|f_i\|_{C^2} \leq 1$, for each $i = 0, 1$.

We remark that by now several extensions of this model have been introduced and studied, starting with the extended model in [26].

Having agreed on a suitable subclass of functions, one might now ask whether there exists a maximal asymptotic approximation rate leading to a notion of optimality. Indeed, such a benchmark result was derived by Donoho in [11].

Theorem 1 [11] *Let* $(m_\lambda)_{\lambda\in\Lambda} \subseteq L^2(\mathbb{R}^2)$. *Under the assumption of polynomial depth search for the representation coefficients used in the N-term approximation, the associated asymptotic approximation rate of some* $f \in \mathscr{E}^2(\mathbb{R}^2)$ *satisfies at best*

$$\|f - f_N\|_2^2 \asymp N^{-2} \quad \text{as } N \to \infty.$$

It is in this sense that a system satisfying this approximation rate is said to deliver *optimally sparse approximations*.

2.3 2D Wavelet Systems

Nowadays, wavelet systems are widely utilized representation systems both for theoretical purposes as well as for engineering applications, for instance, for the decomposition of elliptic operators or for the detection of anomalies in signals. Their success stems from the fact that wavelets deliver optimal sparse approximations for data being governed by isotropic features—which is in particular the case for elliptic operator equations whose solutions may exhibit point singularities (for instance if re-entrant corners are present in the computational domain) as well as in the one-dimensional setting—and from the fast numerical realization of the wavelet transform.

Let us first recall a certain type of wavelet system in $L^2(\mathbb{R}^2)$, obtained by the following tensor product construction, see example [30] for details. Starting with a given multiresolution analysis of $L^2(\mathbb{R})$ with scaling function $\phi^0 \in L^2(\mathbb{R})$ and wavelet $\phi^1 \in L^2(\mathbb{R})$, for every index $e = (e_1, e_2) \in E$, $E = \{0, 1\}^2$, the generators $\psi^e \in L^2(\mathbb{R}^2)$ are defined as the tensor products

$$\psi^e = \phi^{e_1} \otimes \phi^{e_2}.$$

Definition 2 Let $\phi^0, \phi^1 \in L^2(\mathbb{R})$ and $\psi^e \in L^2(\mathbb{R}^2)$, $e \in E$, be defined as above. For fixed sampling parameters $\tau > 1$, $c > 0$, we define the *discrete wavelet system*

$$W(\phi^0, \phi^1; \tau, c) = \left\{\psi^{(0,0)}(\cdot - ck) \;:\; k \in \mathbb{Z}^2\right\} \cup \left\{\tau^j \psi^e(\tau^j \cdot - ck) \;:\; e \in E\backslash\{(0,0)\},\ j \in \mathbb{N}_0,\ k \in \mathbb{Z}^2\right\}.$$

The associated index set is given by

$$\Lambda^w = \left\{((0,0), 0, k) \;:\; k \in \mathbb{Z}^2\right\} \cup \left\{(e, j, k) \;:\; e \in E\backslash\{(0,0)\},\ j \in \mathbb{N}_0,\ k \in \mathbb{Z}^2\right\}.$$

Next, we recall the definition of vanishing moments for univariate wavelets, which says that the associated wavelet system annihilates polynomials up to some degree.

Definition 3 A function $g \in L^2(\mathbb{R})$ is said to possess M vanishing moments, if

$$\int_{\mathbb{R}} g(x)x^k dx = 0, \quad \text{for all} \quad k = 0, \ldots, M-1.$$

It is well known that this property can be characterized by polynomial decay near zero of the associated Fourier transform. For the convenience of the reader, we provide the short proof.

Lemma 2 *Suppose that $g \in L^2(\mathbb{R}) \cap C(\mathbb{R})$ is compactly supported and possesses M vanishing moments. Then*

$$|\hat{g}(\xi)| \lesssim \min(1, |\xi|)^M .$$

Proof First, note that, since g is continuous and compactly supported, $g \in L^1(\mathbb{R})$ and hence $\hat{g}$ is bounded. This shows that the claimed inequality holds for $|\xi| > 1$.

Let now $\xi \in \mathbb{R}$ satisfy $|\xi| \leq 1$. For this, observe that, up to a constant,

$$\int_{\mathbb{R}} g(x)x^k dx = \left(\frac{d}{d\xi}\right)^k \hat{g}(0).$$

Since g possesses M vanishing moments, it follows that all derivatives of order $k < M$ of $\hat{g}$ vanish at 0. Furthermore, since g is compactly supported, its Fourier transform is analytic. Thus

$$|\hat{g}(\xi)| \lesssim |\xi|^M,$$

which proves the claim. □

We now assume that $\phi^0, \phi^1 \in L^2(\mathbb{R})$ satisfy $\widehat{\phi^0}, \widehat{\phi^1} \in C^\infty(\mathbb{R})$ and there are $0 < a$ and $0 < a_1 < a_2$ such that

$$\operatorname{supp} \widehat{\phi^0} \subset [-a, a] \quad \text{and} \quad \operatorname{supp} \widehat{\phi^1} \subset [-a_2, a_2] \backslash [-a_1, a_1].$$

These conditions are fulfilled, for instance, if $\phi^0, \phi^1 \in L^2(\mathbb{R})$ are the generators of a Lemarié-Meyer wavelet system. In this case, it is well known that the associated tensor product wavelets are indeed suboptimal for approximation of anisotropic features modeled by cartoon-like functions.

Theorem 2 *For $f \in \mathscr{E}^2(\mathbb{R}^2)$, the wavelet system $W(\phi^0, \phi^1; \tau, c)$ provides an asymptotic L^2-error of best N-term approximation given by*

$$\|f - f_N\|_2^2 \asymp N^{-1} \quad \text{as } N \to \infty.$$

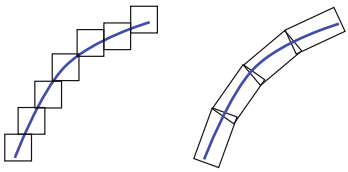


Fig. 3 Approximation of a curve by isotropic-shaped (*1*) and anisotropic-shaped (*2*) elements

3 Directional Representation Systems

The reason for the failure of wavelets to provide optimally sparse approximations of cartoon-like functions is the fact that wavelets are inherently *isotropic* objects and thus not optimally suited for approximating *anisotropic* objects. To overcome this problem, in recent years various directional representation systems were introduced, among which are ridgelets, curvelets, and shearlets, to name just a few. Their main advantage lies in their anisotropic support, which is much better suited to align with curvilinear structures (see Fig. 3), thereby already intuitively promoting a fast error decay of the best N-term approximation.

In this section, we now first introduce the second-generation curvelet system, which was in fact also the first system to provide (almost) optimally sparse approximations of cartoon-like functions (cf. Sect. 3.1). This is followed by a discussion of different versions of shearlets in Sect. 3.2.

3.1 Second-Generation Curvelets

Second-generation curvelets were introduced in 2004 by Candès and Donoho in the seminal work [5]. It is this curvelet system which is today referred to when curvelets are mentioned. The anisotropy of these systems is induced into this system by enforcing a parabolic scaling so that the shape of the support essentially follows the parabolic scaling law "$length^2 \approx width$". Intuitively, this seems a compromise between the isotropic scaling, as utilized for wavelets, and scaling in only one coordinate direction, as utilized for ridgelets. However, the reason is much deeper, since this law is particularly suited for approximating C^2-singularity curves, which is the type of curves our model is based on.

We now describe the original construction. For this, let W and V be two window functions that are both real, nonnegative, C^∞, and supported in $\left(\frac{1}{2}, 2\right)$ and in $(-1, 1)$, respectively. We further require that these windows satisfy

$$\sum_{j\in\mathbb{Z}} W\left(2^j r\right)^2 = 1 \quad \text{for all} \quad r \in \mathbb{R}_+ \quad \text{and} \quad \sum_{\ell\in\mathbb{Z}} V\,(t-\ell)^2 = 1 \quad \text{for all} \quad t \in \left(-\frac{1}{2}, \frac{1}{2}\right).$$

For every scale $j \geq 0$, we now define the functions $\gamma_{(j,0,0)}$ in polar coordinates by

$$\hat{\gamma}_{(j,0,0)}(r,\omega) := 2^{-3j/4} W\left(2^{-j} r\right) V\left(2^{\lfloor j/2 \rfloor} \omega\right).$$

For $j \in \mathbb{Z}$ and $\theta \in \mathbb{T}$, the parabolic scaling matrix A_j and the rotation matrix R_θ are defined by

$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix} \quad \text{and} \quad R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

The definition of curvelets then reads as follows:

$$\gamma_{(j,\ell,k)}(\cdot) := \gamma_{(j,0,0)}\left(R_{\theta_{j,\ell}} \cdot - x_{j,k}\right),$$

where $\theta_{j,\ell} = \ell 2^{-\lfloor j/2 \rfloor} \pi$, $x_{j,k} = A_j^{-1} k$, and $(j,\ell,k) \in \Lambda^0$ with the set of curvelet indices given by

$$\Lambda^0 := \left\{(j,\ell,k) \in \mathbb{Z}^4 : j \geq 0, \quad \ell = -2^{\lfloor j/2 \rfloor - 1}, \ldots, 2^{\lfloor j/2 \rfloor - 1}\right\}. \tag{2}$$

With appropriate modifications for the low-frequency case $j = 0$, for details we refer to [7], the system

$$\Gamma^0 := \left\{\gamma_\lambda : \lambda \in \Lambda^0\right\}$$

constitutes a Parseval frame for $L^2(\mathbb{R}^2)$, which is customarily referred to as the frame of *second generation curvelets*. When identifying frame elements oriented in antipodal directions, this system becomes a frame with real-valued elements.

Let us next discuss the approximation properties of Γ^0 proved in [5]. Ignoring log-like factors, this frame indeed attains the optimal achievable approximation rate for the class of cartoon-like functions $\mathscr{E}^2(\mathbb{R}^2)$. Moreover, this rate is achieved by simple thresholding, which is even more surprising, since this approximation scheme is intrinsically nonadaptive.

Theorem 3 [5] *The second generation curvelet frame Γ^0 provides (almost) optimal sparse approximations of cartoon-like functions $f \in \mathscr{E}^2(\mathbb{R}^2)$, i.e.,*

$$\|f - f_N\|_2^2 \lesssim N^{-2} (\log N)^3 \quad \text{as} \quad N \to \infty, \tag{3}$$

where f_N is the nonlinear N-term approximation obtained by choosing the N largest curvelet coefficients of f.

The implicit constant in (3) only depends on the maximal curvature of the singularity curve of f, the number of corner points, and the minimal opening angle in the corners. In particular, the approximation rate is uniform over all functions whose singularity curve has maximal curvature bounded by a fixed constant.

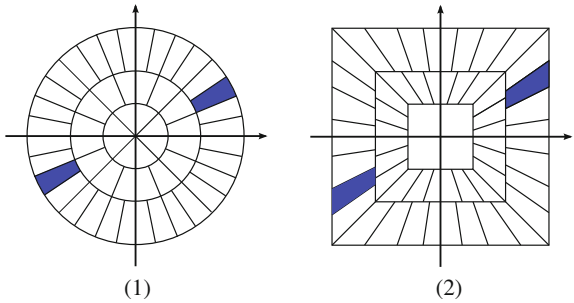


Fig. 4 Frequency tiling induced by a curvelet system (*1*) and a shearlet system (*2*)

Finally, we remark that due to the construction the frame elements of Γ^0 are bandlimited functions. Up to now no constructions of compactly supported curvelets are known.

3.2 Shearlet Systems

Shearlets were introduced in 2006 [19] as the first directional representation system which not only satisfies the same celebrated properties of curvelets, but is also more adapted to the digital realm. In fact, shearlets enable a unified treatment of the continuum and digital setting, which allows implementations faithful to the continuum domain theory. This key property is achieved through utilization of a shearing matrix instead of rotations as a means to parameterize orientation, thereby preserving the structure of the integer grid. The resulting different tilings of frequency domain are illustrated in Fig. 4.

We next introduce a selection of the variety of available shearlet systems, namely bandlimited shearlets (Sect. 3.2.1), the so-called smooth Parseval frames of shearlets (Sect. 3.2.3), and compactly supported shearlets (Sect. 3.2.2). For a more detailed exposition of shearlets than given below, we refer to the book [28].

3.2.1 Bandlimited Shearlets

We first present the classical cone-adapted shearlet construction of bandlimited shearlets presented in [19]. It is worth emphasizing that due to the shearing operator, the frequency domain needs to be split into four cones to ensure an almost uniform treatment of the different directions, which comes naturally for rotation as a means to change the orientation (compare Fig. 4).

First, let $\psi_1,\ \psi_2 \in L^2(\mathbb{R})$ be chosen such that

$$\operatorname{supp} \hat{\psi}_1 \subset \left[-\frac{1}{2}, -\frac{1}{16}\right] \cup \left[\frac{1}{16}, \frac{1}{2}\right], \quad \operatorname{supp} \hat{\psi}_2 \subset [-1, 1],$$

$$\sum_{j \geq 0} \left|\hat{\psi}_1\left(2^{-j}\omega\right)\right|^2 = 1 \quad \text{for} \quad |\omega| \geq \frac{1}{8},$$

and

$$\sum_{\ell=-2^{\lfloor j/2 \rfloor}}^{2^{\lfloor j/2 \rfloor}} \left| \hat{\psi}_2 \left(2^{\lfloor j/2 \rfloor} \omega + \ell \right) \right|^2 = 1 \quad \text{for} \quad |\omega| \leq 1.$$

Then the classical mother shearlet ψ is defined by

$$\hat{\psi}(\xi) := \hat{\psi}_1(\xi_1) \hat{\psi}_2 \left(\frac{\xi_2}{\xi_1} \right).$$

For $j, \ell \in \mathbb{Z}$ let now the parabolic scaling matrix A_j and the shearing matrix S_ℓ be defined by

$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix} \quad \text{and} \quad S_\ell := \begin{pmatrix} 1 & \ell \\ 0 & 1 \end{pmatrix}.$$

Further, for a domain $\Omega \subset \mathbb{R}^2$ let us define the space

$$L^2(\Omega)^\vee := \left\{ f \in L^2(\mathbb{R}^2) : \ \operatorname{supp} \widehat{f} \subset \Omega \right\}.$$

It was then shown in [19] that the system

$$\Sigma^0 := \left\{ 2^{3j/4} \psi \left(S_\ell A_j \cdot - k \right) : \ j \geq 0, \ \ell = -2^{\lfloor j/2 \rfloor}, \ldots, 2^{\lfloor j/2 \rfloor}, k \in \mathbb{Z}^2 \right\}$$

constitutes a Parseval frame for the Hilbert space $L^2(\mathscr{C})^\vee$ on the frequency cone

$$\mathscr{C} := \left\{ \xi : \ |\xi_1| \geq \frac{1}{8}, \ \frac{|\xi_2|}{|\xi_1|} \leq 1 \right\}.$$

By reversing the coordinate axes, also a Parseval frame Σ^1 for $L^2\left(\mathscr{C}'\right)^\vee$, where

$$\mathscr{C}' := \left\{ \xi : \ |\xi_2| \geq \frac{1}{8}, \ \frac{|\xi_1|}{|\xi_2|} \leq 1 \right\},$$

can be constructed. Finally, we can consider a Parseval frame

$$\Phi := \left\{ \phi(\cdot - k) : \ k \in \mathbb{Z}^2 \right\}$$

for the Hilbert space $L^2\left(\left[-\frac{1}{8}, \frac{1}{8} \right]^2 \right)^\vee$. Combining those systems, we obtain the *bandlimited shearlet frame*

$$\Sigma := \Sigma^0 \cup \Sigma^1 \cup \Phi.$$

In [20], it was shown that bandlimited shearlet frames achieve (almost) optimal sparse approximations for elements of $\mathscr{E}^2(\mathbb{R}^2)$, similar to curvelets and in fact even with the *same* log-like factor.

Theorem 4 [20] *The bandlimited shearlet frame* Σ *provides (almost) optimal sparse approximations of cartoon-like functions* $f \in \mathscr{E}^2(\mathbb{R}^2)$*, i.e.,*

$$\|f - f_N\|_2^2 \lesssim N^{-2}(\log N)^3 \quad \text{as } N \to \infty,$$

where f_N *is the nonlinear* N*-term approximation obtained by choosing the* N *largest shearlet coefficients of* f*.*

3.2.2 Smooth Parseval Frames of Shearlets

Following [22], a slight modification of the bandlimited shearlet construction, namely by carefully glueing together boundary elements along the seamlines with angle $\pi/4$, yields a Parseval frame with smooth and well-localized elements.

3.2.3 Compactly Supported Shearlets

In 2011, compactly supported shearlets were introduced by one of the authors and her collaborators in [27]. Currently known constructions of compactly supported shearlets involve separable generators, i.e.,

$$\psi(x_1, x_2) := \psi_1(x_1)\psi_2(x_2), \quad \widetilde{\psi}(x_1, x_2) := \psi(x_2, x_1). \tag{4}$$

with a wavelet ψ_1 and a scaling function ψ_2. Following [27], the cone-adapted discrete shearlet system is then defined as follows, where $A_j := \operatorname{diag}(2^j, 2^{j/2})$ as before and $\widetilde{A}_j := \operatorname{diag}(2^{j/2}, 2^j)$.

Definition 4 For some fixed sampling parameter $c > 0$, the *cone-adapted discrete shearlet system* $SH\big(\phi, \psi, \widetilde{\psi}; c\big)$ generated by $\phi, \psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$ is defined by

$$SH\big(\phi, \psi, \widetilde{\psi}; c\big) = \Phi(\phi; c) \cup \Psi(\psi; c) \cup \widetilde{\Psi}(\widetilde{\psi}; c),$$

where

$$\begin{aligned}
\Phi(\phi; c) &= \{\sigma_k = \phi(\cdot - k) : k \in c\mathbb{Z}^2\}, \\
\Psi(\psi; c) &= \big\{\sigma_{j,\ell,k} = 2^{3j/4}\psi(S_\ell A_j \cdot - k) : j \geq 0, |\ell| \leq \lceil 2^{j/2} \rceil, k \in c\mathbb{Z}^2\big\}, \\
\widetilde{\Psi}(\widetilde{\psi}; c) &= \big\{\widetilde{\sigma}_{j,\ell,k} = 2^{3j/4}\widetilde{\psi}(S_\ell^T \widetilde{A}_j \cdot - k) : j \geq 0, |\ell| \leq \lceil 2^{j/2} \rceil, k \in c\mathbb{Z}^2\big\}.
\end{aligned}$$

Under certain assumptions on $c, \psi, \widetilde{\psi}$ this shearlet system forms a frame with controllable frame bounds [24].

In [27], it was shown that compactly supported shearlet frames, under assumptions on the separable behavior and the directional vanishing moments of the generators, also achieve (almost) optimal sparse approximations for elements of $\mathscr{E}^2(\mathbb{R}^2)$.

Theorem 5 [27] *Let $c > 0$ and let $\phi, \psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported. Suppose that, in addition, for all $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$, the shearlet ψ satisfies*

(i) $|\hat{\psi}(\xi)| \leq C_1 \min(1, |\xi_1|^{\alpha}) \min(1, |\xi_1|^{-\gamma}) \min(1, |\xi_2|^{-\gamma})$ *and*
(ii) $\left|\frac{\partial}{\partial \xi_2} \hat{\psi}(\xi)\right| \leq |h(\xi_1)| \left(1 + \frac{\xi_2}{\xi_1}\right)^{-\gamma}$,

where $\alpha > 5$, $\gamma \geq 4$, $h \in L^1(\mathbb{R})$, and C_1 is a constant, and suppose that the shearlet $\widetilde{\psi}$ satisfies (i) and (ii) with the roles of ξ_1 and ξ_2 reversed. Further, suppose that $SH(\phi, \psi, \widetilde{\psi}; c)$ forms a frame for $L^2(\mathbb{R}^2)$. Then the shearlet frame $SH(\phi, \psi, \widetilde{\psi}; c)$ provides (almost) optimal sparse approximations of cartoon-like functions $f \in \mathscr{E}^2(\mathbb{R}^2)$, i.e.,

$$\|f - f_N\|_2^2 \lesssim N^{-2} (\log N)^3 \quad \text{as } N \to \infty,$$

where f_N is the nonlinear N-term approximation obtained by choosing the N largest shearlet coefficients of f.

With this theorem we end our presentation of directional representation systems, although there do exist more constructions. It is a striking fact that the three presented examples all exhibit the same approximation behavior, although they are construction-wise quite different. The framework of parabolic molecules, which we will present in the subsequent sections, will reveal the fundamental common ingredients in these systems which ensure (almost) optimal sparse approximations of cartoon-like functions.

4 Parabolic Molecules

The concept of parabolic molecules took shape by distilling the essential principles which underly many of the newly constructed directional representation systems, in particular, curvelets and shearlets. It provides a framework which comprises many of these classic systems, and allows the design of new constructions with predefined approximation properties.

Moreover, the approximation properties of some new system are usually proven more or less from scratch. By adopting the higher level viewpoint of time-frequency localization, the parabolic molecule framework is very general and independent of specific constructions. This has the advantage that it enables a unified treatment for many systems. In particular, it can be used to establish approximation results for many systems simultaneously.

A system of parabolic molecules consists of functions obtained from a set of generators via parabolic dilations, rotations, and translations. Similar to curvelets,

each function in a system of parabolic molecules is therefore naturally associated with a certain scale, orientation, and spatial location.

A central feature of the framework, which explains the terminology "molecules," is the concept of variable generators: In order to gain flexibility the generators are allowed to vary, as long as they obey a prescribed time-frequency localization. At the heart of this is the fundamental observation that it is foremost the time-frequency localization which determines the approximation properties and performance of a system.

A nice side effect of this less rigid construction principle is the fact that the strict vanishing moment conditions, usually imposed on the generators of classical constructions, can be relaxed without changing the asymptotic approximation behavior of the system. It suffices to require the moments to vanish asymptotically at high scales.

4.1 Definition of Parabolic Molecules

Let us now delve into the details of the framework of parabolic molecules. A system of parabolic molecules is a family of functions $(m_\lambda)_{\lambda\in\Lambda}$ obtained from a set of generators via parabolic dilations, rotations, and translations. Each function m_λ is therefore associated with a unique point in the parameter space $\mathbb{P}$, sometimes also referred to as phase space, given by

$$\mathbb{P} := \mathbb{R}_+ \times \mathbb{T} \times \mathbb{R}^2,$$

where a point $p = (s, \theta, x) \in \mathbb{P}$ specifies a scale $2^s \in \mathbb{R}_+$, an orientation $\theta \in \mathbb{T}$, and a location $x \in \mathbb{R}^2$.

The relation between the index λ of a molecule m_λ and its location $(s_\lambda, \theta_\lambda, x_\lambda)$ in the parameter space $\mathbb{P}$ is described via so-called parameterizations.

Definition 5 A *parameterization* consists of a pair (Λ, Φ_Λ), where Λ is a discrete index set and Φ_Λ is a mapping

$$\Phi_\Lambda : \Lambda \to \mathbb{P}, \quad \lambda \mapsto (s_\lambda, \theta_\lambda, x_\lambda)\,,$$

which associates with each $\lambda \in \Lambda$ a *scale* s_λ, a *direction* θ_λ, and a *location* $x_\lambda \in \mathbb{R}^2$.

By using parameterizations, the actual indices of the molecules can be decoupled from their associated locations in $\mathbb{P}$. This gives the freedom to assign generic indices to the molecules, a feature that is essential to include systems into the framework, whose constructions are based on different principles, for example shearlet-like and curvelet-like systems. Another benefit of this approach is that a parameterization does not have to sample phase space in a regular fashion. The only property it needs to satisfy for our results to be applicable is consistency as defined below in Sect. 6.2.

Before defining parabolic molecules we fix the following notation. As defined in Sect. 3, let R_θ denotes the rotation matrix by an angle θ, and A_j is the parabolic scaling matrix associated with $j \geq 0$.

Definition 6 Let Λ be a parameterization. A family $(m_\lambda)_{\lambda\in\Lambda}$ of functions $m_\lambda \in L^2(\mathbb{R}^2)$ is called a *family of parabolic molecules* of order (R, M, N_1, N_2) if it can be written as

$$m_\lambda(x) = 2^{3s_\lambda/4} a^{(\lambda)} \left(A_{s_\lambda} R_{\theta_\lambda} (x - x_\lambda)\right)$$

such that

$$\left|\partial^\beta \hat{a}^{(\lambda)}(\xi)\right| \lesssim \min\left(1, 2^{-s_\lambda} + |\xi_1| + 2^{-s_\lambda/2}|\xi_2|\right)^M \langle|\xi|\rangle^{-N_1} \langle\xi_2\rangle^{-N_2} \tag{5}$$

for all $|\beta| \leq R$. The implicit constants shall be uniform over $\lambda \in \Lambda$.

Remark 1 To simplify notation we did not explicitly refer to the utilized parameterization Φ_Λ.

Note that a system of parabolic molecules $(m_\lambda)_{\lambda\in\Lambda}$ is generated by parabolically scaling, rotating, and translating a set of generators $(a^{(\lambda)})_{\lambda\in\Lambda}$. In contrast to many classical constructions, where the set of generators is usually small, each molecule is allowed to have its own individual generator. We only require these generators to uniformly obey a prescribed time-frequency localization.

Recall that for convenience the time-frequency conditions in the definition are formulated on the Fourier side. Thus, the number R actually describes the spatial localization, M the number of directional (almost) vanishing moments, and N_1, N_2 describe the smoothness of an element m_λ.

According to the definition, the frequency support of a parabolic molecule is concentrated in a parabolic wedge associated to a certain orientation, and in the spatial domain its essential support lies in a rectangle with parabolic aspect ratio. For illustration purposes, the approximate frequency support of two parabolic molecules at different scales and orientations is depicted in Fig. 5.

Changing into polar coordinates, we obtain the representation

$$\hat{m}_\lambda(r,\varphi) = 2^{-3s_\lambda/4} \hat{a}^{(\lambda)} \left(2^{-s_\lambda} r\cos(\varphi+\theta_\lambda), 2^{-s_\lambda/2} r\sin(\varphi+\theta_\lambda)\right) \exp\left(2\pi i \langle x_\lambda, \xi\rangle\right),$$

which directly implies the estimate

$$\left|\hat{m}_\lambda(\xi)\right| \lesssim 2^{-2s_\lambda/4} \min\left(1, 2^{-s_\lambda}(1+r)\right)^M \left\langle 2^{-s_\lambda} r\right\rangle^{-N_1} \langle 2^{-s_\lambda/2} r\sin(\varphi+\theta_\lambda)\rangle^{-N_2}.$$

4.2 Index Distance

An essential ingredient for the theory is the fact that the parameter space $\mathbb{P}$ can be equipped with a natural (pseudo-)metric. It was first introduced by Smith [29], albeit

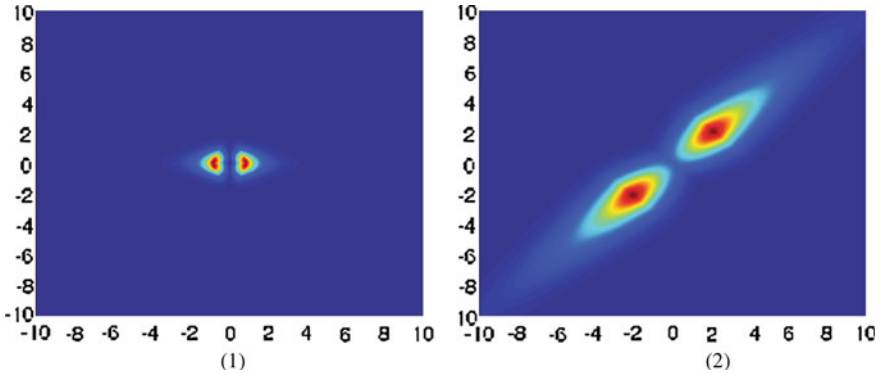


Fig. 5 *1* The weight function $\min\left(1, 2^{-s_\lambda} + |\xi_1| + 2^{-s_\lambda/2}|\xi_2|\right)^M \langle|\xi|\rangle^{-N_1} \langle\xi_2\rangle^{-N_2}$ for $s_\lambda = 3$, $M = 3$, $N_1 = N_2 = 2$. *2* Approximate frequency support of a corresponding molecule $\hat{m}_\lambda$ with $\theta_\lambda = \pi/4$

in a different context, and is therefore sometimes termed as the *Hart-Smith pseudo metric*. Later it was also used in [3].

Definition 7 Following [3, 29], we define for two indices λ, μ the *index distance*

$$\omega(\lambda, \mu) := 2^{|s_\lambda - s_\mu|}\left(1 + 2^{s_{\lambda_0}} d(\lambda, \mu)\right),$$

and

$$d(\lambda, \mu) := |\theta_\lambda - \theta_\mu|^2 + |x_\lambda - x_\mu|^2 + |\langle e_\lambda, x_\lambda - x_\mu\rangle|.$$

where $\lambda_0 = \operatorname{argmin}(s_\lambda, s_\mu)$ and $e_\lambda = (\cos(\theta_\lambda), \sin(\theta_\lambda))^\top$.

Remark 2 The notation $\omega(\lambda, \mu)$ is a slight abuse of notation, since ω is acting on $\mathbb{P}$. Therefore it should read as

$$\omega(\Phi_\Lambda(\lambda), \Phi_\Delta(\mu))$$

for indices $\lambda \in \Lambda, \mu \in \Delta$ with associated parameterizations $\Phi_\Lambda, \Phi_\Delta$. In order not to overload the notation, we stick with the shorter but slightly less accurate definition.

Remark 3 We also mention that there is a slight inaccuracy in the above definition. Real-valued curvelets or shearlets are not associated with an angle but with a ray, i.e., θ and $\theta + \pi$ need to be identified. This is not reflected in the above definition. The "correct" definition should assume that $|\theta_\lambda| \leq \frac{\pi}{2} \in \mathbb{P}^1$, the projective line. Therefore, it should read as

$$d(\lambda, \mu) := |\{\theta_\lambda - \theta_\mu\}|^2 + |x_\lambda - x_\mu|^2 + |\langle\theta_\lambda, x_\lambda - x_\mu\rangle|$$

with $\{\varphi\}$ being the projection of φ onto $\mathbb{P}^1 \cong (-\pi/2, \pi/2]$. However, for our results it will make no difference which definition is used. Thus we employ Definition 7, which avoids additional technicalities.

Note that the Hart-Smith pseudo metric is not a distance in the strict sense, e.g., we have $\omega(\lambda, \lambda) = 1 \neq 0$. As we shall see later, it somehow measures the correlation of a pair of parabolic molecules associated to the corresponding points in $\mathbb{P}$. The following proposition, whose proof can be found in [3], collects some of its properties.

Proposition 1 [3] *For indices λ, μ, ν we have*

(i) *Symmetry:* $\omega(\lambda, \mu) \asymp \omega(\mu, \lambda)$.
(ii) *Triangle Inequality:* $d(\lambda, \mu) \leq C\,(\omega(\lambda, \nu) + \omega(\nu, \mu))$ *for some constant* $C > 0$.
(iii) *Composition: For every integer* $N > 0$ *and some positive constant* C_N *it holds*

$$\sum_{\nu} \omega(\lambda, \nu)^{-N} \omega(\nu, \mu)^{-N} \leq C_N \omega(\lambda, \mu)^{-N-1}.$$

4.3 Decay of the Cross-Gramian

Given two systems $(m_\lambda)_{\lambda\in\Lambda}$ and $(p_\mu)_{\mu\in\Delta}$ of parabolic molecules; we are interested in the magnitudes of the cross-correlations $|\langle m_\lambda, p_\mu\rangle|$. A fast decay will be the key to, for instance, transferring sparse approximation properties from one system of parabolic molecules to another.

The following theorem establishes a relation to the index distance on $\mathbb{P}$. It states that a high distance of two indices can be interpreted as a low cross-correlation of the associated molecules. The proof is quite technical and we refer to [17] for the details.

Theorem 6 [17] *Let $(m_\lambda)_{\lambda\in\Lambda}$, $(p_\mu)_{\mu\in\Delta}$ be two systems of parabolic molecules of order (R, M, N_1, N_2) with*

$$R \geq 2N, \quad M > 3N - \frac{5}{4}, \quad N_1 \geq N + \frac{3}{4}, \quad N_2 \geq 2N.$$

Then

$$\left|\langle m_\lambda, p_\mu\rangle\right| \lesssim \omega\left((s_\lambda, \theta_\lambda, x_\lambda), (s_\mu, \theta_\mu, x_\mu)\right)^{-N}.$$

This result shows that the Gramian matrix between two systems of parabolic molecules satisfies a strong off-diagonal decay property and is in that sense very close to a diagonal matrix. In Sect. 6 we will present several immediate applications of this result, most notably for the approximation properties of parabolic molecules.

5 Examples of Parabolic Molecules

Before going deeper into the theory of parabolic molecules and further exploring their properties, we pause for a while and give some examples for illustration. This will give evidence about the versatility of the concept. In particular, we show that both rotation-based and shear-based constructions fit well into the framework. It will also be proven that earlier constructions, which also employ the "molecule" concept, can be viewed as subclasses of the more general parabolic molecules.

5.1 Curvelet-Like Systems

We begin with the review of curvelet-like systems, i.e., constructions based on rotation. Due to their similar construction principles, it may not come as a surprise that second-generation curvelets are instances of parabolic molecules. It is also easily verified that curvelet molecules as defined in [3] fall into this framework.

5.1.1 Second-Generation Curvelets

We start by specifying the parameterization, which we utilize for fitting second-generation curvelets into the framework of parabolic molecules.

Definition 8 Let

$$\Lambda^0 := \left\{(j,\ell,k) \in \mathbb{Z}^4 \ : \ j \geq 0,\ \ell = -2^{\lfloor j/2 \rfloor -1}, \dots, 2^{\lfloor j/2 \rfloor -1}\right\},$$

be the curvelet index from (2) and define $\Phi^0 : \Lambda^0 \to \mathbb{P}$ by

$$\Phi^0(j,\ell,k) := (j, \ell 2^{-\lfloor j/2 \rfloor}\pi, R_{-\theta_\lambda} A_{-s_\lambda} k).$$

Then (Λ^0, Φ^0) is called the *canonical parameterization*.

We next prove that the frame Γ^0 of second-generation curvelets as defined in Sect. 3.1 forms a system of parabolic molecules of arbitrary order.

Proposition 2 [17] *The second-generation curvelet frame Γ^0 constitutes a system of parabolic molecules of arbitrary order associated with the canonical parameterization.*

Proof Let $\lambda \in \Lambda^0$. Due to rotation invariance, we may restrict ourselves to the case $\theta_\lambda = 0$. Therefore, denoting $\gamma_j := \gamma_{(j,0,0)}$, it is sufficient to prove that the function

$$a^{(\lambda)}(\cdot) := 2^{-3s_\lambda/4}\gamma_j\left(A_{s_\lambda}^{-1}\cdot\right)$$

satisfies (5) for (R, M, N_1, N_2) arbitrary. For this, first note that

$$\hat{a}^{(\lambda)}(\cdot) = 2^{3s_\lambda/4}\hat{\gamma}_j\left(A_{s_\lambda}\cdot\right).$$

The function $\hat{a}^{(\lambda)}$, together with all its derivatives, has compact support in a rectangle away from the ξ_1-axis. Therefore, it only remains to show that, on its support, the function $\hat{a}^{(\lambda)}$ has bounded derivatives, with a bound independent of j. But this follows from elementary arguments, using $r = \sqrt{\xi_1^2 + \xi_2^2}$, $\omega = \arctan(\xi_2/\xi_1)$, which yields

$$\hat{a}^{(\lambda)}(\xi) = \hat{\gamma}_{(j,0,0)}\left(A_j\xi\right) = W\left(\alpha_j(\xi)\right)V\left(\beta_j(\xi)\right),$$

$$\alpha_j(\xi) := 2^{-j}\sqrt{2^{2j}\xi_1^2 + 2^j\xi_2^2} \quad \text{and} \quad \beta_j(\xi) := 2^{j/2}\arctan\left(\frac{\xi_2}{2^{j/2}\xi_1}\right).$$

By a straightforward calculation, all derivatives of α_j and β_j are bounded on the support of $\hat{a}^{(\lambda)}$ and uniformly in j. The proposition is proved. □

5.1.2 Hart Smith's Parabolic Frame

Historically, the first instance of a decomposition into parabolic molecules can be found in Hart Smith's work on Fourier Integral Operators and Wave Equations [29]. This frame, as well as its dual, again forms a system of parabolic molecules of arbitrary order associated with the canonical parameterization. We refer to [1, 29] for the details of the construction which is essentially identical to the curvelet construction, with primal and dual frame being allowed to differ. The same discussion as above for curvelets also shows that this system is a special instance of the framework of parabolic molecules.

5.1.3 Borup and Nielsen's Construction

Another very similar construction has been given in [2]. In this paper, the focus has been on the study of associated function spaces. Again, it is straightforward to prove that this system constitutes a system of parabolic molecules of arbitrary order associated with the canonical parameterization.

5.1.4 Curvelet Molecules

The final concept of parabolic molecules had many predecessors. In [3] the authors also employed the idea of molecules and introduced the notion of *curvelet molecules*. It proved to be a useful concept for showing sparsity properties of wave propagators. Let us first give their exact definition.

Definition 9 Let Λ^0 be the canonical parameterization. A family $(m_\lambda)_{\lambda\in\Lambda^0}$ is called a *family of curvelet molecules* of regularity R if it can be written as

$$m_\lambda(x) = 2^{3s_\lambda/4} a^{(\lambda)}\left(A_{s_\lambda} R_{\theta_\lambda}\,(x - x_\lambda)\right)$$

such that, for all $|\beta| \le R$ and each $N = 0, 1, 2, \ldots,$

$$|\partial^\beta a^{(\lambda)}(x)| \lesssim \langle x\rangle^{-N}$$

and, for all $M = 0, 1, 2, \ldots,$

$$|\hat{a}^{(\lambda)}(\xi)| \lesssim \min\left(1, 2^{-s_\lambda} + |\xi_1| + 2^{-s_\lambda/2}|\xi_2|\right)^M.$$

This definition is similar to our definition of parabolic molecules, however, with two crucial differences: First, (5) allows for arbitrary rotation angles and is therefore more general. Curvelet molecules, on the other hand, are only defined for the canonical parameterization Λ^0 (which, in contrast to our definition, is not sufficiently general to also cover shearlet-type systems). Second, the decay conditions analogous to our condition (5) are more restrictive in the sense that it requires infinitely many nearly vanishing moments.

In fact, the following result can be proven using similar arguments as for Proposition 2.

Proposition 3 [17] *A system of curvelet molecules of regularity R constitutes a system of parabolic molecules of order $(\infty, \infty, R/2, R/2)$.*

5.2 Shearlet-Like Systems

It is perhaps not surprising that curvelets and their relatives described above fall into the framework of parabolic molecules. However, we next show that even shearlets as a very different directional representation system are examples of parabolic molecules. In this regard, we draw the reader's attention to the parameterization chosen for fitting shearlets into this framework.

5.2.1 Shearlet Molecules

Shearlet molecules as introduced in [17] provide a framework for shearlet-like systems in the spirit of curvelet molecules. For their definition, we require the index set

$$\Lambda^\sigma := \left\{(\varepsilon, j, \ell, k) \in \mathbb{Z}_2 \times \mathbb{Z}^4 : \varepsilon \in \{0, 1\},\ j \ge 0,\ \ell = -2^{\lfloor j/2\rfloor}, \ldots, 2^{\lfloor j/2\rfloor}\right\} \quad (6)$$

and generating functions $\phi, \psi_{j,\ell,k}, \widetilde{\psi}_{j,\ell,k} \in L^2(\mathbb{R}^2)$, for $(j, \ell, k) \in \Lambda^\sigma$. The associated shearlet system

$$\Sigma := \left\{\sigma_\lambda : \lambda \in \Lambda^\sigma\right\},$$

is then defined by setting $\sigma_{(\varepsilon,0,0,k)}(\cdot) = \phi(\cdot - k)$ and for $j \geq 1$:

$$\begin{aligned}\sigma_{(0,j,\ell,k)}(\cdot) &= 2^{3j/4}\psi_{j,\ell,k}\left(A_j S_{\ell,j}\cdot - k\right),\\ \sigma_{(1,j,\ell,k)}(\cdot) &= 2^{3j/4}\widetilde{\psi}_{j,\ell,k}\left(\widetilde{A}_j S_{\ell,j}^T\cdot - k\right).\end{aligned}$$

Here, $S_{\ell,j}$ denotes the shearing matrix

$$S_{\ell,j} := \begin{pmatrix} 1 & \ell 2^{-\lfloor j/2\rfloor} \\ 0 & 1 \end{pmatrix}.$$

We proceed to define shearlet molecules of order (R, M, N_1, N_2), which is a generalization of shearlets adapted to parabolic molecules, in particular including the classical shearlet molecules introduced in [21], see Sect. 5.2.5.

Definition 10 We call Σ, a system of *shearlet molecules* of order (R, M, N_1, N_2), if the functions $\psi_{j,\ell,k}$ satisfy

$$|\partial^\beta \hat{\psi}_{j,\ell,k}(\xi_1, \xi_2)| \lesssim \min\left(1, 2^{-j} + |\xi_1| + 2^{-j/2}|\xi_2|\right)^M \langle|\xi|\rangle^{-N_1}\langle\xi_2\rangle^{-N_2} \tag{7}$$

and

$$|\partial^\beta \hat{\phi}(\xi_1, \xi_2)| \lesssim \langle|\xi|\rangle^{-N_1}\langle\xi_2\rangle^{-N_2} \tag{8}$$

for every $\beta \in \mathbb{N}^2$ with $|\beta| \leq R$, and if the functions $\widetilde{\psi}_{j,\ell,k}$ satisfy (7) with the roles of ξ_1 and ξ_2 reversed.

Remark 4 In our proofs, it is nowhere required that the directional parameter ℓ runs between $-2^{\lfloor j/2\rfloor}$ and $-2^{\lfloor j/2\rfloor}$. Indeed, ℓ running in any discrete interval $-C2^{\lfloor j/2\rfloor}$, $\ldots$, $C2^{\lfloor j/2\rfloor}$ would yield the exact same results, as a careful inspection of our arguments shows. Likewise, in certain shearlet constructions, the translational sampling runs not through $k \in \mathbb{Z}^2$, but through $\tau\mathbb{Z}^2$ with $\tau > 0$ a sampling constant. Our results are also valid for this case with similar proofs. The same remark applies to all curvelet-type constructions.

Now we can show the main result of this section, namely that shearlet systems with generators satisfying (7) and (8) are actually instances of parabolic molecules associated with a specific shearlet-adapted parametrization $(\Lambda^\sigma, \Phi^\sigma)$. This result shows that the concept of parabolic molecules is indeed a unification of in particular curvelet and shearlet systems.

Proposition 4 [17] *Assume that the shearlet system Σ constitutes a system of shearlet molecules of order (R, M, N_1, N_2). Then Σ forms a system of parabolic*

molecules of order (R, M, N_1, N_2)*, associated to the parameterization* $(\Lambda^\sigma, \Phi^\sigma)$*, where with* $A_j^0 = A_j$*,* $A_j^1 = \widetilde{A}_j$*,* $S_{\ell,j}^0 = S_{\ell,j}$*,* $S_{\ell,j}^1 = S_{\ell,j}^T$ *the map* Φ^σ *is given by*

$$\Phi^\sigma(\lambda) = (s_\lambda, \theta_\lambda, x_\lambda) := \left(j, \varepsilon\pi/2 + \arctan(-\ell 2^{-\lfloor j/2 \rfloor}), \left(S_{\ell,j}^\varepsilon\right)^{-1} \left(A_j^\varepsilon\right)^{-1} k \right).$$

Proof We confine the discussion to $\varepsilon = 0$, the other case being the same. Further, we suppress the subscripts j, ℓ, k in our notation. We need to show that

$$a^{(\lambda)}(\cdot) := \psi\left(A_{s_\lambda} S_{\ell,s_\lambda} R_{\theta_\lambda}^T A_{-s_\lambda} \cdot\right)$$

satisfies (5). We first observe that the Fourier transform of $a^{(\lambda)}$ is given by

$$\hat{a}^{(\lambda)}(\cdot) = \hat{\psi}\left(A_{-s_\lambda} S_{\ell,s_\lambda}^{-T} R_{\theta_\lambda}^T A_{s_\lambda} \cdot\right),$$

and the matrix $S_{\ell,s_\lambda}^{-T} R_{\theta_\lambda}^T$ has the form

$$S_{\ell,s_\lambda}^{-T} R_{\theta_\lambda}^T = \begin{pmatrix} \cos(\theta_\lambda) & \sin(\theta_\lambda) \\ 0 & -\ell 2^{-\lfloor s_\lambda/2 \rfloor} \sin(\theta_\lambda) + \cos(\theta_\lambda) \end{pmatrix} =: \begin{pmatrix} u & v \\ 0 & w \end{pmatrix}.$$

We next claim that the quantities u and w are uniformly bounded from above and below, independent of j, ℓ. To prove this claim, consider the functions

$$\tau(x) := \cos(\arctan(x)) \quad \text{and} \quad \rho(x) := x \sin(\arctan(x)) + \cos(\arctan(x)),$$

which are bounded from above and below on $[-1, 1]$, as elementary arguments show. In fact, this boundedness holds on any compact interval. We have

$$u = \tau\left(-\ell 2^{\lfloor s_\lambda/2 \rfloor}\right) \quad \text{and} \quad w = \rho\left(-\ell 2^{\lfloor s_\lambda/2 \rfloor}\right).$$

Since we are only considering indices with $\varepsilon = 0$, we have $\left|-\ell 2^{\lfloor s_\lambda/2 \rfloor}\right| \leq 1$, which now implies uniform upper and lower boundedness of the quantities u, w. Hence, there exist constants $0 < \delta_u \leq \Delta_u < \infty$ and $0 < \delta_w \leq \Delta_w < \infty$ such that for all j, ℓ it holds

$$\delta_u \leq u \leq \Delta_u \quad \text{and} \quad \delta_w \leq w \leq \Delta_w.$$

Observing that the matrix $A_{-s_\lambda} R_{\theta_\lambda}^T S_{\ell,s_\lambda}^{-T} A_{s_\lambda}$ has the form

$$\begin{pmatrix} u & 2^{-s_\lambda/2} v \\ 0 & w \end{pmatrix},$$

and by using the upper boundedness of u, v, w, and the chain rule, for any $|\beta| \leq R$, we obtain

$$|\partial^\beta \hat{a}^{(\lambda)}(\xi)| \lesssim \sup_{|\gamma|\leq R} \left|\partial^\gamma \hat{\psi}\left(\begin{pmatrix} u & 2^{-s_\lambda/2}v \\ 0 & w \end{pmatrix}\xi\right)\right| \lesssim \left(|\xi_1| + 2^{-s_\lambda/2}|\xi_2|\right)^M.$$

For the last estimate we utilized the moment estimate for $\hat{\psi}$, which is given by (7). This proves the moment property required in (5).

Finally, we need to show the decay of $\partial^\beta \hat{a}^{(\lambda)}$ for large frequencies ξ. Again, due to the fact that u, v, w are bounded from above and u, w from below, and utilizing the large frequency decay estimate in (7), we can estimate

$$\begin{aligned} |\partial^\beta \hat{a}^{(\lambda)}(\xi)| &\lesssim \sup_{|\gamma|\leq R} \left|\partial^\gamma \hat{\psi}\left(\begin{pmatrix} u & 2^{-s_\lambda/2}v \\ 0 & w \end{pmatrix}\xi\right)\right| \\ &\lesssim \left\langle \left|\begin{pmatrix} u & 2^{-s_\lambda/2}v \\ 0 & w \end{pmatrix}\xi\right| \right\rangle^{-N_1} \langle w\xi_2\rangle^{-N_2} \\ &\lesssim \langle|\xi|\rangle^{-N_1} \langle\xi_2\rangle^{-N_2}. \end{aligned}$$

The statement is proven. □

In the remainder of this section, we examine the main shearlet constructions which are known today and show that they indeed fit into the framework of parabolic molecules.

5.2.2 Classical Shearlets

For the bandlimited shearlet system Σ defined in Sect. 3.2.1, the following results can be shown using Proposition 4.

Proposition 5 [17] *The system $\Sigma := \Sigma^0 \cup \Sigma^1 \cup \Phi$ constitutes a shearlet frame which is a system of parabolic molecules of arbitrary order.*

It is also straightforward to check that the related Parseval frame constructed in [22] constitutes a system of parabolic molecules of arbitrary order.

5.2.3 Bandlimited Shearlets with Nice Duals

The bandlimited shearlet frame Σ as described above suffers from the fact that its dual frames are unknown. In particular, it is not known whether, in general, there exists a dual frame which also forms a system of parabolic molecules. In particular for applications, such a construction is however required. For general frames Σ of parabolic molecules, it can be shown that the canonical dual frame Σ' constitutes a

system of parabolic molecules of lower order [16]. However, the result of that paper is mostly of a qualitative nature and in particular it is difficult to compute the order of the dual frame for a given construction. In [15], this problem was successfully resolved by carefully gluing together the two bandlimited frames associated with the two frequency cones. The result in this paper in fact provides a construction of shearlet frames Σ with a dual frame Σ' such that both Σ and Σ' form systems of parabolic molecules of arbitrary order.

5.2.4 Compactly Supported Shearlets

Again by using the general result Proposition 4, it can be shown that the compactly supported shearlets as introduced in Sect. 3.2.3 also constitute a system of parabolic molecules, this time with the order being dependent in a more delicate way on the chosen generators.

Proposition 6 [17] *Assume that $\psi_1 \in C^{N_1}$ is a compactly supported wavelet with $M + R$ vanishing moments, and $\psi_2 \in C^{N_1+N_2}$ is also compactly supported. Then, with ψ and $\widetilde{\psi}$ defined by (4), the associated shearlet system Σ constitutes a system of parabolic molecules of order (R, M, N_1, N_2).*

We remark that several assumptions on the generators ψ, $\widetilde{\psi}$ could be weakened, for instance, the separability of the shearlet generators is not crucial for the arguments of the associated proof. More precisely, neither compact support nor bandlimitedness is necessary.

5.2.5 Shearlet Molecules of Guo and Labate [21]

In [21] the results of [3] are established for shearlets instead of curvelets. A crucial tool in the proof is the introduction of a certain type of shearlet molecules that are similar to curvelet molecules discussed above, but tailored to the shearing operation rather than rotations.

Definition 11 Let Λ^σ be the shearlet index set as in (6) and A_j^ε, $S_{\ell,j}^\varepsilon$ be defined as in Proposition 4. A family $(m_\lambda)_{\lambda\in\Lambda^\sigma}$ is called a *family of shearlet molecules* of regularity R, if it can be written as

$$m_\lambda(x) = 2^{3s_\lambda/4} a^{(\lambda)} \left(A_j^\varepsilon S_{\ell,j}^\varepsilon x - k \right),$$

such that, for all $|\beta| \le R$ and each $N = 0, 1, 2, \ldots,$

$$|\partial^\beta a^{(\lambda)}(x)| \lesssim \langle x \rangle^{-N}$$

and, for all $M = 0, 1, 2, \ldots,$

$$|\hat{a}^{(\lambda)}(\xi)| \lesssim \min\left(1, 2^{-s_\lambda} + |\xi_1| + 2^{-s_\lambda/2}|\xi_2|\right)^M.$$

By the results in [21], the shearlet molecules defined therein satisfy the inequality (7) with the choice of parameters $(R, M, N_1, N_2) = (\infty, \infty, R/2, R/2)$. Therefore, in view of Proposition 4, shearlet molecules of regularity R as defined in [21] form systems of parabolic molecules of order $(\infty, \infty, R/2, R/2)$.

Proposition 7 [17] *A system of shearlet molecules of regularity R constitutes a system of parabolic molecules of order* $(\infty, \infty, R/2, R/2)$.

6 Sparse Approximation with Parabolic Molecules

This section is devoted to one prominent application of the framework of parabolic molecules, and, in particular, the result of the decay of the cross-Gramian (Theorem 6), namely to sparse approximation behavior. This result also shows that the viewpoint of time-frequency localization as adopted by the framework of parabolic molecules provides the right angle to view questions of approximation behavior.

After introducing a measure for determining similar sparsity behavior, two main results are presented: First, it is shown that any two systems of parabolic molecules that are consistent, in a certain sense made precise later of sufficiently high order, exhibit the same approximation behavior. Second, by linking an arbitrary system to the curvelet frame, we obtain a "stand-alone result" in the sense of sufficient conditions on the order of a system of parabolic molecules for providing (almost) optimally sparse approximations of cartoon-like functions.

6.1 Sparsity Equivalence

In light of Lemma 1, two frames should possess similar sparse approximation behavior, provided that the corresponding coefficient sequences have the same sparsity. This gave rise to the notion of sparsity equivalence from Grohs and Kutyniok [17], which is a useful tool to compare such behavior. It is based on the close connection between the best N-term approximation rate of a frame and the ℓ_p-(quasi-)norm of the associated coefficient sequence.

Definition 12 Let $(m_\lambda)_{\lambda\in\Lambda}$ and $(p_\mu)_{\mu\in\Delta}$ be systems of parabolic molecules and let $0 < p \leq 1$. Then $(m_\lambda)_{\lambda\in\Lambda}$ and $(p_\mu)_{\mu\in\Delta}$ are *sparsity equivalent in* ℓ_p, if

$$\left\| \left(\langle m_\lambda, p_\mu \rangle\right)_{\lambda\in\Lambda,\mu\in\Delta} \right\|_{\ell_p\to\ell_p} < \infty.$$

Intuitively, systems of parabolic molecules being in the same sparsity equivalence class have similar approximation properties. This will subsequently be elaborated more deeply.

6.2 Consistency of Parameterizations

The next goal will be to find conditions that ensure that two systems of parabolic molecules are sparsity equivalent. It seems clear from an intuitive viewpoint that this requires some "consistency" of the associated parameterizations. The next definition provides the correct notion for making this mathematically precise.

Definition 13 Two parameterizations (Λ, Φ_Λ) and (Δ, Φ_Δ) are called *k-consistent*, for $k > 0$, if

$$\sup_{\lambda\in\Lambda} \sum_{\mu\in\Delta} \omega(\lambda, \mu)^{-k} < \infty \quad \text{and} \quad \sup_{\mu\in\Delta} \sum_{\lambda\in\Lambda} \omega(\lambda, \mu)^{-k} < \infty.$$

In combination with Theorem 6, consistency is the essential tool to decide whether two frames of parabolic molecules are sparsity equivalent. We emphasize that although the original definition of systems of parabolic molecules does not require those systems to form a frame in the context of approximation theory, however, the frame property becomes important.

The following result states a sufficient condition for sparsity equivalence.

Theorem 7 [17] *Two frames $(m_\lambda)_{\lambda\in\Lambda}$ and $(p_\mu)_{\mu\in\Delta}$ of parabolic molecules of order (R, M, N_1, N_2) with k-consistent parameterizations for some $k > 0$, are sparsity equivalent in ℓ_p, $0 < p \leq 1$, if*

$$R \geq 2\frac{k}{p}, \quad M > 3\frac{k}{p} - \frac{5}{4}, \quad N_1 \geq \frac{k}{p} + \frac{3}{4}, \quad \text{and} \quad N_2 \geq 2\frac{k}{p}.$$

Proof By Schur's test, a well-known result from operator theory, we have

$$\left\| \left(\langle m_\lambda, p_\mu \rangle \right)_{\lambda\in\Lambda, \mu\in\Delta} \right\|_{\ell_p \to \ell_p} \leq \max \left(\sup_{\mu\in\Delta} \sum_{\lambda\in\Lambda} |\langle m_\lambda, p_\mu \rangle|^p, \sup_{\lambda\in\Lambda} \sum_{\mu\in\Delta} |\langle m_\lambda, p_\mu \rangle|^p \right)^{1/p}.$$

By Theorem 6, this implies that

$$\left\| \left(\langle m_\lambda, p_\mu \rangle \right)_{\lambda\in\Lambda, \mu\in\Delta} \right\|_{\ell_p \to \ell_p} \lesssim \max \left(\sup_{\mu\in\Delta} \sum_{\lambda\in\Lambda} \omega(\lambda, \mu)^{-k}, \sup_{\lambda\in\Lambda} \sum_{\mu\in\Delta} \omega(\lambda, \mu)^{-k} \right)^{1/p}.$$

But the term on the right-hand side is finite, due to the k-consistency of the parameterizations (Λ, Φ_Λ) and (Δ, Φ_Δ). This proves that $(m_\lambda)_{\lambda\in\Lambda}$ and $(p_\mu)_{\mu\in\Delta}$ are sparsity equivalent in ℓ_p. □

Thus, as long as the parameterizations are consistent, the sparsity equivalence can be controlled by the order of the molecules.

In the remainder, we fix the frame of second-generation curvelets Γ^0 from Sect. 3.1 as a reference frame. Recall that with respect to the canonical parameterization $(\Lambda^0, \Phi_{\Lambda^0})$, this frame constitutes a system of parabolic molecules justifying the following definition.

Definition 14 A parameterization (Λ, Φ_Λ) is called *k-admissible*, for $k > 0$, if it is k-consistent with the canonical parameterization $(\Lambda^0, \Phi_{\Lambda^0})$.

Before stating our main results, it seems natural to ask whether the curvelet and shearlet parameterizations are k-admissible. This is the content of the next two lemmata.

Lemma 3 [17] *The canonical parameterization $(\Lambda^0, \Phi_{\Lambda^0})$ is k-admissible for all $k > 2$.*

Proof Writing $s_\mu = j'$ in the definition of $\omega(\mu, \lambda)$, we need to prove that

$$\sum_{j\in\mathbb{Z}_+} \sum_{\lambda\in\Lambda^0, s_\lambda = j} 2^{-k|j-j'|} \left(1 + 2^{\min(j,j')} d(\mu, \lambda)\right)^{-k} < \infty. \tag{9}$$

By [3, Eq. (A.2)], for any q, we have

$$\sum_{\lambda\in\Lambda^0, s_\lambda = j} (1 + 2^q d(\mu, \lambda))^{-2} \lesssim 2^{2(j-q)_+}. \tag{10}$$

Hence, for each $k > 2$, (9) can be estimated by

$$\sum_{j\geq 0} 2^{-k|j-j'|} 2^{2|j-j'|} < \infty,$$

which finishes the proof. □

Lemma 4 [17] *The shearlet parameterization $(\Lambda^\sigma, \Phi^\sigma)$ is k-admissible for $k > 2$.*

Proof The proof follows the same arguments as the proof of Lemma 3, except deriving the analog to (10), i.e.,

$$\sum_{\lambda\in\Lambda^\sigma, s_\lambda = j} (1 + 2^q d(\mu, \lambda))^{-2} \lesssim 2^{2(j-q)_+}, \quad \text{for any } q \text{ and } \mu \in \Lambda^0, \tag{11}$$

requires a bit more work.

Without loss of generality, we assume that $\theta_\mu = 0$ and $x_\mu = 0$. Also, we only restrict ourselves to the case $\varepsilon = 0$, the other case being exactly the same. In the case $q > j$, the term on the left-hand side of (11) can be bounded by a uniform constant.

Thus, it remains to deal with the $j \geq q$. Now we use the fact that, whenever $|\ell| \lesssim 2^{-j/2}$, we have

$$\left|\arctan\left(-\ell 2^{-\lfloor j/2 \rfloor}\right)\right| \gtrsim \left|\ell 2^{-\lfloor j/2 \rfloor}\right| \quad \text{and} \quad |S_\ell^{-1} A_{-j} k| \gtrsim |A_{-j} k|,$$

to estimate (11) by

$$\sum_\ell \sum_k \left(1 + 2^q \left(\left|\ell 2^{-\lfloor j/2 \rfloor}\right|^2 + \left|2^{-\lfloor j/2 \rfloor} k_2\right|^2 + \left|2^{-j} k_1 - \ell 2^{-\lfloor j/2 \rfloor} k_2 2^{-\lfloor j/2 \rfloor}\right|\right)\right)^{-2}.$$

This can be interpreted as a Riemann sum and is bounded (up to a constant) by the corresponding integral

$$\int_{\mathbb{R}^2} \frac{dx}{2^{-3j/2}} \int_{\mathbb{R}} \frac{dy}{2^{-j/2}} \left(1 + 2^q (y^2 + x_2^2 + |x_1 - x_2 y|)\right)^{-2},$$

compare [3, Eq. (A.3)]. This integral is bounded by $C \times 2^{2(j-q)}$ as can be seen by the substitution $x_1 \to 2^q x_1$, $x_2 \to 2^{q/2} x_2$, $y \to 2^{q/2} y$. This yields (11), which completes the proof. □

6.3 Sparse Approximations

The next theorem now states the central fact that any system of parabolic molecules of sufficiently high order, whose parameterization is k-admissible, is sparsity equivalent to the second-generation curvelet frame from Sect. 3.1. This theorem can be interpreted as a means to transfer sparse approximation results from one system of parabolic molecules to another, which is also the key to Theorem 9.

Theorem 8 [17] *Assume that $0 < p \leq 1$, (Λ, Φ_Λ) is a k-admissible parametrization, and $\Gamma^0 = (\gamma_\lambda)_{\lambda \in \Lambda^0}$ the tight frame of bandlimited curvelets. Further, assume that $(m_\lambda)_{\lambda \in \Lambda}$ is a system of parabolic molecules associated with Λ of order (R, M, N_1, N_2) such that*

$$R \geq 2\frac{k}{p}, \quad M > 3\frac{k}{p} - \frac{5}{4}, \quad N_1 \geq \frac{k}{p} + \frac{3}{4}, \quad N_2 \geq 2\frac{k}{p}.$$

Then $(m_\lambda)_{\lambda \in \Lambda}$ is sparsity equivalent in ℓ_p to Γ^0.

Recall that it was shown by Donoho in [11] (cf. Theorem 1) that (under natural conditions) the optimally achievable decay rate of the approximation error for the class $\mathscr{E}^2(\mathbb{R}^2)$ is given by

$$\|f - f_N\|_2^2 \asymp N^{-2}, \quad \text{as } N \to \infty.$$

As discussed before, in [5, 20, 24] rotation-based as well as shear-based systems were constructed, which attain this rate up to a log factor. Since these systems are instances of parabolic molecules with consistent parameterizations, their similar approximation behavior is no coincidence, as we will see in the next result.

Theorem 9 [17] *Assume that $(m_\lambda)_{\lambda\in\Lambda}$ is a system of parabolic molecules of order (R, M, N_1, N_2) with respect to the parameterization (Λ, Φ_Λ) such that*

(i) *$(m_\lambda)_{\lambda\in\Lambda}$ constitutes a frame for $L^2(\mathbb{R}^2)$,*
(ii) *(Λ, Φ_Λ) is k-admissible for every $k > 2$,*
(iii) *it holds that*

$$R \geq 6, \quad M > 9 - \frac{5}{4}, \quad N_1 \geq 3 + \frac{3}{4}, \quad N_2 \geq 6.$$

Then the frame $(m_\lambda)_{\lambda\in\Lambda}$ possesses an almost best N-term approximation rate of order $N^{-1+\varepsilon}$, $\varepsilon > 0$ arbitrary, for the cartoon image class $\mathscr{E}^2(\mathbb{R}^2)$.

We remark that condition (ii) holds in particular for the shearlet parameterization. Hence this result allows a simple derivation of the results in [20, 24] from Candès [5]. In fact, Theorem 9 provides a systematic way to, in particular, prove results on sparse approximation of cartoon-like functions. It moreover enables us to provide a very general class of systems of parabolic molecules that optimally sparsely approximate cartoon-like functions by using the known result for curvelets.

7 Outlook and Further Generalizations

Finally, we discuss some possible extensions and directions for future research.

- *Higher Dimensional Setting.* A general framework such as parabolic molecules would also be of benefit for higher dimensional functions, in particular for the three-dimensional setting which then includes videos with time as third dimension. The model of cartoon-like functions was already extended to this situation in [26]. Then, in [12], a general framework of parabolic molecules for functions in $L^2(\mathbb{R}^3)$ was introduced allowing, in particular, a similar result on the cross-Gramian of two systems of 3D parabolic molecules. We expect that the 3D framework now indicates a natural extension to higher dimensional settings.

- *General Scaling Matrix.* Another key question concerns the inclusion of other types of scaling laws: Can the framework of parabolic molecules be extended to also include, in particular, wavelets and ridgelets as well as newer hybrid constructions such as [26] or [23]? In the parabolic molecule framework the degree of anisotropic scaling is confined to parabolic scaling, but one approach to cover more scaling laws consists in the introduction of a parameter $\alpha \in [0, 1]$, which measures the degree of anisotropy. More precisely, one then considers scaling matrices of the type $\mathrm{diag}(a, a^{\alpha})$ for $\alpha \in [0, 1]$, $\alpha = 0$ corresponding to ridgelets, $\alpha = \frac{1}{2}$ to curvelets and shearlets, and $\alpha = 1$ to wavelets. First results using this approach to introduce an extension of parabolic molecules coined *α-molecules* have been derived in [18].
- *Continuum Setting.* It would be highly desirable to also introduce such a framework for the continuum setting, i.e., with continuous parameter sets, adapted to the continuous shearlet and curvelet transform [6, 14, 25]. This would, for instance, allow the transfer of characterization results of microlocal smoothness spaces between different representation systems.

References

1. Andersson, F., de Hoop, M., Smith, H., Uhlmann, G.: A multi-scale approach to hyperbolic evolution equations with limited smoothness. Comm. PDE **33**, 988–1017 (2008)
2. Borup, L., Nielsen, M.: Frame decompositions of decomposition spaces. J. Fourier Anal. Appl. **13**, 39–70 (2007)
3. Candès, E.J., Demanet, L.: The curvelet representation of wave propagators is optimally sparse. Comm. Pure Appl. Math. **58**, 1472–1528 (2002)
4. Candès, E.J., Donoho, D.L.: Ridgelets: a key to higher-dimensional intermittency? Phil. Trans. R. Soc. Lond. A. **357**, 2495–2509 (1999)
5. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with C^2 singularities. Comm. Pure Appl. Math. **56**, 219–266 (2004)
6. Candès, E.J., Donoho, D.L.: Continuous curvelet transform: I. Resolution of the wavefront set. Appl. Comput. Harmon. Anal. **19**, 162–197 (2005)
7. Candès, E.J., Donoho, D.L.: Continuous curvelet transform: II. Discretization and frames. Appl. Comput. Harmon. Anal. **19**, 198–222 (2005)
8. Casazza, P.G., Kutyniok, G. (eds.): Finite Frames: Theory and Applications. Birkhäuser, Boston (2012)
9. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser, Bostan (2003)
10. DeVore, R.A.: Nonlinear approximation. Acta Numerica **7**, 51–150 (1998)
11. Donoho, D.L.: Sparse components of images and optimal atomic decomposition. Constr. Approx. **17**, 353–382 (2001)
12. Flinth, A.: 3D parabolic molecules. Bachelor's thesis, Technische Universität Berlin (2013)
13. Gribonval, R., Nielsen, M.: Nonlinear approximation with dictionaries. I. Direct estimates. J. Fourier Anal. Appl. **10**, 51–71 (2004)
14. Grohs, P.: Continuous shearlet frames and resolution of the wavefront set. Monatsh. Math. **164**, 393–426 (2011)
15. Grohs, P.: Bandlimited shearlet frames with nice duals. J. Comput. Appl. Math. **244**, 139–151 (2013)
16. Grohs, P.: Intrinsic localization of anisotropic frames. Appl. Comput. Harmon. Anal. **35**, 264–283 (2013)

17. Grohs P., Kutyniok, G.: Parabolic molecules. Found. Comput. Math. **14**(2), 299–337 (2014)
18. Grohs, P., Keiper, S., Kutyniok, G., Schäfer, M.: α-Molecules. Preprint (2013)
19. Guo, K., Kutyniok, G., Labate, D.: Sparse multidimensional representations using anisotropic dilation and shear operators. Wavelets and Splines, pp. 189–201 (2005). Nashboro Press, Athens (2006)
20. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math Anal. **39**, 298–318 (2007)
21. Guo, K., Labate, D.: Representation of Fourier integral operators using shearlets. J. Fourier Anal. Appl. **14**, 327–371 (2008)
22. Guo, K., Labate, D.: The construction of smooth parseval frames of shearlets. Math. Model Nat. Phenom. **8**, 82–105 (2013)
23. Keiper, S.: A flexible shearlet transform—sparse approximation and dictionary learning. Bachelor's thesis, Technische Universität Berlin (2012)
24. Kittipoom, P., Kutyniok, G., Lim, W.-Q.: Construction of compactly supported shearlet. Constr. Approx. **35**, 21–72 (2012)
25. Kutyniok, G., Labate, D.: Resolution of the wavefront set using continuous shearlets. Trans. Amer. Math. Soc. **361**, 2719–2754 (2009)
26. Kutyniok, G., Lemvig, J.: Optimally sparse approximations of 3D functions by compactly supported shearlet frames. SIAM J. Math. Anal. **44**, 2962–3017 (2012)
27. Kutyniok, G., Lim, W.-Q.: Compactly supported shearlets are optimally sparse. J. Approx. Theor. **163**, 1564–1589 (2011)
28. Kutyniok, G., Labate, D. (eds.): Shearlets: Multiscale Analysis for Multivariate Data. Birkhäuser, Boston (2012)
29. Smith, H.: A parametrix construction for wave equations with $C^{1,1}$-coefficients. Ann. Inst. Fourier **48**, 797–835 (1998)
30. Wojtaszczyk, P.: A Mathematical Introduction to Wavelets. Cambridge University Press, Cambridge (1997)

Microlocal Analysis of Singularities from Directional Multiscale Representations

Kanghui Guo, Robert Houska and Demetrio Labate

Abstract The classical wavelet transform is a remarkably effective tool for the analysis of pointwise regularity of functions and distributions. During the last decade, the emergence of a new generation of multiscale representations has extended the classical wavelet approach leading to the introduction of a class of generalized wavelet transforms—most notably the shearlet transform—which offers a much more powerful framework for microlocal analysis. In this paper, we show that the shearlet transform enables a precise geometric characterization of the set of singularities of a large class of multidimensional functions and distributions, going far beyond the capabilities of the classical wavelet transform. This paper generalizes and extends several results that previously appeared in the literature and provides the theoretical underpinning for advanced applications from image processing and pattern recognition including edge detection, shape classification, and feature extraction.

Keywords Analysis of singularities · Continuous wavelet transform · Edge detection · Shearlets · Wavefront set · Wavelets

K. Guo (✉)
Missouri State University, Springfield, MO 65804, USA
e-mail: KanghuiGuo@MissouriState.edu

R. Houska · D. Labate
Department of Mathematics, University of Houston, Houston, TX 77204, USA
e-mail: rhouska@math.uh.edu

D. Labate
e-mail: dlabate@math.uh.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_10,

1 Introduction

How do you detect the location of a jump discontinuity in a function? One possible approach consists in using as probes a collection of well-localized functions of the form $\psi_{a,t}(x) = a^{-n/2}\psi(a^{-1}(x-t))$, $a > 0$, $t \in \mathbb{R}^n$, where $\psi \in L^2(\mathbb{R}^n)$. We assume that ψ is chosen such that $\hat{\psi} \in C_c^\infty(\mathbb{R}^n)$, with $0 \notin \operatorname{supp} \hat{\psi}$. Since ψ has rapid decay in space domain, the functions $\psi_{a,t}$ are mostly concentrated around t, with the size of the essential support controlled by the scaling parameter a. We can then analyze the local regularity of a function or distribution f via the mapping

$$f \to \langle f, \psi_{a,t}\rangle, \quad a > 0,\ t \in \mathbb{R}^n.$$

To illustrate this approach, let us consider as a prototype of a jump discontinuity the one-dimensional Heaviside function $f(x) = 1$ if $x \geq 0$ and $f(x) = 0$ otherwise. Using the Plancherel theorem and the distributional Fourier transform of f, a direct calculation using the analyzing functions $\psi_{a,t}$ with $n = 1$ shows that[1]

$$\begin{aligned}\langle f, \psi_{a,t}\rangle &= \langle \hat{f}, \hat{\psi}_{a,t}\rangle \\ &= \sqrt{a}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\hat{\psi}(a\xi)}\,e^{-2\pi i\xi t}\,d\xi \\ &= \sqrt{a}\int_{\mathbb{R}} \frac{1}{2\pi i\xi}\,\overline{\hat{\psi}(a\xi)}\,e^{-2\pi i\xi t}\,d\xi \\ &= \sqrt{a}\int_{\mathbb{R}} \hat{\gamma}(\eta)\,e^{-2\pi i\eta\frac{t}{a}}\,d\eta,\end{aligned}$$

where $\hat{\gamma}(\eta) = \frac{1}{2\pi i\eta}\,\overline{\hat{\psi}(\eta)}$. If $t = 0$, the calculation above shows that $|\langle f, \psi_{a,t}\rangle| \approx \sqrt{a}$, provided that $\int \hat{\gamma}(\eta)d\eta \neq 0$. On the other hand, if $t \neq 0$, an application of the Inverse Fourier Transform theorem yields that $\langle f, \psi_{a,t}\rangle = \sqrt{a}\,\gamma(-t/a)$. Since $\hat{\gamma} \in C_c^\infty(\mathbb{R})$, γ has rapid decay in space domain, implying that $\langle f, \psi_{a,t}\rangle$ decays rapidly to 0, as $a \to 0$; that is, for any $N \in \mathbb{N}$, there is a constant $C_N > 0$ such that $|\langle f, \psi_{a,t}\rangle| \leq C_N a^N$, as $a \to 0$.

In summary, *the elements* $\langle f, \psi_{a,t}\rangle$ *exhibit rapid asymptotic decay, as* $a \to 0$, *for all* $t \in \mathbb{R}$ *except at the location of the singularity* $t = 0$, *where* $\langle f, \psi_{a,t}\rangle$ *behaves as* $O(\sqrt{a})$.

The mapping $f \to \langle f, \psi_{a,t}\rangle$ is the classical *continuous wavelet transform* and this simple example illustrates its ability to detect local regularity information about functions and distributions through its asymptotic decay at fine scales (cf. [16–18, 22]).

[1] Note that the distributional Fourier transform of f is $\hat{f}(\xi) = \frac{1}{2}\delta(\xi) + \frac{1}{2\pi i}\text{p.v.}\frac{1}{\xi}$, but the term $\frac{1}{2}\delta(\xi)$ gives no contribution in the computation for $\langle f, \psi_{a,t}\rangle$ since $\hat{\psi}(0) = 0$.

The generalization of the example above to higher dimensions is straightforward. Let us consider the two-dimensional Heaviside function $H(x_1, x_2) = \chi_{\{x_1>0\}}(x_1, x_2)$ and let us proceed as in the example above. Using the analyzing functions $\psi_{a,t}$ with $n = 2$ and denoting $t = (t_1, t_2) \in \mathbb{R}^2$ we have:

$$\begin{aligned}
\langle H, \psi_{a,t}\rangle &= \langle \hat{H}, \hat{\psi}_{a,t}\rangle \\
&= a \int_{\mathbb{R}^2} \hat{H}(\xi_1, \xi_2)\, \overline{\hat{\psi}(a\xi_1, a\xi_2)}\, e^{-2\pi i(\xi_1 t_1 + \xi_2 t_2)}\, d\xi_1\, d\xi_2 \\
&= a \int_{\mathbb{R}^2} \frac{\delta(\xi_2)}{2\pi i \xi_1}\, \overline{\hat{\psi}(a\xi_1, a\xi_2)}\, e^{-2\pi i(\xi_1 t_1 + \xi_2 t_2)}\, d\xi_1\, d\xi_2 \\
&= a \int_{\mathbb{R}} \frac{1}{2\pi i \xi_1}\, \overline{\hat{\psi}(a\xi_1, 0)}\, e^{-2\pi i \xi_1 t_1}\, d\xi_1 \\
&= a \int_{\mathbb{R}} \hat{\gamma}(\eta)\, e^{-2\pi i \eta \frac{t_1}{a}}\, d\eta,
\end{aligned}$$

where $\hat{\gamma}(\eta) = \frac{1}{2\pi i \eta}\, \overline{\hat{\psi}(\eta, 0)}$. A similar argument to the one above shows that the elements $\langle H, \psi_{a,t}\rangle$ exhibit rapid asymptotic decay, as $a \to 0$, at all $t \in \mathbb{R}^2$ except at the location of the singularity $t_1 = 0$, where $\langle H, \psi_{a,t}\rangle$ behaves as $O(a)$, provided that $\int \hat{\gamma}(\eta) d\eta \neq 0$.

However, even though the continuous wavelet transform is able to identify the location of the singularities also in this case, the result of this second example is not completely satisfactory since it provides no information about the orientation of the singularity line. In dimensions larger than one, when the singularity points are supported on a curve or on a higher dimensional manifold, it is useful not only to detect the singularity location but also to capture its *geometry*, such as the orientation of a discontinuity curve or boundary.

As a matter of fact, it is possible to overcome this limitation by introducing generalized versions of the continuous wavelet transform that are more capable of dealing with directional information. The idea of considering generalized (discrete or continuous) wavelet transforms with improved directional capabilities has a long history, going back to the steerable filters [8, 23] introduced for the analysis of discrete data and to the notion of directional wavelets [1]. More recently, starting with the introduction of ridgelets [2] and curvelets [3, 4], a new generation of more flexible and powerful multiscale transforms has emerged, which has led to several successful discrete applications in signal and image processing. Among such more recent generalizations of the wavelet transform, the shearlet transform [9, 20] is especially remarkable since it combines a simple mathematical structure that is derived from the general framework of affine systems together with a special ability to capture the geometry of the singularity sets of multidimensional functions and distributions. For example, in the case of the two-dimensional Heaviside function, the continuous shearlet transform is able to determine both the location and the orientation of the discontinuity line. More generally, by extending and generalizing several results

derived previously by two of the authors, in this paper we show that the continuous shearlet transform provides a precise geometric description of the set of discontinuities of a large class of multivariate functions and distributions. These results provide the theoretical underpinning for improved algorithms for image analysis and feature extraction, cf. [25].

The rest of the paper is organized as follows. In Sect. 2, we recall the definition of the continuous shearlet transform; in Sect. 3, we present the shearlet analysis of jump discontinuities in the two-dimensional case; in Sect. 4, we illustrate the generalization of the shearlet approach to other types of singularities.

2 The Continuous Shearlet Transform

To define the continuous shearlet transform, we recall first the definition of the "generalized" continuous wavelet transform associated with the affine group on $\mathbb{R}^n$.

2.1 Wavelet Transforms

The *affine group* $\mathscr{A}$ on $\mathbb{R}^n$ consists of the pairs $(M, t) \in GL_n(\mathbb{R}) \times \mathbb{R}^n$, with group operation $(M, t) \cdot (M', t') = (MM', t + Mt')$. The *affine systems* generated by $\psi \in L^2(\mathbb{R}^n)$ are obtained from the action of the quasi-regular representation of $\mathscr{A}$ on ψ and are the collections of functions of the form

$$\left\{\psi_{M,t}(x) = |\det M|^{-\frac{1}{2}} \psi(M^{-1}(x - t)) : \ (M, t) \in \mathscr{A}\right\}.$$

Let $\Lambda = \{(M, t) : M \in G, t \in \mathbb{R}^n\} \subset \mathscr{A}$, where G is a subset of $GL_n(\mathbb{R})$. If there is an *admissible* function $\psi \in L^2(\mathbb{R}^n)$ such that any $f \in L^2(\mathbb{R}^n)$ can be recovered via the reproducing formula

$$f = \int_{\mathbb{R}^n} \int_G \langle f, \psi_{M,t} \rangle \, \psi_{M,t} \, d\lambda(M) \, dt,$$

where λ is a measure on G, then such ψ is a *continuous wavelet* associated with Λ and the mapping

$$f \to \mathscr{W}_\psi f(M, t) = \langle f, \psi_{M,t} \rangle, \quad (M, t) \in \Lambda,$$

is the *continuous wavelet transform* with respect to Λ. Depending on the choice of G and ψ, there is a variety of continuous wavelet transforms [21, 24]. The simplest case is $G = \{a I : a > 0\}$, where I is the identity matrix. In this situation, we obtain the classical continuous wavelet transform

$$\mathscr{W}_\psi f(a,t) = a^{-n/2} \int_{\mathbb{R}^n} f(x)\, a^{-1}\, \overline{\psi(a^{-1}(x-t))}\, dx,$$

which was used in Sect. 1 for $n = 1, 2$. Note that, in this case, the dilation group G is an isotropic since the dilation factor a acts in the same way for each coordinate direction. It is reasonable to expect that, by choosing more general dilation groups G, one obtains wavelet transforms with more interesting geometric properties.

2.2 The Shearlet Transform

The continuous shearlet transform is the continuous wavelet transform associated with a special subgroup $\mathscr{A}_S$ of $\mathscr{A}$ called the *shearlet group* (cf. [6, 7, 19, 20]). For a fixed $\beta = (\beta_1, \ldots, \beta_{n-1})$, where $0 < \beta_i < 1$, $1 \le i < n-1$, $\mathscr{A}_S$ consists of the elements (M_{as}, t), where

$$M_{as} = \begin{pmatrix} a & -a^{\beta_1} s_1 & \ldots & -a^{\beta_{n-1}} s_{n-1} \\ 0 & a^{\beta_1} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & a^{\beta_{n-1}} \end{pmatrix},$$

$a > 0$, $s = (s_1, \ldots, s_{n-1}) \in \mathbb{R}^{n-1}$, and $t \in \mathbb{R}^n$. Note that each matrix M_{as} is the product of the matrices $B_s A_a$, where

$$A_a = \begin{pmatrix} a & 0 & \ldots & 0 \\ 0 & a^{\beta_1} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & a^{\beta_{n-1}} \end{pmatrix}, \quad B_s = \begin{pmatrix} 1 & -s_1 & \ldots & -s_{n-1} \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix},$$

where A_a is an anisotropic dilation matrix and B_s is a nonexpanding matrix called a *shear matrix*. Hence, for an appropriate admissible function $\psi \in L^2(\mathbb{R}^n)$ and $\beta = (\beta_1, \ldots, \beta_{n-1})$, where $0 < \beta_i < 1$, the continuous shearlet transform is the mapping

$$f \to \langle f, \psi_{M_{as},t} \rangle, \quad (M_{as}, t) \in \mathscr{A}_S.$$

The analyzing elements $\psi_{M_{as},t}$ are called *shearlets* and are the affine functions

$$\psi_{M_{as},t}(x) = |\det M_{as}|^{-\frac{1}{2}} \psi(M_{as}^{-1}(x-t)).$$

In the following we will show that, thanks to the geometric and analytic properties of shearlets, the continuous shearlet transform enables a very precise description of jump discontinuities of functions of several variables. For example, if $f = \chi_S$, where $S \subset \mathbb{R}^n$, $n = 2, 3$, is a bounded region with piecewise smooth boundary, the continuous shearlet transform provides a characterization of the location and orientation of the boundary set through its asymptotic decay at fine scales.

2.3 The Shearlet Transform ($n = 2$)

Before applying the shearlet framework in dimensions $n = 2$, we need to specify the definition of the continuous shearlet transform that will be needed for our analysis.

For appropriate admissible functions $\psi^{(h)}, \psi^{(v)} \in L^2(\mathbb{R}^2)$, a fixed $0 < \beta < 1$, and matrices

$$M_{as} = \begin{pmatrix} a & -a^\beta s \\ 0 & a^\beta \end{pmatrix}, \quad N_{as} = \begin{pmatrix} a^\beta & 0 \\ -a^\beta s & a \end{pmatrix},$$

we define the *horizontal* and *vertical (continuous) shearlets* by

$$\psi^{(h)}_{a,s,t}(x) = |\det M_{as}|^{-\frac{1}{2}} \psi^{(h)}(M_{as}^{-1}(x - t)), \quad a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2,$$

and

$$\psi^{(v)}_{a,s,t}(x) = |\det N_{as}|^{-\frac{1}{2}} \psi^{(v)}(N_{as}^{-1}(x - t)), \quad a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2,$$

respectively. To ensure a more uniform covering of the range of directions through the shearing variable s, rather than using a single shearlet system where s ranges over $\mathbb{R}$, it will be convenient to use the two systems of shearlets defined above and let s range over a bounded interval.

To define our admissible functions $\psi^{(h)}, \psi^{(v)}$, for $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ let

$$\hat{\psi}^{(h)}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}), \quad \hat{\psi}^{(v)}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_2)\, \hat{\psi}_2(\tfrac{\xi_1}{\xi_2}), \tag{1}$$

where

$$\int_0^\infty |\hat{\psi}_1(a\omega)|^2 \frac{da}{a} = 1, \text{ for a.e. } \omega \in \mathbb{R}, \text{ and } \operatorname{supp} \hat{\psi}_1 \subset [-2, -\tfrac{1}{2}] \cup [\tfrac{1}{2}, 2]; \tag{2}$$

$$\|\psi_2\|_2 = 1 \text{ and } \operatorname{supp} \hat{\psi}_2 \subset [-\tfrac{\sqrt{2}}{4}, \tfrac{\sqrt{2}}{4}].$$

Observe that, in the frequency domain, a shearlet $\psi^{(h)}_{a,s,t}$ has the form:

$$\hat{\psi}^{(h)}_{a,s,t}(\xi_1, \xi_2) = a^{\frac{1+\beta}{2}} \hat{\psi}_1(a\,\xi_1)\, \hat{\psi}_2(a^{\beta-1}(\tfrac{\xi_2}{\xi_1} - s))\, e^{-2\pi i \xi \cdot t}. \tag{3}$$

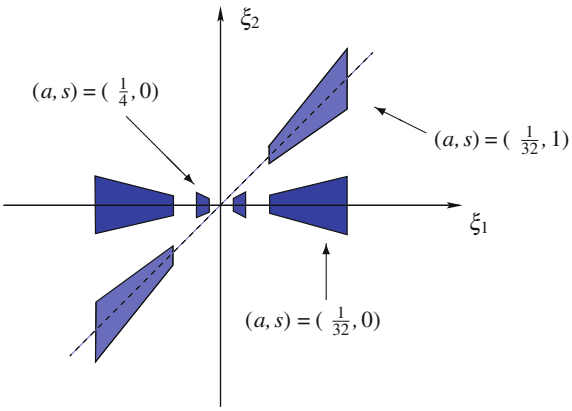


Fig. 1 Supports of the shearlets $\hat{\psi}^{(h)}_{ast}$ (in the frequency domain) for different values of a and s

This shows each function $\hat{\psi}^{(h)}_{a,s,t}$ has support:

$$\operatorname{supp}\hat{\psi}^{(h)}_{a,s,t} \subset \left\{(\xi_1,\xi_2) : \xi_1 \in \left[-\tfrac{2}{a},-\tfrac{1}{2a}\right] \cup \left[\tfrac{1}{2a},\tfrac{2}{a}\right],\ |\tfrac{\xi_2}{\xi_1} - s| \le a^{1-\beta}\right\}.$$

That is, its frequency support is a pair of trapezoids, symmetric with respect to the origin, oriented along a line of slope s. The support becomes increasingly thin as $a \to 0$. This is illustrated in Fig. 1. The shearlets $\psi^{(v)}_{a,s,t}$ have similar properties, with frequency supports oriented along lines of slopes $\frac{1}{s}$.

For $0 < a < \frac{1}{4}$ and $|s| \le \frac{3}{2}$, each system of continuous shearlets spans a subspace of $L^2(\mathbb{R}^2)$ consisting of functions having frequency supports in one of the horizontal or vertical cones defined in the frequency domain by

$$\mathscr{P}^{(h)} = \{(\xi_1,\xi_2) \in \mathbb{R}^2 : |\xi_1| \ge 2 \text{ and } |\tfrac{\xi_2}{\xi_1}| \le 1\},$$
$$\mathscr{P}^{(v)} = \{(\xi_1,\xi_2) \in \mathbb{R}^2 : |\xi_1| \ge 2 \text{ and } |\tfrac{\xi_2}{\xi_1}| > 1\}.$$

More precisely, the following proposition, which is a generalization of a result in [19], shows that the horizontal and vertical shearlets form a continuous reproducing system for the spaces of L^2 functions whose frequency support is contained in $\mathscr{P}^{(h)}$ and $\mathscr{P}^{(h)}$, respectively.

Proposition 1 *Let $\psi^{(h)}$ and $\psi^{(v)}$ be given by (1) with $\hat{\psi}_1$ and $\hat{\psi}_2$ satisfying (2) and (3), respectively. Let*

$$L^2(\mathscr{P}^{(h)})^\vee = \{f \in L^2(\mathbb{R}^2) :\ \operatorname{supp}\hat{f} \subset \mathscr{P}^{(h)}\},$$

with a similar definition for $L^2(\mathscr{P}^{(v)})^\vee$. We have the following:

(i) *For all $f \in L^2(\mathscr{P}^{(h)})^\vee$,*

$$f = \int_{\mathbb{R}^2} \int_{-2}^{2} \int_{0}^{\frac{1}{4}} \langle f, \psi^{(h)}_{a,s,t} \rangle \, \psi^{(h)}_{a,s,t} \, \frac{da}{a^3} \, ds \, dt.$$

(ii) *For all* $f \in L^2(\mathscr{P}^{(v)})^\vee$,

$$f = \int_{\mathbb{R}^2} \int_{-2}^{2} \int_{0}^{\frac{1}{4}} \langle f, \psi^{(v)}_{a,s,t} \rangle \, \psi^{(v)}_{a,s,t} \, \frac{da}{a^3} \, ds \, dt.$$

The equalities are understood in the L^2 *sense.*

Note that $\frac{da}{a^3}\, ds\, dt$ is the left Haar measure of the shearlet group $\mathscr{A}_S$.

Using the horizontal and vertical shearlets, we define the *(fine-scale) continuous shearlet transform* on $L^2(\mathbb{R}^2)$ as the mapping

$$f \in L^2(\mathbb{R}^2 \setminus [-2,2]^2)^\vee \to \mathscr{SH}_\psi f(a,s,t), \quad a \in (0, \tfrac{1}{4}], s \in [-\infty, \infty], t \in \mathbb{R}^2,$$

given by

$$\mathscr{SH}_\psi f(a,s,t) = \begin{cases} \mathscr{SH}^{(h)}_\psi f(a,s,t) = \langle f, \psi^{(h)}_{a,s,t} \rangle, & \text{if } |s| \le 1 \\ \mathscr{SH}^{(v)}_\psi f(a, \frac{1}{s}, t) = \langle f, \psi^{(v)}_{a,s,t} \rangle, & \text{if } |s| > 1. \end{cases}$$

In this expression, it is understood that the limit value $s = \pm\infty$ is defined and that $\mathscr{SH}_\psi f(a, \pm\infty, t) = \mathscr{SH}^{(v)}_\psi f(a, 0, t)$.

The term *fine-scale* refers to the fact that this shearlet transform is only defined for the scale variable $a \in (0, 1/4]$, corresponding to "fine scales". In fact, as it is clear from Proposition 1, the shearlet transform $\mathscr{SH}_\psi f$ defines an isometry on $L^2(\mathbb{R}^2 \setminus [-2,2]^2)^\vee$, the subspace of $L^2(\mathbb{R}^2)$ of functions with frequency support away from $[-2,2]^2$, but not on $L^2(\mathbb{R}^2)$. This is not a limitation since our method for the geometric characterization of singularities will require to derive asymptotic estimates as a approaches 0.

3 Shearlet Analysis of Jump Discontinuities in Dimension $n = 2$

To introduce the main ideas associated with the shearlet-based analysis of singularities, let us examine first the two-dimensional Heaviside function which was considered in Sect. 1. Using the Plancherel theorem and denoting $t = (t_1, t_2) \in \mathbb{R}^2$, when $|s| < 1$ we have

$$\begin{aligned} \mathscr{SH}_\psi H(a,s,t) &= \langle H, \psi^{(h)}_{a,s,t} \rangle \\ &= \int_{\mathbb{R}^2} \hat{H}(\xi_1, \xi_2) \, \overline{\hat{\psi}^{(h)}_{a,s,t}}(\xi_1, \xi_2) \, d\xi_1 \, d\xi_2 \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^2} \frac{\delta_2(\xi_1,\xi_2)}{2\pi i\xi_1}\,\overline{\hat{\psi}^{(h)}_{a,s,t}}(\xi_1,\xi_2)\,d\xi_1\,d\xi_2\\
&= \int_{\mathbb{R}} \frac{1}{2\pi i\xi_1}\,\overline{\hat{\psi}_{a,s,t}}(\xi_1,0)\,d\xi_1\\
&= a^{\frac{1+\beta}{2}} \int_{\mathbb{R}} \frac{1}{2\pi i\xi_1}\,\overline{\hat{\psi}_1}(a\,\xi_1)\,\overline{\hat{\psi}_2}(a^{\beta-1}s)\,e^{2\pi i\xi_1 t_1}\,d\xi_1\\
&= a^{\frac{1+\beta}{2}}\,\overline{\hat{\psi}_2}(a^{\beta-1}s) \int_{\mathbb{R}} \hat{\gamma}(\eta)\,e^{2\pi i\eta\frac{t_1}{a}}\,d\eta,
\end{aligned}
$$

where $\hat{\gamma}(\eta) = \frac{1}{2\pi i\eta}\overline{\hat{\psi}_1}(\eta)$. Hence, using the same argument from the introduction, under the assumption that $\hat{\psi}_1 \in C_c^\infty(\mathbb{R})$ we have that $\mathcal{SH}_\psi H(a,s,t)$ exhibits rapid asymptotic decay, as $a \to 0$, for all $(t_1,t_2) \in \mathbb{R}^2$ when $t_1 \neq 0$. If $t_1 = 0$ and $s \neq 0$, the term $\overline{\hat{\psi}_2}(a^{\beta-1}s)$ will vanish as $a \to 0$, due to the support assumptions on $\hat{\psi}_2$. Finally, if $t_1 = 0$ and $s = 0$, we have that

$$
\mathcal{SH}_\psi H(a,0,(0,t_2)) = a^{\frac{1+\beta}{2}}\,\overline{\hat{\psi}_2}(0) \int_{\mathbb{R}} \hat{\gamma}(\eta)\,d\eta.
$$

Hence, provided that $\hat{\psi}_2(0) \neq 0$ and $\int_{\mathbb{R}} \hat{\gamma}(\eta)\,d\eta \neq 0$, we have the estimate

$$
\mathcal{SH}_\psi H(a,0,(0,t_2)) = O(a^{\frac{1+\beta}{2}}).
$$

A similar computation shows that $\mathcal{SH}_\psi H(a,s,t)$ exhibits rapid asymptotic decay, as $a \to 0$, for all $|s| > 1$. In summary, under appropriate assumptions on ψ_1 and ψ_2, the continuous shearlet transform of H decays rapidly, asymptotically for $a \to 0$, for all t and s, unless t is on the discontinuous line and s corresponds to the normal direction of the discontinuous line at t.

The same properties of the continuous shearlet transform observed on the two-dimensional Heaviside function can be extended to any function of the form $f = \chi_S$ where $S \subset \mathbb{R}^2$ is a compact region whose boundary, denoted by ∂S, is a simple piecewise smooth curve, of finite length L. To define the normal orientation to the boundary curve ∂S, let $\alpha(t)$, $0 \le t \le L$ be a parameterization of ∂S. Let $p_0 = \alpha(t_0)$ and let $s_0 = \tan(\theta_0)$ with $\theta_0 \in (-\frac{\pi}{2}, \frac{\pi}{2})$. We say that s_0 *corresponds to the normal direction of* ∂S at p_0 if $(\cos\theta_0, \sin\theta_0) = \pm\mathbf{n}\,(t_0)$.

The following theorem generalizes a result proved originally in [10] for the special case $\beta = \frac{1}{2}$.

Theorem 1 *Let ψ_1, ψ_2 be chosen such that*

- $\hat{\psi}_1 \in C_c^\infty(\mathbb{R})$, $\operatorname{supp}\hat{\psi}_1 \subset [-2,-\frac{1}{2}] \cup [\frac{1}{2},2]$, *is odd, nonnegative on* $[\frac{1}{2},2]$ *and it satisfies* $\displaystyle\int_0^\infty |\hat{\psi}_1(a\xi)|^2\,\frac{da}{a} = 1$, *for a.e.* $\xi \in \mathbb{R}$; (4)

- $\hat{\psi}_2 \in C_c^\infty(\mathbb{R})$, $\operatorname{supp}\hat{\psi}_2 \subset [-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}]$, *is even, nonnegative, decreasing in* $[0, \frac{\sqrt{2}}{4})$, *and* $\|\psi_2\|_2 = 1$. (5)

Let $\frac{1}{3} < \beta < 1$. *For* $B = \chi_S$, *where* $S \subset \mathbb{R}^2$ *is a compact set whose boundary* ∂S *is a simple piecewise smooth curve, the following holds:*

(i) *If* $p \notin \partial S$ *then, for all* $s \in \mathbb{R}$,

$$\lim_{a\to 0^+} a^{-N}\, \mathcal{SH}_\psi B(a, s, p) = 0, \quad \text{for all } N > 0.$$

(ii) *If* $p_0 \in \partial S$ *is a regular point,* s_0 *corresponds to the normal direction of* ∂S *at* p_0 *and* $s \neq s_0$, *then*

$$\lim_{a\to 0^+} a^{-N}\, \mathcal{SH}_\psi B(a, s, p_0) = 0, \quad \text{for all } N > 0.$$

(iii) *If* $p_0 \in \partial S$ *is a regular point,* s_0 *corresponds to the normal direction of* ∂S *at* p_0 *and* $s = s_0$, *then*

$$\infty > \lim_{a\to 0^+} a^{-\frac{1+\beta}{2}}\, \mathcal{SH}_\psi B(a, s_0, p_0) \neq 0.$$

That is, if $p_0 \in \partial S$, the continuous shearlet transform decays rapidly, asymptotically for $a \to 0$, unless $s = s_0$ corresponds to the normal direction of ∂S at p_0, in which case

$$\mathcal{SH}_\psi B(a, s_0, p_0) = O(a^{\frac{1+\beta}{2}}), \text{ as } a \to 0.$$

Theorem 1 generalizes to the case of functions of the form $f = \chi_S$ where $S \subset \mathbb{R}^2$ and the boundary curve ∂S contains corner points. In this case, if p_0 is a corner point and s corresponds to one of the normal directions of ∂S at p_0, then the continuous shearlet transform has a decay rate of order $O(a^{\frac{1+\beta}{2}})$, as $a \to 0$, similar to the situation of regular points. For other values of s, however, the asymptotic decay rate depends both on the tangent and the curvature at p_0 (cf. [10]).

Theorem 1 was originally proved in [10] for the case $\beta = 1/2$ and its proof was successively simplified and streamlined in [13]. In the following section, we sketch the main ideas of the proof, highlighting how to extend the proof from [13] to the case $\beta \neq 1/2$.

3.1 Proof of Theorem 1 (Sketch)

The argument used for the two-dimensional Heaviside function cannot be extended to this case directly since this would require an explicit expression of the Fourier transform of the function $B = \chi_S$. Instead, we can apply the divergence theorem that allows us to express the Fourier transform of B as a line integral over ∂S:

$$\begin{aligned}\hat{B}(\xi) = \widehat{\chi_S}(\xi) &= \int_S e^{-2\pi i \xi \cdot x}\, dx \\ &= -\frac{1}{2\pi i \|\xi\|^2} \int_{\partial S} e^{-2\pi i \langle \xi, x\rangle}\, \xi \cdot \mathbf{n}\,(x)\, d\sigma(x),\end{aligned} \tag{6}$$

for all $\xi \neq 0$, where ∂S is the boundary of S, $\mathbf{n}$ is the unit outward normal to S, and σ is one-dimensional Hausdorff measure on $\mathbb{R}^2$.

Hence, using (6), we have that

$$\begin{aligned}\mathcal{SH}_\psi B(a,s,p) &= \langle B, \psi^{(d)}_{a,s,p}\rangle \\ &= \langle \hat{B}, \hat{\psi}^{(d)}_{a,s,p}\rangle \\ &= \int_{\mathbb{R}^2} \hat{B}(\xi)\, \overline{\hat{\psi}^{(d)}_{a,s,p}(\xi)}\, d\xi \\ &= -\frac{1}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(d)}_{a,s,p}(\xi)}}{\|\xi\|^2} \int_{\partial S} e^{-2\pi i \xi \cdot x}\, \xi \cdot \mathbf{n}\,(x)\, d\sigma(x)\, d\xi,\end{aligned} \tag{7}$$

where the superscript in $\psi^{(d)}_{a,s,p}$ is either $d = h$, when $|s| \leq 1$, or $d = v$, when $|s| > 1$.

One can observe that the asymptotic decay of the shearlet transform $\mathcal{SH}_\psi B(a,s,p)$, as $a \to 0$, is only determined by the values of the boundary ∂S which are "close" to p. Hence, for $\varepsilon > 0$, let $D(\varepsilon, p)$ be the ball in $\mathbb{R}^2$ of radius ε and center p, and $D^c(\varepsilon, p) = \mathbb{R}^2 \setminus D(\varepsilon, p)$. Using (7), we can write the shearlet transform of B as

$$\mathcal{SH}_\psi B(a,s,p) = I_1(a,s,p) + I_2(a,s,p),$$

where

$$I_1(a,s,p) = -\frac{1}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(d)}_{a,s,p}(\xi)}}{\|\xi\|^2} \int_{\partial S \cap D(\varepsilon,p)} e^{-2\pi i \xi \cdot x}\, \xi \cdot \mathbf{n}\,(x)\, d\sigma(x)\, d\xi, \tag{8}$$

$$I_2(a,s,p) = -\frac{1}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(d)}_{a,s,p}(\xi)}}{\|\xi\|^2} \int_{\partial S \cap D^c(\varepsilon,p)} e^{-2\pi i \xi \cdot x}\, \xi \cdot \mathbf{n}\,(x)\, d\sigma(x)\, d\xi. \tag{9}$$

The Localization Lemma below (whose assumptions are satisfied by the shearlet generator function in Theorem 1) shows that I_2 has rapid asymptotic decay at fine scales. For its proof, we need the following "repeated integration by parts" lemma whose proof follows easily from induction and the standard integration by parts

result. Note that this version of the Localization Lemma is more general than the one that appeared in [10, 13], since it does not assume a special form of the function ψ.

Lemma 1 *Let $N \in \mathbb{Z}^+ = \{1, 2, 3, \dots\}$ and let $f, g \in C^N(\mathbb{R})$ be such that $f^{(n)} g^{(N-1-n)}$ vanishes at ∞, for all $n = 0, \dots, N-1$, and $f^{(n)} g^{(N-n)} \in L^1(\mathbb{R})$, for all $n = 0, \dots, N$. Then,*

$$\int_{\mathbb{R}} f(x) g^{(N)}(x)\, dx = (-1)^N \int_{\mathbb{R}} f^{(N)}(x) g(x)\, dx.$$

Lemma 2 (Localization Lemma) *Fix $p \in \mathbb{R}^2$ and let $N \in \mathbb{Z}^+$. Suppose that*

(i) $\hat{\psi}^{(d)} \in C^N(\mathbb{R}^2)$, *for $d = h, v$;*
(ii) $\partial^\omega \hat{\psi}^{(d)} \in L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$, *for all $0 \le |\omega| \le N-1$ and $d = h, v$;*
(iii) $\partial^\omega \hat{\psi}^{(d)} / r_d^{N+1-|\omega|} \in L^1(\mathbb{R}^2)$, *for all $0 \le |\omega| \le N$ and $d = h, v$, where*

$$r_d(\xi) = \begin{cases} |\xi_1|, & \text{if } d = h \\ |\xi_2|, & \text{if } d = v. \end{cases}$$

Then, there exists a constant $0 < C < \infty$ such that

$$|I_2(a, s, p)| \le C\, a^{N\beta + (1-\beta)/2},$$

for all a and s.

Proof Fix $0 < a \le 1/4$ and $s \in \mathbb{R}$. We may assume that $s \le 1$ and $d = h$. Substituting for $\hat{\psi}^{(h)}_{a,s,p}$ and using (9), the change of variable $\eta_1 = a\xi_1$ and $\eta_2 = a^\beta \xi_2 - a^\beta s \xi_1$, and some algebraic manipulation, we have

$$\begin{aligned}
&I_2(a, s, p) \\
&= \frac{-a^{(1+\beta)/2}}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(h)}(a\xi_1, a^\beta \xi_2 - a^\beta s \xi_1)}}{\|\xi\|^2} \int_{\partial S \cap D^c(\varepsilon, p)} e^{-2\pi i \xi \cdot (x-p)}\, \xi \cdot \mathbf{n}(x)\, d\sigma(x)\, d\xi \\
&= \frac{-a^{-(1+\beta)/2}}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(h)}(\eta)}}{a^{-2}\eta_1^2 + (a^{-\beta}\eta_2 + a^{-1} s \eta_1)^2} \int_{\partial S \cap D^c(\varepsilon, p)} e^{-2\pi i (a^{-1}\eta_1, a^{-\beta}\eta_2 + a^{-1} s \eta_1) \cdot (x-p)} \\
&\quad \times (a^{-1}\eta_1, a^{-\beta}\eta_2 + a^{-1} s \eta_1) \cdot \mathbf{n}(x)\, d\sigma(x)\, d\eta \\
&= \frac{-a^{(1-\beta)/2}}{2\pi i} \int_{\mathbb{R}^2} \frac{\overline{\hat{\psi}^{(h)}(\eta)}}{\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2} \int_{\partial S \cap D^c(\varepsilon, p)} e^{-2\pi i a^{-1}\eta_1 [(x_1 - p_1) + s(x_2 - p_2)]} \\
&\quad \times (\eta_1, a^{1-\beta}\eta_2 + s\eta_1) \cdot \mathbf{n}(x) e^{-2\pi i a^{-\beta}\eta_2 (x_2 - p_2)}\, d\sigma(x)\, d\eta. \qquad (10)
\end{aligned}$$

Note also that

$$\int_{\mathbb{R}^2} \left| \frac{\overline{\hat{\psi}^{(h)}(\eta)}}{\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2} \right| \int_{\partial S \cap D^c(\varepsilon, p)} \left| (\eta_1, a^{1-\beta}\eta_2 + s\eta_1) \cdot \mathbf{n}(x) \right|$$

$$\times \left| e^{-2\pi i a^{-1}\eta_1[(x_1-p_1)+s(x_2-p_2)]} e^{-2\pi i a^{-\beta}\eta_2(x_2-p_2)} \right| d\,\sigma(x)\,d\eta$$
$$\leq \int_{\mathbb{R}^2} \frac{|\hat{\psi}^{(h)}(\eta)|}{\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2} \int_{\partial S \cap D^c(\varepsilon,p)} \|(\eta_1, a^{1-\beta}\eta_2 + s\eta_1)\| \|\mathbf{n}(x)\|\, d\,\sigma(x)\,d\eta$$
$$\leq \sigma(\partial S) \int_{\mathbb{R}^2} \frac{|\hat{\psi}^{(h)}(\eta)|}{r_h(\eta)}\, d\eta < \infty, \tag{11}$$

where, in the last inequality, we have used properties (ii) and (iii) in the statement of this lemma.

Choose $\delta > 0$ not depending on s and disjoint Borel measurable subsets $E_q \subset \mathbb{R}^2$, for $q = 1, 2$, satisfying

$$E_q \subset \{x \in \mathbb{R}^2 : |(x_q - p_q) + s_q(x_2 - p_2)| \geq \delta\} \text{ and } E_1 \cup E_2 = \partial S \cap D^c(\varepsilon, p), \tag{12}$$

where $s_1 = s$ and $s_2 = 0$. Then, using (10), (11), and the Fubini–Tonelli theorem, it follows that

$$I_2(a,s,p) = \frac{-a^{(1-\beta)/2}}{2\pi i} \sum_{q=1,2} \int_{E_q} \int_{\mathbb{R}^2} f_a(x,\eta) \times e^{-2\pi i a^{-1}\eta_1[(x_1-p_1)+s(x_2-p_2)]} e^{-2\pi i a^{-\beta}\eta_2(x_2-p_2)}\, d\eta\, d\,\sigma(x), \tag{13}$$

where $f_a : \partial S \times \mathbb{R}^2 \to \mathbb{C}$ is defined by

$$f_a(x,\eta) = \frac{(\eta_1, a^{1-\beta}\eta_2 + s\eta_1)\cdot \mathbf{n}(x)}{\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2} \overline{\hat{\psi}^{(h)}(\eta)}$$

for a.e. (x,η). We require the following claim, whose proof is a straightforward application of induction and the quotient rule.

For each $q \in \{1,2\}$ and $n \in \{0,\ldots,N\}$, there exists $L_n^q \in \mathbb{Z}^+$ and, for each $l = 1,\ldots,L_n^q$, there exist $\gamma_l^{qn} \geq 0$, $c_l^{qn} \in L^\infty(S,\sigma)$ not depending on a, η, or s, a monomial $m_l^{qn} : \mathbb{R}^2 \to \mathbb{R}$, and a multi-index ω_l^{qn} with $|\omega_l^{qn}| \leq n$ and $|\omega_l^{qn}| = \deg(m_l^{qn}) - 2^{n+1} + n + 1$ such that

$$\frac{\partial^n}{\partial \eta_q^n} f_a(x,\eta) = \sum_{l=1}^{L_n^q} \frac{a^{\gamma_l^{qn}} c_l^{qn}(x) m_l^{qn}(\eta_1, a^{1-\beta}\eta_2 + s\eta_1)}{(\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2)^{2^n}} \overline{\partial^{\omega_l^{qk}} \hat{\psi}^{(h)}(\eta)},$$

for a.e. (x,η). We are using monomial in the strict sense (i.e., $\eta_1\eta_2$ is a monomial but $-\eta_1\eta_2$ and $2\eta_1\eta_2$ are not).

If $q \in \{1,2\}$, choose r such that $\{q,r\} = \{1,2\}$. If $m : \mathbb{R}^2 \to \mathbb{R}$ is a monomial and $\gamma \in \mathbb{R}$, then, by switching to spherical coordinates, it is clear that $|m(\eta)|/\|\eta\|^\gamma \leq 1/\|\eta\|^{\gamma - \deg(m)}$, for all $\eta \neq 0$. Using this and the claim, if $n \in \{0,\ldots,N\}$, we have

$$\begin{aligned}\left|\frac{\partial^n}{\partial\eta_q^n} f_a(x,\eta)\right| &\le \sum_{l=1}^{L_n^q} \|c_l^{qn}\|_\infty \left|\frac{m_l^{qn}(\eta_1, a^{1-\beta}\eta_2 + s\eta_1)}{(\eta_1^2 + (a^{1-\beta}\eta_2 + s\eta_1)^2)^{2^n}}\right| |\partial^{\omega_l^{qn}} \hat{\psi}^{(h)}(\eta)| \\ &\le \sum_{l=1}^{L_n^q} \|c_l^{qn}\|_\infty \frac{|\partial^{\omega_l^{qn}} \hat{\psi}^{(h)}(\eta)|}{\|(\eta_1, a^{1-\beta}\eta_2 + s\eta_1)\|^{k+1-|\omega_l^{qn}|}} \qquad (14) \\ &\le \sum_{l=1}^{L_n^q} \|c_l^{qn}\|_\infty \frac{|\partial^{\omega_l^{qn}} \hat{\psi}^{(h)}(\eta)|}{r_h(\eta)^{n+1-|\omega_l^{qn}|}},\end{aligned}$$

for a.e. (x, η). The second inequality, together with the claim and property (ii) of $\psi^{(h)}$, implies that $\frac{\partial^n}{\partial\eta_q^n} f_a(x,\cdot)$ vanishes at ∞, for $n = 0, \ldots, N-1$ and σ-a.e. x. The third inequality, together with the claim and properties (ii) and (iii) of $\psi^{(h)}$ implies that $\frac{\partial^n}{\partial\eta_q^n} f_a(x,\cdot) \in L^1(\mathbb{R}^2)$, for $n = 0, \ldots, N$ and σ-a.e. x.

Using the observations of the previous paragraph, the Fubini–Tonelli theorem, Lemma 1, (12), (14), the claim, and property (i) of $\psi^{(h)}$, we obtain

$$\begin{aligned}&\left|\int_{E_q}\int_{\mathbb{R}^2} f_a(x,\eta) e^{-2\pi i a^{-1}\eta_1[(x_1-p_1)+s(x_2-p_2)]} e^{-2\pi i a^{-\beta}\eta_2(x_2-p_2)}\, d\eta\, d\sigma(x)\right| \\ &\le \int_{E_q}\int_{\mathbb{R}} \left|\int_{\mathbb{R}} f_a(x,\eta) e^{-2\pi i a^{-\beta_q}\eta_q[(x_q-p_q)+s_q(x_2-p_2)]}\, d\eta_q\right| d\eta_r\, d\sigma(x) \\ &= \int_{E_q}\int_{\mathbb{R}}\left|\int_{\mathbb{R}} f_a(x,\eta) \frac{\partial^N}{\partial\eta_q^N}\left(\frac{e^{-2\pi i a^{-\beta_q}\eta_q[(x_q-p_q)+s_q(x_2-p_2)]}}{\left(-2\pi i a^{-\beta_q}[(x_q-p_q)+s_q(x_2-p_2)]\right)^N}\right) d\eta_q\right| d\eta_r\, d\sigma(x) \\ &= \int_{E_q}\int_{\mathbb{R}}\left|\int_{\mathbb{R}} \frac{\partial^N}{\partial\eta_q^N} f_a(x,\eta) \frac{e^{-2\pi i a^{-\beta_q}\eta_q[(x_q-p_q)+s_q(x_2-p_2)]}}{\left(-2\pi i a^{-\beta_q}[(x_q-p_q)+s_q(x_2-p_2)]\right)^N} d\eta_q\right| d\eta_r\, d\sigma(x) \\ &\le \frac{a^{N\beta_q}}{(2\pi\delta)^N}\int_{E_q}\int_{\mathbb{R}^2}\left|\frac{\partial^N}{\partial\eta_q^N} f_a(x,\eta)\right| d\eta\, d\sigma(x) \\ &\le \frac{\sigma(\partial S) a^{N\beta_q}}{(2\pi\delta)^N}\sum_{l=1}^{L_N^q} \|c_l^{qN}\|_\infty \|\partial^{\omega_l^{qN}} \hat{\psi}^{(h)} / r_h^{N+1-|\omega_l^{qN}|}\|_1,\end{aligned}$$

where $\beta_1 = 1$ and $\beta_2 = \beta$. The lemma follows from the claim, property (iii) of $\psi^{(h)}$, the above inequality, and (13). □

For the analysis of the term I_1, we will use a local approximation of the curve ∂S.

Let $\alpha(t)$ be the boundary curve ∂S, with $0 \le t \le L$, and $p \in \partial S$. Without loss of generality, we may assume that $L > 1$ and $p = (0,0) = \alpha(1)$. We can write the boundary curve near p as $\mathscr{C} = \partial S \cap D(\varepsilon, (0,0))$, where

$$\mathscr{C} = \{\alpha(t) : 1 - \varepsilon \le t \le 1 + \varepsilon\}.$$

Rather than using the arclength representation of $\mathscr{C}$, we can also write $\mathscr{C} = \{(G(u), u),\ -\varepsilon \le u \le \varepsilon\}$, where $G(u)$ is a smooth function. Since $p = (0, 0)$, then $G(0) = 0$. Hence, we define the quadratic approximation of ∂S near $p = (0, 0)$ by $\partial S_0 = (G_0(u), u)$, where G_0 is the Taylor polynomial of degree 2 of G centered at the origin, given by $G_0(u) = G'(0)u + \frac{1}{2}G''(0)u^2$. Accordingly, we define $B_0 = \chi_{S_0}$, where S_0 is obtained by replacing the curve ∂S in $B = \chi_S$ with the quadratic curve ∂S_0 near the point $p = (0, 0)$.

The following lemma, which is a generalization from [13], shows that to derive the estimates of Theorem 1 it is sufficient to replace the set B with set B_0, since this produces a "low-order" error. Note that this approximation result only holds for $\frac{1}{3} < \beta < 1$, that is, when the anisotropic scaling factor of the dilation matrices is not too high. The argument provided below does not extend to smaller values of β. Possibly this restriction could be removed by considering a higher order polynomial approximation for the boundary curve ∂S, but this would make the rest of the proof of Theorem 1 significantly more involved.

Lemma 3 *Let $\frac{1}{3} < \beta < 1$. For any $|s| \le \frac{3}{2}$, we have*

$$\lim_{a \to 0^+} a^{-\frac{1+\beta}{2}} \left| \mathscr{SH}_\psi B(a, s, 0)) - \mathscr{SH}_\psi B_0(a, s, 0) \right| = 0.$$

Proof Let $p = (0, 0) \in \partial S$. Since we assume $|s| \le \frac{3}{2}$, we need to use the system of "horizontal" shearlets only.

Let γ be chosen such that $\frac{1+\beta}{4} < \gamma < \beta$ (this can be satisfied for $\frac{1}{3} < \beta < 1$) and assume that a is sufficiently small, so that $a^\gamma \ll 1$. A direct calculation shows that

$$\begin{aligned} \left| \mathscr{SH}_\psi B(a, s, 0) - \mathscr{SH}_\psi B_0(a, s, 0) \right| &\le \int_{\mathbb{R}^2} |\psi^{(h)}_{a,s,0}(x)|\, |\chi_S(x) - \chi_{S_0}(x)|\, dx \\ &= T_1(a) + T_2(a), \end{aligned}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ and

$$\begin{aligned} T_1(a) &= a^{-\frac{1+\beta}{2}} \int_{D(a^\gamma, (0,0))} |\psi^{(h)}(M_{as}^{-1}x)|\, |\chi_S(x) - \chi_{S_0}(x)|\, dx, \\ T_2(a) &= a^{-\frac{1+\beta}{2}} \int_{D^c(a^\gamma, (0,0))} |\psi^{(h)}(M_{as}^{-1}x)|\, |\chi_S(x) - \chi_{S_0}(x)|\, dx. \end{aligned}$$

Observe that:

$$T_1(a) \le C\, a^{-\frac{1+\beta}{2}} \int_{D(a^\gamma, (0,0))} |\chi_S(x) - \chi_{S_0}(x)|\, dx.$$

To estimate the above quantity, it is enough to compute the area between the regions S and S_0. Since G_0 is the Taylor polynomial of G of degree 2, we have

$$T_1(a) \le C\, a^{-\frac{1+\beta}{2}} \int_{|x|<a^\gamma} |x|^3 dx \le C\, a^{4\gamma-\frac{1+\beta}{2}}.$$

Since $\gamma > \frac{1+\beta}{4}$, the above estimate shows that $T_1(a) = o(a^{\frac{1+\beta}{2}})$.

The assumptions on the generator function $\psi^{(h)}$ of the shearlet system $\psi^{(h)}$ imply that, for each $N > 0$, there is a constant $C_N > 0$ such that $|\psi(x)| \le C_N\,(1+|x|^2)^{-N}$. Also note that $(M_{as})^{-1} = A_a^{-1}\, B_s^{-1}$, where $B_s^{-1} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ and $A_a^{-1} = \begin{pmatrix} a^{-1} & 0 \\ 0 & a^{-\beta} \end{pmatrix}$. It is easy to verify that, for all $|s| \le \frac{3}{2}$, there is a constant $C_0 > 0$ such that $\|B_s^{-1}x\|^2 \ge C_0\|x\|^2$, or $(x_1+sx_2)^2 + x_2^2 \ge C_0(x_1^2+x_2^2)$, for all $x \in \mathbb{R}^2$. Thus, for $a < 1$, we can estimate $T_2(a)$ as:

$$\begin{aligned}
T_2(a) &\le C\, a^{-\frac{1+\beta}{2}} \int_{D^c(a^\gamma,(0,0))} |\psi^{(h)}(M_{as}x)|\, dx \\
&\le C_N\, a^{-\frac{1+\beta}{2}} \int_{D^c(a^\gamma,(0,0))} \left(1 + (a^{-1}(x_1+sx_2))^2 + (a^{-\beta}x_2)^2\right)^{-N} dx \\
&\le C_N\, a^{-\frac{1+\beta}{2}} \int_{D^c(a^\gamma,(0,0))} \left((a^{-\beta}(x_1+sx_2))^2 + (a^{-\beta}x_2)^2\right)^{-N} dx \\
&\le C_N\, a^{2\beta N-\frac{1+\beta}{2}} \int_{D^c(a^\gamma,(0,0))} (x_1^2+x_2^2)^{-N}\, dx \\
&= C_N\, a^{2\beta N-\frac{1+\beta}{2}} \int_{a^\gamma}^{\infty} r^{1-2N}\, dr \\
&= C_N\, a^{2N(\beta-\gamma)}\, a^{2\gamma-\frac{1+\beta}{2}},
\end{aligned}$$

where the constant C_0 was absorbed in C_N. Since $\gamma < \beta$ and N can be chosen arbitrarily large, it follows that $T_2(a) = o(a^{\frac{1+\beta}{2}})$. □

The proof of Theorem 1 can now be completed using Lemmata 2 and 3, following the arguments from [13].

4 Shearlet Analysis of General Singularities

The shearlet analysis of singularities extends beyond the case of functions of the form χ_S considered in the previous sections. The results presented below illustrate the shearlet analysis of singularities of rather general functions.

As a first case, we will examine the case of "general" functions of two variables containing jump discontinuities. Let S be a bounded open subset of $\mathbb{R}^2$ and

assume that its boundary ∂S is generated by a C^3 curve that can be parameterized as $(\rho(\theta)\cos\theta, \rho(\theta)\sin\theta)$ where $\rho(\theta) : [0, 2\pi) \to [0, 1]$ is a radius function. We will consider functions of the form $f\,\chi_S$, where f is a smooth function. Note that this model is a special case of the class of cartoon-like images, where the set ∂S describes the edge of an object. Similar image models are commonly used, for example, in the variational approach to image processing (cf. [5, Chap. 3]).

We have the following result, which is a refinement from an observation in [14]:

Theorem 2 *Let ψ_1, ψ_2, β be chosen as in Theorem 1. Let $B = f\,\chi_S$, where $S \subset \mathbb{R}^2$ is a bounded region whose boundary ∂S is a simple C^3 curve and $f \in C^\infty(\mathbb{R}^2)$. Then we have the following results:*

(i) *If $p \notin \partial S$ then, for all $s \in \mathbb{R}$,*

$$\lim_{a\to 0^+} a^{-N}\, \mathcal{SH}_\psi B(a, s, p) = 0, \quad \text{for all } N > 0.$$

(ii) *If $p_0 \in \partial S$ is a regular point, s_0 corresponds to the normal direction of ∂S at p_0 and $s \neq s_0$, then*

$$\lim_{a\to 0^+} a^{-N}\, \mathcal{SH}_\psi B(a, s, p_0) = 0, \quad \text{for all } N > 0.$$

(iii) *If $p_0 \in \partial S$ is a regular point, s_0 corresponds to the normal direction of ∂S at p_0, $s = s_0$ and $f(p) \neq 0$, then*

$$\lim_{a\to 0^+} a^{-\frac{1+\beta}{2}}\, \mathcal{SH}_\psi B(a, s_0, p_0) \neq 0.$$

For simplicity of notation, we will prove Theorem 2 in the special case, where $\beta = \frac{1}{2}$. The case of general $\frac{1}{3} < \beta < 1$ can be easily derived from here. For the proof, we need first the following lemma (where we assume $\beta = \frac{1}{2}$).

Lemma 4 *Let $S \subset \mathbb{R}^2$ be a bounded region whose boundary ∂S is a simple C^3 curve. Assume that $p_0 \in \partial S$ is a regular point and P_S is a polynomial with $P_S(p_0) = 0$. For any $N > 0$, we have*

(i) $\lim_{a\to 0} a^{-N} \langle P_S\, \chi_S, \psi^{(h)}_{a,s,p_0}\rangle = 0, \quad s \neq \pm\mathbf{n}(p_0),$

(ii) $\lim_{a\to 0} a^{-\frac{5}{4}} \langle P_S\, \chi_S, \psi^{(h)}_{a,s,p_0}\rangle = C, \quad s = \pm\mathbf{n}(p_0),$

where C is a finite real number.

Proof We only prove the lemma when P_S is a polynomial of degree 2, since the same argument works for a polynomial of degree >2. Without loss of generality, we may assume $p_0 = (0, 0)$ and that near p_0, we have that $\partial S = \{(g(u), u),\ -\varepsilon < u < \varepsilon\}$, where $g(u) = Au^2 + Bu$. Also we may write $s = \tan\theta_0$ with $\theta_0 = 0$.

Recall that, by the divergence theorem,

$$\widehat{\chi_S}(\rho,\theta) = -\frac{1}{2\pi i\rho}\int_{\partial S} e^{-2\pi i\rho\Theta(\theta)\cdot x}\,\Theta(\theta)\cdot\mathbf{n}(x)\,d\sigma(x)$$
$$= -\frac{1}{2\pi i\rho}\int_0^L e^{-2\pi i\rho\,\Theta(\theta)\cdot\boldsymbol{\alpha}(t)}\,\Theta(\theta)\cdot\mathbf{n}(t)\,dt.$$

Since $P_S(0) = 0$, we can write $P_S(x)$ as $A_1x_1 + A_2x_2 + A_3x_1^2 + A_4x_1x_2 + A_5x_2^2$. Let $P_S(\frac{i}{2\pi}D)$ be the differential operator obtained from the polynomial $P_S(x)$ by replacing x_1 with $\frac{i}{2\pi}\frac{\partial}{\partial\xi_1}$ and x_2 with $\frac{i}{2\pi}\frac{\partial}{\partial\xi_2}$.

A direct computation gives that

$$\frac{\partial}{\partial\xi_1}\left(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))\right)$$
$$= a\hat{\psi}_1{}'(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)) - \frac{\xi_2}{\xi_1^2}a^{-\frac{1}{2}}\hat{\psi}_1(a\xi_1)\hat{\psi}_2{}'(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)),$$

and

$$\frac{\partial^2}{\partial\xi_1^2}\left(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))\right)$$
$$= a^2\hat{\psi}_1{}''(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)) - a^{\frac{1}{2}}\frac{\xi_2}{\xi_1^2}\hat{\psi}_1{}'(a\xi_1)\hat{\psi}_2{}'(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))$$
$$+a^{-\frac{1}{2}}\frac{2\xi_2}{\xi_1^3}\hat{\psi}_1(a\xi_1)\hat{\psi}_2{}'(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)) - a^{\frac{1}{2}}\frac{\xi_2}{\xi_1^2}\hat{\psi}_1{}'(a\xi_1)\hat{\psi}_2{}'(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))$$
$$+(a^{-\frac{1}{2}}\frac{\xi_2}{\xi_1^2})^2\hat{\psi}_1(a\xi_1)\hat{\psi}_2{}''(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)).$$

Using these expressions, we obtain that

$$\langle P_S\,\chi_S, \psi^{(h)}_{a,s,p}\rangle = \langle\chi_S, P_S\,\psi^{(h)}_{a,s,p}\rangle$$
$$= \langle\widehat{\chi_S}, \widehat{P_S\,\psi^{(h)}_{a,s,p}}\rangle$$
$$= \langle\widehat{\chi_S}, P_S(\frac{i}{2\pi}D)\widehat{(\psi^{(h)}_{a,s,p})}\rangle$$
$$= \sum_{m=1}^{5} J_m(a,s,p),$$

where, using $p = (0,0)$,

$$J_1(a,s,0) = \frac{A_1 i}{2\pi}\langle\widehat{\chi_S}, \frac{\partial}{\partial\xi_1}(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)))\rangle,$$

$$J_2(a,s,0) = \frac{A_2 i}{2\pi}\langle \widehat{\chi_S}, \frac{\partial}{\partial \xi_2}(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1} - s)))\rangle,$$

$$J_3(a,s,0) = -\frac{A_3}{(2\pi)^2}\langle \widehat{\chi_S}, \frac{\partial^2}{\partial \xi_1^2}(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1} - s)))\rangle,$$

$$J_4(a,s,0) = -\frac{A_4}{(2\pi)^2}\langle \widehat{\chi_S}, \frac{\partial^2}{\partial \xi_1 \partial \xi_2}(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1} - s)))\rangle,$$

$$J_5(a,s,0) = -\frac{A_5}{(2\pi)^2}\langle \widehat{\chi_S}, \frac{\partial^2}{\partial \xi_2^2}(\hat{\psi}_1(a\xi_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1} - s)))\rangle.$$

Since $s \neq \pm \mathbf{n}(p_0)$, by integration by parts, it is easy to see that for each $N > 0$, we have $\widehat{\chi_S}(a^{-1}\rho, \theta) = O(a^N)$, as $a \to 0$, uniformly for all ρ and θ. For each J_m, let $\xi = \rho\,\Theta(\theta)$ and $a\rho = \rho'$, by the Localization Lemma (Lemma 2) we see that $J_m = O(a^N)$ for $m = 1, 2, 3, 4, 5$ and this proves part (i).

For $s = \pm \mathbf{n}(p_0)$, let us first examine the term J_1. By the Localization Lemma, we can assume that J_1 has the following expression:

$$\begin{aligned}
&J_1(a,s,0)\\
&= \frac{A_1 i}{2\pi}\int_{R^2} \widehat{\chi_S}(\xi)\frac{\partial}{\partial \xi_1}\left(\hat{\psi}_1(a\xi_1)\,\hat{\psi}_2(a^{-\frac{1}{2}}(\tfrac{\xi_2}{\xi_1}))\right) d\xi\\
&= -\frac{a^{\frac{3}{4}}A_1}{(2\pi)^2}\int_0^\infty \int_0^{2\pi}\int_{-\varepsilon}^{\varepsilon} e^{-2\pi i\rho\,\Theta(\theta)\cdot(g(u),u)}\,\Theta(\theta)\cdot\mathbf{n}(u)\,du\\
&\times\left(a\hat{\psi}_1{}'(a\rho\cos\theta)\hat{\psi}_2(a^{-\frac{1}{2}}\tan\theta) - \frac{a^{-\frac{1}{2}}\tan\theta}{\rho\cos\theta}\hat{\psi}_1(a\rho\cos\theta)\hat{\psi}_2{}'(a^{-\frac{1}{2}}\tan\theta)\right)d\theta\,d\rho\\
&= -\frac{a^{-\frac{1}{4}}A_1}{(2\pi)^2}\int_0^\infty \int_0^{2\pi}\int_{-\varepsilon}^{\varepsilon} e^{-2\pi i a^{-1}\rho\,\Theta(\theta)\cdot(g(u),u)}\,\Theta(\theta)\cdot\mathbf{n}(u)\,du\\
&\quad\times\left(a\hat{\psi}_1{}'(\rho\cos\theta)\hat{\psi}_2(a^{-\frac{1}{2}}\tan\theta) - \frac{a^{\frac{1}{2}}\tan\theta}{\rho\cos\theta}\hat{\psi}_1(\rho\cos\theta)\hat{\psi}_2{}'(a^{-\frac{1}{2}}\tan\theta)\right)d\theta\,d\rho\\
&= J_{11}(a,s,0) + J_{12}(a,s,0),
\end{aligned}$$

where

$$\begin{aligned}
J_{11}(a,s,0) = -\frac{a^{-\frac{1}{4}}A_1}{(2\pi)^2}\int_0^\infty \int_0^{2\pi}\int_{-\varepsilon}^{\varepsilon} e^{-2\pi i a^{-1}\rho\,\Theta(\theta)\cdot(g(u),u)}\,\Theta(\theta)\cdot\mathbf{n}(u)\,du\\
\times\, a\,\hat{\psi}_1{}'(\rho\cos\theta)\,\hat{\psi}_2(a^{-\frac{1}{2}}\tan\theta)\,d\theta\,d\rho,
\end{aligned}$$

$$J_{12}(a,s,0) = \frac{a^{-\frac{1}{4}}A_1}{(2\pi)^2}\int_0^\infty \int_0^{2\pi}\int_{-\varepsilon}^{\varepsilon} e^{-2\pi i a^{-1}\rho\,\Theta(\theta)\cdot(g(u),u)}\,\Theta(\theta)\cdot\mathbf{n}(u)\,du$$

$$\times \frac{a^{\frac{1}{2}} \tan\theta}{\rho\cos\theta}\, \hat{\psi}_1(\rho\cos\theta)\, \hat{\psi}_2{}'(a^{-\frac{1}{2}}\tan\theta)\, d\theta\, d\rho.$$

Then, similar to part (iii) of the proof of Theorem 1, we examine the oscillatory integrals J_{11} and J_{12} depending on the behavior of the phase $\Theta(\theta)\cdot(g(u),u)$. As in the proof of Theorem 1, part (iii), there are two cases to consider depending on $A=0$ or $A\neq 0$ (recall that $g(u)=Au^2+Bu$). In either case, we break up the interval $[0,2\pi]$ into $[-\frac{\pi}{2},\frac{\pi}{2}]\bigcup(\frac{\pi}{2},\frac{3\pi}{2}]$ and let $t=a^{-\frac{1}{2}}\tan\theta)$, $u'=a^{-\frac{1}{2}}u$. Thus, we have the following estimates:

Case 1: $A\neq 0$. We will only consider the case $A>0$ since the case $A<0$ is similar. Using the formulas of Fresnel integrals, we have

$$\begin{aligned}
&\lim_{a\to 0^+}(2\pi)^2 2\sqrt{A}a^{-\frac{7}{4}}J_{11}(a,s,0)\\
&=-A_1\sqrt{A}\int_0^\infty \hat{\psi}_1{}'(\rho)\int_{-1}^1 e^{\frac{\pi i\rho}{2A}t^2}\hat{\psi}_2(t)dt\int_{-\infty}^\infty e^{-2\pi i\rho Au^2}du\,d\rho\\
&\quad+A_1\sqrt{A}\int_0^\infty \hat{\psi}_1{}'(\rho)\int_{-1}^1 e^{-\frac{\pi i\rho}{2A}t^2}\hat{\psi}_2(t)dt\int_{-\infty}^\infty e^{2\pi i\rho Au^2}du\,d\rho\\
&=A_1\int_0^\infty \frac{\hat{\psi}_1{}'(\rho)}{\sqrt{\rho}}\int_{-1}^1\Big(\cos(\frac{\pi\rho}{2A}t^2)-\sin(\frac{\pi\rho}{2A}t^2)\Big)\hat{\psi}_2(t)dt\,d\rho\\
&=C_{11},
\end{aligned}$$

where C_{11} is a finite real number.

A similar calculation gives that

$$\begin{aligned}
&\lim_{a\to 0^+}(2\pi)^2 2\sqrt{A}a^{-\frac{7}{4}}J_{12}(a,s,0)\\
&=A_1\sqrt{A}\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 e^{\frac{\pi i\rho}{2A}t^2}t\hat{\psi}_2{}'(t)dt\int_{-\infty}^\infty e^{-2\pi i\rho Au^2}du\,d\rho\\
&\quad+A_1\sqrt{A}\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 e^{-\frac{\pi i\rho}{2A}t^2}t\hat{\psi}_2{}'(t)dt\int_{-\infty}^\infty e^{2\pi i\rho Au^2}du\,d\rho\\
&=A_1\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho^{\frac{3}{2}}}\int_{-1}^1\Big(\cos(\frac{\pi\rho}{2A}t^2)+\sin(\frac{\pi\rho}{2A}t^2)\Big)t\hat{\psi}_2{}'(t)dt\,d\rho\\
&=C_{12},
\end{aligned}$$

where C_{12} is a finite real number.

The same argument applied to the term J_2 gives that

$$\lim_{a\to 0^+}(2\pi)^2 2\sqrt{A}a^{-\frac{5}{4}}J_2(a,s,0)$$

$$
\begin{aligned}
&= A_2\sqrt{A}\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 e^{\frac{\pi i\rho}{2A}t^2}\hat{\psi_2}'(t)dt\int_{-\infty}^\infty e^{-2\pi i\rho Au^2}du\,d\rho\\
&\quad + A_2\sqrt{A}\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 e^{-\frac{\pi i\rho}{2A}t^2}\hat{\psi_2}'(t)dt\int_{-\infty}^\infty e^{2\pi i\rho Au^2}du\,d\rho\\
&= A_2\int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho^{\frac{3}{2}}}\int_{-1}^1 \left(\cos(\frac{\pi\rho}{2A}t^2) - \sin(\frac{\pi\rho}{2A}t^2)\right)\hat{\psi_2}'(t)dt\,d\rho = C_2 = 0,
\end{aligned}
$$

where C_2 is a finite real number and, similarly,

$$
\lim_{a\to 0^+} a^{-\frac{11}{4}}J_3(a,s,0) = C_3,\quad \lim_{a\to 0^+} a^{-\frac{9}{4}}J_4(a,s,0) = C_4,\quad \lim_{a\to 0^+} a^{-\frac{7}{4}}J_5(a,s,0) = C_5,
$$

where C_3, C_4, C_5 are finite real numbers.

In general, for $m = (m_1, m_2) \in \mathbb{N}\times\mathbb{N}$, we have

$$
\lim_{a\to 0^+} a^{-(\frac{3}{4}+m_1+\frac{m_2}{2})}\langle\widehat{\chi S}, \frac{\partial^m}{\partial\xi^m}\left(\hat{\psi}_1(a\xi_1)\,\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))\right)\rangle = C_m,
$$

where C_m is a finite real number for each fixed m. This shows that part (ii) holds for the case $A \neq 0$.

Case 2: $A = 0$. Using an argument similar to the one used in the proof of part (iii) of Theorem 1, we have that

$$
\begin{aligned}
&\lim_{a\to 0^+}(2\pi)^2 2a^{-\frac{7}{4}}\langle\widehat{\chi S}, \frac{\partial}{\partial\xi_1}(\hat{\psi}_1(a\xi_1)\,\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)))\rangle\\
&= \int_0^\infty \hat{\psi}_1'(\rho)\int_{-1}^1 \hat{\psi}_2(t)e^{-2\pi i\rho tu}\,dt\,du\,d\rho - \int_0^\infty \hat{\psi}_1'(\rho)\int_{-1}^1 \hat{\psi}_2(t)e^{2\pi i\rho tu}\,dt\,du\,d\rho\\
&\quad + \int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 t\hat{\psi_2}'(t)e^{-2\pi i\rho tu}\,dt\,du\,d\rho - \int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 t\hat{\psi_2}'(t)e^{2\pi i\rho tu}\,dt\,du\,d\rho\\
&= 0,
\end{aligned}
$$

where we have used the assumption that $\hat{\psi}_1$ is odd and $\hat{\psi}_2$ is even.

Similarly,

$$
\begin{aligned}
&\lim_{a\to 0^+}(2\pi)^2 2a^{-\frac{5}{4}}\langle\widehat{\chi S}, \frac{\partial}{\partial\xi_2}(\hat{\psi}_1(a\xi_1)\,\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s)))\rangle\\
&= \int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 \hat{\psi}_2'(t)e^{-2\pi i\rho tu}\,dt\,du\,d\rho - \int_0^\infty \frac{\hat{\psi}_1(\rho)}{\rho}\int_{-1}^1 \hat{\psi}_2'(t)e^{2\pi i\rho tu}\,dt\,du\,d\rho\\
&= 0.
\end{aligned}
$$

Also in this case, in general, for $m = (m_1, m_2) \in \mathbb{N}\times\mathbb{N}$, we have

$$\lim_{a\to 0^+} a^{-(\frac{3}{4}+m_1+\frac{m_2}{2})}\langle \widehat{\chi_S}, \frac{\partial^m}{\partial \xi^m}\left(\hat{\psi}_1(a\xi_1)\,\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{\xi_2}{\xi_1}-s))\right)\rangle = C_m,$$

where C_m is a finite real number for each fixed m. □

We can now complete the proof of the theorem.

Proof (of Theorem 2) It will be sufficient to consider the horizontal shearlet system $\{\psi^{(h)}_{a,s,p}\}$ since the analysis of the vertical system is similar.

(i) For any $p \notin \partial S$ using the argument from the proof of Lemma 3, one can find the Taylor polynomial P_S of f at p of degree N' such that, for any $N \in \mathbb{N}$,

$$\lim_{a\to 0^+} a^{-N}|\langle P_S\ \chi_S, \psi^{(h)}_{a,s,p}\rangle - \langle f\ \chi_S, \psi^{(h)}_{a,s,p}\rangle| = 0.$$

As in the proof of Lemma 4, we convert $P_S(x)$ into the differential operator $P_S(\frac{i}{2\pi}D)$. Then, by the Localization Lemma 2, it follows that

$$\lim_{a\to 0^+} a^{-N}|\langle P_S\ \chi_S, \psi^{(h)}_{a,s,p}\rangle| = 0.$$

This completes the proof of part (i).

(ii) As in the proof of part (i), we can replace $B = f\ \chi_S$ by the expression $P_s\ \chi_S$. Then part (ii) follows from the argument used in the proof of part (i) of Lemma 4.

(iii) Again we can replace $B = f\ \chi_S$ by $P_s\ \chi_S$. Then, using Lemma 3, we see that near p the boundary curve can be replaced by $(g(u), u)$ where, as in Lemma 4, g is a polynomial of degree 2. Since $P_s(p) = f(p) \neq 0$, Lemma 4 and part (iii) of Theorem 1 imply that

$$\lim_{a\to 0^+} a^{-\frac{3}{4}}\,\mathcal{SH}_\psi B(a, s_0, p_0) = f(0)\lim_{a\to 0^+} a^{-\frac{3}{4}}\,\mathcal{SH}_\psi \chi_S(a, s_0, p_0) \neq 0. \qquad \square$$

As yet another class of two-dimensional singularities, let us consider the case of discontinuities in the derivative. As a prototype of such singularities, let us examine the two-dimensional ramp function $x_1 H(x_1, x_2)$, where H is the two-dimensional Heaviside function defined in Sect. 3. Using a calculation very similar to Sect. 3, we obtain

$$\begin{aligned}
&\mathcal{SH}_\psi(x_1 H)(a, s, t) = \langle x_1 H, \psi_{a,s,t}\rangle \\
&= -\frac{1}{2\pi i}\int_{\mathbb{R}^2} \partial_1\hat{H}(\xi_1, \xi_2)\,\overline{\hat{\psi}_{a,s,t}}(\xi_1, \xi_2)\,d\xi_1\,d\xi_2 \\
&= \frac{1}{2\pi i}\int_{\mathbb{R}^2} \hat{H}(\xi_1, \xi_2)\,\partial_1\overline{\hat{\psi}_{a,s,t}}(\xi_1, \xi_2)\,d\xi_1\,d\xi_2 \\
&= \int_{\mathbb{R}^2} \frac{\delta_2(\xi_1, \xi_2)}{2\pi i\xi_1}\,\partial_1\overline{\hat{\psi}_{a,s,t}}(\xi_1, \xi_2)\,d\xi_1\,d\xi_2 \\
&= \int_{\mathbb{R}} \frac{1}{2\pi i\xi_1}\,\partial_1\overline{\hat{\psi}_{a,s,t}}(\xi_1, \xi_2)|_{\xi_2=0}\,d\xi_1
\end{aligned}$$

$$= a^{\frac{1+\beta}{2}} \overline{\hat{\psi}_2}(a^{\beta-1}s) \int_{\mathbb{R}} \frac{1}{2\pi i \xi_1} \partial_1 \left(\overline{\hat{\psi}_1}(a\,\xi_1)\, e^{2\pi i \xi_1 t_1} \right) d\xi_1$$
$$= a^{\frac{1+\beta}{2}} \overline{\hat{\psi}_2}(a^{\beta-1}s) \int_{\mathbb{R}} \frac{1}{2\pi i \xi_1} \left(a \partial_1 \left(\overline{\hat{\psi}_1} \right)(a\,\xi_1) + 2\pi i t_1 \overline{\hat{\psi}_1}(a\,\xi_1) \right) e^{2\pi i \xi_1 t_1}\, d\xi_1.$$

As in the case of shearlet transform of H, under the assumption that $\hat{\psi}_1 \in C_c^\infty(\mathbb{R})$ it follows that $\mathcal{SH}_\psi(x_1 H)(a, s, t)$ decays rapidly, asymptotically for $a \to 0$, for all (t_1, t_2) when $t_1 \neq 0$, and for $t_1 = 0$, $s \neq 0$. On the other hand, if $t_1 = 0$ and $s = 0$ we have:

$$\mathcal{SH}_\psi(x_1 H)(a, s, t) = a^{\frac{3+\beta}{2}} \overline{\hat{\psi}_2}(0) \int_{\mathbb{R}} \frac{1}{2\pi i \xi_1} \partial_1 \left(\overline{\hat{\psi}_1} \right)(a\,\xi_1)\, d\xi_1.$$

Provided that $\hat{\psi}_2(0) \neq 0$ and that the integral on the right-hand side of the equation above is nonzero, it follows that $\mathcal{SH}_\psi(x_1 H)(a, s, t) = O(a^{\frac{3+\beta}{2}})$.

This result suggests that, under appropriate assumptions on ψ_1 and ψ_2, the analysis of Sect. 3 extends to singularities that behave locally as the ramp function. The complete discussion of this problem is beyond the scope of this paper.

Finally, we remark that the analysis of singularities using the continuous shearlet transform extends "naturally" to the 3D setting. In particular, one can derive a characterization result of jump discontinuities, which follows rather closely the analysis we presented in the 2D setting even though not all arguments from the 2D case carry over to this case (cf. [11, 12]). However, the analysis of the irregular boundary points and other types of singularities is more involved and only partial results are currently available in the references cited above.

Acknowledgments The authors are partially supported by NSF grant DMS 1008900/1008907. DL is also partially supported by NSF grant DMS 1005799.

References

1. Antoine, J.-P., Murenzi, R.: Two-dimensional directional wavelets and the scale-angle representation. Signal Process. **52**, 259–281 (1996)
2. Candès, E.J., Donoho, D.L.: Ridgelets: a key to higher-dimensional intermittency? Philos. Trans. R. Soc. Lond. A **357**, 2495–2509 (1999)
3. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with C^2 singularities. Commun. Pure Appl. Math. **56**, 219–266 (2004)
4. Candès, E.J., Donoho, D.L.: Continuous curvelet transform. I: Resolution of the wavefront set. Appl. Comput. Harmon. Anal. **19**, 162–197 (2005)
5. Chan, T., Shen, J.: Image Processing and Analysis. SIAM, Philadelphia (2005)
6. Dahlke, S., Kutyniok, G., Maass, P., Sagiv, C., Stark, H.-G., Teschke, G.: The uncertainty principle associated with the continuous shearlet transform. Int. J. Wavelets Multiresolut. Inf. Process. **6**, 157–181 (2008)

7. Dahlke, S., Steidl, G., Teschke, G.: The continuous shearlet transform in arbitrary space dimensions. J. Fourier Anal. Appl. **16**(3), 340–364 (2009)
8. Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 891–906 (1991)
9. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal. **39**, 298–318 (2007)
10. Guo, K., Labate, D.: Characterization and analysis of edges using the continuous shearlet transform. SIAM J. Imaging Sci. **2**, 959–986 (2009)
11. Guo, K., Labate, D.: Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. Appl. Comput. Harmon. Anal. **30**, 231–242 (2011)
12. Guo, K., Labate, D.: Characterization of piecewise-smooth surfaces using the 3D continuous shearlet transform. J. Fourier Anal. Appl. **18**, 488–516 (2012)
13. Guo, K., Labate, D.: Analysis and identification of multidimensional singularities using the continuous shearlet transform. In: Kutyniok, G., et al. (eds.) Shearlets, pp. 69–103. Birkhuser/Springer, New York (2012)
14. Guo, K., Labate, D., Lim, W.: Edge analysis and identification using the continuous shearlet transform. Appl. Comput. Harmon. Anal. **27**, 24–46 (2009)
15. Herz, C.S.: Fourier transforms related to convex sets. Ann. Math. **75**, 81–92 (1962)
16. Holschneider, M.: Wavelets: Analysis Tool. Oxford University Press, Oxford (1995)
17. Jaffard, S., Meyer, Y.: Wavelet methods for pointwise regularity and local oscillations of functions. Memoirs AMS **123**(587), 1–110 (1996)
18. Jaffard, S.: Pointwise smoothness, two-microlocalization and wavelet coefficients. Publications Matematiques **35**, 155–168 (1991)
19. Kutyniok, G., Labate, D.: Resolution of the wavefront set using continuous shearlets. Trans. Am. Math. Soc. **361**, 2719–2754 (2009)
20. Labate, D., Lim, W., Kutyniok, G., Weiss, G.: Sparse multidimensional representation using shearlets. Wavelets XI (San Diego, CA, 2005). In: Proceedings of SPIE, vol. 5914, pp. 254–262, Bellingham, WA (2005)
21. Laugesen, R.S., Weaver, N., Weiss, G., Wilson, E.: A characterization of the higher dimensional groups associated with continuous wavelets. J. Geom. Anal. **12**, 89–102 (2001)
22. Meyer, Y.: Wavelets and operators. Cambridge studies in advanced mathematics, vol. 37. Cambridge University Press, Cambridge (1992)
23. Perona, P.: Steerable-scalable kernels for edge detection and junction analysis. Image Vis. Comput. **10**, 663–672 (1992)
24. Weiss, G., Wilson, E.: The mathematical theory of wavelets. In: Proceeding of the NATO-ASI Meeting, Harmonic Analysis 2000: A Celebration, Kluwer (2001)
25. Yi, S., Labate, D., Easley, G.R., Krim, H.: A shearlet approach to edge analysis and detection. IEEE Trans. Image Process. **18**, 929–941 (2009)

Barycentric Interpolation

Kai Hormann

Abstract This survey focusses on the method of barycentric interpolation, which ties up to the ideas that August Ferdinand Möbius published in his seminal work "Der barycentrische Calcul" in 1827. For univariate data, it leads to a special kind of rational interpolation which is guaranteed to have no poles and favorable approximation properties. We further discuss how to extend this idea to bivariate data, both for scattered data and for data given at the vertices of a polygon.

Keywords Rational interpolation · Barycentric coordinates · Approximation order · Lebesgue constant

1 Introduction

Consider a system of $n + 1$ particles, located at $x_0, \ldots, x_n$ and with masses $w_0, \ldots, w_n$. It is then well known from physics that the *centre of mass* or *barycentre* of this particle system is the unique point x which satisfies

$$\sum_{i=0}^{n} w_i (x - x_i) = 0,$$

that is,

$$x = \frac{\sum_{i=0}^{n} w_i x_i}{\sum_{i=0}^{n} w_i}.$$

K. Hormann (✉)
Faculty of Informatics, Università della Svizzera italiana, Via Giuseppe Buffi 13,
6904 Lugano, Switzerland
e-mail: kai.hormann@usi.ch

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_11,

The idea of barycentric interpolation stems from this concept, by asking the question: given a fixed set of distinct locations or *nodes* $x_0, \dots, x_n$ and an arbitrary point x, do there exist some masses or *weights* $w_0, \dots, w_n$, such that x is the barycentre of the corresponding particle system? Consequently, we are interested in functions $w_0(x), \dots, w_n(x)$, such that

$$x = \frac{\sum_{i=0}^{n} w_i(x) x_i}{\sum_{i=0}^{n} w_i(x)}. \tag{1}$$

Möbius [24] was probably the first to answer this question in full generality. He showed that for particle systems in $\mathbb{R}^m$ such weights always exist[1] for any $x \in \mathbb{R}^m$, as long as the number of particles is greater than the dimension, that is, for $n \geq m$. Möbius called the weights $w_0(x), \dots, w_n(x)$ the *barycentric coordinates* of x with respect to $x_0, \dots, x_n$.

It is clear that barycentric coordinates are *homogeneous* in the sense that they can be multiplied with a common nonzero scalar and still satisfy (1). In the context of barycentric interpolation we therefore assume without loss of generality that the barycentric coordinates sum to one for any x. We further demand that they are 1 at the corresponding node and 0 at all other nodes. The resulting *barycentric basis functions* $b_i : \mathbb{R}^m \to \mathbb{R}$, $i = 0, \dots, n$ are then characterized by the three properties:

$$\text{Partition of unity:} \qquad \sum_{i=0}^{n} b_i(x) = 1, \tag{2a}$$

$$\text{Barycentric property:} \qquad \sum_{i=0}^{n} b_i(x) x_i = x, \tag{2b}$$

$$\text{Lagrange property:} \qquad b_i(x_j) = \delta_{ij}, \tag{2c}$$

where (2b) is equivalent to (1) because of (2a). Möbius observed that these barycentric basis functions are unique in the special case $n = m$, when the nodes $x_0, \dots, x_n$ can be considered the vertices of an m-simplex, and he gave an explicit formula for b_i in this case, which reveals that b_i is a linear function.

Let us now consider data $f_0, \dots, f_n$ corresponding to the nodes $x_0, \dots, x_n$ and possibly sampled from some function $f : \mathbb{R}^m \to \mathbb{R}$, that is, $f_i = f(x_i)$ for $i = 0, \dots, n$. The *barycentric interpolant* of this data is then given by

$$F(x) = \sum_{i=0}^{n} b_i(x) f_i. \tag{3}$$

It follows from (2c) that the function $F : \mathbb{R}^m \to \mathbb{R}$ interpolates the data f_i at x_i for $i = 0, \dots, n$, and from (2a) and (2b) that this kind of interpolation reproduces

[1] Note that at least one of the $w_i(x)$ must be negative if x is outside the *convex hull* of the nodes $x_0, \dots, x_n$, which is physically impossible and motivates to call the w_i weights rather than masses.

linear functions. That is, if the data $f_0, \ldots, f_n$ are sampled from a linear polynomial $f \in \Pi_1$, where Π_d denotes the space of polynomials with degree at most d, then $F = f$.

Vice versa, if an interpolation operator reproduces linear functions, then its cardinal basis functions clearly satisfy the three conditions in (2). Therefore, many classical interpolation methods, including interpolation with splines, radial basis functions, and subdivision schemes, just to name a few, could be called barycentric. However, we prefer to use the term *barycentric interpolation* whenever simple closed-form expressions for the barycentric basis functions b_i exist, so that evaluating the interpolant (3) is efficient.

In this survey, we review recent progress in the construction of such barycentric basis functions and the related interpolants. We mainly focus on the univariate setting in Sect. 2, but also summarize some results on scattered data interpolation in two variables in Sect. 3. The special case of barycentric interpolation at the vertices of a polygon in $\mathbb{R}^2$ is only briefly discussed in Sect. 4, as more details can be found in [12].

2 Univariate Barycentric Interpolation

Suppose we are given two distinct nodes $x_0, x_1 \in \mathbb{R}$. Then it is clear that the two functions $b_0, b_1 : \mathbb{R} \to \mathbb{R}$ with

$$b_0(x) = \frac{x_1 - x}{x_1 - x_0} \quad \text{and} \quad b_1(x) = \frac{x - x_0}{x_1 - x_0}$$

are barycentric basis functions[2] with respect to x_0 and x_1, that is, these functions satisfy the three conditions in (2). Therefore, the barycentric interpolant to the data f_0 and f_1, associated with x_0 and x_1, is the linear function

$$F_1(x) = \frac{x_1 - x}{x_1 - x_0} f_0 + \frac{x - x_0}{x_1 - x_0} f_1. \tag{4}$$

In order to generalize this approach to more than two nodes, we first rewrite $F_1(x)$ as

$$F_1(x) = \frac{(x - x_1) f_0 - (x - x_0) f_1}{-(x_1 - x_0)} = \frac{(x - x_1) f_0 - (x - x_0) f_1}{(x - x_1) - (x - x_0)},$$

[2] Since $n = m = 1$, these are the unique barycentric basis functions, according to Möbius [24].

and then, after dividing numerator and denominator both by $(x - x_0)(x - x_1)$, as

$$F_1(x) = \frac{\frac{1}{x-x_0} f_0 - \frac{1}{x-x_1} f_1}{\frac{1}{x-x_0} - \frac{1}{x-x_1}} = \frac{\sum_{i=0}^{1} \frac{(-1)^i}{x-x_i} f_i}{\sum_{i=0}^{1} \frac{(-1)^i}{x-x_i}}. \tag{5}$$

2.1 Berrut's Interpolants

The extension to $n + 1$ distinct nodes in ascending order $x_0 < \cdots < x_n$ with associated data $f_0, \ldots, f_n$ is now as easy as changing the upper bound of summation in (5) from 1 to n, giving the interpolant

$$F_n(x) = \frac{\sum_{i=0}^{n} \frac{(-1)^i}{x-x_i} f_i}{\sum_{i=0}^{n} \frac{(-1)^i}{x-x_i}}. \tag{6}$$

To see that F_n indeed interpolates the data, we multiply numerator and denominator both with

$$\ell(x) = \prod_{i=0}^{n} (x - x_i), \tag{7}$$

so that

$$F_n(x) = \frac{\sum_{i=0}^{n} (-1)^i \prod_{j=0,\, j\neq i}^{n} (x - x_j) f_i}{\sum_{i=0}^{n} (-1)^i \prod_{j=0,\, j\neq i}^{n} (x - x_j)}, \tag{8}$$

and evaluation at $x = x_k$ reveals that

$$F_n(x_k) = \frac{\sum_{i=0}^{n} (-1)^i \prod_{j=0,\, j\neq i}^{n} (x_k - x_j) f_i}{\sum_{i=0}^{n} (-1)^i \prod_{j=0,\, j\neq i}^{n} (x_k - x_j)} = \frac{(-1)^k \prod_{j=0,\, j\neq k}^{n} (x_k - x_j) f_k}{(-1)^k \prod_{j=0,\, j\neq k}^{n} (x_k - x_j)} = f_k.$$

Equation (8) shows that F_n is a rational function of degree at most n over n. This rational interpolant was discovered by Berrut [1], who also shows that F_n does not have any poles in $\mathbb{R}$, because the denominator of (6) does not vanish for any $x \in \mathbb{R} \setminus \{x_0, \ldots, x_n\}$. For example, if $x \in (x_0, x_1)$, then

$$\sum_{i=0}^{n} \frac{(-1)^i}{x - x_i} = \underbrace{\frac{1}{x - x_0}}_{>0} + \underbrace{\frac{1}{x_1 - x} - \frac{1}{x_2 - x}}_{>0} + \underbrace{\frac{1}{x_3 - x} - \cdots}_{>0} > 0,$$

that is, for each negative term $-1/(x_{2i} - x)$ there is a positive term $1/(x_{2i-1} - x)$ such that their sum is positive, because $x_{2i-1} < x_{2i}$. All other cases of x can be treated similarly. Another property of F_n is that it is a barycentric interpolant in case n is odd.

Proposition 1 *Berrut's first interpolant F_n in (6) is barycentric for odd n.*

Proof It is clear that the underlying basis functions

$$b_i(x) = \frac{\frac{(-1)^i}{x-x_i}}{\sum_{j=0}^{n} \frac{(-1)^j}{x-x_j}}, \quad i = 0, \ldots, n. \tag{9}$$

of F_n satisfy conditions (2a) and (2c). Applying the construction of F_1 in (5) to data sampled from the identity function at x_i and x_{i+1} gives $F_1(x) = x$, hence

$$\left(\frac{1}{x - x_i} - \frac{1}{x - x_{i+1}}\right) x = \frac{1}{x - x_i} x_i - \frac{1}{x - x_{i+1}} x_{i+1} \tag{10}$$

for $i = 0, \ldots, n-1$. Adding these equations for $i = 0, 2, \ldots, (n-1)/2$ gives

$$\sum_{i=0}^{n} \frac{(-1)^i}{x - x_i} x = \sum_{i=0}^{n} \frac{(-1)^i}{x - x_i} x_i,$$

which is equivalent to (2b) for the b_i in (9). □

Unfortunately, the trick used in the proof of Proposition 1 to establish condition (2b) does not work for n even, but a slight modification of F_n takes care of it. We just need to weight all but the first and the last terms of the sums in (6) by a factor of 2, giving the interpolant

$$\hat{F}_n(x) = \frac{\frac{1}{x-x_0} f_0 + 2 \sum_{i=1}^{n-1} \frac{(-1)^i}{x-x_i} f_i + \frac{(-1)^n}{x-x_n} f_n}{\frac{1}{x-x_0} + 2 \sum_{i=1}^{n-1} \frac{(-1)^i}{x-x_i} + \frac{(-1)^n}{x-x_n}}. \tag{11}$$

This rational interpolant was also discovered by Berrut [1] and like F_n it does not have any poles in $\mathbb{R}$ [13]. Its advantage, however, is that it is a barycentric interpolant for any n.

Proposition 2 *Berrut's second interpolant $\hat{F}_n$ in (11) is barycentric for any n.*

Proof Multiplying the equations in (10) by $(-1)^i$ and adding them for $i = 0, \ldots, n-1$ gives

$$\left(\frac{1}{x - x_0} + 2 \sum_{i=1}^{n-1} \frac{(-1)^i}{x - x_i} + \frac{(-1)^n}{x - x_n}\right) x = \frac{1}{x - x_0} x_0 + 2 \sum_{i=1}^{n-1} \frac{(-1)^i}{x - x_i} x_i + \frac{(-1)^n}{x - x_n} x_n.$$

Therefore, condition (2b) holds for the basis functions of the interpolant $\hat{F}_n$ in (11), and it is clear that these basis functions also satisfy conditions (2a) and (2c). □

2.2 General Rational Interpolants

Berrut's interpolants F_n and $\hat{F}_n$ are special cases of the general rational function

$$F_\beta(x) = \frac{\sum_{i=0}^n \frac{\beta_i}{x-x_i} f_i}{\sum_{i=0}^n \frac{\beta_i}{x-x_i}} \tag{12}$$

with coefficients $\beta = (\beta_0, \ldots, \beta_n)$, which was introduced and studied by Schneider and Werner [28]. They show that F_β interpolates f_i at x_i as long as $\beta_i \neq 0$, which can be seen immediately after multiplying numerator and denominator both with $\ell(x)$ in (7), similarly to how we did it for F_n above.

Moreover, Berrut and Mittelmann [4] observe that any rational interpolant to the data $f_0, \ldots, f_n$ at $x_0, \ldots, x_n$ can be written in the form (12) for some suitable choice of β. Assume that

$$r(x) = \frac{p(x)}{q(x)}, \quad p \in \Pi_k, \quad q \in \Pi_m$$

is a rational function of degree k over m with $k, m \leq n$ and $r(x_i) = f_i$ for $i = 0, \ldots, n$. Now consider the Lagrange form of p,

$$p(x) = \sum_{i=0}^n \prod_{j=0, j\neq i}^n \frac{x - x_j}{x_i - x_j} p(x_i) = \ell(x) \sum_{i=0}^n \frac{p(x_i)}{(x - x_i)\ell'(x_i)},$$

with $\ell(x)$ as in (7), and the Lagrange form of q,

$$q(x) = \ell(x) \sum_{i=0}^n \frac{q(x_i)}{(x - x_i)\ell'(x_i)},$$

and let

$$\beta_i = \frac{q(x_i)}{\ell'(x_i)}, \quad i = 0, \ldots, n.$$

The interpolation condition of r implies $p(x_i) = q(x_i) f_i$ and substituting this in $p(x)$ as well as β_i both in $p(x)$ and $q(x)$ then gives $r(x)$ in the form (12) after cancelling out the common factor $\ell(x)$. These coefficients β are actually unique up to a common nonzero scaling factor.

An immediate consequence of this observation is that F_β with

$$\beta_i = \frac{1}{\ell'(x_i)} = \prod_{j=0, j\neq i}^{n} \frac{1}{x_i - x_j}, \quad i = 0, \ldots, n,$$

is the interpolating rational function with denominator $q(x) = 1$, that is, the interpolating polynomial of degree n. This special way of writing the interpolating polynomial is called the *(true) barycentric formula*,[3] and it provides a fast and stable algorithm for evaluating the interpolating polynomial, which outperforms even Newton's interpolation formula [5, 18].

Returning to the general rational interpolant F_β in (12), a natural question arises in the context of this survey: how to choose the coefficients β such that F_β is a barycentric interpolant and without poles in $\mathbb{R}$? The coefficients from Berrut's second interpolant as well as those from the interpolating polynomial certainly satisfy both goals, but are there other choices? The answer is positive, but before we go into details, let us review some basic facts.

The first goal can easily be achieved by slightly constraining the coefficients β.

Proposition 3 *If the coefficients $\beta = (\beta_0, \ldots, \beta_n)$ satisfy*

$$\sum_{i=0}^{n} \beta_i = 0, \tag{13}$$

then the interpolant F_β in (12) is barycentric.

Proof As in the proof of Proposition 2, we consider the equations in (10). Multiplying each by $\gamma_i = \sum_{j=0}^{i} \beta_j$ and adding them for $i = 0, \ldots, n-1$ gives

$$\begin{aligned}
&\sum_{i=0}^{n-1} \gamma_i \left(\frac{1}{x - x_i} - \frac{1}{x - x_{i+1}} \right) x = \sum_{i=0}^{n-1} \gamma_i \left(\frac{1}{x - x_i} x_i - \frac{1}{x - x_{i+1}} x_{i+1} \right) \\
\iff \quad &\sum_{i=0}^{n-1} \frac{\gamma_i}{x - x_i} x - \sum_{i=1}^{n} \frac{\gamma_{i-1}}{x - x_i} x = \sum_{i=0}^{n-1} \frac{\gamma_i}{x - x_i} x_i - \sum_{i=1}^{n} \frac{\gamma_{i-1}}{x - x_i} x_i \\
\iff \quad &\frac{\gamma_0}{x - x_0} x + \sum_{i=1}^{n-1} \frac{\gamma_i - \gamma_{i-1}}{x - x_i} x - \frac{\gamma_{n-1}}{x - x_n} x = \frac{\gamma_0}{x - x_0} x_0 + \sum_{i=1}^{n-1} \frac{\gamma_i - \gamma_{i-1}}{x - x_i} x_i - \frac{\gamma_{n-1}}{x - x_n} x_n \\
\iff \quad &\sum_{i=0}^{n} \frac{\beta_i}{x - x_i} x = \sum_{i=0}^{n} \frac{\beta_i}{x - x_i} x_i,
\end{aligned}$$

where the last equivalence stems from the identities $\gamma_0 = \beta_0$, $\gamma_i - \gamma_{i-1} = \beta_i$, and $\gamma_{n-1} = -\beta_n$ by (13). This shows that condition (2b) holds for the basis functions

[3] According to Henrici [17], this terminology goes back to Rutishauser [27] and is justified because the interpolating polynomial reproduces linear functions for $n \geq 1$ and therefore is a barycentric interpolant.

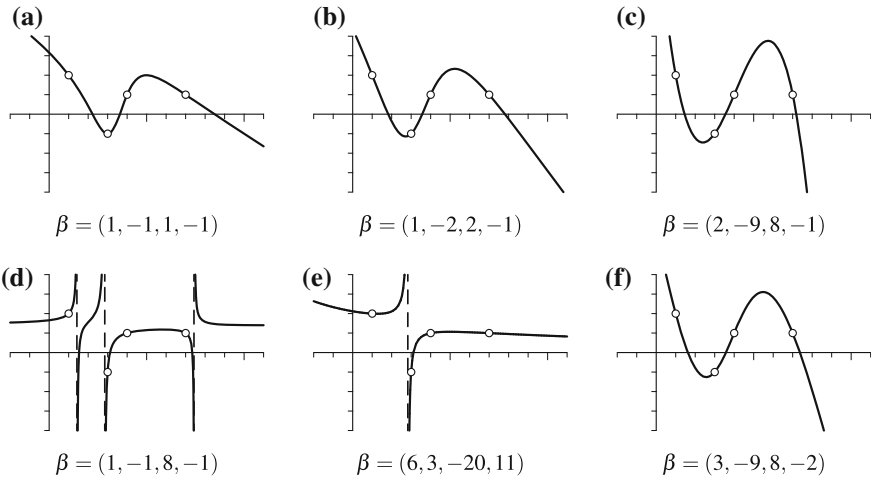


Fig. 1 Some examples of the rational interpolant F_β in (12) to the data $(f_0, f_1, f_2, f_3) = (2, -1, 1, 1)$ at the nodes $(x_0, x_1, x_2, x_3) = (1, 3, 4, 7)$ for different choices of β: **a** Berrut's first interpolant; **b** Berrut's second interpolant; **c** interpolating cubic polynomial; **d**, **e** two examples of rational interpolants with poles in $\mathbb{R}$; **f** Floater–Hormann interpolant for $d = 1$

of the interpolant $\hat{F}_\beta$ in (11), and it is clear that these basis functions also satisfy conditions (2a) and (2c). □

As for the second goal, the absence of poles, Schneider and Werner [28] derive a necessary condition: the coefficients β_i need to have alternating sign, that is, $\beta_i \beta_{i+1} < 0$ for $i = 0, \ldots, n-1$. But as the examples in Fig. 1 illustrate, this is not a sufficient condition. Schneider and Werner [28] also show that F_β has an odd number of poles in the open interval (x_i, x_{i+1}) if β_i and β_{i+1} have the same sign. However, this is all that is known so far and deriving further conditions remains a challenging open problem.

2.3 Floater–Hormann Interpolants

A set of coefficients β, which is different from the special cases above, is

$$\beta_i = \frac{(-1)^i}{x_{i+1} - x_i} + \frac{(-1)^i}{x_i - x_{i-1}}, \quad i = 1, \ldots, n-1,$$

and

$$\beta_0 = \frac{1}{x_1 - x_0}, \quad \beta_n = \frac{(-1)^n}{x_n - x_{n-1}}.$$

These coefficients clearly satisfy the condition of Proposition 3, but there is another way to show that the corresponding rational interpolant F_β is barycentric. To this end, let us rewrite the numerator of F_β as

$$\begin{aligned}\sum_{i=0}^{n} \frac{\beta_i}{x-x_i} f_i &= \sum_{i=0}^{n-1} \frac{(-1)^i f_i}{(x-x_i)(x_{i+1}-x_i)} + \sum_{i=1}^{n} \frac{(-1)^i f_i}{(x-x_i)(x_i-x_{i-1})} \\ &= \sum_{i=0}^{n-1} \frac{(-1)^i f_i}{(x-x_i)(x_{i+1}-x_i)} + \sum_{i=0}^{n-1} \frac{(-1)^{i+1} f_{i+1}}{(x-x_{i+1})(x_{i+1}-x_i)} \\ &= \sum_{i=0}^{n-1} \frac{(-1)^{i+1}}{(x-x_i)(x-x_{i+1})} \cdot \frac{(x_{i+1}-x) f_i + (x-x_i) f_{i+1}}{x_{i+1}-x_i}.\end{aligned}$$

Remembering (4), we recognize the term

$$\pi_i(x) = \frac{(x_{i+1}-x) f_i + (x-x_i) f_{i+1}}{x_{i+1}-x_i}$$

as the linear interpolant to the data f_i and f_{i+1} at x_i and x_{i+1}. Introducing the functions

$$\lambda_i(x) = \frac{(-1)^{i+1}}{(x-x_i)(x-x_{i+1})}, \quad i = 0, \ldots, n-1,$$

we can now write the numerator of F_β as

$$\sum_{i=0}^{n} \frac{\beta_i}{x-x_i} f_i = \sum_{i=0}^{n-1} \lambda_i(x) \pi_i(x)$$

and the denominator as

$$\sum_{i=0}^{n} \frac{\beta_i}{x-x_i} = \sum_{i=0}^{n-1} \lambda_i(x). \tag{14}$$

It then turns out that the rational interpolant F_β is an affine combination of the local linear interpolants π_i,

$$F_\beta(x) = \sum_{i=0}^{n-1} \mu_i(x) \pi_i(x), \tag{15}$$

with weight functions

$$\mu_i(x) = \frac{\lambda_i(x)}{\sum_{j=0}^{n-1} \lambda_j(x)}, \qquad i = 0, \ldots, n-1,$$

which clearly sum to one. Now, if the data is sampled from the function $f(x) = x$, that is, $f_i = x_i$ for $i = 0, \ldots, n$, then $\pi_i(x) = x$ for $i = 0, \ldots, n-1$ and $F_\beta(x) = x$ by (15), which confirms F_β to be a barycentric interpolant.

With the denominator of F_β written as in (14), it is also easy to see that F_β has no poles in $\mathbb{R}$. If $x \in (x_0, x_1)$, then

$$\sum_{i=0}^{n} \frac{\beta_i}{x - x_i} = \underbrace{\frac{1}{(x - x_0)(x_1 - x)}}_{>0} + \underbrace{\frac{1}{(x_1 - x)(x_2 - x)} - \frac{1}{(x_2 - x)(x_3 - x)}}_{>0} + \cdots > 0,$$

similar to our consideration for the denominator of Berrut's first interpolant F_n above, and analyzing the other cases shows that the denominator of F_β does not vanish for any $x \in \mathbb{R} \setminus \{x_0, \ldots, x_n\}$.

In the same way that F_β in (15) is an affine combination of local *linear* interpolants, Berrut's first interpolant F_n in (6) can be seen as an affine combination of local *constant* interpolants with the b_i in (9) as weight functions, which also indicates why it does not reproduce linear functions in general.

Equipped with this new point of view, it is now straightforward to design barycentric rational interpolants which reproduce polynomials up to some general degree $d \le n$. Let us denote the unique polynomials of degree at most d that interpolate the data $f_i, \ldots, f_{i+d}$ at $x_i, \ldots, x_{i+d}$ by $\pi_i^d \in \Pi_d$ for $i = 0, \ldots, n-d$ and consider their affine combination

$$F_n^d(x) = \sum_{i=0}^{n-d} \mu_i^d(x) \pi_i^d(x) \tag{16}$$

for certain weight functions $\mu_i^d(x)$. Looking at the weight functions in the constant and linear case, the obvious generalization is

$$\mu_i^d(x) = \frac{\lambda_i^d(x)}{\sum_{j=0}^{n-d} \lambda_j^d(x)}, \quad i = 0, \ldots, n-d, \tag{17}$$

with

$$\lambda_i^d(x) = \frac{(-1)^{i+d}}{(x - x_i) \cdots (x - x_{i+d})}, \quad i = 0, \ldots, n-d.$$

The functions F_n^d in (16) were introduced by Floater and Hormann [13], who also show that they do not have any poles in $\mathbb{R}$, using similar arguments as above. Multiplying numerator and denominator of F_n^d with $\ell(x)$ in (7), it is clear that this function is rational of degree n over $n-d$ and that it interpolates the data $f_0, \ldots, f_n$ at

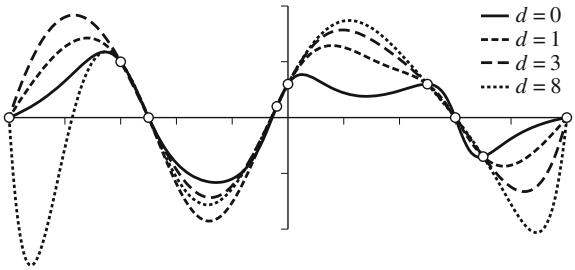


Fig. 2 Comparison of several Floater–Hormann interpolants to data at 9 irregularly distributed nodes, including Berrut's first interpolant ($d = 0$) and the interpolating polynomial ($d = 8$)

$x_0, \ldots, x_n$. Therefore, it must be possible to convert F_n^d into the general barycentric form (12) and Floater and Hormann [13] derive that

$$\beta_i = (-1)^i \sum_{j=\max(i-d,0)}^{\min(i,n-d)} \prod_{k=j,k\neq i}^{j+d} \frac{1}{|x_i - x_k|}, \qquad i = 0, \ldots, n. \tag{18}$$

is the correct choice of coefficients β. As F_n^d clearly reproduces polynomials up to degree d by construction, it is a barycentric interpolant, as long as $d \geq 1$.

This family of Floater–Hormann interpolants nicely closes the gap between Berrut's first interpolant $F_n = F_n^0$ and the interpolating polynomial F_n^n and the barycentric form allows us to efficiently evaluate F_n^d with $O(n)$ operations. In this regard, note that the coefficients β in (18) do not depend on the data. Hence, they can be computed once for a specific set of nodes $x_0, \ldots, x_n$ and then be used to interpolate any data $f_0, \ldots, f_n$ given at these nodes. This also shows that the rational interpolant F_n^d depends linearly on the data. Some examples of F_n^d for different values of d are shown in Fig. 2.

In the special case of equidistant nodes $x_i = x_0 + ih$, $i = 0, \ldots, n$ with spacing $h > 0$, the coefficients in (18), after multiplying them by $d!h^d$, simplify to [13]

$$\beta_i = (-1)^i \sum_{j=\max(i-d,0)}^{\min(i,n-d)} \binom{d}{i-j}, \qquad i = 0, \ldots, n. \tag{19}$$

Ignoring the sign and assuming $n \geq 2d$, the first few sets of these coefficients are

$$\begin{aligned} d = 0: &\quad 1, 1, \ldots, 1, 1, \\ d = 1: &\quad 1, 2, 2, \ldots, 2, 2, 1, \\ d = 2: &\quad 1, 3, 4, 4, \ldots, 4, 4, 3, 1, \\ d = 3: &\quad 1, 4, 7, 8, 8, \ldots, 8, 8, 7, 4, 1, \end{aligned}$$

and we recognize that F_n^1 is identical to Berrut's second interpolant $\hat{F}_n$ in the case of equidistant nodes, but not in general, as shown in Fig. 1b, f.

2.4 Approximation Properties

The Floater–Hormann interpolants F_n^d in (16) have some remarkable approximation properties, both with respect to the approximation order and to the Lebesgue constant. On the one hand, the approximation order of the interpolant F_n^d is essentially $O(h^{d+1})$, where

$$h = \max_{0\le i\le n-1}(x_{i+1} - x_i) \tag{20}$$

is the maximal distance between neighbouring nodes. On the other hand, the Lebesgue constant of F_n^d grows logarithmically with n for equidistant nodes, which is a setting where polynomial interpolation is known to be very ill-conditioned.

To be more precise, let $[a, b] = [x_0, x_n]$ be the interpolation interval, assume that the data $f_0, \dots, f_n$ is sampled from some function $f \in C^{d+2}[a, b]$, and denote the maximum norm on $[a, b]$ by $\|f\| = \max_{a\le x\le b} |f(x)|$. Floater and Hormann [13] show that for $d \ge 1$ the error between f and the rational interpolant F_n^d satisfies

$$\|F_n^d - f\| \le h^{d+1}(b-a)\frac{\|f^{(d+2)}\|}{d+2}, \tag{21a}$$

if $n - d$ is odd, and if $n - d$ is even, then

$$\|F_n^d - f\| \le h^{d+1}\left((b-a)\frac{\|f^{(d+2)}\|}{d+2} + \frac{\|f^{(d+1)}\|}{d+1}\right). \tag{21b}$$

The key idea of the proof is to note that the weighting functions μ_i^d in (17) are a partition of unity and to remember the Newton form of the error between f and the interpolating polynomial π_i^d [22]. Then,

$$\begin{aligned} f(x) - F_n^d(x) &= \sum_{i=0}^{n-d} \mu_i^d(x)\big(f(x) - \pi_i^d(x)\big) \\ &= \sum_{i=0}^{n-d} \mu_i^d(x) \prod_{j=i}^{i+d}(x - x_j) f[x_i, \dots, x_{i+d}, x] \\ &= \frac{\sum_{i=0}^{n-d} (-1)^{i+d} f[x_i, \dots, x_{i+d}, x]}{\sum_{i=0}^{n-d} \lambda_i^d(x)}, \end{aligned} \tag{22}$$

where $f[x_i, \dots, x_{i+d}, x]$ denotes the divided difference of f at $x_i, \dots, x_{i+d}, x$. The error bounds in (21) then follow after bounding the numerator and the denominator in (22) suitably from above and from below, respectively.

Floater and Hormann [13] also derive similar error bounds for Berrut's first interpolant (i.e., for $d = 0$), but only if the *local mesh ratio* is bounded, that is, if a

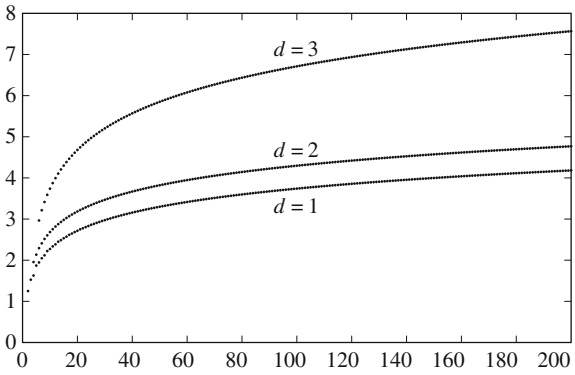


Fig. 3 Numerically computed Lebesgue constants Λ_n^d of the Floater–Hormann interpolants F_n^d at $n+1$ equidistant nodes for $2d \le n \le 200$ and several values of d

constant $R \ge 1$ exists, such that

$$\frac{1}{R} \le \frac{x_{i+1} - x_i}{x_i - x_{i-1}} \le R, \quad i = 1, \ldots, n-1. \tag{23}$$

For equidistant points with mesh ratio $R = 1$, these bounds show that the approximation order of F_n is $O(h)$, which confirms the conjecture of Berrut [1].

Another way to bound the approximation error is by using the *Lebesgue constant* Λ_n^d for the interpolant F_n^d, which is defined as the maximum of the associated *Lebesgue function*

$$\bar{\Lambda}_n^d(x) = \sum_{i=0}^{n} |b_i(x)| = \frac{\sum_{i=0}^{n} \frac{|\beta_i|}{|x-x_i|}}{\left|\sum_{i=0}^{n} \frac{\beta_i}{x-x_i}\right|} \tag{24}$$

with the coefficients β in (18),

$$\Lambda_n^d = \max_{a \le x \le b} \bar{\Lambda}_n^d(x).$$

Since F_n^d reproduces polynomials of degree d by construction, it follows [25] that

$$\|F_n^d - f\| \le (\Lambda_n^d + 1)\|\pi_*^d - f\|,$$

where $\pi_*^d \in \Pi_d$ is the best approximation to f among all polynomials of degree at most d. Moroever, if $\tilde{F}_n^d$ is the Floater–Hormann interpolant to the perturbed data $\tilde{f}_i = f_i + \epsilon_i$, $i = 0, \ldots, n$ with noise $\epsilon = \max\{|\epsilon_0|, \ldots, |\epsilon_n|\}$, then [7]

$$\|\tilde{F}_n^d - F_n^d\| \le \epsilon \Lambda_n^d.$$

Hence, the interpolation process is well conditioned if the Lebesgue constant is small.

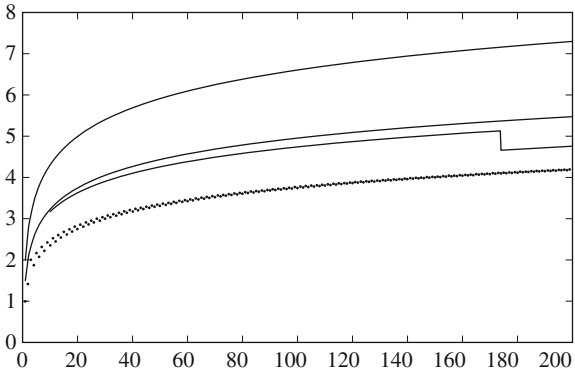


Fig. 4 Numerically computed Lebesgue constants Λ_n^0 of Berrut's first interpolant F_n at $n+1$ equidistant nodes for $1 \le n \le 200$ and the upper bounds (*from top*) by Bos et al. [6], Hormann et al. [20], and Zhang [32]

For the special case of equidistant nodes, Bos et al. [6, 7] show that the Lebesgue constant Λ_n^d for the Floater–Hormann interpolant F_n^d grows only logarithmically with n, as illustrated in Fig. 3, while the Lebesgue constant for polynomial interpolation at such nodes is known to grow exponentially. In particular, they prove that

$$\Lambda_n^d \le \gamma_d(2 + \ln n) \tag{25}$$

with $\gamma_d = 1$ for $d = 0, 1$ and $\gamma_d = 2^{d-1}$ for $d > 1$. The key idea of the proof is to multiply both the numerator and the denominator in (24) with $(x - x_k)(x_{k+1} - x)$ for some $k \in \{0, 1, \ldots, n-1\}$ and to consider $x_k < x < x_{k+1}$. It is then possible to bound the numerator from above and the denominator from below by bounds that do not depend on k, and (25) follows after noticing that $\bar{\Lambda}_n^d(x_i) = 1$ for $i = 0, \ldots, n$.

This initial result has been improved and extended subsequently in various ways. Hormann et al. [20] tighten the upper bound on the Lebesgue constant Λ_n^0 for Berrut's first interpolant $F_n = F_n^0$ to

$$\Lambda_n^0 \le \frac{3}{4}(2 + \ln n)$$

and Zhang [32] further improves it to

$$\Lambda_n^0 \le \frac{1}{1 + \pi^2/24} \ln(n+1) + \begin{cases} 1.47, & \text{if} n \ge 10, \\ 1.00, & \text{if} n \ge 174, \\ 0.99, & \text{if} n \ge 500. \end{cases}$$

Figure 4 shows a visual comparison of these two and the initial bound in (25).

Based on extensive numerical experiments, Ibrahimoglu and Cuyt [21] predict the asymptotic growth rate of the Lebesgue constant Λ_n^d to be

$$\Lambda_n^d \sim \gamma_d \frac{2}{\pi} \ln(n+1)$$

as $n \to \infty$. For $d = 0, 1$ this is identical to the optimal growth rate of the Lebesgue constant for polynomial interpolation [30], which is obtained, for example, by sampling at the extended Chebyshev nodes.

Hormann et al. [20] generalize the upper bound in (25) to the case where the nodes are only quasi-equidistant. That is, they assume the existence of a *global mesh ratio* $M \geq 1$, independent of n, such that

$$\frac{h}{h_*} \leq M$$

with h from (20) and

$$h_* = \min_{0 \leq i \leq n-1} (x_{i+1} - x_i),$$

and then show

$$\Lambda_n^d \leq \tilde{\gamma}_d (2 + M \ln n)$$

with $\tilde{\gamma}_0 = \frac{3}{4} M$ and $\tilde{\gamma}_d = 2^{d-1} M^d$ for $d \geq 1$.

Finally, Bos et al. [8] prove that the Lebesgue constant Λ_n^0 of Berrut's first interpolant grows logarithmically with n for the very general class of well-spaced nodes. A family $X = (X_n)_{n \in \mathbb{N}}$ of sets of nodes $X_n = \{x_0, \ldots, x_n\}$ is called *well-spaced* if each X_n the local mesh ratio is bounded as in (23) for some $R \geq 1$ and if

$$\frac{x_{k+1} - x_k}{x_{k+1} - x_j} \leq \frac{C}{k + 1 - j}, \quad j = 0, \ldots, k, \quad k = 0, \ldots, n-1, \tag{26}$$

$$\frac{x_{k+1} - x_k}{x_j - x_k} \leq \frac{C}{j - k}, \quad j = k+1, \ldots, n, \quad k = 0, \ldots, n-1, \tag{27}$$

for some $C \geq 1$, where both constants R and C must be independent of n. Under these assumptions,

$$\Lambda_n^0 \leq (R + 1)(1 + 2C \ln n).$$

This definition of well-spaced nodes includes equidistant nodes (with $R = C = 1$), *extended Chebyshev nodes*

$$x_i = \frac{\cos \frac{(2i+1)\pi}{2n+2}}{\cos \frac{\pi}{2n+2}}, \quad i = 0, \ldots, n$$

(with $R = 2$ and $C = \pi^2/2$), and *Chebyshev–Gauss–Lobatto* or *Clenshaw–Curtis nodes*

$$x_i = \cos\frac{k\pi}{n}, \quad i = 0, \ldots, n$$

(with $R = 9\pi/2$ and $C = 2\pi$). In general, nodes are well spaced as long as they do not cluster too heavily, but they are allowed to cluster anywhere in the interpolation interval, not just toward its ends, and still the Lebesgue constant Λ_n^0 is guaranteed to grow only logarithmically.

2.5 Conclusion

We have seen in the previous sections that the rational Floater–Hormann interpolants F_n^d provide a promising alternative to other univariate interpolation methods, so let us quickly summarize their advantages. More details regarding recent extensions and applications of rational Floater–Hormann interpolants can be found in [3].

Compared to classical rational interpolation, F_n^d is guaranteed to not have any poles in $\mathbb{R}$, which is important in many applications. Moreover, interpolation with F_n^d is linear in the data and does not require to solve a linear system.

The advantage over polynomial interpolation is that interpolation with F_n^d is stable for a larger class of nodes, and in particular for equidistant nodes, where polynomial interpolation can be infeasible even for rather small $n \approx 20$.

Spline interpolation is probably the closest competitor, because approximation error and convergence rate of F_n^d are similar to those of spline interpolation with (odd) degree d, and this carries over to the approximation of derivatives. Berrut et al. [2] show that

$$\|(F_n^d)^{(k)} - f^{(k)}\| \le Ch^{d+1-k}$$

for $k = 1, 2$ and f being sufficiently smooth, where the constant C may depend on the local mesh ratio (23) of the nodes, and they conjecture that a similar approximation result holds for $k \ge 3$. The advantage over spline interpolation is that F_n^d is infinitely smooth, while the interpolating spline is only $d-1$ times continuously differentiable.

However, the favorable properties of the rational interpolant F_n^d may disappear if d is chosen incorrectly. On the one hand, small values of d lead to very stable interpolation, but rather low approximation order. On the other hand, large values of d guarantee good convergence rates, but the interpolation process may become unstable for equidistant nodes, because the Lebesgue constant Λ_n^d grows exponentially in d for fixed n, which is not too surprising, as F_n^d approaches the polynomial interpolant as $d \to n$. In practice, it is recommended [26, Chap. 3.4.1] to start with small values of d, say $d = 3$ and then try larger values to get better results. For the case when f is analytic in a domain that contains the interpolation interval, Güttel and Klein [16] suggest an algorithm for choosing an optimal value of d.

3 Bivariate Barycentric Interpolation

The main idea behind the construction of the univariate rational barycentric interpolants in Sect. 2 can also be generalized to the bivariate setting. To this end, let $X = \{x_1, \ldots, x_n\}$ be a set of n distinct nodes in $\mathbb{R}^2$ with associated data $f_1, \ldots, f_n$.

The classical *Shepard interpolant* [29]

$$S(x) = \sum_{i=1}^{n} \omega_i(x) f_i$$

with

$$\omega_i(x) = \frac{\frac{1}{\|x - x_i\|^\alpha}}{\sum_{j=1}^{n} \frac{1}{\|x - x_j\|^\alpha}}, \quad i = 1, \ldots, n$$

for some $\alpha > 0$ can be seen as a convex combination of local constant interpolants with weight functions $\omega_i(x)$. Like Berrut's first interpolant, S does not reproduce linear functions in general and so it is not a barycentric interpolant.

To construct the simplest bivariate barycentric interpolant, we consider a triangulation $T = \{t_1, \ldots, t_m\}$ of the nodes X with triangles $t_j = [x_{j_1}, x_{j_2}, x_{j_3}]$. Analogously to (15) we then define

$$F(x) = \sum_{j=1}^{m} \mu_j(x) \pi_j(x), \tag{28}$$

where π_j is the local linear interpolant to the data given at the vertices of the triangle t_j and

$$\mu_j(x) = \frac{\lambda_j(x)}{\sum_{k=1}^{m} \lambda_k(x)}, \quad j = 1, \ldots, m,$$

are some weight functions that sum to one. Little [23] suggests to let

$$\lambda_j(x) = \frac{1}{\|x - x_{j_1}\|^2 \|x - x_{j_2}\|^2 \|x - x_{j_3}\|^2}, \quad j = 1, \ldots, m, \tag{29}$$

which guarantees F to interpolate f_i at x_i and avoids the occurrence of poles, because the common denominator of the weight functions μ_j is positive. Since this *triangular Shepard interpolant* F reproduces linear functions by construction, it clearly is a barycentric interpolant.

Little [23] observes that the triangular Shepard interpolant surpasses Shepard's interpolant in aesthetic behavior, because it does not suffer from flat spots at the

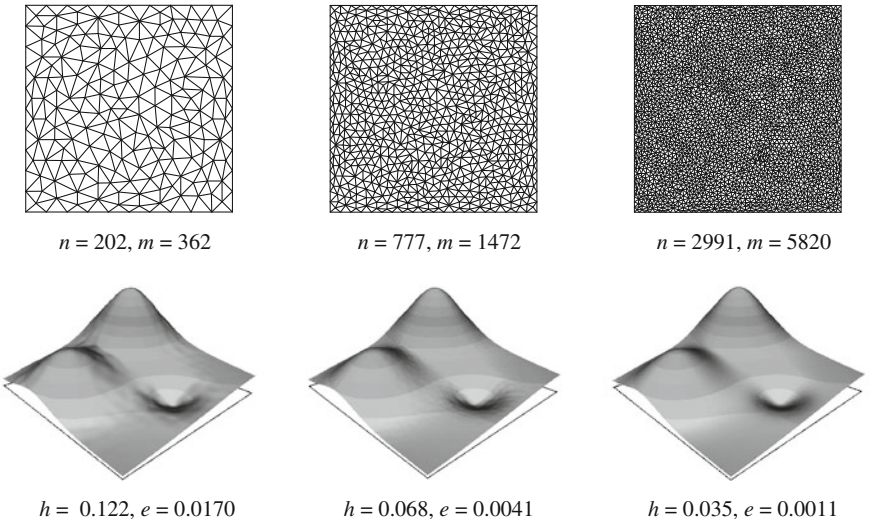


Fig. 5 Examples of triangular Shepard interpolants to data sampled from Franke's test function at n uniformly distributed nodes and with respect to the Delaunay triangulation of the nodes with m triangles and maximum edge length h. The approximation error e decreases roughly by a factor of 4 as h decreases by a factor of 2

nodes and is generally "smoother". But he also notices that it requires the choice of an appropriate triangulation T. One possible choice is to take the *Delaunay triangulation* [9] of X and Fig. 5 shows some examples for this choice and data sampled from Franke's classical test function [15]. In these examples, the approximation error seems to be $O(h^2)$, where h is the maximum edge length of the triangles in Dell'Accio et al. [10] prove that the triangular Shepard interpolant has indeed quadratic approximation order for a very general class of triangulations, which includes the Delaunay triangulation.

While this construction can easily be extended to the multivariate setting and generalized to barycentric interpolants with arbitrary reproduction degree by taking convex combinations of higher order local polynomial interpolants with suitable weighting functions, it lacks two essential properties from the univariate interpolants. On the one hand, the degree of the bivariate rational interpolant is roughly twice the degree of the univariate analogue, because of the squared distances between x and the nodes in the denominator of λ_j in (29). The univariate setting allows us to take signed distances instead, which makes it harder to avoid poles but keeps the degree of the rational interpolant low. On the other hand, an equivalent of the elegant barycentric form in (12) is not known for the triangular Shepard interpolant, and its evaluation is therefore slightly less efficient.

4 Barycentric Interpolation Over Polygons

A very special case of bivariate interpolation occurs if the data $f_1, \ldots, f_n$ is given as the vertices $x_1, \ldots, x_n$ of a planar polygon Ω. In this setting, let us consider the n triangles $t_i = [x_{i-1}, x_i, x_{i+1}]$ for $i = 1, \ldots, n$, where the vertices are indexed cyclically (i.e., $x_{n+1} = x_1$ and $x_0 = x_n$); see Fig. 6.

As for the triangular Shepard interpolant, we then define F as in (28) with $m = n$, except that we replace the functions λ_j in (29) by

$$\lambda_i(x) = \varphi(r_i(x)) \frac{C_i}{A_{i-1}(x) A_i(x)}, \quad i = 1, \ldots, n,$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}$ is an arbitrary function, $r_i(x) = \|x - x_i\|$ is the distance between x and x_i, C_i is the signed area of t_i and $A_{i-1}(x)$, $A_i(x)$ are the signed areas of the triangles $[x, x_{i-1}, x_i]$, $[x, x_i, x_{i+1}]$, respectively; see Fig. 6.

Denoting by $B_i(x)$ the signed area of the triangle $[x, x_{i-1}, x_{i+1}]$ and remembering that $A_i(x)$, $-B_i(x)$, and $A_{i-1}(x)$ are homogeneous barycentric coordinates of x with respect to t_i, we can write the linear interpolant to the data given at the vertices of t_i as

$$\pi_i(x) = \frac{A_i(x) f_{i-1} - B_i(x) f_i + A_{i-1}(x) f_{i+1}}{A_{i-1}(x) - B_i(x) + A_i(x)}.$$

Since $C_i = A_{i-1}(x) - B_i(x) + A_i(x)$, we then have

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_i(x) \pi_i(x) &= \sum_{i=1}^{n} \varphi(r_i(x)) \left(\frac{1}{A_{i-1}(x)} f_{i-1} - \frac{B_i(x)}{A_{i-1}(x) A_i(x)} f_i + \frac{1}{A_i(x)} f_{i+1} \right) \\
&= \sum_{i=1}^{n} \left(\frac{\varphi(r_{i+1}(x))}{A_i(x)} - \frac{\varphi(r_i(x)) B_i(x)}{A_{i-1}(x) A_i(x)} + \frac{\varphi(r_{i-1}(x))}{A_{i-1}(x)} \right) f_i \\
&= \sum_{i=1}^{n} w_i(x) f_i,
\end{aligned}$$

where

$$w_i(x) = \frac{\varphi(r_{i+1}(x)) A_{i-1}(x) - \varphi(r_i(x)) B_i(x) + \varphi(r_{i-1}(x)) A_i(x)}{A_{i-1}(x) A_i(x)}, \quad i = 0, \ldots, n.$$

Likewise,

$$\sum_{i=1}^{n} \lambda_i(x) = \sum_{i=1}^{n} w_i(x),$$

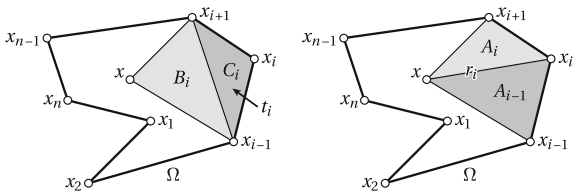


Fig. 6 Notation used for the definition of the barycentric interpolant over a planar polygon Ω with vertices $x_1, \ldots, x_n$

and it turns out that we can rewrite F in terms of the basis functions

$$b_i(x) = \frac{w_i(x)}{\sum_{j=1}^{n} w_j(x)}, \qquad i = 1, \ldots, n, \tag{30}$$

as

$$F(x) = \sum_{i=1}^{n} b_i(x) f_i.$$

Since F reproduces linear functions by construction, it follows that the $b_i(x)$ in (30) satisfy conditions (2a) and (2b), and Floater et al. [14] show that they further satisfy (2c), if the polygon Ω is convex and the function φ has the four properties

Positivity:	$\varphi(r) \geq 0,$
Monotonicity:	$\varphi'(r) \geq 0,$
Convexity:	$\varphi''(r) \geq 0,$
Sublinearity:	$\varphi(r) \geq r\varphi'(r).$

Under these assumptions, it also follows that $b_i(x)$ is positive for any x in the interior of Ω, so that $F(x)$ lies in the convex hull of the data $f_1, \ldots, f_n$, and that the $b_i(x)$ as well as $F(x)$ are linear along the edges of the polygon.

Two examples of functions that satisfy the four conditions above and thus give barycentric basis functions b_i and corresponding barycentric interpolants F are the functions $\varphi_1(r) = 1$ and $\varphi_2(r) = r$. Floater et al. [14] show that the b_i corresponding to φ_1 are the *Wachspress coordinates* [31], which are important in the context of polygonal finite element methods. Instead, φ_2 leads to *mean value coordinates* [11], which turn out to be well defined also for nonconvex and even nested polygons [19] and are used in computer graphics for surface parameterization, image warping, shading, and many other applications. More details on both coordinates can be found in [12].

References

1. Berrut, J.P.: Rational functions for guaranteed and experimentally well-conditioned global interpolation. Comput. Math. Appl. **15**(1), 1–16 (1988)
2. Berrut, J.P., Floater, M.S., Klein, G.: Convergence rates of derivatives of a family of barycentric rational interpolants. Appl. Numer. Math. **61**(9), 989–1000 (2011)
3. Berrut, J.P., Klein, G.: Recent advances in linear barycentric rational interpolation. J. Comput. Appl. Math. **259**(Part A), 95–107 (2014)
4. Berrut, J.P., Mittelmann, H.D.: Lebesgue constant minimizing linear rational interpolation of continuous functions over the interval. Comput. Math. Appl. **33**(6), 77–86 (1997)
5. Berrut, J.P., Trefethen, L.N.: Barycentric Lagrange interpolation. SIAM Rev. **46**(3), 501–517 (2004)
6. Bos, L., De Marchi, S., Hormann, K.: On the Lebesgue constant of Berrut's rational interpolant at equidistant nodes. J. Comput. Appl. Math. **236**(4), 504–510 (2011)
7. Bos, L., De Marchi, S., Hormann, K., Klein, G.: On the Lebesgue constant of barycentric rational interpolation at equidistant nodes. Numer. Math. **121**(3), 461–471 (2012)
8. Bos, L., De Marchi, S., Hormann, K., Sidon, J.: Bounding the Lebesgue constant for Berrut's rational interpolant at general nodes. J. Approx. Theory **169**, 7–22 (2013)
9. Delaunay, B.: Sur la sphère vide. A la mémoire de Georges Voronoï. Bull. Acad. Sci. URSS **6**, 793–800 (1934)
10. Dell'Accio, F., Di Tommaso, F., Hormann, K.: On the approximation order of triangular Shepard interpolation. Department of Mathematics, Università della Calabria, Technical Report (2013)
11. Floater, M.S.: Mean value coordinates. Comput. Aided Geom. Des. **20**(1), 19–27 (2003)
12. Floater, M.S.: Wachspress and mean value coordinates. In: Approximation Theory XIV: San Antonio 2013, Springer Proceedings in Mathematics, pp. 81–101. Springer, New York (2014)
13. Floater, M.S., Hormann, K.: Barycentric rational interpolation with no poles and high rates of approximation. Numer. Math. **107**(2), 315–331 (2007)
14. Floater, M.S., Hormann, K., Kós, G.: A general construction of barycentric coordinates over convex polygons. Adv. Comput. Math. **24**(1–4), 311–331 (2006)
15. Franke, R., Nielson, G.: Smooth interpolation of large sets of scattered data. Int. J. Numer. Methods Eng. **15**(11), 1691–1704 (1980)
16. Güttel, S., Klein, G.: Convergence of linear barycentric rational interpolation for analytic functions. SIAM J. Numer. Anal. **50**(5), 2560–2580 (2012)
17. Henrici, P.: Barycentric formulas for interpolating trigonometric polynomials and their conjugates. Numer. Math. **33**(2), 225–234 (1979)
18. Higham, N.J.: The numerical stability of barycentric Lagrange interpolation. IMA J. Numer. Anal. **24**(4), 547–556 (2004)
19. Hormann, K., Floater, M.S.: Mean value coordinates for arbitrary planar polygons. ACM Trans. Graph. **25**(4), 1424–1441 (2006)
20. Hormann, K., Klein, G., De Marchi, S.: Barycentric rational interpolation at quasi-equidistant nodes. Dolomites Res. Notes Approx. **5**, 1–6 (2012)
21. Ibrahimoglu, B.A., Cuyt, A.: Sharp bounds for Lebesgue constants of barycentric rational interpolation. Department of Mathematics and Computer Science, Universiteit Antwerpen, Techical Report (2013)
22. Isaacson, E., Keller, H.B.: Analysis of Numerical Methods. Dover Publications, Mineola (1994)
23. Little, F.F.: Convex combination surfaces. In: Barnhill, R.E., Boehm, W. (eds.) Surfaces in Computer Aided Geometric Design, pp. 99–107. North-Holland, Amsterdam (1983)
24. Möbius, A.F.: Der barycentrische Calcul. Johann Ambrosius Barth Verlag, Leipzig (1827)
25. Powell, M.J.D.: Approximation Theory and Methods. Cambridge University Press, Cambridge (1981)
26. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes. The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge (2007)

27. Rutishauser, H.: Vorlesungen über numerische Mathematik, Band 1: Gleichungssysteme, Interpolation und Approximation, Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften, Mathematische Reihe, vol. 50. Birkhäuser Verlag, Basel (1976)
28. Schneider, C., Werner, W.: Some new aspects of rational interpolation. Math. Comput. **47**(175), 285–299 (1986)
29. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 23rd ACM National Conference, pp. 517–524. ACM Press, New York (1968)
30. Szabados, J., Vértesi, P.: Interpolation of Functions. World Scientific, Singapore (1990)
31. Wachspress, E.L.: A Rational Finite Element Basis, Mathematics in Science and Engineering, vol. 114. Academic Press, New York (1975)
32. Zhang, R.J.: An improved upper bound on the Lebesgue constant of Berrut's rational interpolation operator. J. Comput. Appl. Math. **255**, 652–660 (2014)

Numerical Determination of Extremal Points and Asymptotic Order of Discrete Minimal Riesz Energy for Regular Compact Sets

Manuel Jaraczewski, Marco Rozgič and Marcus Stiemer

Abstract The asymptotic approximation of continuous minimal s-Riesz energy ($0 \leq s \leq d-2$) by the discrete minimal energy of systems of n-points on compact Ahlfors-David d-regular sets in $\mathbb{R}^d$, $d \geq 2$, is analyzed. In addition, numerical examples are presented, computed via an interior point method for constrained optimization.

Keywords Minimal discrete Riesz energy · Riesz potential · Distributing points on manifolds

1 Introduction

This work is motivated by the hypothesis that recent results on the order of asymptotic approximation of minimal s-Riesz energy by the discrete minimal energy of point systems on the $(d-1)$-dimensional sphere may be extended to classes of more general compact sets in $\mathbb{R}^d$. In this paper, we first give some evidence to this hypothesis by numerically computing the extremal points and the corresponding energy for solid ellipses of different eccentricity as well as for a more complicated set in $\mathbb{R}^3$. To this end, an approach to computing the extremal points based on an interior point method is proposed (Sect. 3), which can quite easily be employed to compute extremal points and discrete minimal energy for large classes of manifolds. Aiming

M. Jaraczewski (✉) · M. Rozgič · M. Stiemer
Helmut Schmidt University, University of the Federal Armed Forces Hamburg,
Holstenhofweg 85, 22034 Hamburg, Germany
e-mail: manuel.jaraczewski@hsu-hh.de

M. Rozgič
e-mail: m.rozgic@hsu-hh.de

M. Stiemer
e-mail: m.stiemer@hsu-hh.de

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_12,

to find a suitable method for analyzing the asymptotic approximation of continuous minimal energy, we study some approximation properties of measures which are constructed by redistributing *(smearing out)* a point-mass continuously in a surrounding ball. If discrete minimal energy is replaced by the continuous energy of such measures two kinds of errors occur: the so-called *diagonal error* and the *local approximation error*. The first is due to the fact that for discrete minimal energy the infinite self-energy of a point mass must be left out, while the second results from the redistribution of the mass. These will be analyzed in Sect. 5. To control both errors, a suitable assumption on the regularity of the considered compact set is required. In this work, Ahlfors-David regularity is proposed as a suitable property. It will be briefly introduced in Sect. 4. For a class of d-Ahlfors-David regular sets Ω, it is analyzed how the asymptotic behavior of the *diagonal error* is related to the densities of the redistributed point measures in Sect. 5. We now begin with a short review of results known for the sphere and the torus and introduce the basic notion in the following Sect. 2.

2 Minimal Riesz Energy and Extremal Points

The *s-Riesz potential* of a point charge located at the origin of $\mathbb{R}^d$ with $d \geq 2$ and $0 \leq s < d$ is defined by

$$R_s(x) := \begin{cases} \|x\|^{-s}, & s > 0, \\ -\log \|x\|, & s = 0, \end{cases}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$. If Ω is a compact subset of $\mathbb{R}^d$, the total Riesz energy can be attributed to any normalized charge distribution ν represented by a Borel measure with total mass $\nu(\Omega) = 1$ by the *energy integral*

$$I_s(\nu) := \int_{\Omega} \int_{\Omega} R_s(x - y) \, \mathrm{d}\nu(x) \mathrm{d}\nu(y). \tag{1}$$

The set of all Borel measures on Ω with total mass 1 is denoted by $\mathscr{M}(\Omega)$. In the case $d = 3$ and $s = 1$, the Riesz potential coincides with the three-dimensional Newton potential, and for $d = 2$ and $s = 0$ the Riesz potential equals the logarithmic potential in the plane. The *s-Riesz energy* $V_d(s)$ of Ω is defined by

$$V_d(s) := V_d(s, \Omega) := \inf \left\{ I_s(\nu) \ : \ \nu \in \mathscr{M}(\Omega) \right\}.$$

This infimum always exists and is larger than 0 for $s > 0$ and larger than $-\infty$ for $s = 0$, but it may coincide with $+\infty$. A measure μ_e with

$$I_s(\mu_e) = \min\left\{ I_s(\nu) \ : \ \nu \in \mathscr{M}(\Omega) \right\}$$

is called *equilibrium measure*. Potential theory has been intensively studied for a very long time due to its intrinsic relations to many other fields both in physics and in mathematics, see, e.g., [11]. In particular for plane sets, the close connection of logarithmic potentials ($s = 0$) and complex analysis offers an extremely rich theory, see, e.g., [29]. During the past 20 years, an increasing interest in algorithmic and computational aspects of potential theory has arisen, e.g., [32]. This led both to a new interest in classical approaches to discrete minimal energy problems in the complex plane as developed, e.g., by Fekete [9], Menke [22] or Tsuji [33], and to new results. In the last decades' new developments, focus was both imposed on extensions of the plane theory, as, e.g., weighted potentials [32], and on extensions to higher dimensions, see, e.g., [14, 28, 31].

Both in the plane and in higher dimensions, discretization of the energy integral (1) can be achieved with the help of point charges distributed on the set under consideration. As the energy integral (1) in its original form equals ∞ for any point charge, discrete energy functionals are introduced, defined for a vector $P_n := (w_1, \ldots, w_n) \in \Omega^n$ consisting of $n \in \mathbb{N}$ distinct points in Ω by

$$E_{d,s}(P_n) := \sum_{j=1}^{n} \sum_{k=j+1}^{n} R_s(w_j - w_k). \tag{2}$$

The *discrete n-point s-energy of* Ω is consequently defined via

$$\mathscr{E}_{d,s}(n) := \inf_{P_n \in \Omega^n} E_{d,s}(P_n) = \min_{P_n \in \Omega^n} E_{d,s}(P_n). \tag{3}$$

It is well known that the normalized discrete energies $\left(\frac{2}{n^2}\mathscr{E}_{d,s}(n)\right)_{n\in\mathbb{N}}$ of a compact set $\Omega \subseteq \mathbb{R}^d$ converge to the continuous energy $V_d(s)$ as n tends to ∞, e.g., [19, 24]. Among others, discrete minimal energy is connected to coverings of manifolds by balls (*sphere packing*) and minimal energy positions of atoms in crystals, e.g., [1, 10, 37]. Most investigations into minimal discrete energy configurations focus on the sphere or on the surface of a torus as *canonical* manifolds, e.g., [28, 31] for the sphere and [3, 12, 15] for tori surfaces.

It should be mentioned that discrete minimal energy can also be considered in cases where the continuous energy integral (1) fails to converge, i.e., $s \geq d$, e.g., [14, 18]. In this case, local interaction between points dominates over global phenomena, and for $s \to \infty$ minimal energy configurations are given by the midpoints of best packing balls. In this work, however, only the case $0 \leq s < d$ is relevant, and we remain in the realm of potential theory.

For numerical purposes, such as, e.g., approximation of continuous energy and potentials by their discrete counterparts, it is essential to have information on the quality of the approximation provided by discrete expressions. In case that Ω coin-

cides with the unit sphere, i.e., $\Omega = \{x \in \mathbb{R}^d : \|x\| = 1\}$, the asymptotic behavior of the sequence $(\mathcal{E}_{d,s}(n))_{n\in\mathbb{N}}$ is analyzed, e.g., in [14, 18, 28]: From Wagner's work [37], the lower estimates

$$\begin{aligned} \mathcal{E}_{d,s}(n) &\geq \frac{1}{2}V_d(s)n^2 - Cn^{1+\frac{s}{d-1}} \quad \text{for } d-3 < s < d-1 \text{ and} \\ \mathcal{E}_{d,s}(n) &\geq \frac{1}{2}V_d(s)n^2 - Cn^{1+\frac{s}{2+s}} \quad \text{for } d \geq 4 \text{ and } 0 < s \leq d-3 \end{aligned} \tag{4}$$

follow, where C denotes a positive constant that may depend on d and s, but not on n. Employing techniques provided by Rakhmanov et al. [28], Kuijlaars and Saff [18] showed

$$\mathcal{E}_{d,s}(n) \leq \frac{1}{2}V_d(s)n^2 - Cn^{1+\frac{s}{d-1}} \quad \text{for } d \geq 3 \text{ and } 0 < s < d-1 \tag{5}$$

for the unit sphere in $\mathbb{R}^d$. Before, G. Wagner [38] already proved this estimate for the particular case $d \geq 3$ and $0 < s < 2$. For logarithmic energy ($s = 0$ and $d \geq 2$) on the sphere, a sharper lower and upper estimate including a higher order term has recently been proven by Brauchart et al. [4].

3 Computing Extremal Points with an Interior Point Method

In this section, we present a flexible method to compute the extremal points on a large class of compact sets in $\mathbb{R}^d$ and present first numerical results on the discrete minimal energy of some sets.

Optimization methods based on quadratic programming are used, e.g., by Hardin, Saff and Kuijlaars [13, 31] to determine extremal points and the minimal discrete energy on the sphere or the surface of a torus. Minimal energy for more general sets, like a solid cube or its boundary, has been computed in [25] by Rajon et al. By providing rigorous upper and lower bounds, this method leads to reliable values for minimal energy and the related capacities. This method has been extended to weighted s-Riesz energy in the presence of external fields in [26]. In contrast, the presented method here is based on an *interior point method*, more precisely on the efficient implementation IPOPT of this method by Wächter and Biegler [36]. This approach solves the minimization problem (3) under constraints given by the set under consideration. It is easily manageable and works for large numbers of sets and kernel functions. The underlying primal-dual framework [23, Chap. 14] leads to a separation of the objective function, which is purely given by the energy functional $E_{d,s}$, and the geometric constraints given by an implicit representation of the particular set Ω under consideration. Hence, the position of the points distributed over the given compact set Ω need not be parametrized, but are immediately primal variables of the minimization problem, while the geometrical constraints resulting

from the description of the considered set lead to additional dual variables in the corresponding Lagrangian (see below). This does not only allow for a high degree of flexibility, since the energy functional and the geometrical constraints can be altered independently from each other, but also allows for a more efficient numerical treatment compared to a gradient method as well as more insight into the success of the numerical scheme via the duality gap. However, such a formulation of the minimal energy problem requires an optimization method that can deal with a large number of constraints. Below we will briefly sketch this approach. Further investigations into the efficiency and reliability of this method represent work in progress.

We consider sets $\Omega \subseteq \mathbb{R}^d$ that can be described by a set of finitely many equations or inequalities of the type

$$\varphi_1(x) = 0\,, \ldots, \varphi_k(x) = 0\,, \quad \psi_1(x) \geq 0\,, \ldots, \psi_\ell(x) \geq 0,$$

where the functions $\varphi_i, \psi_j : \mathbb{R}^d \to \mathbb{R}$, $1 \leq i \leq k$, $1 \leq j \leq \ell$ are assumed to be at least twice continuously differentiable. This general form contains, among other sets, smooth compact manifolds of arbitrary (integer) dimensions $\beta \leq d$ and sets which are the union or intersection of a finite number of such manifolds. Fixing d, s and n we consider $P_n = (w_1, \ldots, w_n) \in \Omega^n$. A set of *extremal points* of order n on Ω, i.e., points $w_1, \ldots, w_n \in \Omega$ minimizing (2), can be determined by solving the *constrained nonlinear optimization problem*

$$\begin{aligned} &\min_{P_n \in \Omega^n} E_{d,s}(P_n) \\ \text{subject to} \quad &\varphi_i(w_\nu) = 0\,, \quad i = 1, \ldots, k\,,\ \nu = 1, \ldots, n, \\ &\psi_j(w_\nu) \geq 0\,, \quad j = 1, \ldots, \ell\,,\ \nu = 1, \ldots, n. \end{aligned}$$

Here, $E_{d,s}$ is the *objective function* as given in (2) and the constraints ensure the extremal points to be located in Ω. The usually nonlinear inequalities may be rendered into equalities by subtracting positive *slack variables* $\zeta_j \in \mathbb{R}$, $j = 1, \ldots, \ell\, n$, from each inequality, yielding the following reformulation

$$\min_{P_n \in \Omega^n} E_{d,s}(P_n) \tag{6a}$$

$$\text{subject to} \quad c(P_n, \zeta) = 0\,, \tag{6b}$$

$$\zeta_j \geq 0\,, \quad j = 1, \ldots, \ell\, n\,. \tag{6c}$$

Here, $c : \Omega^n \times \mathbb{R}^{\ell n} \to \mathbb{R}^{n(k+\ell)}$ contains the constraining information given by $\varphi_1, \ldots, \varphi_k$ and $\psi_j(w_\nu) - \zeta_{j+\ell(\nu-1)}$ for $1 \leq j \leq \ell$, $1 \leq \nu \leq n$ and $\zeta := (\zeta_j)_{1 \leq j \leq \ell n}$. We refer to [36, Sect. 3.4] for a more detailed description. In the sequel we use IPOPT, cf. [36], to solve (6). Interior point (or barrier) methods provide a powerful tool for solving nonlinear constrained optimization problems. For an introduction to this field we refer to [23, Chap. 14]. The problem (6) can be transformed to a constrained problem *without* inequality bounds: By converting the bounds into *barrier terms* in

the objective function $E_{d,s}$ we obtain

$$\min_{\substack{P_n \in \Omega^n \\ \zeta \in \mathbb{R}^{\ell n}}} \mathscr{B}(P_n, \zeta, \lambda)\,, \tag{7a}$$

$$\text{subject to} \quad c(P_n, \zeta) = 0 \tag{7b}$$

with the barrier function

$$\mathscr{B}(P_n, \zeta, \lambda) := E_{d,s}(P_n) - \lambda \sum_{j=1}^{\ell n} \log \zeta_j\,,$$

and a barrier parameter $\lambda > 0$. If λ tends to 0, any point fulfilling the *Karush-Kuhn-Tucker conditions* (KKT condition) [16, 17] of problem (7) tends to a KKT point of the original problem (6), see [30] for more details on the relationship of the barrier problem and the original problem. The KKT conditions represent a set of first-order necessary conditions for $w_1, \ldots, w_n$ to be optimal. If additionally *constraint qualifications* are satisfied [5], the KKT conditions become sufficient. A. Wächter and L. Biegler showed global convergence to a local minimum under quite mild but technical assumptions for the IPOPT algorithm. A more detailed discussion is out of the scope of this work, we refer to [34, 35].

Let $A_k := \operatorname{grad} c(P_{n,k}, \zeta_k)$ and $W_k := \Delta \mathscr{L}(P_{n,k}, \zeta_k, \omega_k, z_k)$ represent the Hessian with respect to $\left(P_n^\top, \zeta^\top\right)^\top$ of the Lagrangian

$$\mathscr{L}(P_n, \zeta, \omega, z) := E_{d,s}(P_n) + c(P_n, \zeta)^\top \omega - z^\top \zeta$$

of the original problem (6) in the kth step with the Lagrange multipliers $\omega \in \mathbb{R}^{n(k+\ell)}$ and $z \in \mathbb{R}^{\ell n}$ for Eqs. (6b) and (6c), respectively. Then, IPOPT solves the optimization problem (6) by applying Newton's method to the barrier problem (7). The system to derive a Newton direction in the kth iteration for a fixed barrier parameter λ reads as

$$\begin{pmatrix} W_k & A_k & -\mathrm{Id} \\ A_k^\top & 0 & 0 \\ Z_k & 0 & X_k \end{pmatrix} \begin{pmatrix} d_k^{(P_n,\zeta)} \\ d_k^\omega \\ d_k^z \end{pmatrix} = - \begin{pmatrix} \operatorname{grad} \mathscr{L}(P_{n,k}, \zeta_k, \omega_k, z_k) \\ c(P_{n,k}, \zeta_k) \\ X_k Z_k \mathbf{1} - \lambda \mathbf{1} \end{pmatrix},$$

yielding the *search directions* $d_k^{(P_n,\zeta)}$, d_k^ω and d_k^z, which are scaled with an adequate step size and then added to $(P_{n,k}, \zeta_k)$, ω_k, and z_k, respectively, to obtain the corresponding values in the $(k+1)$th iteration step. Here, X_k is a diagonal matrix representing the vectors $P_{n,k}$ and ζ_k, i.e., $X_k := \operatorname{diag}\left(P_{n,k}^\top, \zeta_k^\top\right)^\top$, Id represents the identity matrix of adequate size and $\operatorname{grad} \mathscr{L}(P_{n,k}, \zeta_k, \omega_k, z_k)$ the gradient of the Lagrangian with respect to $\left(P_{n,k}^\top, \zeta_k^\top\right)^\top$. Finally, $Z_k := \operatorname{diag}(z_k)$ represents the La-

grange multiplier z_k and $\mathbf{1} := (1, \ldots, 1)^\top$. For details about how the step size for the obtained Newton direction is computed within IPOPT we refer to [36].

After each solution of (7) with a current value for the barrier parameter λ, the barrier parameter is decreased (see [36] for the particular algorithm to find a new λ) and IPOPT continues with a further barrier problem based on the approximated solution of the previous one. To solve the KKT system the IPOPT solver requires information about the first and second derivatives of $E_{d,s}$ and c to derive search directions proceeding toward the minimal energy. It should be mentioned that the objective function $E_{d,s}$ is *not convex*. The number of local minima of the objective function on the unit sphere for instance (ignoring rotations and reflections) grows exponentially with n (at least for certain subsequences of integers) [8, 27]. Thus, by optimization only *relative* minima (and maxima) can be derived. Nevertheless, as we will see below, the numerical results match the analytical for the asymptotics of minimal energy, where those are available, i.e., in case of the sphere.

To compute some numerical examples by solving the constrained optimization problem in (6) with IPOPT, a MATLAB [20] interface is used, providing all necessary derivative information of the objective function $E_{d,s}$ and the functions representing the constraints c. Due to the expected large number of local minima, various randomly chosen points on the set Ω are used as starting points for the above described optimization scheme. Hence, some information about the set of local extrema of the energy functional on Ω is gained as well as some confidence that a value close to the global minimum has numerically been detected by the optimization method. In a first numerical study, we have implemented the above described optimization procedure to compute the discrete n-point s-energy for different compact sets Ω in the case $d = 3$ and $s = 1$ (Newtonian energy). Note, that in this case the equilibrium measure is concentrated on the boundary of Ω, see [19]. Hence, the results of computations carried out for solids, as reported on below, can be compared with results known for their boundary (e.g., results computed for a solid ball can be compared with those known for the sphere). However, a solid is treated differently by the optimization method than its boundary. Consequently, it is important to separate between both. To validate the implementation, we first consider the solid unit ball (set Ω_1, see Table 1), and check if we can reproduce the known results for the unit sphere. Then, solid ellipsoids with different eccentricities are considered (sets Ω_2 and Ω_3 in Table 1), and finally a more complicated set, namely the union of a solid ball and a solid ellipsoid (Ω_4 in Table 1).

We solved the minimization problem for a 64 point configuration on the sets $\Omega_1, \ldots, \Omega_4$ with 160 randomly chosen starting configurations. All computations were performed with a stop criterion tolerance of 10^{-5} or a maximum number of 3000 iterations. By studying the evolution of the (normalized) discrete energy $\frac{2}{64^2} E_{3,1}(P_{64})$ during the iteration process of the optimization algorithm, nearly the same behavior is observed for all sets. In Fig. 1 the evolution of $\frac{2}{64^2} E_{3,1}(P_{64})$ is presented for different starting values on the solid unit ball Ω_1 and on the set Ω_4, respectively. Here, the evolution is presented for ten different starting configurations which are exemplary for all the executed instances. It becomes obvious that all com-

Table 1 Description of sets for which extremal points and discrete energies are numerically computed

Name	Set	Semi-axes
Ω_1	$\left\{x \in \mathbb{R}^3 \,:\, \|x\| \leq 1\right\}$	$a = b = c = 1$
Ω_2	$\left\{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 \,:\, x_1^2 + \frac{x_2^2}{2} + \frac{x_3^2}{2} \leq 1\right\}$	$a = 1, b = c = 2^{1/2}$
Ω_3	$\left\{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 \,:\, x_1^2 + \frac{x_2^2}{10} + \frac{x_3^2}{10} \leq 1\right\}$	$a = 1, b = c = 10^{1/2}$
Ω_4	$\Omega_1 \cup \left\{(x_1, x_2, x_3)^\top \in \mathbb{R}^3 \,:\, 2x_1^2 + \frac{x_2^2}{2} + \frac{x_3^2}{2} \leq 1\right\}$	–

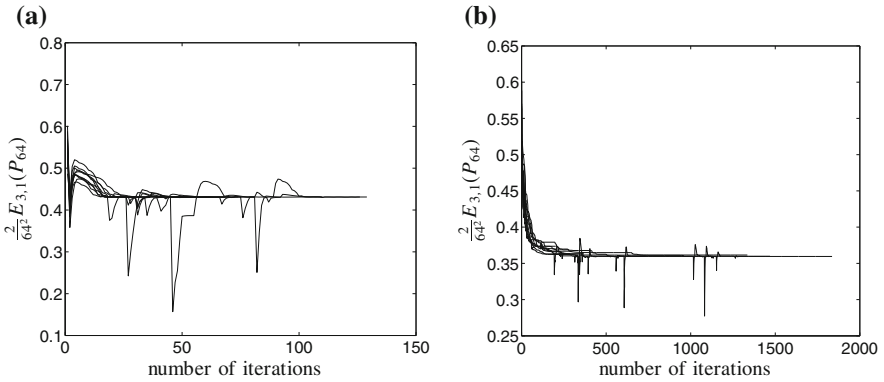


Fig. 1 Behavior of discrete Newtonian energy $\frac{2}{64^2}E_{3,1}(P_{64})$ for a 64-point configuration on the sets Ω_1 and Ω_4. **a** Discrete Newtonian energy for the set Ω_1. **b** Discrete Newtonian energy for the set Ω_4

putations approximatively lead to the same minimal value. The downwardly pointing peaks in Fig. 1 occurring during the iteration indicate constraint violations: At these stages the optimization procedure yields points with a lower energy, which violate the set of constraints given by the implicit description of the set Ω. By exploiting special *feasibility restoration* techniques (see [36]) IPOPT can then pursue the optimization with feasible point configurations that are consequently higher in energy. This may eventually lead to convergence to a feasible point configuration with locally minimal energy. Moreover, the average number of iterations needed to obtain the minimal energy within the scope of the stop criterion mentioned above, is much higher in the case of the nondifferentiable set Ω_4 than for the smooth set Ω_1.

In Fig. 2 the computed minimal discrete energies of 160 optimization procedures with different starting values for the sets Ω_1 and Ω_4 are shown. The observed behavior here is again exemplary for the obtained minimal discrete energies for all regarded sets $\Omega_1, \ldots, \Omega_4$. For Ω_4 a small number (seven) of the 160 performed optimization instances did not converge within the given tolerances. Further the obtained minima spread in a larger range as can be seen in Fig. 2b. Nevertheless, a huge number

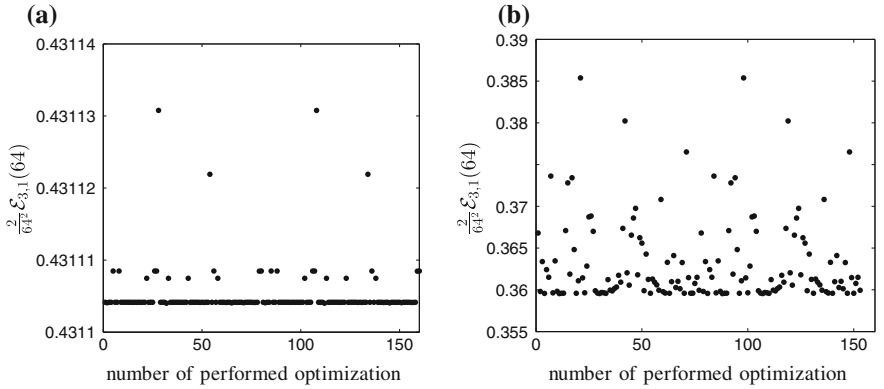


Fig. 2 Convergence of IPOPT. Derived minimal energy for 160 different computations on the solid unit ball and 153 successful optimization instances of the sets Ω_1 and Ω_4. **a** Discrete Newtonian energy for the set Ω_1. **b** Discrete Newtonian energy for the set Ω_4

of computations terminated close to the same optimal value. The fact that IPOPT converged for the majority of performed instances for the non differentiable set Ω_4 is quite remarkable. Since the majority of computations terminated at nearly the same minimal discrete energy value, some confidence in the assumption that the derived extremal values can be considered being close to the *global extremum* is gained.

In Fig. 3 the value of $\frac{2}{n^2}\mathscr{E}_{3,1}(n)$ is computed for certain numbers of points $4 \leq n \leq 128$. For better visualization of the trend and for an estimation of the continuous minimal energy

$$V_d(s) = \lim_{n\to\infty} \frac{2}{n^2}\mathscr{E}_{d,s}(n)$$

a function of the form $x \mapsto a + bx^{-c}$ has been fitted to the data with $a, b, c \in \mathbb{R}$ determined by a least-squares fit.

In case of the solid ball, which is expected to have the same discrete energy as the unit sphere, the theoretical result is reproduced very well: By direct computation one obtains $\lim_{n\to\infty} \frac{2}{n^2}\mathscr{E}_{3,1}(n) = \frac{1}{2}$. As it is shown in [18] and [37], the error $\left|\frac{2}{n^2}\mathscr{E}_{3,1}(n) - \frac{1}{2}\right|$ is of the order $\mathbf{O}(n^{-\frac{1}{2}})$ $(n \to \infty)$, which matches the numerical results displayed in Fig. 3. Finally, comparing the asymptotic behavior of $\frac{2}{n^2}\mathscr{E}_{3,1}(n)$ for the sets Ω_2, Ω_3 and Ω_4 with the results for the unit sphere, which can be represented by the results for Ω_1 (cf. [19]), the same asymptotic behavior of $\frac{2}{n^2}\mathscr{E}_{3,1}(n)$ for all these sets seems to occur. This puts some emphasis on the hypothesis of the *universality* of the asymptotics within a larger class of sets Ω. It is particularly remarkable that the lack of smoothness of Ω_4 does not seem to influence the convergence rate as far as this can be deduced from the computed data. A computation of the minimal energy for the different sets yields

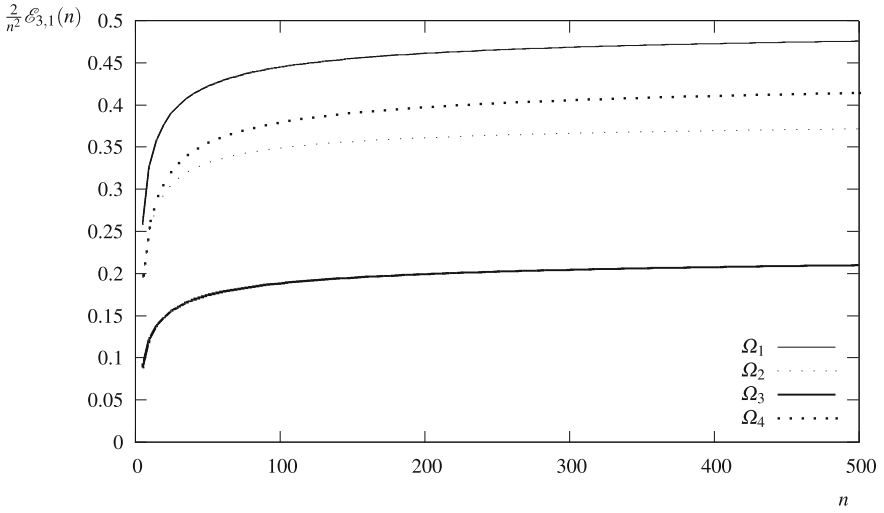


Fig. 3 Discrete Newtonian energy $\frac{2}{n^2}\mathscr{E}_{3,1}(n)$ for different manifolds

$$V_d(s) = \lim_{n\to\infty} \frac{2}{n^2}\mathscr{E}_{3,1}(n) \approx \begin{cases} \frac{1}{2}, & \Omega_1, \\ 0.388375, & \Omega_2, \\ 0.235602, & \Omega_3, \\ 0.449464, & \Omega_4. \end{cases}$$

4 Ahlfors-David Regularity

As a class in which we want to derive asymptotic estimates like (4) and (5), we consider the class of Ahlfors-David regular sets, cf. [21].

Definition 1 A compact set $D \subseteq \mathbb{R}^d$ is called *q-regular* (*Ahlfors-David regular*, [6, 7]), $0 < q < \infty$, if there exists a Borel (outer) measure σ on $\mathbb{R}^d$ and a constant $\eta \in [1, \infty)$, depending on D, such that $\sigma(\mathbb{R}^d \setminus D) = 0$ and

$$r^q \le \sigma\left(B_r(x)\right) \le \eta\, r^q \quad \text{for all } x \in D\,,\ 0 < r < \operatorname{diam}(D)\,,\quad r < \infty\,. \tag{8}$$

Here, $\operatorname{diam}(D)$ denotes the diameter of D and $B_r(x) := \left\{y \in \mathbb{R}^d \,:\, \|x - y\| \le r\right\}$ is the closed ball with radius $r > 0$ centered around $x \in D$. An equivalent definition can be given with the restriction of the q-dimensional Hausdorff measure to D instead of general Borel measures. Then, r^q on the left-hand side of (8) has to be replaced by $\frac{r^q}{\eta}$. It should be pointed out that Ahlfors-David q-regularity is not a

conventional regularity property, but rather a scale-invariant way of expressing the fact that D has Hausdorff-dimension q. From the technical point of view it provides a convenient way to derive estimates on measures defined on D from metric properties of the set. The class of d-regular sets in $\mathbb{R}^d$ contains all images of closed balls under a *bi-Lipschitz* mapping [7]. Since any bi-Lipschitz image of a q-regular set is still q-regular, see, e.g., [21], and since this is obviously true for $q = d$ and a closed ball in $\mathbb{R}^d$, d-regularity follows for bi-Lipschitz images of closed balls. Similarly, each set having a continuously differentiable boundary is $(d-1)$-Ahlfors-David regular.

We will assume d-regularity of the given compact set $\Omega \subseteq \mathbb{R}^d$ itself, i.e., $D = \Omega$ or, alternatively that it bounds a (compact) d-regular set D, i.e., $D = \partial\Omega$. This means, that we assume

$$\frac{r^d}{\eta} \leq \sigma\left(B_r(x)\right) \leq \eta\, r^d \quad \text{for all } x \in D,\ 0 < r < \operatorname{diam}(D)\,, \quad r < \infty$$

for fixed $\eta > 1$ with σ being the d-dimensional Lebesgue measure. Geometrically speaking, this means that there exists a constant $1 \geq \lambda = \frac{C}{\eta} > 0$ (with $C > 0$) such that any intersection of D with a closed ball $B_r(x)$ with $0 < r < \operatorname{diam}(D)$ centered at a point $x \in D$ possesses a d-dimensional Lebesgue measure $\sigma(D \cap B_r(x))$ of at least $\lambda\,\sigma(B_r(x))$:

$$\lambda\,\sigma\left(B_r(x)\right) \leq \sigma\left(D \cap B_r(x)\right).$$

5 Asymptotics of Discrete Minimal Energy

As it is shown in [2], the bounds in (4) and (5) are asymptotically optimal for the sphere. Our aim is to compute similar bounds for larger classes of sets Ω with an appropriate regularity. In case of the sphere, techniques for such an analysis have been provided, e.g., in [31] or [37]. Since $\mathscr{E}_{d,s}(n) \leq \frac{1}{2} V_d\,(s)\, n^2$ is true for all $n \in \mathbb{N}$ and any compact set Ω, it is only required to analyze how much the continuous minimal energy exceeds the normalized discrete one depending on $n \in \mathbb{N}$. But first, we repeat the argument from [19, p. 161] in the following Lemma 1 for the reader's convenience.

Lemma 1 *Let $\Omega \subseteq \mathbb{R}^d$ be compact. Then*

$$\mathscr{E}_{d,s}(n) \leq \frac{1}{2} V_d\,(s)\, n^2 \quad \textit{for all } n \in \mathbb{N}.$$

Proof (Lemma 1) By definition we have $\mathscr{E}_{d,s}(n) \leq E_{d,s}(P_n)$ for any vector $P_n = (x_1, \ldots x_n)$ of n distinct points in Ω^n. Integrating this inequality $\binom{n}{2}$ times over $\Omega \times \Omega$ with respect to $\mathrm{d}\mu_e(x_j)\mathrm{d}\mu_e(x_k)$ with the equilibrium measure μ_e on Ω we obtain

$$\mathscr{E}_{d,s}(n) \le \binom{n}{2} \int\limits_{\Omega} \int\limits_{\Omega} R_s(x-y)\, \mathrm{d}\mu_e(x) \mathrm{d}\mu_e(y) = \binom{n}{2} V_d(s) \le \frac{1}{2} V_d(s)\, n^2. \qquad \square$$

To analyze how much the continuous minimal energy exceeds the normalized discrete n-point energy on sets possessing an adequate regularity property, we consider a certain class of measures μ_n which approximate a distribution of point masses of size $\frac{1}{n}$ in the nth extremal points, such that the following properties hold:

1. For any $n \in \mathbb{N}$, the measure μ_n is absolutely continuous with respect to the d-dimensional Lebesgue measure σ. Let κ_n denote the corresponding L^1-density function, i.e., $\mathrm{d}\mu_n(x) = \kappa_n(x) \mathrm{d}\,\sigma(x)$ on Ω.
2. The support of μ_n is restricted to the intersection $\Omega \cap \bigcup_{j=1}^{n} B_j$ of Ω with the union of closed balls

$$B_j := \left\{ x \in \mathbb{R}^d \ : \ \|x - w_j\| \le \varepsilon \right\},$$

centered about the nth extremal points $w_1, \ldots, w_n$, and each with radius

$$\varepsilon := \sqrt[d]{\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{d/2} n}} \quad (\Gamma : \text{Gamma function}). \tag{9}$$

3. We have $\mu_n(B_j) = \frac{1}{n}$ for all $j = 1, \ldots, n$, i.e.,

$$\int\limits_{B_j \cap \Omega} \kappa_n(x)\, \mathrm{d}\,\sigma(x) = \frac{1}{n}.$$

Definition 2 A measure μ_n with the above properties will be denoted a *locally redistributed point mass*. The set of all functions κ being the density of such a measure will be denoted by $\mathscr{D}_n$.

The existence of such measures depends on regularity properties of the set Ω. For a set Ω which is d-regular in the sense of Ahlfors-David, we always have $\mathscr{D}_n \neq \emptyset$ (see Theorem 1). If $\mathscr{D}_n \neq \emptyset$, there is a $\mu \in \mathscr{M}(\Omega)$ such that $I_s(\mu) = ||\kappa||_{L^1} < \infty$, and hence, $V_d(s) < \infty$. The following lemma points out, how locally redistributed point masses are related to asymptotic estimates for minimal energy.

Lemma 2 *Let $\Omega \subseteq \mathbb{R}^d$ be compact with $\mathscr{D}_n \neq \emptyset$. Then, for any $n \in \mathbb{N}$, the estimate*

$$V_d(s) \le \frac{2}{n^2} \mathscr{E}_{d,s}(n) + \inf_{\kappa \in \mathscr{D}_n} \left\{ c_\kappa^{\text{approx}}(n) + c_\kappa^{\text{diag}}(n) \right\}$$

holds, with

$$c_\kappa^{\mathrm{diag}}(n) := \sum_{j=1}^{n} \int_{B_j}\int_{B_j} R_s(x-y)\kappa(x)\kappa(y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y),$$

$$c_\kappa^{\mathrm{approx}}(n) := \sum_{\substack{j=1\\k\neq j}}^{n}\sum_{k=1}^{n} \int_{B_j}\int_{B_k} \big[R_s(x-y) - R_s(w_j-w_k)\big]\kappa(x)\kappa(y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y).$$

Definition 3 The numbers $c_\kappa^{\mathrm{diag}}(n)$ in Lemma 2 are called the *diagonal error* and the numbers $c_\kappa^{\mathrm{approx}}(n)$ are denoted *local approximation error* of the energy-approximation by extremal points.

Proof (Lemma 2) Since any function in $\mathscr{D}_n$ defines a measure in $\mathscr{M}(\Omega)$, and from the definition of the continuous minimal energy we immediately conclude

$$\begin{aligned} V_d(s) \le\ & \sum_{\substack{j=1\\k\neq j}}^{n}\sum_{k=1}^{n} \int_{B_j}\int_{B_k} R_s(x-y)\kappa(x)\kappa(y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) \\ & + \sum_{j=1}^{n} \int_{B_j}\int_{B_j} R_s(x-y)\kappa(x)\kappa(y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) \\ =\ & \frac{1}{n^2}\sum_{\substack{j=1\\k\neq j}}^{n}\sum_{k=1}^{n} R_s(w_j-w_k) + c_\kappa^{\mathrm{diag}}(n) \\ & + \sum_{\substack{j=1\\k\neq j}}^{n}\sum_{k=1}^{n} \int_{B_j}\int_{B_k} \big[R_s(x-y) - R_s(w_j-w_k)\big]\kappa(x)\kappa(y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) \\ =\ & \frac{2}{n^2}\mathscr{E}_{d,s}(n) + c_\kappa^{\mathrm{diag}}(n) + c_\kappa^{\mathrm{approx}}(n) \end{aligned}$$

for any $\kappa \in \mathscr{D}_n$. □

This estimate is true for any vector of points on Ω if $\mathscr{E}_{d,s}(n)$ is replaced by the discrete energy of the particular point system. An upper estimate for $V_d(s)$ results if it is applied to the extremal points as done in the preceding lemma. With regard to the desired estimate on the asymptotic behavior of the discrete minimal energy, the question arises if functions κ exists such that $c_\kappa^{\mathrm{diag}}(n)$ and $c_\kappa^{\mathrm{approx}}(n)$ possess the sharp error order for $n \to \infty$. Such an estimate, and hence the construction of a corresponding κ, is not independent of the position of the extremal points, since the local approximation error $c_\kappa^{\mathrm{approx}}(n)$ depends on the position of the extremal points. The diagonal error $c_\kappa^{\mathrm{diag}}(n)$ in contrast is independent of the extremal point

configuration. To get a first idea on the quality of the estimates on $\mathscr{E}_{d,s}(n)$ that can be obtained by this approach we concentrate on the construction of estimates on the diagonal error via appropriate density functions κ in this work. In the following Lemma 3 the diagonal error will be analyzed in a more abstract setting and then concretized to Ahlfors-David regular sets in Theorem 1.

Lemma 3 *Let $\Omega \subseteq \mathbb{R}^d$ ($d \geq 2$) be compact with $\mathscr{D}_n \neq \emptyset$. Furthermore, let $0 < s \leq d-2$. Then, for any $\kappa \in \mathscr{D}_n$*

$$c_\kappa^{\mathrm{diag}} \leq C(d,s)\|\kappa\|_\infty n^{\frac{s}{d}-1}$$

with the least upper bound $\|\kappa\|_\infty$ of κ on B_j. For $s = 0$, we have

$$c_\kappa^{\mathrm{diag}} \leq C(d,0)\|\kappa\|_\infty \frac{\log n}{n}$$

for any $\kappa \in \mathscr{D}_n$. The constant $C(d,s)$ depends on d and s, but not on n.

Proof (Lemma 3) Due to $0 \leq s \leq d-2$, the locally defined potential

$$U_{s,B_j}^{\mu_n}(y) := \int_{B_j} R_s(x-y)\kappa(x)\,\mathrm{d}\,\sigma(x)$$

of the measure μ_n on B_j with density $\kappa|_{B_j\cap\Omega}, \kappa \in \mathscr{D}_n$ with respect to the Lebesgue-measure σ is a *(super)harmonic* function on B_j. Hence,

$$c_\kappa^{\mathrm{diag}}(n) = \sum_{j=1}^{n} \int_{B_j} U_{s,B_j}^{\mu_n}(y)\kappa(y)\,\mathrm{d}\,\sigma(y) \leq \sum_{j=1}^{n} U_{s,B_j}^{\mu_n}(w_j).$$

To estimate $U_{s,B_j}^{\mu_n}(w_j)$, we now treat the cases $s = 0$ and $s > 0$ separately. First we tackle the case $s > 0$. Using spherical coordinates (in $\mathbb{R}^d$), we obtain

$$\begin{aligned}
U_{s,B_j}^{\mu_n}(w_j) &\leq \int_{B_j} \kappa(x)\|x-w_j\|^{-s}\,\mathrm{d}\,\sigma(x) \\
&\leq 2\pi\|\kappa\|_\infty \int_0^\varepsilon \int_0^\pi \dots \int_0^\pi \rho^{d-s-1}\sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2) \\
&\qquad \dots \sin(\varphi_{d-2})\,\mathrm{d}\varphi_{d-2}\dots\mathrm{d}\varphi_1\mathrm{d}\rho \\
&= 2\pi\|\kappa\|_\infty \frac{\varepsilon^{d-s}}{d-s} \int_0^\pi \dots \int_0^\pi \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2) \\
&\qquad \dots \sin(\varphi_{d-2})\,\mathrm{d}\varphi_{d-2}\dots\mathrm{d}\varphi_1.
\end{aligned}$$

With

$$I_m := \int_0^\pi \sin^m(t)\,\mathrm{d}t \quad (m \in \mathbb{N})$$

and by inserting ε according to Eq. (9), we obtain

$$\begin{aligned} U^{\mu_n}_{s,B_j}(w_j) &\le 2\pi\|\kappa\|_\infty \frac{\varepsilon^{d-s}}{d-s} I_0 \dots I_{d-2} \\ &= \widetilde{C}(d,s)\|\kappa\|_\infty \frac{1}{d-s}\left(\sqrt[d]{\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{d/s} n}}\right)^{d-s} = C(d,s)\|\kappa\|_\infty n^{\frac{s}{d}-1}. \end{aligned} \tag{10}$$

Here, $\widetilde{C}(d,s) := I_0 \cdot \ldots \cdot I_{d-2}$ with

$$I_{2m} = \frac{(2m-1)(2m-3)\dots 3\cdot 1}{2m(2m-2)\dots 4\cdot 2}\pi, \quad I_{2m+1} = \frac{2m(2m-2)\dots 4\cdot 2}{(2m+1)(2m-1)\dots 5\cdot 3}2,$$

is a constant depending on d and s but not on n.

In case of $s = 0$ (logarithmic potential) the integral over B_j is computed as follows:

$$\begin{aligned} &\int_{B_j} -\kappa(x)\log\|x-w_j\|\,\mathrm{d}\sigma(x) \\ &\qquad \le \|\kappa\|_\infty \int_{\rho=0}^{\varepsilon}\int_{\varphi_1=0}^{\pi}\dots\int_{\varphi_{d-2}=0}^{\pi}\int_{\theta=0}^{2\pi}(-\log\rho)\rho^{d-1}\sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2) \\ &\qquad\qquad \dots \sin(\varphi_{d-2})\,\mathrm{d}\theta\mathrm{d}\varphi_{d-2}\dots\mathrm{d}\varphi_1\mathrm{d}\rho \\ &\qquad = C(d,0)\|\kappa\|_\infty \frac{\log n}{n}, \end{aligned} \tag{11}$$

yielding the claimed estimate for all cases $s = 0$ and $d \ge 2$. □

The above lemma reduces the analysis of the asymptotic behavior of the discrete energy on constructing a sequence of suitable functions κ_n on Ω, such that both local approximation error and diagonal error remain in the desired order, which is $n^{\frac{s}{d-1}-1}$ for $s > 0$ according to the results known for the sphere, since demanding a corresponding estimate for the diagonal error requires that $\|\kappa_n\|_\infty$ must not grow too fast. On the other hand, functions κ_n localized in a very small area around the center w_j of B_j provide a smaller local approximation error. For d-regular sets in the sense of Ahlfors-David, we can state the following theorem:

Theorem 1 *Let $\Omega \subseteq \mathbb{R}^d$ ($d \geq 2$) be compact and d-regular in the sense of Ahlfors-David, such that $\lambda\, \sigma(B_r(x)) \leq \sigma(\Omega \cap B_r(x))$ holds with some $\lambda > 0$ for all $0 < r < \operatorname{diam}(\Omega)$. Then $\mathscr{D}_n \neq \emptyset$ and the following holds:*

1. *Functions constant on each B_j, $1 \leq j \leq n$ are contained in $\mathscr{D}_n$, and we obtain*

$$c_{\kappa_n}^{\mathrm{diag}} \leq \begin{cases} \frac{C(d,s)}{\lambda} n^{\frac{s}{d}-1}, & s > 0, \\ \frac{C(d,0)}{\lambda} \frac{\log n}{n}, & s = 0, \end{cases}$$

   for each $n \in \mathbb{N}$.
2. *For $s > 0$ there is $\kappa_n \in \mathscr{D}_n$ with*

$$c_{\kappa_n}^{\mathrm{diag}} \leq \frac{C(d,s)}{\lambda} n^{\frac{s}{d-1}-1}$$

   and $\|\kappa_n\|_\infty = \frac{C}{\lambda} n^{\frac{s}{d(d-1)}}$

All constants depend on s and d, but not on n.

The meanings of these estimates are as follows:

1. A constant κ_n is the worst case for the approximation error and the best for the diagonal error. In case of d-regular sets we can obtain a diagonal error that is below the sharp error asymptotics for the sphere.
2. A d-regular set permits a mild growth of κ_n and simultaneously an asymptotic behavior of the diagonal error that matches the error asymptotics in case of the sphere. Hence, the question arises if the growth of κ_n suffices to have a local approximation error in the desired magnitude.

Theorem 1 does also hold if Ω itself is not a d-regular set, but bounding such a set. In this case we can replace Ω by the d-regular set D that is bounded by it. The reason for this is that for (super) harmonic potentials ($0 \leq s \leq d-2$) the equilibrium measure is concentrated on the outer boundary, i.e., we have $V_d(s, D) = V_d(s, \Omega)$.

Proof (Theorem 1) Ahlfors-David d-regularity implies $\sigma(B_j \cap \Omega) \geq \frac{\lambda}{n} > 0$ with a constant $\lambda = \lambda(\Omega)$. Hence, $B_j \cap \Omega$ is no null set for the Lebesgue measure σ. By setting

$$\kappa_n(x) = \frac{1}{n\, \sigma(B_j \cap \Omega)}, \quad x \in B_j \cap \Omega,$$

we, hence, define a function $\kappa_n \in \mathscr{D}_n$ that is constant on each B_j. The d-regularity further implies

$$\|\kappa_n\|_\infty \leq \frac{1}{n\lambda\, \sigma(B_j)} \leq \frac{1}{\lambda}$$

with a constant $0 < \lambda \leq 1$, which—inserted into estimate in Lemma 3—yields

$$c_{\kappa_n}^{\mathrm{diag}} \leq \begin{cases} \frac{C(d,s)}{\lambda} n^{\frac{s}{d}-1}, & s > 0, \\ \frac{C(d,0)}{\lambda} \frac{\log n}{n}, & s = 0, \end{cases}$$

as stated in the first part of the theorem.

For the second part of the theorem, we choose $0 < \delta < \varepsilon$ such that the closed ball $B_\delta(w_j)$ about w_j with radius δ possesses d-dimensional Lebesgue measure $\sigma(B_\delta(w_j)) = n^{-\frac{s}{d(d-1)}-1}$. Due to d-regularity, neither $B_\delta(w_j)$ nor B_j is a Lebesgue-null set, and hence we can define

$$k_n(x) = \begin{cases} \frac{1-\alpha}{n\,\sigma(B_\delta(w_j)\cap\Omega)}, & x \in B_\delta(w_j), \\ \frac{\alpha}{n\,\sigma((B_j\setminus B_\delta(w_j))\cap\Omega)}, & x \in B_j \setminus B_\delta(w_j), \end{cases}$$

where $0 < \alpha < \lambda \leq 1$ is chosen that the maximum of κ_n is attained on $B_\delta(w_j)$, which may only be critical for a finite number of small n. For such a κ_n we have $\kappa_n \in \mathscr{D}_n$ and, again employing the d-regularity of Ω,

$$\|\kappa_n\|_\infty \leq \frac{1-\alpha}{\lambda} n^{\frac{s}{d(d-1)}}.$$

Hence, by Lemma 3

$$c_{\kappa_n}^{\mathrm{diag}} \leq \frac{C(d,s)}{\lambda} n^{\frac{s}{d(d-1)}+\frac{s}{d}-1} = \frac{C(d,s)}{\lambda} n^{\frac{s}{d-1}-1}. \qquad \square$$

With regard to the estimate in Lemma 2, we end the paper with the following question:

Problem 1 If Ω is a d-regular set in the sense of Ahlfors-David, is there for any $n \in \mathbb{N}$ a $\kappa_n \in \mathscr{D}_n$ such that $\|\kappa_n\|_\infty \leq C n^{\frac{s}{d(d-1)}}$ and

$$c_{\kappa_n}^{\mathrm{approx}} \leq \frac{C(d,s)}{\lambda} n^{\frac{s}{d-1}-1}?$$

Moreover, if this is not the case, which sort of regularity assumptions on Ω are required to have such an estimate?

Acknowledgments The authors wish to thank the two anonymous referees for their insightful comments and suggestions, which helped to improve the paper. Marco Rozgič would like to express his gratitude toward the German Research Foundation (DFG) for support and funding under the contract PAK 343/2.

References

1. Bondarenko, A.V., Hardin, D.P., Saff, E.B.: Minimal N-point diameters and f-best-packing constants in $\mathbb{R}^d$. ArXiv e-prints 1204.4403 (2012)
2. Brauchart, J.S.: About the second term of the asymptotics for optimal Riesz energy on the sphere in the potential-theoretical case. Integr. Transform Spec. Funct. **17**, 321–328 (2006)
3. Brauchart, J.S., Hardin, D.P., Saff, E.B.: Riesz energy and sets of revolution in $\mathbb{R}^3$, functional analysis and complex analysis. Contemp. Math. **481**, 47–57 (2009)
4. Brauchart, J.S., Hardin, D.P., Saff, E.B.: The next-order term for optimal Riesz and logarithmic energy asymptotics on the sphere, recent advances in orthogonal polynomials, special functions, and their applications. Contemp. Math. **578**, 31–61 (2012)
5. Conn, A., Gould, N., Toint, P.: Trust Region Methods. SIAM, Philadelphia (2000)
6. David, G., Semmes, S.: Analysis of and on Uniformly Rectifiable Sets, Mathematical Surveys and Monographs, vol. 38, American Mathematical Society, Providence, RI (1993)
7. David, G., Semmes, S.: Fractured Fractals and Broken Dreams: Self-Similar Geometry Through Metric and Measure. Oxford Lecture Series in Mathematics and its Applications, Oxford University Press on Demand, New York (1997)
8. Erber, T., Hockney, G.: Complex systems: equilibrium configurations of N equal charges on a sphere ($2 \leq N \leq 112$). Adv. Chem. Phys. **98**, 495–594 (1997). doi:10.1002/9780470141571.ch5
9. Fekete, M.: Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. Math. Z. **17**(1), 228–249 (1923). doi:10.1007/BF01504345. http://dx.doi.org/10.1007/BF01504345
10. Frank, F.C., Kasper, J.S.: Complex alloy structures regarded as sphere packings. I. Definitions and basic principles. Acta Crystallogr. **11**(3), 184–190 (1958). doi:10.1107/S0365110X58000487. http://dx.doi.org/10.1107/S0365110X58000487
11. Frank, P., von Mises, R.: Die Differential-und Integralgleichungen der Mechanik und Physik. Rosenberg, New York (1943)
12. Giomi, L., Bowick, M.J.: Defective ground states of toroidal crystals. Phys. Rev. E. **78**, 010,601 (2008). doi:10.1103/PhysRevE.78.010601. http://link.aps.org/doi/10.1103/PhysRevE.78.010601
13. Hardin, D.P., Saff, E.B.: Discretizing manifolds via minimum energy points. Notes AMS **51**, 647–662 (2004)
14. Hardin, D.P., Saff, E.B.: Minimal Riesz energy point configurations for rectifiable d-dimensional manifolds. Adv. Math. **193**, 174–204 (2005)
15. Hardin, D.P., Saff, E.B., Stahl, H.: Support of the logarithmic equilibrium measure on sets of revolution in $\mathbb{R}^3$. J. Math. Phys. **48**(2), 022901 (2007). doi:10.1063/1.2435084. http://link.aip.org/link/?JMP/48/022901/1
16. Karush, W.: Minima of functions of several variables with inequalities as side constraints. Ph.D. thesis, Master's thesis, Deptartment of Mathematics, University of Chicago (1939)
17. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: Proceedings of the Second Berkeley Symposium on Mathematica Statistics and Probability, 1950, pp. 481–492. University of California Press, Berkeley and Los Angeles (1951)
18. Kuijlaars, A.B.J., Saff, E.B.: Asymptotics for minimal discrete energy on the sphere. Trans. Am. Math. Soc. **350**(2), 523–538 (1998)
19. Landkof, N.S.: Foundations of Modern Potential Theory. Springer, Berlin (1972). http://opac.inria.fr/record=b1078433
20. MATLAB: version 8.2.0.701 (R2013b).: The MathWorks Inc., Natick, Massachusetts (2013)
21. Mattila, P., Saaranen, P.: Ahlfors-David regular sets and bilipschitz maps. Ann. Acad. Sci. Fenn. **34**, 487–502 (2009)
22. Menke, K.: Extremalpunkte und konforme Abbildung. Math. Ann. **195**, 292–308 (1972). doi:10.1007/BF01423615. http://dx.doi.org/10.1007/BF01423615
23. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research, Springer, New York (1999)

24. Pólya, G., Szegő, G.: Über den transfiniten Durchmesser (Kapazitätskonstante) von ebenen und räumlichen Punktmengen. J. Reine Angew. Math. **165**, 4–49 (1931). doi:10.1515/crll.1931.165.4
25. Rajon, Q., Ransford, T., Rostand, J.: Computation of capacity via quadratic programming. J. Math. Pure Appl. **94**, 398–413 (2010). doi:10.1016/j.matpur.2010.03.004
26. Rajon, Q., Ransford, T., Rostand, J.: Computation of weighted capacity. J. Approx. Theory **162**(6), 1187–1203 (2010). doi:10.1016/j.jat.2009.12.010http://dx.doi.org/10.1016/j.jat.2009.12.010
27. Rakhmanov, E., Saff, E., Zhou, Y.: Electrons on the sphere. In: Ali, R.M. (ed.) Computational Methods and Function Theory, pp. 111–127. World Scientific, Singapore (1995)
28. Rakhmanov, E.A., Saff, E.B., Zhou, Y.M.: Minimal discrete energy on the sphere. Math. Res. Lett. **1**, 647–662 (1994)
29. Ransford, T.: Potential Theory in the Complex Plane. London Mathematical Society Student Texts. Cambridge University Press, Cambridge (1995)
30. Rozgič, M., Jaraczewski, M., Stiemer, M.: Inner point methods: on the necessary conditions of various reformulations of a constrained optimization problem. Technical report, Helmut Schmidt University—University of the Federal Armed Forces Hamburg (2013, in preparation)
31. Saff, E., Kuijlaars, A.: Distributing many points on a sphere. Math. Intell. **19**, 5–11 (1997). doi:10.1007/BF03024331. http://dx.doi.org/10.1007/BF03024331
32. Saff, E.B., Totik, V.: Logarithmic Potentials with External Fields. Springer, Berlin (1997). http://opac.inria.fr/record=b1093628
33. Tsuji, M.: Potential Theory in Modern Function Theory, 2nd edn. Chelsea Publishing Company, New York (1975)
34. Wächter, A., Biegler, L.T.: Line search filter methods for nonlinear programming: local convergence. SIAM J. Optim. **16**(1), 32–48 (2005)
35. Wächter, A., Biegler, L.T.: Line search filter methods for nonlinear programming: motivation and global convergence. SIAM J. Optim. **16**(1), 1–31 (2005). doi:10.1137/S1052623403426556. http://dx.doi.org/10.1137/S1052623403426556
36. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**, 25–57 (2006)
37. Wagner, G.: On means of distances on the surface of a sphere (lower bounds). Pac. J. Math. **144**(2), 389–398 (1990). doi:10.2140/pjm.1990.144.389
38. Wagner, G.: On means of distances on the surface of a sphere. II: upper bounds. Pac. J. Math. **154**(2), 381–396 (1992). doi:10.2140/pjm.1992.154.381

Eigenvalue Sequences of Positive Integral Operators and Moduli of Smoothness

T. Jordão, V. A. Menegatto and Xingping Sun

Abstract We utilize moduli of smoothness and K-functionals as new tools in the arena of estimating the decay rates of eigenvalue sequences associated with some commonly used positive integral operators on spheres. This approach is novel and effective. We develop two readily verifiable and implementable conditions for the kernels of the integral operators under which favorable decay rates of eigenvalue sequences are derived. The first one (based on spherical mean operators) is an enhancement of the classical Hölder condition. The second one, works seamlessly with the Laplace-Beltrami operators and can be applied directly to Bessel potential kernels.

Keywords Sphere · Decay rates · Positive integral operators · Fourier coefficients · Moduli of smoothness

1 Introduction

In this paper, we study the decay rates of eigenvalues for certain types of kernel operators on spheres. Research of this nature can trace its own origin to at least the year 1912. To give readers a historical perspective, we begin by summarizing the

T. Jordão (✉) · V. A. Menegatto
Departamento de Matemática-ICMC-USP, Universidade de São Paulo,
São Carlos, SP 13560-970, Brazil
e-mail: thsjordao@gmail.com

V. A. Menegatto
e-mail: menegatt@icmc.usp.br

X. Sun
Department of Mathematics, Missouri State University, 901 S. National Ave.,
Springfield, MO 65804, USA
e-mail: xsun@missouristate.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_13,

main results in this research area. A function $K \in L^2([0,1]^2)$ gives rise to a compact operator $\mathscr{L}_K$ from $L^2([0,1])$ to itself as described by the following equation:

$$\mathscr{L}_K(f)(x) = \int_{[0,1]} K(x,y) f(y)\, dy, \quad f \in L^2([0,1]), \quad x \in [0,1].$$

We will refer to K as the generating kernel for the operator $\mathscr{L}_K$. In most contexts, the association of K with $\mathscr{L}_K$ is obvious. We will then simply call K the kernel and $\mathscr{L}_K$ the operator. If we make the following symmetry assumption:

$$K(x,y) = \overline{K(y,x)}, \quad \text{for almost all } (x,y) \in [0,1]^2,$$

then the operator $\mathscr{L}_K$ is self-adjoint, and therefore has an eigenvalue sequence $\{\lambda_n\}$ approaching zero. The eigenvalues can be conveniently arranged in decreasing order according to their modulus:

$$|\lambda_1| \geq |\lambda_2| \geq \cdots,$$

in which the number of appearances of each eigenvalue is equal to its algebraic multiplicity.

Weyl [22] proved that, if $K \in C^\ell\left([0,1]^2\right)$, then we have

$$\lambda_n = o\left(n^{-(\ell+1/2)}\right),$$

as $n \to \infty$. Here $C^\ell\left([0,1]^2\right)$ denotes the Banach space of all functions whose partial derivatives up to order ℓ are continuous on $[0,1]^2$. If, in addition, the operator is positive, then Reade [16] established the faster decay rate of the eigenvalues

$$\lambda_n = O\left(n^{-(\ell+1)}\right),$$

as $n \to \infty$. The operator $\mathscr{L}_K$ is *positive* if and only if the kernel K is *positive definite* in the sense that for each $f \in L^2([0,1])$, we have

$$\int_0^1 \int_0^1 K(x,y) f(x) \overline{f(y)}\, dxdy \geq 0.$$

In a separate paper, Reade [17] considered kernels of Hölder class, i.e., symmetric kernels K for which there is a constant C, independent of x and y, such that

$$|K(x,y) - K(x',y')| \leq C(|x-x'|^r + |y-y'|^r),$$

where $0 < r < 1$ is a prescribed constant. Reade showed that if K satisfies the above inequality and is positive definite, then

$$\lambda_n = O\left(n^{-(1+r)}\right)$$

as $n \to \infty$. Remarkably, Kühn [12] made a sweeping generalization. For a multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, let D^α stand for $\partial^{|\alpha|}/(\partial x_1^{\alpha_1}\, \partial x_2^{\alpha_2} \cdots \partial x_m^{\alpha_m})$. The difference operators Δ_h, $h \in \mathbb{R}^m$, are defined by

$$\Delta_h f(x) = f(x+h) - f(x), \quad x \in \mathbb{R}^m,$$

and Δ_h^ℓ, $\ell \in \mathbb{N}$, are defined iteratively. For $0 < r = \kappa + s < \infty$, with $\kappa \in \mathbb{Z}_+$, $0 < s \le 1$, the Hölder class $\mathscr{C}^r(\mathbb{R}^m)$ consists of all functions $f : \mathbb{R}^m \to \mathbb{C}$ such that

$$\|f\|_{\mathscr{C}^r(\mathbb{R}^m)} := \sum_{|\alpha| \le \kappa} \sup_{x \in \mathbb{R}^m} |D^\alpha f(x)| + \sum_{|\alpha| = \kappa} \sup_{h, x \in \mathbb{R}^m} \frac{|\Delta_h^2 D^\alpha f(x)|}{\|h\|^r} < \infty.$$

For a subset $\Omega \subset \mathbb{R}^m$, one defines

$$\mathscr{C}^r(\Omega) = \{f : \Omega \to \mathbb{C} : \text{ there exists } g \in \mathscr{C}^r(\mathbb{R}^m) \text{ with } f = g|_\Omega\},$$

and for $f \in \mathscr{C}^r(\Omega)$, one defines

$$\|f\|_{\mathscr{C}^r(\Omega)} = \inf\{\|g\|_{\mathscr{C}^r(\mathbb{R}^d)} : \ f = g|_\Omega\}.$$

Let M be a compact m-dimensional C^∞-manifold. One defines $\mathscr{C}^r(M)$ to be the class of all continuous functions f on M that are locally in $\mathscr{C}^r$, i.e., for each chart (of M) $\Phi : U \to \mathbb{R}^m$, $f \circ \Phi^{-1} \in \mathscr{C}^r(\Phi(U))$. Choosing charts $\Phi_i : U_i \to \mathbb{R}^m$, $1 \le i \le N$, such that $M = \cup_{i=1}^N U_i$ and $\Phi_i(U_i)$ are bounded C^∞-domains of $\mathbb{R}^m$, one defines

$$\|f\|_{\mathscr{C}^r(M)} = \sup_{1 \le i \le N} \|f \circ \Phi_i^{-1}\|_{\mathscr{C}^r(\Phi_i(U_i))}.$$

The kernel classes Kühn had investigated are $\mathscr{C}^{r,0}(M)$ consisting of all continuous functions $K : M \times M \to \mathbb{C}$ such that

$$K(\cdot, y) \in \mathscr{C}^r(M), \quad \text{for each fixed } y \in M,$$

and

$$\|K\|_{\mathscr{C}^{r,0}(M)} := \sup_{y \in M} \|K(\cdot, y)\|_{\mathscr{C}^r(M)} < \infty.$$

Kühn [12] proved the following result:

Theorem 1 *Let M be a compact m-dimensional C^∞-manifold equipped with a finite Lebesgue-type measure μ. For each $0 < r < \infty$ and every positive definite kernel $K \in \mathscr{C}^{r,0}(M)$, it holds that*

$$\lambda_n = O(n^{-(r/m+1)}).$$

Both Reade and Kühn had given examples of operators whose eigenvalue sequences have the desirable decay rates but the kernels do not have any extra smoothness, which implies in a certain sense that these types of estimates are best possible. Powerful and general as Kühn's result is, the conditions posed on the kernels are not necessarily easy to verify. This problem becomes more acute when the underlying manifold is the m-dimensional sphere, arguably the most simple and useful compact m-dimensional manifold. Verifying these conditions is a demanding job. Sometimes this task can be outright burdensome. In addition, the supremal norms permeated the process,[1] which has practically excluded some of the most useful kernels, such as those that resemble the Bessel potential kernels.[2] One such kernel is

$$\sin^{2r}(\theta/2), \quad \theta = \cos^{-1} x \cdot y, \quad x, y \in \mathbb{R}^{m+1}, \ |x| = |y| = 1.$$

Here $r > -1/2, r \neq 0$ is a prescribed constant. This kernel has shown promising capacities and utilities for several approximation problems on Euclidean spaces and spherical domains; see [6] and [11].

In this paper, we work on the m-dimensional sphere embedded in the $(m+1)$-dimensional Euclidean space. We develop new and readily verifiable conditions on spherical kernels under which the operators' eigenvalue sequences have the desirable decay rates. Precisely, we give two such conditions; see Eqs. (3) and (4). In establishing these conditions, we introduce new tools, moduli of smoothness and K-functionals of fractional order. To wit, the tools themselves are not new but the idea of utilizing them in this context is. The effectiveness of this approach not only empowers us to derive the results in the current paper, but also has opened a door for future research in identifying new smoothness conditions on kernels so that operators' eigenvalue sequences decay exponentially, or at other nonalgebraic rates.

An outline of the paper is as follows: In Sect. 2, we introduce notation and state our main results. In Sect. 3, we describe the technical machinery that will be needed in the proofs of the main results. The details of the proofs will be given in Sect. 4.

2 Notation and Results

Let S^m $(m \geq 2)$ denote the unit sphere in $\mathbb{R}^{m+1}$ endowed with its usual surface measure σ_m. For $1 \leq p \leq \infty$, we denote $L^p(S^m) := L^p(S^m, \sigma_m)$ the usual L_p space on S^m consisting of all functions $f : S^m \to \mathbb{C}$ satisfying $\int_{S^m} |f(x)|^p \, d\sigma_m(x) < \infty$. A function (kernel) $K \in L^2(S^m \times S^m, \sigma_m \times \sigma_m)$ induces a compact operator $\mathscr{L}_K$

[1] Kühn and his co-author studied integrated Hölder conditions in Cobos and Kühn [4].

[2] Here we call K a Bessel potential kernel if K can reproduce a Bessel potential Sobolev space on spheres. We refer readers to Mhaskar et al. [15] for technical details in this regard.

from $L^2(S^m)$ to itself as described by the following equation:

$$\mathscr{L}_K(f)(x) = \int_{S^m} K(x, y) f(y)\, d\underset{m}{\sigma}(y), \quad f \in L^2(S^m), \quad x \in S^m. \tag{1}$$

A kernel K is called *zonal* if there exists a function $\phi : [-1, 1] \to \mathbb{C}$ such that

$$K(x, y) = \phi(x \cdot y), \quad (x, y) \in S^m \times S^m,$$

where $x \cdot y$ denotes the usual dot product of x and y. Schoenberg [20] characterized all the continuous zonal positive definite kernels $\phi(x \cdot y)$ in the following way:

$$\phi(x \cdot y) = \sum_{k=0}^{\infty} a_k \sum_{j=1}^{d_k^{(m)}} Y_{k,j}(x) Y_{k,j}(y).$$

Here $a_k \geq 0$, and $\sum_{k=0}^{\infty} a_k d_k < \infty$. The set $\{Y_{k,j}\}_{j=1}^{d_k^{(m)}}$ is an orthonormal basis for the space of all the homogeneous harmonic polynomials of degree k whose dimension is

$$d_k^{(m)} = \frac{m + 2k - 1}{k} \binom{m + k - 2}{k - 1} \asymp k^{m-1}.$$

Characterizations of positive definite kernels on spheres of various generalities are also accessible in the literature for which we refer the interested readers to Bochner [2] and Stewart [21].

Our study in this paper concerns kernels of the form:

$$K(x, y) = \sum_{k=0}^{\infty} \sum_{j=1}^{d_k^{(m)}} \alpha_{k,j} Y_{k,j}(x) Y_{k,j}(y), \quad \sum_{k=0}^{\infty} d_k^{(m)} \sum_{j=1}^{d_k^{(m)}} \alpha_{k,j} < \infty.$$

We make two basic assumptions on these kernels:

(A) (Positivity) The expansion coefficients are nonnegative, i.e., $\alpha_{k,j} \geq 0$.
(B) (Monotonicity) The expansion coefficients are monotone decreasing with respect to k, i.e., $\alpha_{k+1,j} \leq \alpha_{k,j'}$, $1 \leq j \leq d_{k+1}^{(m)}$, $1 \leq j' \leq d_k^{(m)}$.

Assumption (A) assures that the operator $\mathscr{L}_K$ is positive and has a uniquely defined square root operator $\mathscr{L}_K^{1/2}$ whose generating kernel $K_{1/2}$ is given by

$$K_{1/2}(x, y) = \sum_{k=0}^{\infty} \sum_{j=1}^{d_k^{(m)}} \alpha_{k,j}^{1/2} Y_{k,j}(x) Y_{k,j}(y).$$

Both $\mathscr{L}_K$ and $\mathscr{L}_K^{1/2}$ are self-adjoint positive operators. The kernel K of the original operator $\mathscr{L}_K$ can be recovered from the kernel $K_{1/2}$ of its square root operator $\mathscr{L}_K^{1/2}$ by the integral relation,

$$\int_{S^m} K_{1/2}(x, y)K_{1/2}(w, x)\, d\sigma_m(x) = K(w, y), \quad y, w \in S^m. \tag{2}$$

A quick reference to Eq. (1) shows that the spherical harmonics $Y_{k,j}, k = 0, 1, \ldots, j = 1, \ldots, d_k^{(m)}$, are all eigenvectors of the operator $\mathscr{L}_K$ (the associated eigenvalues are $\alpha_{k,j}$). Since they form an orthonormal basis of $L^2(S^m)$, Assumption (B) gives an eigenvalue ordering that is suitable for our analysis. It is worth noting that the $Y_{k,j}$'s are also the eigenvectors of the Laplace-Betrami operator.[3] The associated eigenvalues are $-k(k + m - 1)$.

Various forms of Hölder conditions have been proposed and studied in the literature (see [3, 4, 7, 13, 16, 19]). This is not surprising, as the sphere is rich with symmetrical structures for us to explore and utilize. Our first goal in this paper is to continue the path traveled by the authors in Castro and Menegatto [3]. We say that a kernel K satisfies the (B, β)-Hölder condition if there exist a fixed $\beta \in (0, 2]$ and a function B in $L^1(S^m)$ such that

$$|S_t(K(y, \cdot))(x) - K(y, x)| \leq B(y)t^\beta, \quad x, y \in S^m, \quad t \in (0, \pi). \tag{3}$$

Here, S_t stands for the usual *shifting operator*[4] defined by the formula

$$S_t f(x) = \frac{1}{R_m(t)} \int_{R_x^t} f(y)\, d\sigma_r(y), \quad x \in S^m, \quad f \in L^p(S^m), \quad t \in (0, \pi),$$

in which $d\sigma_r(y)$ is the volume element of the ring $R_x^t := \{y \in S^m : d_m(x, y) = t\}$ and $R_m(t)$ is its total volume.

Theorem 2 *If $\mathscr{L}_K$ is a positive integral operator induced by the kernel K satisfying the (B, β)-Hölder condition, then it holds that*

$$\lambda_n(\mathscr{L}_K) = O(n^{-1-\beta/m}), \quad (n \to \infty).$$

---

[3] The Laplace-Betrami operator is the restriction to the sphere S^m of the classical Laplace operator

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_{m+1}^2}$$

in the Euclidean space $\mathbb{R}^{m+1}$.

[4] The shifting operator here can be considered as the restriction to S^m of the spherical mean operator in $\mathbb{R}^{m+1}$.

It is straightforward to verify that the (B, β)-Hölder condition is weaker, and therefore more easily satisfied than many other forms of such conditions studied in the literature. As such, Theorem 2 gives an improvement over the previously known results.

Our second goal is to generalize the result of Theorem 2.5 in Castro and Menegatto [3] so that it will also work with Laplace-Beltrami derivatives of fractional orders. To this end, we need to introduce more notation. For a positive real number r, we write $\mathscr{D}^r(g)$ to denote the *fractional derivative of order r* of a function $g \in L^p(S^m)$ [5, 10]. We denote by $\mathscr{Y}_k(g)$ the orthogonal projection of a function $g \in L^2(S^m)$ onto the space $\mathscr{H}_k^m$ of homogeneous harmonic polynomials of degree k in $m+1$ dimensions (restricted to S^m), and define $\mathscr{D}^r(g)$ via the following Fourier expansion:

$$\mathscr{D}^r(g) \sim \sum_{k=0}^{\infty} (k(k+m-1))^{r/2} \mathscr{Y}_k(g). \tag{4}$$

The space of Bessel potentials $W_p^r(S^m)$ on S^m [18] is defined by

$$W_p^r(S^m) := \{g \in L^p(S^m) : \|g\|_p + \|\mathscr{D}^r(g)\|_p < \infty\}.$$

For convenience, we write $\|g\|_{W_p^r} := \|g\|_p + \|\mathscr{D}^r(g)\|_p$, $g \in W_p^r(S^m)$, and observe that $(W_p^r(S^m), \|\cdot\|_{W_p^r})$ is a Banach space. We remark that the fractional derivative of order r as defined in (4) is also called fractional Laplace-Beltrami operator (of order r) in Dai and Xu [5]. That may be justified by the fact that such concept coincides with the usual Laplace-Beltrami derivative of order r when r is a positive integer [10, 14]. If K is a kernel from $L^2(S^m \times S^m, \sigma_m \times \sigma_m)$ and z is fixed in S^m, then we write K^z to denote the function $\cdot \mapsto K(\cdot, z)$. We will use the symbol $\mathscr{D}^{r,0}K$ to stand for the action of the fractional derivative operator only applied to the first group of variables.

Theorem 3 *Let $\mathscr{L}_K$ be a positive integral operator and assume that, for a fixed $r > 0$, all K^z belong to $W_2^{2r}(S^m)$. If the integral operator generated by $\mathscr{D}^{2r,0}K$ is trace-class, then*

$$\lambda_n(\mathscr{L}_K) = O(n^{-1-2r/m}), \quad (n \to \infty).$$

3 Estimating Sums of Fourier Coefficients

In this section, we review concepts and background materials that we will need in the proofs of our main results. These include difference operators, moduli of smoothness, and the associated K-functionals. Some of Ditzian's recent results concerning spherical type Hausdorff-Young inequalities play an important role here, which we will highlight in the sequel. Pertinent references are [5, 8, 9, 18].

If r is a positive real number, the *difference operator of order r* (with step t) Δ_t^r is given by the formula

$$\Delta_t^r(f) := (I - S_t)^{r/2}(f) = \sum_{k=0}^{\infty} (-1)^k \binom{r/2}{k} S_t^k(f), \quad f \in L^p(S^m),$$

where I denotes the identity operator. Since the shifting operator satisfies [1]

$$\mathscr{Y}_k(S_t(f)) = \frac{P_k^{(m-1)/2}(\cos t)}{P_k^{(m-1)/2}(1)} \mathscr{Y}_k(f), \quad f \in L^p(S^m), \quad k \in \mathbb{Z}_+,$$

where $P_k^{(m-1)/2}$ is the Gegenbauer polynomial of degree k and index $(m-1)/2$, the following Fourier expansion holds

$$\Delta_t^r(f) \sim \sum_{k=0}^{\infty} \left(1 - \frac{P_k^{(m-1)/2}(\cos t)}{P_k^{(m-1)/2}(1)}\right)^{r/2} \mathscr{Y}_k(f), \quad f \in L^p(S^m), \quad t \in (0, \pi).$$

The difference operator is the main object in the definition of the *rth-order modulus of smoothness* (with step t) of a function f in $L^p(S^m)$:

$$\omega_r(f, t)_p := \sup\{\|\Delta_s^r(f)\|_p : s \in (0, t]\}.$$

The last definition we want to introduce is that of K-functional associated to the space W_p^r. For $r > 0$ and $t > 0$, it is given by

$$K_r(f, t)_p := \inf\{\|f - g\|_p + t^r \|g\|_{W_p^r} : g \in W_p^r(S^m)\}. \tag{5}$$

For $f \in L^p(S^m)$, it is known that $\omega_r(f, t)_p$ and $K_r(f, t)_p$ are equivalent [18]:

$$K_r(f, t)_p \approx \omega_r(f, t)_p, \quad t \in (0, \pi). \tag{6}$$

Another interesting property involving the K-functional, a realization theorem for $K_r(f, t)_p$ [9], is given in the lemma below (for the sake of easy referencing). In its statement, the multiplier operator η_t depends upon a fixed function η in $C^\infty[0, \infty)$ possessing the following features: $\eta = 1$ in $[0, 1]$, $\eta = 0$ in $[2, \infty)$ and $\eta(s) \leq 1$, $s \in (1, 2)$. The action of the operator η_t itself is defined by the formula

$$\eta_t(f) = \sum_{k=1}^{\infty} \eta(tk) \mathscr{Y}_k(f), \quad f \in L^p(S^m).$$

Lemma 1 *If $r > 0$ and $f \in L^p(S^m)$, then*

$$\|f - \eta_t(f)\|_p + t^r \|\eta_t(f)\|_{W_p^r} \approx K_r(f, t)_p, \quad t \in (0, \pi).$$

The Fourier coefficients of a function $f \in L^p(S^m)$ with respect to the basis $\{Y_{k,j} : j = 1, 2, \ldots, d_k^{(m)};\ k = 0, 1, \ldots\}$ of $L^2(S^m)$ are defined by

$$c_{k,j}(f) := \int_{S^m} f(y)\overline{Y_{k,j}(y)}\, d\sigma_m(y), \quad j = 1, 2, \ldots, d_k^{(m)}; \quad k = 0, 1, \ldots.$$

In the remainder of the section, we provide estimates for the sums

$$s_k(f) := \sum_{j=1}^{d_k^{(m)}} |c_{k,j}(f)|^2, \quad k = 0, 1, \ldots. \tag{7}$$

The following lemma is proved in Ditzian [9]. We include it here for completeness.

Lemma 2 *($1 \le p \le 2$) If f belongs to $L^p(S^m)$ and q is the conjugate exponent of p, then*

$$\left\{\sum_{k=1}^{\infty} (d_k^{(m)})^{(2-q)/2q} [s_k(f)]^{q/2}\right\}^{1/q} \le a(p)\|f\|_p,$$

in which $a(p)$ is a positive constant depending upon p (and m).

Theorem 4 *If f belong to $L^p(S^m)$ ($1 \le p \le 2$) and q is the conjugate exponent of p, then for each fixed $r > 0$, there exists a constant c_p for which*

$$\left\{\sum_{k=1}^{\infty} (d_k^{(m)})^{(2-q)/2q} (\min\{1, tk\})^{rq} [s_k(f)]^{q/2}\right\}^{1/q} \le c_p \omega_r(f, t)_p, \quad t \in (0, \pi). \tag{8}$$

Ditzian [9] proved this theorem for the special case in which r is a positive integer. Ditzian also mentioned that the same proof can be slightly modified to work for the general case. Because the result of the above theorem plays an important role in the derivation of our main results, we include a full proof here.

Proof Due to the equivalence (6) and Lemma 1, it suffices to prove that

$$\left\{\sum_{k=1}^{\infty} (d_k^{(m)})^{(2-q)/2q} (\min\{1, tk\})^{rq} [s_k(f)]^{q/2}\right\}^{1/q} \le a_p \left(\|f - \eta_t(f)\|_p + t^r \|\eta_t(f)\|_{W_p^r}\right),$$

where a_p is a constant depending upon p. Clearly,

$$s_k(f) = \sum_{j=1}^{d_k^{(m)}} |c_{k,j}(f - \eta_t(f)) + c_{k,j}(\eta_t(f))|^2 \le 2^2 s_k(f - \eta_t(f)) + 2^2 s_k(\eta_t(f)).$$

Writing $S_{t,r,q}(f)$ to denote the left-hand side of (8), we have

$$S_{t,r,q}(f) \leq 2\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}(\min\{1,tk\})^{rq}[s_k(f-\eta_t(f))]^{q/2}\right\}^{1/q} + 2\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}(\min\{1,tk\})^{rq}[s_k(\eta_t(f))]^{q/2}\right\}^{1/q}.$$

We use S_1 and S_2 to denote, respectively, the two terms on the right-hand side of the above inequality. We have

$$S_1 \leq 2\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}[s_k(f-\eta_t(f))]^{q/2}\right\}^{1/q}.$$

We then apply Lemma 2 to obtain

$$S_1 \leq 2a(p)\|f-\eta_t(f)\|_p.$$

Similarly, we estimate S_2,

$$\begin{aligned} S_2 &\leq 2t^r\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}k^{rq}[s_k(\eta_t(f))]^{q/2}\right\}^{1/q} \\ &\leq 2t^r\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}\left[(k(k+m-1))^r s_k(\eta_t(f))\right]^{q/2}\right\}^{1/q} \\ &\leq 2t^r\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}\left[s_k(\mathscr{D}^r(\eta_t(f)))\right]^{q/2}\right\}^{1/q}. \end{aligned}$$

Applying Lemma 2 once again, we deduce that

$$S_2 \leq 2t^r\|\mathscr{D}^r(\eta_t(f))\|_p \leq 2a(p)t^r\|\eta_t(f)\|_{W_p^r}.$$

Thus,

$$S_{t,r,q}(f) \leq 2a(p)\left[\|f-\eta_t(f)\|_p + t^r\|\eta_t(f)\|_{W_p^r}\right],$$

and the proof is complete. □

We conclude this section by bringing the shifting operator into the inequality presented in the above theorem. Its derivation requires an additional equivalence

$$\|S_t(f)-f\|_p \approx \omega_2(f,t)_p, \quad t \in (0,\pi)$$

proved in Ditzian [9].

Corollary 1 *(1 ≤ p ≤ 2) If f belongs to $L^p(S^m)$ and q is the conjugate exponent of p, then there exists a constant c_p for which*

$$\left\{\sum_{k=1}^{\infty}(d_k^{(m)})^{(2-q)/2q}(min\{1,tk\})^{2q}[s_k(f)]^{q/2}\right\}^{1/q} \le c_p\|S_t(f)-f\|_p, \quad t\in(0,\pi).$$

4 Proofs of the Main Results

Our task in this section is to prove both Theorems 2 and 3. To present the proofs in an easily followed fashion, we will first prove a few additional technical results which we organize in a succession of lemmas. We remind readers that the kernels K we are dealing with satisfy Assumptions (A) and (B) in Sect. 2. It follows that for each $z \in S^m$, the Fourier coefficients of the function K^z are

$$c_{k,j}(K^z) = \alpha_{k,j}\overline{Y_{k,j}(z)}, \quad j = 1,2,\ldots,d_k^{(m)}, \quad k = 0,1,\ldots.$$

It also follows that the kernel $K_{1/2}$ of the square root of the integral operator $\mathscr{L}_K$ has the expansion:

$$K_{1/2} \sim \sum_{k=0}^{\infty}\sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}^{1/2}Y_{k,j}\otimes\overline{Y_{k,j}}.$$

That is,

$$c_{k,j}(K_{1/2}^z) = \alpha_{k,j}^{1/2}\,\overline{Y_{k,j}(z)}, \quad j = 1,2,\ldots,d_k^{(m)}, \quad k = 0,1,\ldots,$$

which implies that

$$s_k(K_{1/2}^z) = \sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}|Y_{k,j}(z)|^2, \quad z\in S^m, \quad k = 0,1,\ldots.$$

Integrating on both sides of this equation yields the following result.

Lemma 3 *Under the notations and conditions stated above, the following formula holds:*

$$\int_{S^m} s_k(K_{1/2}^z)\,d\sigma_m(z) = \sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}, \quad k = 0,1,\ldots.$$

The action of the fractional derivative on $K_{1/2}^z$ can be expressed by

$$\mathscr{D}^r(K^z_{1/2}) \sim \sum_{k=0}^{\infty}\sum_{j=1}^{d_k^{(m)}} \alpha_{k,j}^{1/2}(k(k+m-1))^{r/2}\overline{Y_{k,j}(z)}\,Y_{k,j}.$$

It follows that $|\mathscr{D}^r(K^z_{1/2})|^2$ has the Fourier expansion:

$$\sum_{k=0}^{\infty}\sum_{p=0}^{\infty}\sum_{j=1}^{d_k^{(m)}}\sum_{i=1}^{d_k^{(m)}} \alpha_{k,i}^{1/2}\alpha_{p,j}^{1/2}(k(k+m-1))^{r/2}(p(p+m-1))^{r/2}\overline{Y_{k,j}(z)}\,Y_{p,i}(z)Y_{k,j}\otimes\overline{Y_{p,i}}.$$

Since the set of all the spherical harmonics forms an orthonormal basis of $L^2(S^m)$, we have

$$\|\mathscr{D}^r(K^z_{1/2})\|_2^2 = \sum_{k=0}^{\infty}\sum_{j=1}^{d_k^{(m)}} \alpha_{k,j}(k(k+m-1))^r|Y_{k,j}(z)|^2.$$

The following lemma is an immediate implication of the above discussion.

Lemma 4 *The following equation holds true:*

$$\|\mathscr{D}^r(K^z_{1/2})\|_2^2 = \mathscr{D}^{2r}K^z(z) = \mathscr{D}^{2r,0}K(z,z), \quad z \in S^m.$$

Next, we derive an estimate for $\|S_t(K^z_{1/2}) - K^z_{1/2}\|_2^2$, where K satisfies the (B,β)-Hölder condition.

Lemma 5 *If K satisfies the (B,β)-Hölder condition, then*

$$\int_{S^m} \|S_t(K^z_{1/2}) - K^z_{1/2}\|_2^2\, d\sigma_m(z) \le 2\|B\|_1 t^{\beta}, \quad z \in S^m, \quad t \in (0,\pi).$$

Proof Fix z and t. We have

$$\begin{aligned}\|S_t(K^z_{1/2}) - K^z_{1/2}\|_2^2 = &\int_{S^m} S_t(K_{1/2}(\cdot,z))(y)S_t(K_{1/2}(z,\cdot))(y)\,d\sigma_m(y)\\ &-\int_{S^m} S_t(K_{1/2}(\cdot,z))(y)K_{1/2}(z,y)\,d\sigma_m(y)\\ &-\int_{S^m} S_t(K_{1/2}(z,\cdot))(y)K_{1/2}(y,z)]\,d\sigma_m(y)\\ &+\int_{S^m} K_{1/2}(y,z)K_{1/2}(z,y)\,\sigma_m(y).\end{aligned}$$

Writing

$$I_t^z := \int_{S^m} \|S_t(K_{1/2}^z) - K_{1/2}^z\|_2^2 \, d\sigma_m(z),$$

integrating on both sides of the above equation, and using (2), we get

$$I_t^z = \int_{S^m} \frac{1}{R_m(t)} \int_{R_y^t} (S_t(K(x,\cdot))(y) - K(x,y)) \, d\sigma_m(x) \, d\sigma_m(y) + \int_{S^m} (K(y,y) - S_t(K(y,\cdot))(y)) \, d\sigma_m(y).$$

Since K satisfies the (B,β)-Hölder condition, applying Inequality (3), we obtain

$$\begin{aligned} I_t^z &\le \int_{S^m} \frac{1}{R_m(t)} \int_{R_y^t} B(x) t^\beta \, d\sigma_r(x) \, d\sigma_m(y) + \int_{S^m} B(y) t^\beta \, d\sigma_m(y) \\ &= \int_{S^m} S_t(B)(y) t^\beta \, d\sigma_m(y) + \int_{S^m} B(y) t^\beta \, d\sigma_m(y) \\ &= t^\beta \left(\|S_t(B)\|_1 + \|B\|_1 \right). \end{aligned}$$

The result of the lemma then follows from the fact that $\|S_t(B)\|_1 = 1$ (see [1]). □

We are now ready to prove the main results in this paper.

Proof of Theorem 3 Applying Theorem 4 to the function $K_{1/2}^z$ (for the case $p = q = 2$), we have

$$\sum_{k=0}^{\infty} (\min\{1, tk\})^{2r} s_k(K_{1/2}^z) \le c_p \left[\omega_r(K_{1/2}^z, t)_2 \right]^2, \quad z \in S^m, \quad t \in (0, \pi).$$

Since $K_{1/2}^z \in W_2^r$, Lemma 3.8 in Rustamov [18] asserts the existence of a constant $C_1 > 0$ (independent of both $K_{1/2}^z$ and t) so that

$$\omega_r(K_{1/2}^z, t)_2 \le C_1 t^r \|\mathscr{D}^r(K_{1/2}^z)\|_2, \quad z \in S^m, \quad t \in (0, \pi).$$

Hence, we have

$$\sum_{k=0}^{\infty} (\min\{1, tk\})^{2r} s_k(K_{1/2}^z) \le c_p C_1^2 t^{2r} \|\mathscr{D}^r(K_{1/2}^z)\|_2^2, \quad z \in S^m, \quad t \in (0, \pi).$$

Integrating on both sides of the above inequality with respect to z (against the measure $\sigma_m(z)$), we have

$$\sum_{k=0}^{\infty}(\min\{1,tk\})^{2r}\left(\int_{S^m} s_k(K_{1/2}^z)\,d\sigma_m(z)\right) \le c_p C_1^2 t^{2r}\int_{S^m}\|\mathscr{D}^r(K_{1/2}^z)\|_2^2\,d\sigma_m(z),\quad t\in(0,\pi).$$

Since $D^{2r,0}K$ is a trace-class kernel, the result of Lemma 4 asserts that $c_p C_1^2\|\mathscr{D}^r(K_{1/2}^z)\|_2^2$ is a nonnegative constant. Denoting this constant by C_2 and invoking Lemma 3, we obtain

$$\sum_{k=0}^{\infty}(\min\{1,tk\})^{2r}\sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}\le C_2 t^{2r},\quad t\in(0,\pi).$$

Letting $t=1/n$ in the above inequality, we get

$$\sum_{k=0}^{\infty}(\min\{1,k/n\})^{2r}\sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}\le C_2 n^{-2r},\quad n=1,2,\ldots.$$

All the summands in the left-hand side of the above inequality are nonnegative. Dropping those terms with index $k<n$, we derive the following inequality:

$$\sum_{k=n}^{\infty}\sum_{j=1}^{d_k^{(m)}}\alpha_{k,j}\le C_2 n^{-2r},\quad n=1,2,\ldots.$$

It implies that

$$d_n^m\sum_{k=n}^{\infty}\alpha_k\le\sum_{k=n}^{\infty}d_k^{(m)}\alpha_k\le C_2 n^{-2r},\quad n=1,2,\ldots,$$

in which $\alpha_k:=\min\{\alpha_{k,j}: j=1,2,\ldots,d_k^{(m)}\}$, $k=0,1,\ldots$. Using the equivalence $d_n^m\asymp n^{m-1}$ as $n\to\infty$, we arrive at

$$n^{m-1}\sum_{k=n}^{\infty}\alpha_k\le C_3C_2n^{-2r},\quad n=1,2,\ldots,$$

for some $C_3>0$, that is,

$$\sum_{k=n}^{\infty}\alpha_k\le C_3 n^{-2r-m+1},\quad n=1,2,\ldots.$$

Next, observe that

$$n^{2r+m}\alpha_n = n^{2r+m-1}\sum_{k=n}^{2n-1}\alpha_n \le n^{2r+m-1}\sum_{k=n}^{\infty}\alpha_k \le C_3, \quad n = 1, 2, \ldots,$$

or, equivalently, $\alpha_k = O(n^{-2r-m})$, as $n \to \infty$. Returning to our original notation for the eigenvalues of $\mathscr{L}_K$ and recalling that $\{\lambda_n(\mathscr{L}_K)\}$ decreases to 0, we have that $\alpha_n = \lambda_{d_n^{(m+1)}}(\mathscr{L}_K)$, $n = 1, 2, \ldots$. In particular,

$$\lambda_{d_n^{(m+1)}}(\mathscr{L}_K) = O(n^{-2r-m}), \quad (n \to \infty).$$

Therefore, the decay in the statement of the theorem follows. □

Proof of Theorem 2 Many steps in this proof are essentially repetitions of those in Theorem 4. Applying the inequality in Corollary 1 to the function $K_{1/2}^z$ for the case $p = q = 2$ and $r = 2$, we get

$$\sum_{k=1}^{\infty}(\min\{1, tk\})^4 s_k(K_{1/2}^z) \le c_2\|S_t(K_{1/2}^z) - K_{1/2}^z\|_2^2, \quad z \in S^m, \quad t \in (0, \pi).$$

Repeating the same procedure used in the first half of the proof of Theorem 3, we obtain the inequality

$$\sum_{k=0}^{\infty}(\min\{1, tk\})^4 \sum_{j=1}^{d_k^{(m)}}\alpha_{k,j} \le c_2\int_{S^m}\|S_t(K_{1/2}^z) - K_{1/2}^z\|_2^2\, d\sigma_m(z), \quad z \in S^m, \quad t \in (0, \pi).$$

Since K satisfies the (B, β)-Hölder condition, the result of Lemma 5 asserts that

$$\sum_{k=0}^{\infty}(\min\{1, tk\})^4 \sum_{j=1}^{d_k^{(m)}}\alpha_{k,j} \le 2c_2\|B\|_1 t^{\beta}, \quad t \in (0, \pi).$$

Repeating the procedure used in the second half of the proof of Theorem 3 leads us to the conclusion of the theorem we are proving. □

Acknowledgments The first author was partially supported by FAPESP, grant # 2012/25097-4. Part of this research was done while the first author visited the Department of Mathematics at Missouri State University. She is thankful to many people in the host institution for their hospitality. We are grateful to two anonymous referees for their suggestions that have enhanced the exposition of the paper.

References

1. 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) (German)
2. Bochner, S.: Hilbert distances and positive definite functions. Ann. Math. **42**(2), 647–656 (1941)
3. Castro, M.H., Menegatto, V.A.: Eigenvalue decay of positive integral operators on the sphere. Math. Comput. **81**(280), 2303–2317 (2012)
4. Cobos, F., Kühn, T.: Eigenvalues of integral operators with positive definite kernels satisfying integrated Hölder conditions over metric compacta. J. Approx. Theory **63**(1), 39–55 (1990)
5. Dai, F., Xu, Y.: Approximation Theory and Harmonic Analysis on Spheres and Balls. Springer Monographs in Mathematics. Springer, New York (2013)
6. Dai, F., Ditzian, Z.: Combinations of multivariate averages. J. Approx. Theory **131**(2), 268–283 (2004)
7. Dikmen, C.M., Reade, J.B.: Factorisation of positive definite operators. Arch. Math. (Basel) **91**(4), 339–343 (2008)
8. Ditzian, Z.: Smoothness of a function and the growth of its fourier transform or its Fourier coefficients. J. Approx. Theory **162**(5), 980–986 (2010)
9. Ditzian, Z.: Relating smoothness to expressions involving Fourier coefficient or to a Fourier transform. J. Approx. Theory **164**(10), 1369–1389 (2012)
10. Jordão, T., Menegatto, V.A.: Weighted Fourier-Laplace transforms in reproducing kernel Hilbert spaces on the sphere. J. Math. Anal. Appl. **411**(2), 732–741 (2014)
11. Jordão, T., Sun, X.: A general type of spherical mean operators and K-functionals of fractional orders, submitted
12. Kühn, T.: Eigenvalues of integral operators with smooth positive definite kernels. Arch. Math. (Basel) **49**(6), 525–534 (1987)
13. Menegatto, V.A., Oliveira, C.P.: Eigenvalue and singular value estimates for integral operators: a unifying approach. Math. Nachr. **258**(17–18), 2222–2232 (2012)
14. Menegatto, V.A., Piantella, A.C.: Old and new on the Laplace-Beltrami derivative. Numer. Funct. Anal. Optim. **32**(3), 309–341 (2011)
15. Mhaskar, H.N., Narcowich, F.J., Prestin, J., Ward, J.D.: l^p Bernstein estimates and approximation by spherical basis functions. Math. Comput. **79**, 1647–1679 (2010)
16. Reade, J.B.: Eigenvalues of Lipschitz kernels. Math. Proc. Cambridge Philos. Soc. **93**(1), 135–140 (1983a)
17. Reade, J.B.: Eigenvalues of positive definite kernels. SIAM J. Math. Anal. **14**(1), 152–157 (1983b)
18. Rustamov, Kh.P.: On the approximation of functions on a sphere. Izv. Ross. Akad. Nauk Ser. Mat. **57**(5), 127–148 (1993) (translation in Russian Acad. Sci. Izv. Math. **43**(2), 311–329 (1994))
19. Samko, S.G., Vakulov, B.G.: On equivalent norms in fractional order function spaces of continuous functions on the unit sphere. Fract. Calc. Appl. Anal. **3**(4), 401–433 (2000)
20. Schoenberg, I.J.: Positive definite functions on spheres. Duke Math. J. **9**, 96–108 (1942)
21. Stewart, J.: Positive definite functions and generalizations, a historical survey. Rocky Mt. J. Math. **6**(3), 409–434 (1976)
22. Weyl, H.: Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). Math. Ann. **71**(4), 441–479 (1912) (German)

Reconstructing Multivariate Trigonometric Polynomials from Samples Along Rank-1 Lattices

Lutz Kämmerer

Abstract The approximation of problems in d spatial dimensions by trigonometric polynomials supported on known more or less sparse frequency index sets $I \subset \mathbb{Z}^d$ is an important task with a variety of applications. The use of rank-1 lattices as spatial discretizations offers a suitable possibility for sampling such sparse trigonometric polynomials. Given an arbitrary index set of frequencies, we construct rank-1 lattices that allow a stable and unique discrete Fourier transform. We use a component-by-component method in order to determine the generating vector and the lattice size.

Keywords Mutlivariate trigonometric approximation · Lattice rule · Rank-1 lattice · Component-by-component (CBC) · Fast Fourier transform

1 Introduction

Given a spatial dimension $d \in \mathbb{N}$, we consider Fourier series of sufficiently smooth functions $f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}$ mapping the d-dimensional torus $[0, 1)^d$ into the complex numbers $\mathbb{C}$, where $\hat{f}_{\mathbf{k}} \in \mathbb{C}$ are the Fourier coefficients. A sequence $\left(\hat{f}_{\mathbf{k}}\right)_{\mathbf{k}\in\mathbb{Z}^d}$ with a finite number of nonzero elements specifies a trigonometric polynomial. We call the index set of the nonzero elements the frequency index set of the corresponding trigonometric polynomial. For a fixed index set $I \subset \mathbb{Z}^d$ with a finite cardinality $|I|$, $\Pi_I = \mathrm{span}\{\mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}} : \mathbf{k} \in I\}$ is called the space of trigonometric polynomials with frequencies supported on I.

Assuming the index set I is of finite cardinality and a suitable discretization in frequency domain for approximating functions, e.g., functions of specific smoothness,

L. Kämmerer (✉)
Faculty of Mathematics, Technische Universität Chemnitz, 09107 Chemnitz, Germany
e-mail: kaemmerer@mathematik.tu-chemnitz.de

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_14,

cf. [5, 8], we are interested in evaluating the corresponding trigonometric polynomials at sampling nodes and reconstructing the Fourier coefficients $\left(\hat{f}_{\mathbf{k}}\right)_{\mathbf{k}\in I}$ from sample values. Accordingly, we consider (sparse) multivariate trigonometric polynomials

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in I} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}$$

and assume the frequency index set I is given.

For different specific index sets I there has been done some related work using rank-1 lattices as spatial discretizations [4, 7]. A multivariate trigonometric polynomial evaluated at all nodes of a rank-1 lattice essentially simplifies to a one-dimensional fast Fourier transform (FFT) of the length of the cardinality of the rank-1 lattice, cf. [6]. Allowing for some oversampling one can find a rank-1 lattice, which even allows the reconstruction of the trigonometric polynomial from the samples at the rank-1 lattice nodes. A suitable strategy to search for such reconstructing rank-1 lattices can be adapted from numerical integration. In particular, a modification of the component-by-component constructions of lattice rules based on various weighted trigonometric degrees of exactness described in Cools et al. [3] allows one to find adequate rank-1 lattices in a relatively fast way. We already showed the existence and upper bounds on the cardinality of reconstructing rank-1 lattices for hyperbolic crosses as index sets, cf. [4].

In this paper, we generalize these results considering arbitrary frequency index sets I instead of symmetric hyperbolic crosses and suggest some strategies for determining reconstructing rank-1 lattices even for frequency index sets I containing gaps. To this end, we present corresponding component-by-component (CBC) algorithms, where the frequency index set I is the only input.

In Sect. 2, we introduce the necessary notation and specify the relation between exact integration of trigonometric polynomials and reconstruction of trigonometric polynomials using rank-1 lattices. Section 3 contains the main results, i.e., a component-by-component algorithm searching for reconstructing rank-1 lattices for given frequency index sets I and given rank-1 lattice sizes M. In detail, we determine conditions on M guaranteeing the existence of a reconstructing rank-1 lattice of size M for the frequency index set I. The proof of this existence result describes a component-by-component construction of a corresponding generating vector $\mathbf{z} \in \mathbb{N}^d$ of the rank-1 lattice, such that we obtain directly a component-by-component algorithm. In Sect. 4, we give some simple improvements of the component-by-component construction, such that the corresponding algorithms automatically determine suitable rank-1 lattice sizes. Accordingly, the only input is the frequency index set I here. Finally, we give some specific examples and compare the results of our different algorithms in Sect. 5.

2 Rank-1 Lattices

For given $M \in \mathbb{N}$ and $\mathbf{z} \in \mathbb{N}^d$ we define the *rank-1 lattice*

$$\Lambda(\mathbf{z}, M) := \{\mathbf{x}_j = \frac{j\mathbf{z}}{M} \bmod 1,\ j = 0, \ldots, M-1\}$$

as discretization in the spatial domain. Following [6], the evaluation of the trigonometric polynomial $f \in \Pi_I$ with frequencies supported on I simplifies to a one-dimensional discrete Fourier transform (DFT), i.e.,

$$f(\mathbf{x}_j) = \sum_{\mathbf{k}\in I} \hat{f}_{\mathbf{k}} \mathrm{e}^{2\pi \mathrm{i} j \mathbf{k}\cdot\mathbf{z}} = \sum_{l=0}^{M-1} \left(\sum_{\mathbf{k}\cdot\mathbf{z}\equiv l \pmod M} \hat{f}_{\mathbf{k}} \right) \mathrm{e}^{2\pi \mathrm{i} \frac{jl}{M}}.$$

We evaluate f at all nodes $\mathbf{x}_j \in \Lambda(\mathbf{z}, M)$, $j = 0, \ldots, M-1$, by the precomputation of all $\hat{g}_l := \sum_{\mathbf{k}\cdot\mathbf{z}\equiv l \pmod M} \hat{f}_{\mathbf{k}}$ and a one-dimensional (inverse) FFT in $\mathcal{O}(M \log M + d|I|)$ floating point operations, cf. [2], where $|I|$ denotes the cardinality of the frequency index set I.

As the fast evaluation of trigonometric polynomials at all sampling nodes $\mathbf{x}_j$ of the rank-1 lattice $\Lambda(\mathbf{z}, M)$ is guaranteed, we draw our attention to the reconstruction of a trigonometric polynomial f with frequencies supported on I using function values at the nodes $\mathbf{x}_j$ of a rank-1 lattice $\Lambda(\mathbf{z}, M)$. We consider the corresponding Fourier matrix $\mathbf{A}$ and its adjoint $\mathbf{A}^*$,

$$\mathbf{A} := \left(\mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}\right)_{\mathbf{x}\in\Lambda(\mathbf{z},M),\mathbf{k}\in I} \in \mathbb{C}^{M\times|I|} \quad \text{and } \mathbf{A}^* := \left(\mathrm{e}^{-2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}\right)_{\mathbf{k}\in I,\ \mathbf{x}\in\Lambda(\mathbf{z},M)} \in \mathbb{C}^{|I|\times M},$$

in order to determine necessary and sufficient conditions on rank-1 lattices $\Lambda(\mathbf{z}, M)$ allowing for a unique reconstruction of all Fourier coefficients of $f \in \Pi_I$. Note that we assume to run through the sets I and $\Lambda(\mathbf{z}, M)$ in some fixed order whenever we use $\mathbf{k} \in I$ or $\mathbf{x} \in \Lambda(\mathbf{z}, M)$ as running index of matrices or vectors. Hence, the reconstruction of the Fourier coefficients $\hat{\mathbf{f}} = (\hat{f}_{\mathbf{k}})_{\mathbf{k}\in I} \in \mathbb{C}^{|I|}$ from sampling values $\mathbf{f} = (f(\mathbf{x}))_{\mathbf{x}\in\Lambda(\mathbf{z},M)} \in \mathbb{C}^M$ can be realized by solving the normal equation $\mathbf{A}^*\mathbf{A}\hat{\mathbf{f}} = \mathbf{A}^*\mathbf{f}$, which is equivalent to solve the least squares problem

$$\text{find } \hat{\mathbf{f}} \in \mathbb{C}^{|I|} \text{ such that } \|\mathbf{A}\hat{\mathbf{f}} - \mathbf{f}\|_2 \to \min,$$

cf. [1]. Assuming $\mathbf{f} = (f(\mathbf{x}))_{\mathbf{x}\in\Lambda(\mathbf{z},M)}$ being a vector of sampling values of the trigonometric polynomial $f \in \Pi_I$, the vector $\mathbf{f}$ belongs to the range of $\mathbf{A}$ and we can find a possibly nonunique solution $\hat{\mathbf{f}}$ of $\mathbf{A}\hat{\mathbf{f}} = \mathbf{f}$. We compute a unique solution of the normal equation, iff the Fourier matrix $\mathbf{A}$ has full column rank.

Lemma 1 *Let $I \subset \mathbb{Z}^d$ of finite cardinality and $\Lambda(\mathbf{z}, M)$ a rank-1 lattice be given. Then two distinct columns of the corresponding Fourier matrix $\mathbf{A}$ are orthogonal or equal, i.e., $(\mathbf{A}^*\mathbf{A})_{\mathbf{h},\mathbf{k}} \in \{0, M\}$ for $\mathbf{h}, \mathbf{k} \in I$.*

Proof The matrix $\mathbf{A}^*\mathbf{A}$ contains all scalar products of two columns of the Fourier matrix $\mathbf{A}$, i.e., $(\mathbf{A}^*\mathbf{A})_{\mathbf{h},\mathbf{k}}$ is the scalar product of column $\mathbf{k}$ with column $\mathbf{h}$ of the Fourier matrix $\mathbf{A}$. We obtain

$$(\mathbf{A}^*\mathbf{A})_{\mathbf{h},\mathbf{k}} = \sum_{j=0}^{M-1} \left(\mathrm{e}^{2\pi\mathrm{i}\frac{(\mathbf{k}-\mathbf{h})\cdot\mathbf{z}}{M}}\right)^j = \begin{cases} M, & \text{for } \mathbf{k}\cdot\mathbf{z} \equiv \mathbf{h}\cdot\mathbf{z} \,(\mathrm{mod}\, M), \\ \frac{\mathrm{e}^{2\pi\mathrm{i}(\mathbf{k}-\mathbf{h})\cdot\mathbf{z}}-1}{\mathrm{e}^{2\pi\mathrm{i}\frac{(\mathbf{k}-\mathbf{h})\cdot\mathbf{z}}{M}}-1} = 0, & \text{else.} \end{cases}$$ □

According to Lemma 1 the matrix $\mathbf{A}$ has full column rank, iff

$$\mathbf{k}\cdot\mathbf{z} \not\equiv \mathbf{h}\cdot\mathbf{z} \,(\mathrm{mod}\, M), \quad \text{for all} \quad \mathbf{k} \neq \mathbf{h};\ \mathbf{k},\mathbf{h} \in I, \tag{1}$$

or, equivalently,

$$\mathbf{k}\cdot\mathbf{z} \not\equiv 0 \,(\mathrm{mod}\, M), \quad \text{for all} \quad \mathbf{k} \in \mathscr{D}(I)\backslash\{\mathbf{0}\} \tag{2}$$

with $\mathscr{D}(I) := \{\mathbf{h} = \mathbf{l}_1 - \mathbf{l}_2 \colon \mathbf{l}_1, \mathbf{l}_2 \in I\}$. We call the set $\mathscr{D}(I)$ *difference set* of the frequency index set I. Furthermore, we name a rank-1 lattice $\Lambda(\mathbf{z}, M)$ ensuring (1) and (2) *reconstructing rank-1 lattice* for the index set I. In particular, condition (2) ensures the exact integration of all trigonometric polynomials $g \in \Pi_{\mathscr{D}(I)}$ applying the lattice rule given by $\Lambda(\mathbf{z}, M)$, i.e., the identity $\int_{\mathbb{T}^d} g(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{1}{M}\sum_{j=0}^{M-1} g(\mathbf{x}_j)$ holds for all $g \in \Pi_{\mathscr{D}(I)}$, cf. [9]. Certainly, $f \in \Pi_I$ and $\mathbf{k} \in I$ implies that $f\mathrm{e}^{-2\pi\mathrm{i}\mathbf{k}\cdot\circ} \in \Pi_{\mathscr{D}(I)}$ and we obtain

$$\frac{1}{M}\sum_{j=0}^{M-1} f\left(\frac{j\mathbf{z}}{M}\right)\mathrm{e}^{-2\pi\mathrm{i}j\frac{\mathbf{k}\cdot\mathbf{z}}{M}} = \int_{\mathbb{T}^d} f(\mathbf{x})\mathrm{e}^{-2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\mathrm{d}\mathbf{x} =: \hat{f}_{\mathbf{k}},$$

where the right equality is the usual definition of the Fourier coefficients.

Another fact, which comes out of Lemma 1, is that the matrix $\mathbf{A}$ fulfills $\mathbf{A}^*\mathbf{A} = M\mathbf{I}$ in the case of $\Lambda(\mathbf{z}, M)$ being a reconstructing rank-1 lattice for I. The normalized normal equation simplifies to

$$\hat{\mathbf{f}} = \frac{1}{M}\mathbf{A}^*\mathbf{A}\hat{\mathbf{f}} = \frac{1}{M}\mathbf{A}^*\mathbf{f},$$

and in fact we reconstruct the Fourier coefficients of $f \in \Pi_I$ applying the lattice rule

$$\hat{f}_{\mathbf{k}} = \frac{1}{M}\sum_{j=0}^{M-1} f(\mathbf{x}_j)\mathrm{e}^{-2\pi\mathrm{i}\frac{j\mathbf{k}\cdot\mathbf{z}}{M}} = \frac{1}{M}\sum_{j=0}^{M-1} f(\mathbf{x}_j)\mathrm{e}^{-2\pi\mathrm{i}\frac{jl}{M}}$$

for all $\mathbf{k} \in I$ and $l \equiv \mathbf{k}\cdot\mathbf{z} \,(\mathrm{mod}\, M)$. In particular, one computes all Fourier coefficients using one one-dimensional FFT and the unique inverse mapping of $\mathbf{k} \mapsto \mathbf{k}\cdot\mathbf{z} \bmod M$. The corresponding complexity is given by $\mathscr{O}(M\log M + d|I|)$.

Up to now, we wrote about reconstructing rank-1 lattices without saying how to get them. In the following section, we prove existence results and give a first algorithm in order to determine reconstructing rank-1 lattices.

3 A CBC Construction of Reconstructing Rank-1 Lattices

A reconstructing rank-1 lattice for the frequency index set I is characterized by (1) and (2), respectively. Similar to the construction of rank-1 lattices for the exact integration of trigonometric polynomials of specific trigonometric degrees, see [3], we are interested in existence results and suitable construction algorithms for reconstructing rank-1 lattices. In order to prepare the theorem of this section, we define the projection of an index set $I \subset \mathbb{Z}^d$ on $\mathbb{Z}^s$, $d \geq s \in \mathbb{N}$,

$$I_s := \{(k_j)_{j=1}^s : \mathbf{k} = (k_j)_{j=1}^d \in I\}. \tag{3}$$

Furthermore, we call a frequency index set $I \subset \mathbb{Z}^d$ *symmetric to the origin* iff $I = \{-\mathbf{k} : \mathbf{k} \in I\}$, i.e., $\mathbf{k} \in I$ implies $-\mathbf{k} \in I$ for all $\mathbf{k} \in I$.

Theorem 1 *Let $s \in \mathbb{N}$, $d \geq s \geq 2$, $\tilde{I} \subset \mathbb{Z}^d$ be an arbitrary d-dimensional set of finite cardinality that is symmetric to the origin, and M be a prime number satisfying*

$$M \geq \frac{|\{\mathbf{k} \in \tilde{I}_s : \mathbf{k} = (\mathbf{h}, h_s), \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z} \backslash \{0\}\}|}{2} + 2.$$

Additionally, we assume that each nonzero element of the set of the s-th component of $\tilde{I}_s$ and M are coprime, i.e., $M \nmid l$ for all $l \in \{h_s \in \mathbb{Z} \backslash \{0\} : \mathbf{k} = (\mathbf{h}, h_s) \in \tilde{I}_s, \mathbf{h} \in \tilde{I}_{s-1}\}$, and that there exists a generating vector $\mathbf{z}^* \in \mathbb{N}^{s-1}$ that guarantees

$$\mathbf{h} \cdot \mathbf{z}^* \not\equiv 0 \,(\mathrm{mod}\, M) \quad \text{for all} \quad \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\}.$$

Then there exists at least one $z_s^* \in \{1, \ldots, M-1\}$ such that

$$(\mathbf{h}, h_s) \cdot (\mathbf{z}^*, z_s^*) \not\equiv 0 \,(\mathrm{mod}\, M) \quad \text{for all} \quad (\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\}.$$

Proof We adapt the proof of [3, Theorem 1]. Let us assume that

$$\mathbf{h} \cdot \mathbf{z}^* \not\equiv 0 \,(\mathrm{mod}\, M) \quad \text{for all} \quad \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\}.$$

Basically, we determine an upper bound of the number of elements $z_s \in \{1, \ldots, M-1\}$ with

$$(\mathbf{h}, h_s) \cdot (\mathbf{z}^*, z_s) \equiv 0 \,(\mathrm{mod}\, M) \quad \text{for at least one} \quad (\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{0\}$$

or, equivalent,

$$\mathbf{h} \cdot \mathbf{z}^* \equiv -h_s z_s \,(\mathrm{mod}\, M) \quad \text{for at least one} \quad (\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{0\}.$$

Similar to Cools et al. [3] we consider three cases:

$h_s = 0$: With $(\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\}$ we have $\mathbf{0} \neq \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\}$. Consequently, $\mathbf{h} \cdot \mathbf{z}^* \equiv -0 z_s \,(\mathrm{mod}\, M)$ never holds because of $\mathbf{h} \cdot \mathbf{z}^* \not\equiv 0 \,(\mathrm{mod}\, M)$ for all $\mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\}$.

$\mathbf{h} = \mathbf{0}$: We consider $z_s \in \{1, \dots, M-1\}$. We required M being prime, so z_s and M are coprime. Due to $(\mathbf{h}, h_s) \in \tilde{I} \backslash \{\mathbf{0}\}$, we obtain $h_s \neq 0$ and we assumed M and $h_s \neq 0$ are coprime. Consequently, we realize $z_s h_s \neq 0$ and $z_s h_s$ and M are relatively prime. So $\mathbf{0}\mathbf{z}^* \equiv -h_s z_s \,(\mathrm{mod}\, M)$ never holds for $(\mathbf{0}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\}$ and $z_s \in \{1, \dots, M-1\}$.

else: Since $0 \neq h_s$ and M are coprime and $\mathbf{h} \cdot \mathbf{z}^* \not\equiv 0 \,(\mathrm{mod}\, M)$, there is at most one $z_s \in \{1, \dots, M-1\}$ that fulfills $\mathbf{h} \cdot \mathbf{z}^* \equiv -h_s z_s \,(\mathrm{mod}\, M)$. Due to the symmetry of the considered index set $\{(\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\} : \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z} \backslash \{0\}\}$ we have to count at most one z_s for the two elements $(\mathbf{h}, h_s)$ and $-(\mathbf{h}, h_s)$.

Hence, we have at most

$$\frac{|\{(\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\} : \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z} \backslash \{0\}\}|}{2} \tag{4}$$

elements of $\{1, \dots, M-1\}$ with

$$\mathbf{h} \cdot \mathbf{z}^* \equiv -h_s z_s \,(\mathrm{mod}\, M) \quad \text{for at least one} \quad (\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\}.$$

If the candidate set $\{1, \dots, M-1\}$ for z_s^* contains more elements than (4) we can determine at least one z_s^* with

$$\mathbf{h} \cdot \mathbf{z}^* \not\equiv -h_s z_s^* \,(\mathrm{mod}\, M) \quad \text{for all} \quad (\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\}.$$

Consequently, the number of elements in $\{1, \dots, M-1\}$ with

$$|\{1, \dots, M-1\}| \geq \frac{|\{(\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\} : \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z} \backslash \{0\}\}|}{2} + 1$$

and M is prime guarantees that there exists such a z_s^*. Since, we assumed M being prime and

$$\begin{aligned} M &= |\{1, \dots, M-1\}| + 1 \\ &\geq \frac{|\{(\mathbf{h}, h_s) \in \tilde{I}_s \backslash \{\mathbf{0}\} : \mathbf{h} \in \tilde{I}_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z} \backslash \{0\}\}|}{2} + 2 \end{aligned}$$

we can find at least one z_s by testing out all possible candidates $\{1, 2, \dots, M-1\}$. □

Theorem 1 outlines one step of a component-by-component construction of a rank-1 lattice, guaranteeing the exact integration of trigonometric polynomials with frequencies supported on index sets $\tilde{I}$ which are symmetric to the origin.

We obtain this symmetry of the difference sets $\mathscr{D}(I)_s$

$$\mathbf{h} \in \mathscr{D}(I)_s \Rightarrow \exists \mathbf{k}_1, \mathbf{k}_2 \in I_s : \mathbf{h} = \mathbf{k}_1 - \mathbf{k}_2 \Rightarrow -\mathbf{h} = \mathbf{k}_2 - \mathbf{k}_1 \in \mathscr{D}(I)_s.$$

So, our strategy is to apply Theorem 1 to the difference set $\mathscr{D}(I)_s$ of the frequency index set I_s for all $2 \leq s \leq d$. In order to use Theorem 1, we have to find sufficient conditions on rank-1 lattices of dimension $d = 1$ guaranteeing that $hz_1 \not\equiv 0 \pmod{M}$ for all $h \in \mathscr{D}(I)_1 \backslash \{0\}$.

Lemma 2 *Let $I \subset \mathbb{Z}$ be a one-dimensional frequency index set of finite cardinality and M be a prime number satisfying $M \geq |I|$. Additionally, we assume M and h being coprime for all $h \in \mathscr{D}(I)\backslash\{0\}$. Then we can uniquely reconstruct the Fourier coefficients of all $f \in \Pi_I$ applying the one-dimensional lattice rule given by $\Lambda(1, M)$.*

Algorithm 1 Component-by-component lattice search

Input:	$M \in \mathbb{N}$ prime	cardinality of rank-1 lattice
	$I \subset \mathbb{Z}^d$	frequency index set

$\mathbf{z} = \emptyset$
for $s = 1, \ldots, d$ **do**
  form the set I_s as defined in (3)
  search for one $z_s \in [1, M-1] \cap \mathbb{Z}$ with $|\{(\mathbf{z}, z_s) \cdot \mathbf{k} \bmod M : \mathbf{k} \in I_s\}| = |I_s|$
  $\mathbf{z} = (\mathbf{z}, z_s)$
end for

Output:	$\mathbf{z} \in \mathbb{Z}^d$	generating vector

Proof Applying the lattice rule given by $\Lambda(1, M)$ to the integrands of the integrals computing the Fourier coefficient $\hat{f}_k$, $k \in I$, of $f \in \Pi_I$, we obtain

$$\begin{aligned}\frac{1}{M}\sum_{j=0}^{M-1} f\left(\frac{j}{M}\right) \mathrm{e}^{-2\pi \mathrm{i}\frac{kj}{M}} &= \frac{1}{M}\sum_{j=0}^{M-1}\sum_{h\in I} \hat{f}_h \mathrm{e}^{2\pi\mathrm{i}\frac{hj}{M}} \mathrm{e}^{-2\pi\mathrm{i}\frac{kj}{M}} \\ &= \frac{1}{M}\sum_{h\in I} \hat{f}_h \sum_{j=0}^{M-1} \mathrm{e}^{2\pi\mathrm{i}\frac{(h-k)j}{M}} = \hat{f}_k = \int_0^1 f(x)\mathrm{e}^{-2\pi\mathrm{i}kx}\mathrm{d}x\end{aligned}$$

due to $h - k \in \mathscr{D}(I)\backslash\{0\}$ and M are coprime. □

We summarize the results of Theorem 1 and Lemma 2 and figure out the following

Corollary 1 *Let $I \subset \mathbb{Z}^d$ be an arbitrary d-dimensional index set of finite cardinality and M be a prime number satisfying*

$$M \geq \max\left(|I_1|, \max_{s=2,\dots,d} \frac{|\{\mathbf{k} \in \mathscr{D}(I)_s : \mathbf{k} = (\mathbf{h}, h_s), \mathbf{h} \in \mathscr{D}(I)_{s-1} \backslash \{\mathbf{0}\} \text{ and } h_s \in \mathbb{Z}\backslash\{0\}\}|}{2} + 2\right).$$

In addition we assume that $M \nmid l$ for all $l \in \{k = \mathbf{e}_s \cdot \mathbf{h} : \mathbf{h} \in \mathscr{D}(I), s = 1, \dots, d\}\backslash\{0\}$, where $\mathbf{e}_s \in \mathbb{N}^d$ is a d-dimensional unit vector with $e_{s,j} = \begin{cases} 0, & \text{for } j \neq s \\ 1, & \text{for } j = s. \end{cases}$. Then there exists a rank-1 lattice of cardinality M that allows the reconstruction of all trigonometric polynomials with frequencies supported on I by sampling along the rank-1 lattice. Furthermore, once we determined a suitable M the proof of Theorem 1 verifies that we can find at least one appropriate generating vector component-by-component. Algorithm 1 indicates the corresponding strategy.

Algorithm 1 is already specified in [4, Algorithm 3] for hyperbolic crosses as frequency index set. Both algorithms do not differ. In contrast to [4], we simply allow arbitrary frequency index sets I as input, now. Only for reasons of clarity and comprehensibility, we stated the algorithm in this paper again.

Algorithm 2 Lattice size decreasing

Input:	$I \subset \mathbb{Z}^d$	frequency index set
	$M_{\max} \in \mathbb{N}$	cardinality of rank-1 lattice
	$\mathbf{z} \in \mathbb{N}^d$	$\Lambda(\mathbf{z}, M_{\max})$ is reconstructing rank-1 lattice for I

for $j = |I|, \dots, M_{\max}$ **do**
  if $|\{\mathbf{z} \cdot \mathbf{k} \bmod j : \mathbf{k} \in I\}| = |I|$ **then**
    $M_{\min} = j$
    break
  end if
end for

Output:	$M_{\min}$	reduced lattice size

Once one has discovered a reconstructing rank-1 lattice $\Lambda(\mathbf{z}, M)$ for the index set I, the condition

$$\mathbf{k} \cdot \mathbf{z} \neq \mathbf{h} \cdot \mathbf{z}, \quad \text{for all} \quad \mathbf{k} \neq \mathbf{h};\ \mathbf{k}, \mathbf{h} \in I,$$

holds and one can ask for $M' < M$ fulfilling

$$\mathbf{k} \cdot \mathbf{z} \not\equiv \mathbf{h} \cdot \mathbf{z} \,(\mathrm{mod}\, M'), \quad \text{for all} \quad \mathbf{k} \neq \mathbf{h};\ \mathbf{k}, \mathbf{h} \in I.$$

For a fixed frequency index set I and a fixed generating vector $\mathbf{z}$, we assume the rank-1 lattice $\Lambda(\mathbf{z}, M_{\max})$ being a reconstructing rank-1 lattice. Then, Algorithm 2 computes the smallest lattice size M' guaranteeing the reconstruction property of the rank-1 lattice $\Lambda(\mathbf{z}, M')$.

Finally, we give a simple upper bound on the cardinality of the difference set $\mathscr{D}(I)$ depending on the cardinality of I

$$|\mathscr{D}(I)| = |\{\mathbf{k} - \mathbf{h} \colon \mathbf{k}, \mathbf{h} \in I\}| = |\{\mathbf{k} - \mathbf{h} \colon \mathbf{k}, \mathbf{h} \in I, \mathbf{k} \neq \mathbf{h}\} \cup \{\mathbf{0}\}| \leq |I|(|I| - 1) + 1.$$

According to this and applying Bertrand's postulate, the prime number M from Corollary 1 is bounded from above by $|\mathscr{D}(I)|$ and $|I|^2$, provided that $|\mathscr{D}(I)| > 4$ and $|I| \geq 4$, respectively.

Remark 1 In [4] we considered frequency index sets I of the specific hyperbolic cross type. Since hyperbolic crosses are a subset of the d-dimensional box $[-(|I| - 1)/2, (|I| - 1)/2]^d \cap \mathbb{Z}^d$ and we necessarily have $|I| \leq M$, we obtain the difference set $\mathscr{D}(I)$ is contained in the box $[-M + 1, M - 1] \cap \mathbb{Z}^d$ and thus we know a priori that the prime M and the components of the elements of the difference set $\mathscr{D}(I)$ are coprime. The additional assumption of the coprimality of each number $l \in \{k = \mathbf{e}_s \cdot \mathbf{h} : \mathbf{h} \in \mathscr{D}(I), s = 1, \ldots, d\} \backslash \{0\}$ and the prime number M is essential in order to generalize the result of [4, Theorem 3.2] in Corollary 1.

4 Improvements

There are two serious problems concerning Corollary 1. In general, the computational costs of determining the cardinality of the difference sets $\mathscr{D}(I)_s$, $2 \leq s \leq d$, has a complexity of $\Omega(d|I|^2)$ and, maybe, the minimal M satisfying the assumptions of Corollary 1 is far away from a best possible reconstructing rank-1 lattice size. Accordingly, we are interested in somehow good estimations of the reconstructing rank-1 lattice size for the index set I.

In this section, we present another strategy to find reconstructing rank-1 lattices. We search for rank-1 lattices using a component-by-component construction determining the generating vectors $\mathbf{z} \in \mathbb{Z}^d$ *and* suitable rank-1 lattice sizes $M \in \mathbb{N}$.

Theorem 2 *Let $d \in \mathbb{N}$, $d \geq 2$, and $I \subset \mathbb{Z}^d$ of finite cardinality $|I| \geq 2$ be given. We assume that $\Lambda(\mathbf{z}, M)$ with $\mathbf{z} = (z_1, \ldots, z_{d-1})^\top$ is a reconstructing rank-1 lattice for the frequency index set $I_{d-1} := \{(h_s)_{s=1}^{d-1} : \mathbf{h} \in I\}$. Then the rank-1 lattice $\Lambda((z_1, \ldots, z_{d-1}, M)^\top, MS)$ with*

$$S := \min\{m \in \mathbb{N} : |\{h_d \bmod m : \mathbf{h} \in I\}| = |\{h_d : \mathbf{h} \in I\}|\}$$

is a reconstructing rank-1 lattice for I.

Proof We assume the rank-1 lattice $\Lambda((z_1, \ldots, z_{d-1})^\top, M)$ is a reconstructing rank-1 lattice for I_{d-1} and $\Lambda((z_1, \ldots, z_{d-1}, M)^\top, MS)$ is not a reconstructing rank-1 lattice for I, i.e., there exist at least two different elements $(\mathbf{h}, h_d), (\mathbf{k}, k_d) \in I$, $(\mathbf{h}, h_d) \neq (\mathbf{k}, k_d)$, such that

$$\mathbf{h} \cdot \mathbf{z} + h_d M \equiv \mathbf{k} \cdot \mathbf{z} + k_d M \pmod{MS}.$$

We distinguish three different possible cases of $(\mathbf{h}, h_d), (\mathbf{k}, k_d) \in I$, $(\mathbf{h}, h_d) \neq (\mathbf{k}, k_d)$:

- $\mathbf{h} = \mathbf{k}$ and $h_d \neq k_d$
  We consider the corresponding residue classes

$$0 \equiv \mathbf{k} \cdot \mathbf{z} + k_d M - \mathbf{h} \cdot \mathbf{z} - h_d M \equiv (k_d - h_d) M \,(\mathrm{mod}\, MS)$$

  and obtain $S|(k_d - h_d)$, i.e., $k_d \equiv h_d \,(\mathrm{mod}\, S)$. Thus, we determine the cardinality $|\{h_d \,\mathrm{mod}\, S : \mathbf{h} \in I\}| < |\{h_d : \mathbf{h} \in I\}|$, which is in contradiction to the definition of S.
- $\mathbf{h} \neq \mathbf{k}$ and $h_d = k_d$
  Accordingly, we calculate

$$0 \equiv \mathbf{k} \cdot \mathbf{z} + k_d M - \mathbf{h} \cdot \mathbf{z} - h_d M \equiv (\mathbf{k} - \mathbf{h}) \cdot \mathbf{z} \,(\mathrm{mod}\, MS)$$

  and obtain $MS|(\mathbf{k}-\mathbf{h}) \cdot \mathbf{z}$ and $M \mid (\mathbf{k}-\mathbf{h}) \cdot \mathbf{z}$ as well. According to that, we obtain $\mathbf{h} \cdot \mathbf{z} \equiv \mathbf{k} \cdot \mathbf{z} \,(\mathrm{mod}\, M)$, which is in contradiction to the assumption $\Lambda(\mathbf{z}, M)$ is a reconstructing rank-1 lattice for I_{d-1}.
- $\mathbf{h} \neq \mathbf{k}$ and $h_d \neq k_d$
  Due to $\Lambda(\mathbf{z}, M)$ is a reconstructing rank-1 lattice for I_{d-1} we have

Algorithm 3 Component-by-component lattice search (unknown lattice size M)

Input:	$I \subset \mathbb{Z}^d$	frequency index set
	$M_1 = \min\{m \in \mathbb{N} : \lvert\{k_1 \,\mathrm{mod}\, m : \mathbf{k} \in I\}\rvert = \lvert\{k_1 : \mathbf{k} \in I\}\rvert\}$	
	$z_1 = 1$	
	for $s = 2, \ldots, d$ **do**	
	$S = \min\{m \in \mathbb{N} : \lvert\{k_s \,\mathrm{mod}\, m : \mathbf{k} \in I\}\rvert = \lvert\{k_s : \mathbf{k} \in I\}\rvert\}$	
	$\mathbf{z} = (\mathbf{z}, z_s)$	
	$z_s = M_{s-1}$	
	form the set I_s as defined in (3)	
	search for $M_s = \min\{m \in \mathbb{N} : \lvert\{\mathbf{z} \cdot \mathbf{k} \,\mathrm{mod}\, m : \mathbf{k} \in I_s\}\rvert = \lvert I_s\rvert\} \leq S M_{s-1}$ using Algorithm 2	
	end for	
Output:	$\mathbf{z} \in \mathbb{N}^d$	generating vector
	$\mathbf{M} \in \mathbb{N}^d$	rank-1 lattice sizes for dimension $s = 1, \ldots, d$

$$0 \not\equiv \mathbf{k} \cdot \mathbf{z} - \mathbf{h} \cdot \mathbf{z} \,(\mathrm{mod}\, M).$$

Thus, we can find uniquely specified $a_{\mathbf{k},\mathbf{h}} \in \mathbb{Z}$ and $b_{\mathbf{k},\mathbf{h}} \in \{1, \ldots, M-1\}$ such that $\mathbf{k} \cdot \mathbf{z} - \mathbf{h} \cdot \mathbf{z} = a_{\mathbf{k},\mathbf{h}} M + b_{\mathbf{k},\mathbf{h}}$. We calculate

$$0 \equiv \mathbf{k} \cdot \mathbf{z} + k_d M - \mathbf{h} \cdot \mathbf{z} - h_d M \equiv (a_{\mathbf{k},\mathbf{h}} + k_d - h_d) M + b_{\mathbf{k},\mathbf{h}} \,(\mathrm{mod}\, MS)$$

Algorithm 4 Component-by-component lattice search (unknown lattice size M, improved)

Input: $I \subset \mathbb{Z}^d$ frequency index set

$M_1 = \min\{m \in \mathbb{N} : |\{k_1 \bmod m : \mathbf{k} \in I\}| = |\{k_1 : \mathbf{k} \in I\}|\}$
$z_1 = 1$
for $s = 2, \ldots, d$ **do**
  $S = \min\{m \in \mathbb{N} : |\{k_s \bmod m : \mathbf{k} \in I\}| = |\{k_s : \mathbf{k} \in I\}|\}$
  form the set I_s as defined in (3)
  search for the smallest $z_s \in [1, M_{s-1}] \cap \mathbb{Z}$ with $|\{(\mathbf{z}, z_s) \cdot \mathbf{k} \bmod SM_{s-1} : \mathbf{k} \in I_s\}| = |I_s|$
  $\mathbf{z} = (\mathbf{z}, z_s)$
  search for $M_s = \min\{m \in \mathbb{N} : |\{\mathbf{z} \cdot \mathbf{k} \bmod m : \mathbf{k} \in I_s\}| = |I_s|\}$ using Algorithm 2
end for

Output: $\mathbf{z} \in \mathbb{N}^d$ generating vector
$\mathbf{M} \in \mathbb{N}^d$ rank-1 lattice sizes for dimension $s = 1, \ldots, d$

and obtain $MS|(a_{\mathbf{k},\mathbf{h}} + k_d - h_d)M + b_{\mathbf{k},\mathbf{h}}$. As a consequence, we deduce $M \mid b_{\mathbf{k},\mathbf{h}}$, which is in conflict with $b_{\mathbf{k},\mathbf{h}} \in \{1, \ldots, M-1\}$.

Extending the reconstructing rank-1 lattice $\Lambda(\mathbf{z}, M)$ for I_{d-1} to $\Lambda((\mathbf{z}, M), MS)$ with S as defined above, we actually get a reconstructing rank-1 lattice for the frequency index set $I \subset \mathbb{Z}^d$. □

In addition to the strategy provided by Theorem 2 and the corresponding Algorithm 3, we bring the following heuristic into play. We assume small components of the vector $\mathbf{z}$ being better than large ones. Therefore, we tune Algorithm 3 and additionally search for the smallest possible component z_s fulfilling

$$|\{(\mathbf{z}, z_s) \cdot \mathbf{h} \bmod SM_{s-1} : \mathbf{h} \in I_s\}| = |I_s|.$$

Due to Theorem 2 the integer M_{s-1} is an upper bound for the minimal z_s we can find. Algorithm 4 indicates the described strategy in detail. Algorithms 3 and 4 provide deterministic strategies to find reconstructing rank-1 lattices for a given index set I. We would like to point out that in both algorithms the only input we need is the frequency index set I.

5 Numerical Examples

Our numerical examples treat frequency index sets of the type

$$I_{p,N}^d := \left\{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_p \leq N\right\},$$

where $\|\cdot\|_p$ is the usual p–(quasi–)norm

$$\|\mathbf{k}\|_p := \begin{cases} \left(\sum_{s=1}^{d} |k_s|^p\right)^{1/p} & \text{for} \quad 0 < p < \infty \\ \max_{s=1,\dots,d} |k_s| & \text{for} \quad p = \infty. \end{cases}$$

In particular, trigonometric polynomials with frequencies supported on the index sets $I^d_{p,N}$ are useful in order to approximate functions of periodic Sobolev spaces $H^{\alpha,p}(\mathbb{T}^d)$ of isotropic smoothness

$$H^{\alpha,p}(\mathbb{T}^d) := \{f : \mathbb{T}^d \to \mathbb{C} | \sum_{\mathbf{k}\in\mathbb{Z}^d} \max(1, \|\mathbf{k}\|_p)^\alpha |\hat{f}_\mathbf{k}|^2 < \infty\},$$

where $\alpha \in \mathbb{R}$ is the smoothness parameter. In Kühn et al. [5], detailed estimates of the approximation error for $p = 1, 2$ are given. Furthermore, tractability results are specified therein.

According to Kühn et al. [5], our examples deal with $p = 1$, $p = 2$, and, in addition, $p = 1/2$, $p = \infty$, see Fig. 1 for illustrations in dimension $d = 2$. We construct corresponding frequency index sets $I^d_{p,N}$ and apply Algorithms 1, 3, and 4 in order to determine reconstructing rank-1 lattices. We have to determine suitable rank-1 lattice sizes M for using Algorithm 1. For this, we compute the minimal prime number M_{Cor1} fulfilling Corollary 1. Since this computation is of high costs, we only apply Algorithm 1 to frequency index sets $I^d_{p,N}$ of cardinalities not larger than 20,000. We apply Algorithm 1 using the lattice size M_{Cor1} and the frequency index set $I^d_{p,N}$ as input. With the resulting generating vector, we apply Algorithm 2 in order to determine the reduced lattice size $M_{\text{Alg1+Alg2}}$. Additionally, we use Algorithms 3 and 4 computing rank-1 lattices $\Lambda(\mathbf{z}_{\text{Alg3}}, M_{\text{Alg3}})$ and $\Lambda(\mathbf{z}_{\text{Alg4}}, M_{\text{Alg4}})$, respectively. For reasons of clarity, we present only the rank-1 lattice sizes M_{Cor1}, $M_{\text{Alg1+Alg2}}$, M_{Alg3}, and M_{Alg4} but not the generating vectors $\mathbf{z} \in \mathbb{N}^d$ in our tables.

First, we interpret the results of Table 1. In most cases, the theoretical result of Corollary 1 gives a rank-1 lattice size M_{Cor1} which is much larger than the rank-1 lattice sizes found by applying the different strategies in practice. For $p = \infty$, all our algorithms determined a rank-1 lattice of best possible cardinalities, i.e., $|I^d_{\infty,N}| = M_{\text{Alg1+Alg2}} = M_{\text{Alg3}} = M_{\text{Alg4}}$. The outputs M_{Alg3} of Algorithm 3 are larger than those of Algorithm 1 in tandem with Algorithm 2 and Algorithm 4, with a few exceptions. Considering the nonconvex frequency index sets $I^d_{\frac{1}{2},N}$, Algorithm 3 brings substantially larger rank-1 lattice sizes M_{Alg3} than the two other approaches. Maybe, we observe the consequences of the missing flexibility in choosing the generating vector in Algorithm 3. Moreover, we observe the equality $M_{\text{Alg1+Alg2}} = M_{\text{Alg4}}$ in all our examples. We would like to point out that Algorithm 1 requires an input lattice size M, which we determined using Corollary 1. However, Algorithm 4 operates without this input.

Since our approach is applicable for frequency index sets with gaps, we also consider frequency index sets $I^{d,\text{even}}_{p,N} := I^d_{p,N} \cap (2\mathbb{Z})^d$. These frequency index sets are suitable in order to approximate functions which are even in each coordinate, i.e., the Fourier coefficients $\hat{f}_\mathbf{k}$ are a priori zero for $\mathbf{k} \in \mathbb{Z}^d \setminus (2\mathbb{Z})^d$, cf. Fig. 1.

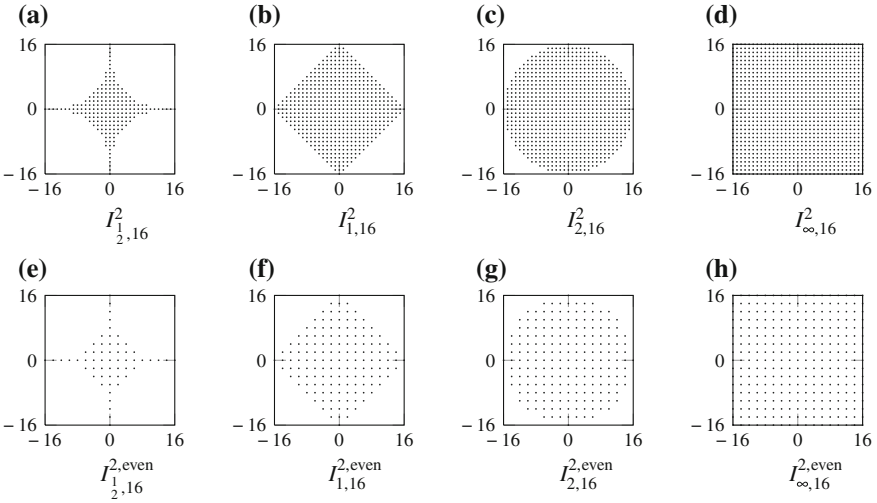


Fig. 1 Two-dimensional frequency index sets $I^2_{p,16}$ and $I^{2,\text{even}}_{p,16}$ for $p \in \{\frac{1}{2}, 1, 2, \infty\}$

Certainly, the gaps of the index sets $I^{d,\text{even}}_{p,N}$ are homogeneously distributed. We stress the fact, that the theoretical results and the algorithms can also be applied to strongly inhomogeneous frequency index sets.

Analyzing the frequency index sets $I^{d,\text{even}}_{p,N}$ in detail, we obtain

$$I^{d,\text{even}}_{p,N} = \{2\mathbf{k} \colon \mathbf{k} \in I^d_{p,N/2}\}.$$

We assume $\Lambda(\mathbf{z}, M)$ being a reconstructing rank-1 lattice for $I^d_{p,N/2}$. Accordingly, we know

$$\mathbf{k}_1 \cdot \mathbf{z} - \mathbf{k}_2 \cdot \mathbf{z} \not\equiv 0 \,(\text{mod}\, M)$$

for all $\mathbf{k}_1 \neq \mathbf{k}_2$, $\mathbf{k}_1, \mathbf{k}_2 \in I^d_{p,N/2}$. We determine $l_{\mathbf{k}_1,\mathbf{k}_2} \in \{1, \ldots, M-1\}$ and $t \in \mathbb{Z}$ such that

$$\mathbf{k}_1 \cdot \mathbf{z} - \mathbf{k}_2 \cdot \mathbf{z} = tM + l_{\mathbf{k}_1,\mathbf{k}_2}$$

and, furthermore,

$$2\mathbf{k}_1 \cdot \mathbf{z} - 2\mathbf{k}_2 \cdot \mathbf{z} = t2M + 2l_{\mathbf{k}_1,\mathbf{k}_2}.$$

This yields

$$2\mathbf{k}_1 \cdot \mathbf{z} - 2\mathbf{k}_2 \cdot \mathbf{z} \equiv 2l_{\mathbf{k}_1,\mathbf{k}_2} \,(\text{mod}\, M), \tag{5}$$

Table 1 Cardinalities of reconstructing rank-1 lattices of index sets $I^d_{p,N}$ found by applying Corollary 1, Algorithm 1 and 2, Algorithm 3, and Algorithm 4

p	N	d	$\lvert I^d_{p,N}\rvert$	M_{Cor1}	$M_{\text{Alg1+Alg2}}$	M_{Alg3}	M_{Alg4}
$\frac{1}{2}$	8	10	1,241	51,679	5,895	16,747	5,895
$\frac{1}{2}$	8	20	4,881	469,841	36,927	172,642	36,927
$\frac{1}{2}$	8	30	10,921	1,654,397	128,370	804,523	128,370
$\frac{1}{2}$	16	5	2,561	122,509	16,680	23,873	16,680
$\frac{1}{2}$	16	10	21,921	–	–	910,271	403,799
$\frac{1}{2}$	16	15	83,081	–	–	9,492,633	3,495,885
$\frac{1}{2}$	32	3	3,529	51,169	17,280	15,529	17,280
$\frac{1}{2}$	32	6	63,577	–	–	1,932,277	1,431,875
$\frac{1}{2}$	64	3	24,993	–	–	113,870	99,758
1	2	10	221	1,361	369	399	369
1	2	20	841	10,723	1,935	2,641	1,935
1	2	30	1,861	36,083	5,664	8,213	5,664
1	4	5	681	4,721	1,175	1,225	1,175
1	4	10	8,361	329,027	36,315	41,649	36,315
1	4	15	39,041	–	–	400,143	340,247
1	8	3	833	2,729	1,113	1,169	1,113
1	8	6	40,081	–	–	126,863	126,738
1	16	3	6,017	21,839	8,497	8,737	8,497
2	2	5	221	1,373	356	353	356
2	2	10	4,541	203,873	21,684	20,013	21,684
2	2	15	25,961	3,865079	259,517	280,795	259,571
2	2	20	87,481	–	–	1,634,299	1,481164
2	4	3	257	809	346	377	346
2	4	6	23,793	496,789	69,065	72,776	69,065
2	8	3	2,109	7,639	2,893	3,050	2,893
2	16	3	17,077	65,309	23,210	23,889	23,210
∞	1	3	27	53	27	27	27
∞	1	6	729	6,257	729	729	729
∞	1	9	19,683	781,271	19,683	19,683	19,683
∞	2	3	125	331	125	125	125
∞	2	6	15,625	236,207	15,625	15,625	15,625

where $2l_{\mathbf{k}_1,\mathbf{k}_2} \in \{2, 4, \ldots, 2M-2\}$. Assuming M being odd, we obtain $2l_{\mathbf{k}_1,\mathbf{k}_2} \not\equiv 0 \pmod{M}$ for all $\mathbf{k}_1 \neq \mathbf{k}_2$, $\mathbf{k}_1, \mathbf{k}_2 \in I^d_{p,N/2}$ and $\Lambda(\mathbf{z}, M)$ is a reconstructing rank-1 lattice for $I^{d,\text{even}}_{p,N}$.

Table 2 Cardinalities of reconstructing rank-1 lattices of index sets $I_{p,N}^{d,\text{even}}$ found by applying Corollary 1, Algorithm 1 and 2, Algorithm 3, and Algorithm 4

p	N	d	$\lvert I_{p,N}^{d,\text{even}}\rvert$	M_{Cor1}	$M_{\text{Alg1+Alg2}}$	M_{Alg3}	M_{Alg4}
$\frac{1}{2}$	16	10	1,241	51,679	5,895	15,345	5,895
$\frac{1}{2}$	16	20	4,881	469,841	36,927	176,225	36,927
$\frac{1}{2}$	16	30	10,921	1,654,397	129,013	763,351	129,013
$\frac{1}{2}$	32	5	2,561	122,509	17,825	23,873	17,825
$\frac{1}{2}$	32	10	21,921	–	–	992,097	403,799
$\frac{1}{2}$	32	15	83,081	–	–	8,848,095	3,495,885
$\frac{1}{2}$	64	3	3,529	51,169	17,689	15,529	17,689
$\frac{1}{2}$	64	6	63,577	–	–	1,932,277	1,431,875
$\frac{1}{2}$	128	3	24,993	–	–	119,159	105,621
1	4	10	221	1,361	369	399	369
1	4	20	841	10,723	1,935	2,641	1,935
1	4	30	1,861	36,083	5,711	8,213	5,711
1	8	5	681	4,721	1,175	1,225	1,175
1	8	10	8,361	329,027	36,315	41,649	36,315
1	8	15	39,041	–	–	400,143	340,247
1	16	3	833	2,729	1,113	1,169	1,113
1	16	6	40,081	–	–	126,863	126,875
1	32	3	6,017	21,839	8,497	8,737	8,497
2	4	5	221	1,373	361	353	361
2	4	10	4541	203,873	22,525	20,013	22,525
2	4	15	25,961	–	–	280,795	259,571
2	4	20	87,481	–	–	1,634,299	1,497,403
2	8	3	257	809	347	13,309	347
2	8	6	23,793	–	–	72,777	69,065
2	16	3	2,109	7,639	2,893	3,063	2,893
2	32	3	17,077	65,309	23,243	23,915	23,243
∞	2	3	27	53	27	27	27
∞	2	6	729	6,257	729	729	729
∞	2	9	19,683	781,271	19,683	19,683	19,683
∞	4	3	125	331	125	125	125
∞	4	6	15,625	236,207	15,625	15,625	15,625

In Table 2 we present the reconstructing rank-1 lattice sizes we found for even frequency index sets. Comparing the two tables, we observe the same odd lattice sizes $M_{\text{Alg1+Alg2}}$ and M_{Alg4} for $I_{p,N/2}^{d}$ and $I_{p,N}^{d,\text{even}}$. In fact the corresponding generating

vectors are also the same. In the case we found even reconstructing lattice sizes for $I^d_{p,N/2}$, we constructed some slightly larger reconstructing rank-1 lattice sizes for $I^{d,\mathrm{even}}_{p,N}$. In these cases, we cannot use the found reconstructing rank-1 lattices for $I^d_{p,N/2}$ in order to reconstruct trigonometric polynomials with frequencies supported on $I^{d,\mathrm{even}}_{p,N}$. The statement in (5) shows the reason for this observation. There exists at least one pair $\mathbf{k}_1, \mathbf{k}_2 \in I^d_{p,N}, \mathbf{k}_1 \neq \mathbf{k}_2$ with $\mathbf{k}_1 \cdot \mathbf{z} - \mathbf{k}_2 \cdot \mathbf{z} \equiv \frac{M}{2} \pmod M$. Consequently, doubling $\mathbf{k}_1$ and $\mathbf{k}_2$ leads to $2\mathbf{k}_1 \cdot \mathbf{z} - 2\mathbf{k}_2 \cdot \mathbf{z} \equiv 0 \pmod M$ and, hence, $\Lambda(\mathbf{z}, M)$ is not a reconstructing rank-1 lattice for $I^{d,\mathrm{even}}_{p,N}$.

The fastest way for determining reconstructing rank-1 lattices is to apply Algorithm 1 with a small and suitable rank-1 lattice size M. As mentioned above, the biggest challenge is to determine this small and suitable rank-1 lattice size M. Consequently, estimating relatively small M using some a priori knowledge about the structure of the frequency index set I or some empirical knowledge, leads to the fastest way to reasonable reconstructing rank-1 lattices. We stress the fact, that this strategy fails if there exists no generating vector $\mathbf{z}$ which can be found using Algorithm 1 with input I and M. For that reason, we presented Algorithms 3 and 4. Both algorithms determine reconstructing rank-1 lattices of reasonable cardinalities M using the frequency index set I as the only input. In particular, the rank-1 lattice sizes M_{Alg4} determined by Algorithm 4 are the same as the rank-1 lattice sizes $M_{\mathrm{Alg1+Alg2}}$ in our examples.

All presented deterministic approaches use Algorithm 2. The computational complexity of Algorithm 2 is bounded by $\mathcal{O}((M_{\max} - |I|)|I|)$. However, some heuristic strategies can decrease the number of loop passes. The disadvantage of this strategy is that one does not find $M_{\min}$ but, maybe, an M with $M_{\min} \leq M \ll M_{\max}$. We do not prefer only one of the presented algorithms because the computational complexity mainly depends on the structure of the specific frequency index set and the specific algorithm which is used. In detail, the number of function calls of Algorithm 2 in connection with the order of magnitude of the corresponding inputs, i.e., the cardinality $|I_s|$ of the frequency index sets I_s , and outputs $M_{\min}$ essentially cause the computational costs of all presented algorithms.

6 Summary

Based on Theorem 1, we determined a lattice size M_{Cor1} guaranteeing the existence of a reconstructing rank-1 lattice for a given arbitrary frequency index set I in Corollary 1. In order to prove this result, we used a component-by-component argument, which leads directly to the component-by-component algorithm given by Algorithm 1, that computes a generating vector $\mathbf{z}$ such that $\Lambda(\mathbf{z}, M)$ is a reconstructing rank-1 lattice for the frequency index set I. Due to difficulties in determining M_{Cor1}, we developed some other strategies in order to compute reconstructing rank-1 lattices. The corresponding Algorithms 3 and 4 are also component-by-component

algorithms. These algorithms compute complete reconstructing rank-1 lattices, i.e., generating vectors $\mathbf{z} \in \mathbb{N}^d$ and lattice sizes $M \in \mathbb{N}$, for a given frequency index set I. All the mentioned approaches are applicable for arbitrary frequency index sets of finite cardinality.

Acknowledgments The author thanks the referees for their careful reading and their valuable suggestions for improvements.

References

1. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
2. Cooley, J.W., Tukey, J.W.: An algorithm for machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965)
3. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing lattice rules based on weighted degree of exactness and worst case error. Computing **87**, 63–89 (2010)
4. Kämmerer, L.: Reconstructing hyperbolic cross trigonometric polynomials by sampling along rank-1 lattices. SIAM J. Numer. Anal. **51**, 2773–2796 (2013). http://dx.doi.org/10.1137/120871183
5. Kühn, T., Sickel, W., Ullrich, T.: Approximation numbers of Sobolev embeddings—sharp constants and tractability. J. Complex. **30**, 95–116 (2013). doi:10.1016/j.jco.2013.07.001
6. Li, D., Hickernell, F.J.: Trigonometric spectral collocation methods on lattices. In: Cheng, S.Y., Shu, C.-W., Tang T. (eds.) Recent Advances in Scientific Computing and Partial Differential Equations, AMS Series in Contemporary Mathematics, vol. 330, pp. 121–132. American Mathematical Society, Providence, RI (2003)
7. Munthe-Kaas, H., Sørevik, T.: Multidimensional pseudo-spectral methods on lattice grids. Appl. Numer. Math. **62**, 155–165 (2012). doi:10.1016/j.apnum.2011.11.002
8. Sickel, W., Ullrich, T.: The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness. East J. Approx. **13**, 387–425 (2007)
9. Sloan, I.H., Joe, S.: Lattice Methods for Multiple Integration. Oxford Science Publications, The Clarendon Press Oxford University Press, New York (1994)

On Nondegenerate Rational Approximation

L. Franklin Kemp

Abstract The total degree algorithm of [3] failed to converge for some delicate approximations because of a wrong Remes one point exchange or an inaccurate eigenvalue (i.e., reference equioscillation error) unless started near an optimal reference. Here we present a modified algorithm that is robust and prove when the revised algorithm's total degree rational ℓ_∞ approximation is optimal for **all** lesser degrees. Detailed examples and their figures show the bounded eigenvalue computations, steps of the Remes one point exchange for reference searching, and degeneracy pyramids of the revised robust algorithm.

Keywords ℓ_∞ Rational approximation · Equioscillation error · Eigenvalue bounds

1 Introduction

Kemp [3] combines orthogonal polynomials, persymmetry, bounded symmetric eigenvalue problems, eigenvalue interlacing, inverse iteration, Rayleigh's quotient, Sturm's root test, and Remes exchange to produce a total degree algorithm that searches for the best discrete nondegenerate rational ℓ_∞ approximation in the $p+q+1$ sets of rational functions of total degree $p+q$: $R(p+q,0)$, $R(p+q-1,1)$, $R(p+q-2,2)$, $\ldots$, $R(1,p+q-1)$, $R(0,p+q)$. It bypasses degenerate approximations (i.e., those with minimax errors equioscillating on references of less than $p+q+2$ points) and approximations with no minimax error thereby avoiding extra searching. The possibility arises that a bypassed degenerate might be better than what the total degree algorithm finds among nondegenerates even though it seems that degenerates should have larger minimax errors than nondegenerates since their minimax errors equioscillate on fewer points (i.e., are less constrained) than nonde-

L. F. Kemp (✉)
Collin College, 2800 E Spring Creek Pkwy, Plano, TX 75074, USA
e-mail: lfkemp@collin.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_15,

generates. Moreover, all [3] test cases never found such a degenerate. In the absence of a counterexample or a proof of impossibility, the best we can offer is a proof of impossibility when the total degree algorithm terminates under the condition that each rational set has a pole free reference over its $p+q+2$ points with equioscillating error that is greater than or equal to the total degree approximation minimax error. Under this termination condition the total degree rational ℓ_∞ approximation is not only the best out of the $p+q+1$ rational sets, but also best over all rational sets of lesser degree. We describe how we compute eigenvalues accurately and list the steps of the Remes one point exchange method for deeper search of references. The detailed examples help to explain the descriptions.

2 Summary of Minimax (l_∞) Rational Approximation

Let X be a finite point set, $I(X)$ its smallest containing interval, and $R(p,q)$ the set of rational functions of form:

$$P(x)/Q(x) = \frac{a_0 + a_1 x + \cdots + a_p x^p}{b_0 + b_1 x + \cdots + b_q x^q}$$

Chebyshev Theorem: $P(x)/Q(x) \in R(p,q)$ is the minimax approximation of $y(x)$ on X iff $\exists$ a **reference** $X_{ref} = \{x_0 < x_1 < \cdots < x_{n_{ref}-1}\} \subseteq X$ such that

$$\mathbf{y} - \mathbf{P}/\mathbf{Q} = \mathbf{s}\lambda, \quad \text{on } X_{ref} \text{ and } 0 \notin Q(I(X))$$

where

$$\begin{aligned}
|\lambda| &= \max_X |y - P/Q|, && \text{(minimax error, (eigenvalue))}\\
s(x_i) &= (-1)^{n_{ref}-i},\ 0 \le i \le n_{ref}-1, && \text{(alternating sign function on } X_{ref})\\
d &= \min(p - \deg P, q - \deg Q), && \text{(\textbf{degeneracy} of } y \text{ for } R(p,q) \text{ if } d>0)\\
n_{ref} &= p+1+q+1-d, && \text{(no. of minimax error alternations)}
\end{aligned}$$

For any d multiply the above dubbed **equioscillating error equation** by Q to get

$$\mathbf{Q}(\mathbf{y} - \mathbf{s}\lambda) - \mathbf{P} = \mathbf{0} \quad \text{on} \quad X_{ref},$$

a linear system of n_{ref} equations (also dubbed the equioscillating error equation) in n_{ref} unknown coefficients of P and Q with parameter λ. It has a nontrivial solution if the determinant of its coefficient matrix, a polynomial of degree $q+1$ in λ, is zero. This polynomial has $q+1$ real roots; therefore, the Chebyshev theorem shows the problem is a finite maximization problem: solves the equioscillating equation for every $X_{ref} \subseteq X$, discard those for which $0 \in Q(I(X))$, and select from the remaining references one with largest error.

3 Degeneracy

$d=3$			$R(0,0)$				
$d=2$		$R(2,0)$	$R(1,1)$	$R(0,2)$			
$d=1$	$R(4,0)$	$R(3,1)$	$R(2,2)$	$R(1,3)$	$R(0,4)$		
$d=0$	$R(6,0)$	$R(5,1)$	$R(4,2)$	$R(3,3)$	$R(2,4)$	$R(1,5)$	$R(0,6)$

Pyramid of Degeneracy Sets for Total Degree $p+q=6$.

Theorem 1 (Degeneracy Theorem) *If P/Q is the minimax approximation of $y(x)$ out of $R(p,q)$ with degeneracy d, then $P/Q \in R(p-d, q-d)$, but $P/Q \notin R(p-d-1, q-d-1)$.*

Proof If $d = p - \deg(P) \le q - \deg(Q)$, then $\deg(P) = p-d$ and $\deg(Q) \le q-d$; hence, $P/Q \in R(p-d, q-d)$, but $P/Q \notin R(p-d-1, q-d-1)$ since $\deg(P) = p-d > p-d-1$. The same argument applies if $d = q - \deg(Q) \le p - \deg(P)$. □

Corollary 1 *y cannot be degenerate for $R(p,0)$ or $R(0,q)$.*

Proof Approximation in $R(p,0)$ is polynomial so y has a minimax P. y cannot be degenerate for $R(0,q)$, $q > 0$ so y either has a minimax $1/Q$ or it does not. □

Corollary 2 *If $p \ge q$, then y for $R(p,q)$ has a minimax P/Q for some degeneracy $d = 0, 1, \ldots, q$.*

Corollary 3 *If $p < q$, then y for $R(p,q)$ has a minimax P/Q for some $d = 0, 1, \ldots, p$ or it does not.*

Corollary 4 *If $R(p_1, q_1) \subset R(p_2, q_2)$ have minimax errors λ_1 and λ_2, respectively, then $\lambda_1 \ge \lambda_2$.*

Proof Minimax P/Q coefficients for $R(p_1, q_1)$ are attainable in $R(p_2, q_2)$. □

4 Total Degree Rational Approximation

Total degree rational approximation means $d = 0$ so it is *nondegenerate* approximation. It seeks the best minimax P/Q on full references of $p+q+2$ points out of up to $p+q+1$ *existing* minimax's from $R(p+q, 0)$, $R(p+q-1, 1)$, $R(p+q-2, 2), \ldots, R(0, p+q)$ sets. It is effective because existence is guaranteed, degeneracy is eliminated so degenerate references are not calculated, all approximations share the same reference, each equioscillating error equation involves the same set of orthogonal polynomials up to degree $p+q$, all equioscillating errors (λ's) have bounds, it visits more references, it may yield the minimax P/Q $\forall \deg P + \deg Q \le p+q$, and for free, it can save references with least ℓ_1 and ℓ_2 norms that it visits [4].

Remark 1 There may be P/Q's with lower error but they are not minimax. For example, the total degree 1 algorithm applied to {(0,1),(1,0),(2,0)} yields minimax $P/Q = (3 - 2x)/4 \in R(1, 0)$ with minimax error .25 while approximations like $1/(1 + 1000x) \in R(0, 1)$ have maximum errors approaching zero, but none equal to zero. There is no minimax $P/Q \in R(0, 1)$. The total degree 1 result is $(3 - 2x)/4$.

Definition 1 X_{ref} is a **pole free** reference if $P/Q \in R(p, q)$ equioscillates on X_{ref} with no root of Q in $I(X_{ref})$.

Lemma 1 *A pole free reference error is a lower bound on the minimax error.*

Proof If X_{ref} is a pole free reference, then its P/Q coefficients only depend on the values from X_{ref} and $y(X_{ref})$ so its error is independent of whether or not Q has a root in $I(X)$. In other words, the error on a pole free reference is the same for $0 \in Q(I(X))$ or $0 \notin Q(I(X))$; hence, every pole free reference is a contender for the minimax error pole free reference when it exists. If all pole free references have $0 \in Q(I(X))$, then there is no pole free reference for a minimax error; hence, no minimax error. □

Theorem 2 (Total Degree Theorem) *If each set in the pyramid base exhibits a pole free reference i.e.,* $0 \notin Q(I(X_{ref}))$*, with equioscillating error greater than that of the set of least minimax error in the base, then the* P/Q *of least minimax error in the base has the least minimax error for* ***all*** *sets of lesser total degree.*

Proof The degeneracy theorem and its corollaries show that all minimax errors (that exist) in sets above the base which includes those interleaved between the pyramid rows are greater than or equal to those in the base. The claim follows from the lemma since any pole free reference error that exceeds another rational's minimax error, its own minimax error also exceeds it. □

Theorem 3 (Convergence Theorem) *The total degree Remes one point exchange algorithm converges to a minimax rational approximation of total degree* $p + q$ *in one of the sets* $R(p + q, 0), R(p + q - 1, 1), \ldots, R(0, p + q)$.

Proof Polynomial approximation (i.e., for $R(p+q, 0)$), always exists and the Remes one point exchange ascent will converge to its minimax error regardless of starting reference. □

Remark 2 If other sets have close enough pole free references to their minimax pole free reference, then the Remes one point exchange method for reference searching will find their unique minimax error.

5 Revised Total Degree Algorithm

5.1 Eigenvalue/Eigenvector Computations

See the example eigenvalue triangles below for understanding.

1. Eigenvalues are computed in ascending order down columns right to left.
2. Eigenvalues at equal bounds are computed first without inverse iteration.
3. One guess suffices for inverse iteration to produce an eigenvalue.
4. Eigenvalues are swapped and sorted to maintain ascent down a column.

5.2 Remes One Point Exchange

In the following steps "ref" means "pole free reference."

1. Start with a ref X_{ref}.
2. Compute X_{ref}'s equioscillating error for each total degree rational.
3. Pick the rational ref X_{ref} with smallest equioscillating error greater than the previous smallest equioscillating error.
4. For this rational find two points in $X - X_{ref}$: One with the greatest error and another with the least error exceeding the equioscillating error.
5. Exchange each point of X_{ref} with the first point until a new ref appears with greater equioscillating error; then do the same for the second point to get a new ref X_{ref} updating all the other rationals with any exchanged refs that increase their equioscillating errors (unflagging a flagged rational).
6. Remove a minimax rational.
7. Flag the rational in Step 5 if its equioscillating error did not increase.
8. Stop if only minimax and flagged rationals remain; else go to Step 3.

6 Examples

6.1 All Total Degree 2 Equioscillating Errors for $|x|$

Figures 1 and 2 show 30 graphs of all total degree 2 rational approximations of $|x|$ on five references of four points in $X = \{-1, -.5, 0, .5, 1\}$ that are the solutions to the 30 equioscillating (EO) error equations. Each reference has 1 approximation with $p = 2, q = 0$, 2 approximations with $p = 1, q = 1$, and 3 approximations with $p = 0, q = 2$ for a total of six; hence, 30 for five references.

From Figs. 1 and 2 it is easy to find the optimum in this small problem. In general the number of approximations of total degree m is ${}_nC_{m+2}(1 + 2 + \cdots + m + m + 1)$

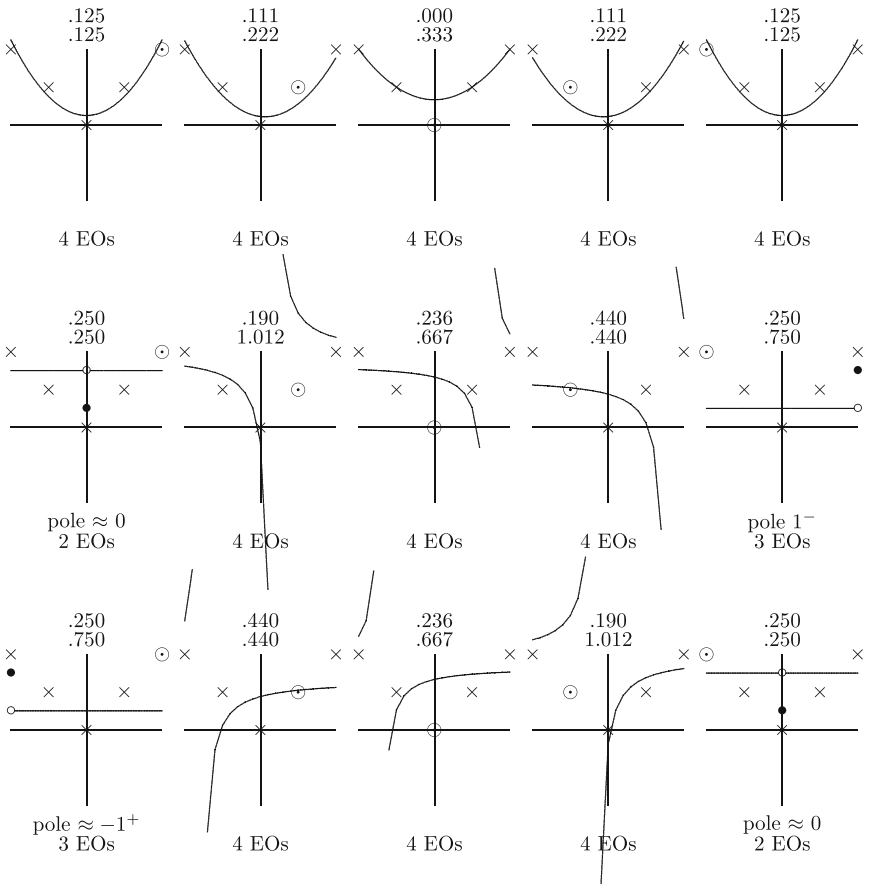


Fig. 1 $R(2, 0)$ and $R(1, 1)$ equioscillating and maximum errors on $X = \{-1. - .5, 0, .5, 1\}$

where n is the number of points of X. For the example $y = y_{12}$ below with $n = 21$ and $m = 6$, the number of approximations is 5697720 which justifies an algorithm for finding the optimum in lieu of exhaustive search. Although approximations of lower total degree are not examined by the total degree algorithm, the optimum over all such will be in one of the rational function sets of total degree when the termination condition of the total degree theorem is satisfied.

Because $R(1, 1)$ has a pole in X for each reference, it does not have a minimax error. By Corollary 2 it has degeneracy $d = 1$ with minimax error .5 in $R(0, 0)$, the midpoint of the greatest and least y values over X.

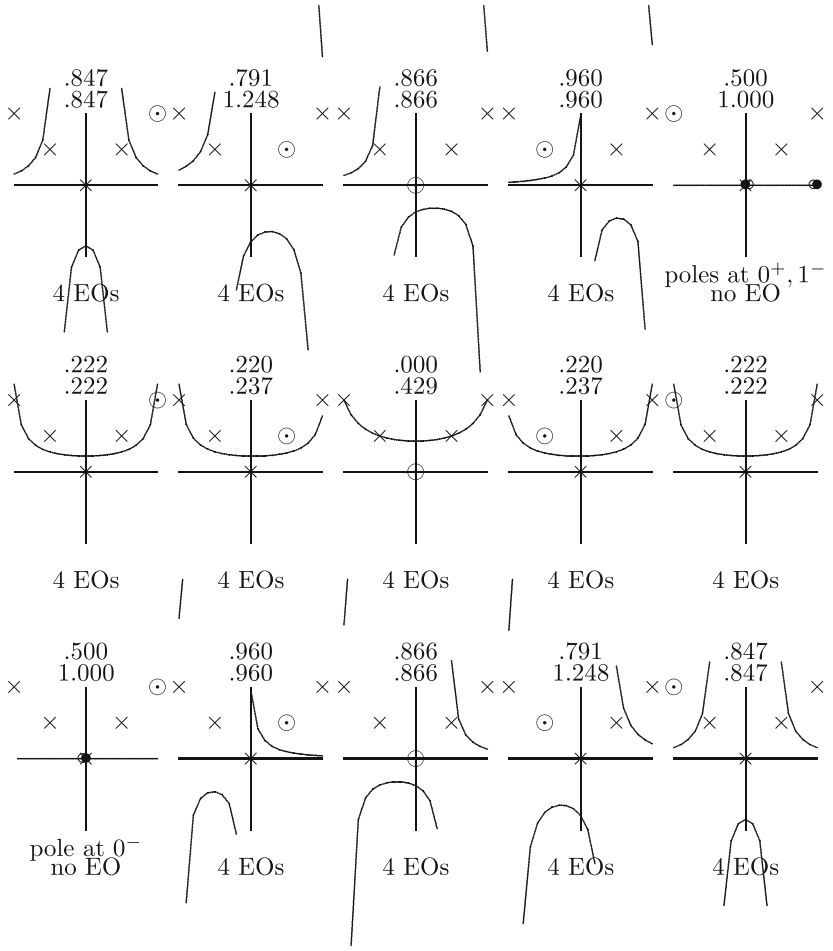


Fig. 2 $R(0, 2)$ equioscillating (EO) and maximum errors on $X = \{-1. - .5, 0, .5, 1\}$

6.2 Steps of the Total Degree 2 Algorithm for $|x|$

The initial eigenvalue (equioscillation error) bounds that appear outside the triangles in Step 1 and Step 2 are the ordered values of $\{s(x_i)y(x_i)|x_i \in X_{\text{ref}}, 0 \le i \le nref - 1\}$ first proved in Kemp [2]. Initial bounds for $X_{ref1} = \{-1. - .5, 0, 1\}$ and $X_{\text{ref2}} = \{-.5, 0, .5, 1\}$ are $\{1. - .5, 0, -1\}$ and $\{.5, 0, .5, -1\}$, respectively. Steps 1 and 2 find $R(0, 2)$'s errors first using initial bounds, then $R(1, 1)$'s errors second using $R(0, 2)$'s errors for bounds. $R(2, 0)$'s error is a direct calculation but bounded by $R(1, 1)$'s errors.

Step 1

$R(2,0)$	$R(1,1)$	$R(0,2)$	
.111	0	**.220**	= λ=equioscillating error on X_{ref}, $0 \notin Q(I(X_{ref}))$
3	0	3	= exchange point of max or min error $\geq \lambda$ on X
.222	.5	**.237**	= max or min error $\geq \lambda$ on X
.222	.5	**.237**	= least max or min error $\geq \lambda$ on X yet met (if -, then
0	0	0	no λ ascent from X_{ref})
1	0	1	
2	0	2	=X_{ref}'s nref indices
4	0	4	
1	0	5	=index of λ in eigenvalue triangle (if -, then $0 \in Q(I(X))$)

-1	-1	-1	-1
-.111	-.440	-.960	-1
1	.190	**-.220**	-.5
0	1	.791	0
0	0	1	1

Step 2

$R(2,0)$	$R(1,1)$	$R(0,2)$
.125	0	**.222**
2	0	2
.125	.5	**.222**
.125	.5	**.222**
1	0	1
2	0	2
3	0	3
4	0	4
1	0	5

Eigenvalues (equioscillating errors) are inside the triangle. Initial eigenvalue bounds are outside. Right values bound left values.

$R(2,0)$	$R(1,1)$	$R(0,2)$	
-1	-1	-1	-1
.125	-.25	-.847	-1
.5	.25	**.222**	0
0	.5	.5	.5
0	0	.5	.5

The total degree theorem fails because $R(1,1)$ does not have a pole free reference with error greater than .125 at Step 2. Nevertheless, .125 is the minimax error for all degrees less than or equal to 2 because $|x|$ has degeneracy $d = 1$ for $R(1,1)$ and minimax .5 since, all 5 references are visited by the total degree algorithm. If X had more than 5 points and the result of the algorithm were the same as that at Step 2, then it is uncertain that $|x|$ is degenerate.

However, if a point, say $(1.5, \pm.0001)$, is added to $\{X, y(X)\}$ to get a $y^+(x) = |x|$ on X which nearly equioscillates four times with error $\approx .5$ on $X \cup \{1.5\}$, then the total degree algorithm produces a bounded rational function in $R(1,1)$ for $y^+(x)$ with minimax error $\approx .5$ on $I(X \cup \{1.5\})$. This rational function restricted to $I(X)$ has error $\approx .5$ greater than the minimax errors associated with $R(2,0)$ and $R(0,2)$.

Moreover, it can be adjusted to come as close to the minimax error of .5 as desired but not with 4 equioscillations on X. There is no such rational function in $R(1, 1)$ with 4 equioscillations on X; therefore, $|x|$ for $R(1, 1)$ has degeneracy $d = 1$.

$R(0,0)$
.7143

$R(2,0)$	$R(1,1)$	$R(0,2)$
.5036	.5	.5
0	0	0
.5036	.5	.5
.5036	.5	.5
0	5	5
5	10	10
10	15	15
15	20	20
1	2	5

$R(1,0)$	$R(0,1)$
.5375	.5
5	5
.5375	.5
.5375	.5
0	10
5	15
20	20
1	2

$R(4,0)$	$R(3,1)$	$R(2,2)$	$R(1,3)$	$R(0,4)$
.2685	.4681	.4657	.4483	.3692
0	20	0	20	1
.2685	.7114	.7002	.6654	.3692
.2685	−.7114	−.6904	−.6654	.3692
0	0	1	0	0
2	1	5	1	1
5	5	10	5	5
10	10	15	10	10
15	15	19	15	15
18	19	20	19	20
1	3	5	9	13

$R(3,0)$	$R(2,1)$	$R(1,2)$	$R(0,3)$
.5009	.5	.5	.5
0	10	10	10
.5009	.5	.5	.5
.5009	.5	.5	.5
0	0	0	0
5	5	5	5
10	10	10	10
15	15	15	15
20	20	20	20
1	3	4	8

$R(6,0)$	$R(5,1)$	$R(4,2)$	$R(3,3)$	$R(2,4)$	$R(1,5)$	$R(0,6)$
.1046	.2085	.1019	.2081	.2360	.2841	.2022
15	20	0	20	19	19	9
.1046	.9467	.1019	.8013	.2360	.3511	.2022
.1046	−.6268	.1019	−.6205	.2360	−.3511	.2022
0	0	0	0	0	0	0
1	1	3	2	2	1	1
3	3	5	5	5	2	3
5	5	7	8	9	5	9
10	10	9	10	11	10	12
15	15	13	12	12	15	15
17	17	15	15	15	18	18
19	19	17	17	19	20	20
1	2	5	9	13	18	25

$R(5,0)$	$R(4,1)$	$R(3,2)$	$R(2,3)$	$R(1,4)$	$R(0,5)$
.2674	.2675	.2759	.2739	.3425	.3283
2	2	5	5	1	1
.2674	.2675	.6429	.6406	.3425	.3283
.2674	.2675	−.6429	−.6406	.3425	.3283
0	0	1	1	0	0
2	2	7	7	1	1
5	5	10	10	5	5
10	10	13	13	10	10
15	15	15	15	15	15
18	18	16	16	19	19
20	20	20	20	20	20
1	2	5	8	12	18

6.3 Total Degree Approximations for $y = y_{12}$ from Kemp [3]

Kemp [3] uses $y_{12}(x)$ from Kaufman [1] which notes that $y_{12}(x)$ on $X = \{0, .05, .1, \ldots, 95., 1\}$ is designed to give degenerate approximations because on the five points with indices {0 5 10 15 20}, $y_{12}(x) - 1/1 + x$ equioscillates with minimax error .5.

The starting references for each total degree rational approximation from 1 to 6 use uniformly distributed indices: {0 10 20}, {0 7 13 20}, {0 5 10 15 20}, {0 4 8 12 16 20}, {0 3 7 10 13 17 20}, and {0 3 6 9 11 14 17 10}, respectively. The total degree theorem's pole free reference error condition is satisfied in each case so the total degree minimax errors are also the best for all lesser degrees than the total degree. They are .7143, .5, .5, .5, .2685, .2674, and .1019, for total degrees from 0 to 6, respectively.

From the two pyramid of degeneracy sets, there are apparently seven $d = 1$ degeneracies: $R(5, 1) \to R(4, 0)$, $R(3, 1) \to R(2, 0)$, $R(2, 2) \to R(1, 1)$, $R(1, 3) \to R(0, 2)$, $R(1, 5) \to R(0, 4)$, $R(3, 2) \to R(2, 1)$, and $R(2, 3) \to R(1, 2)$, and one $d = 2$ degeneracy $R(3, 3) \to R(2, 2) \to R(1, 1)$. In fact, these are true degeneracies.

For example, if three equally spaced points are added to X for $R(3, 3)$ with ordinates $1/1 + x \mp .5 + .0001$ or $1/1 + x \mp .5 - .0001$, then the total degree algorithm produces a bounded rational function in $R(3, 3)$, with minimax error $\approx .5$ on the larger set with eight equioscillations. Note that only three points are needed because $y_{12}(x)$ actually equioscillates five times for $R(1, 1)$. Thus $y_{12}(x)$ has degeneracy $d = 2$ for $R(3, 3)$. To check degeneracy $d = 1$ for $y_{12}(x)$ in $R(2, 2)$, add just one point to obtain the required six equioscillations. Make similar calculations for the other degeneracy $d = 1$ cases. In the case of $R(5, 1)$, the minimax rational function in $R(4, 0)$ with error .2685 serves as the basis curve around which $y_{12}(x)$ equioscillates, not $1/1 + x$. Add two points to X to check degeneracy for $y_{12}(x)$ approximated by $R(5, 1)$.

The upshot is that the total degree algorithm output only hints at degeneracy. When the total degree algorithm fails to find a minimax for a rational set, one can check for degeneracy by finding a minimax of lower degree. With this minimax P/Q, add enough points to X so that P/Q equioscillates as if it were nondegenerate. If the total degree algorithm produces approximately the same minimax error for this larger X, then there is a degeneracy d.

References

1. Kaufman Jr, E.H., Leeming, D.J., Taylor, G.D.: A combined Remes-differential correction algorithm for rational approximation: experimental results. Comp. Math. Appl. **6**(2), 155–160 (1980)
2. Kemp, L.F.: Rational approximation and symmetric eigenvalue bounds. In: Chui, C., Schumaker, L., Ward, J. (eds.) Approximation Theory V, pp. 415–417. Academic Press, New York (1986)
3. Kemp, L.F.: Non-degenerate rational approximation. In: Chui, C., Schumaker, L., Stöckler, J. (eds.) Approximation Theory X, pp. 246–266. Vanderbilt University Press, Nashville, TN (2002). Available at https://www.researchgate.net/profile/L_Franklin_Kemp/publications
4. Powell, M.J.D.: Approximation Theory and Methods, pp. 7–8. Cambridge University Press, New York (1981)

Multivariate C^1-Continuous Splines on the Alfeld Split of a Simplex

Alexei Kolesnikov and Tatyana Sorokina

Abstract Using algebraic geometry methods and Bernstein-Bézier techniques, we find the dimension of C^1-continuous splines on the Alfeld split of a simplex in $\mathbb{R}^n$ and describe a minimal determining set for this space.

Keywords Multivariate spline · Minimal determining set · Alfeld split

1 Introduction

Let $\mathcal{P}^n$ denote the set of polynomials in n variables over $\mathbb{R}$. In approximation theory, a spline is a piecewise-polynomial function defined on a polyhedral domain $\Omega \subset \mathbb{R}^n$ that belongs to a certain smoothness class. More precisely, for a fixed partition Δ of the domain Ω into a finite number of n-dimensional polyhedral subsets σ, a spline space is defined with respect to that partition:

$$S_d^r(\Delta) = \{s \in C^r(\Omega) : s|_\sigma \in \mathcal{P}_d^n \quad \text{for all } \sigma \in \Delta\}.$$

We use the following notation:

$$S^r(\Delta) := \bigcup_{d \geq 0} S_d^r, \quad S(\Delta) := \bigcup_{r \geq 0} S^r(\Delta).$$

A. Kolesnikov (✉) · T. Sorokina
Towson University, 7800 York Road, Towson, MD 21252, USA
e-mail: akolesnikov@towson.edu

T. Sorokina
e-mail: tsorokina@towson.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_16,

Bernstein-Bézier techniques have become a standard tool used to analyze multivariate splines. We assume that the reader is familiar with the concepts of domain points, rings, disks, determining sets, and smoothness conditions, see [2, 3, 7].

To explain the approach of algebraic geometry, let us temporarily suspend the smoothness assumption. A piecewise polynomial function can be written in the form $s = \sum_{\sigma\in\Delta} p_\sigma \cdot \chi_\sigma$, where χ_σ is the characteristic function of the set σ and $p_\sigma \in \mathcal{P}^n$. Thus, the set $S(\Delta)$ can be naturally identified with a subset of the following set:

$$\mathcal{R}_n(\Delta) := \Big\{\{(\sigma, p_\sigma)\}_{\sigma\in\Delta} : p_\sigma \in \mathcal{P}^n\Big\}.$$

The set $\mathcal{R}_n(\Delta)$ has the natural structure of a module over the ring $\mathcal{P}^n$: the sum and scalar multiplication are defined as follows:

$$\{(\sigma, p_\sigma)\}_{\sigma\in\Delta} + \{(\sigma, q_\sigma)\}_{\sigma\in\Delta} := \{(\sigma, p_\sigma + q_\sigma)\}_{\sigma\in\Delta},$$
$$p \cdot \{(\sigma, p_\sigma)\}_{\sigma\in\Delta} := \{(\sigma, p \cdot p_\sigma)\}_{\sigma\in\Delta}.$$

Let $\mathcal{R}_n^r(\Delta)$ be the subset of $\mathcal{R}_n(\Delta)$ that corresponds to $S^r(\Delta)$. This subset is easily seen to be a submodule of $\mathcal{R}_n(\Delta)$. Let σ and σ' be adjacent regions, that is, the regions sharing an $(n-1)$-dimensional face or *facet* $\sigma \cap \sigma'$ located on the hyperplane with the equation $l_{\sigma\cap\sigma'} = 0$. Then the smoothness condition of order r across the facet is given by the smoothness equation

$$p_\sigma - p_{\sigma'} = l_{\sigma\cap\sigma'}^{r+1} \cdot q_{\sigma,\sigma'}$$

for some polynomial $q_{\sigma,\sigma'}$. The key idea behind our approach can be phrased as follows: if two different partitions of Ω give rise to the same set of equations, then the spaces of splines of degree $\le d$ for the two partitions are isomorphic. For a more detailed treatment of spline modules we refer the reader to [5, 8].

The paper is organized as follows. In Sect. 2 we introduce the Alfeld split A_n of a simplex T^n in $\mathbb{R}^n$, and the associated Alfeld pyramid $\hat{A}_n$. We prove that the space of splines $S_d^r(A_n)$ on the Alfeld split is isomorphic to the space of splines $S_d^r(\hat{A}_n)$ on the Alfeld pyramid. In Sect. 3, we construct a determining set for $S_d^1(A_n)$. In Sect. 4, we find the dimension of $S_d^1(\hat{A}_n)$ using induction on the spatial dimension n. Since $S_d^1(\hat{A}_n)$ and $S_d^1(A_n)$ are isomorphic they have the same dimension. We conclude the paper with several remarks in Sect. 5.

2 Splines on the Alfeld Split and Pyramid

Let $T^n := [v_1, \ldots, v_{n+1}]$ be a nondegenerate simplex in $\mathbb{R}^n$, and A_n be its Alfeld split around an interior point v_0 into $n+1$ subsimplices, see Figs. 1 and 2 for the two-dimensional case. We note that the two-dimensional Alfeld split coincides with the Clough–Tocher split, and some authors refer to the n-dimensional Alfeld split as the Clough–Tocher split as well. This is inaccurate since there exists an

Fig. 1 Simplex T^2

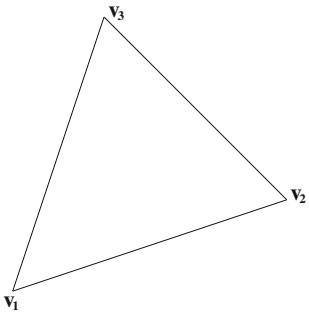
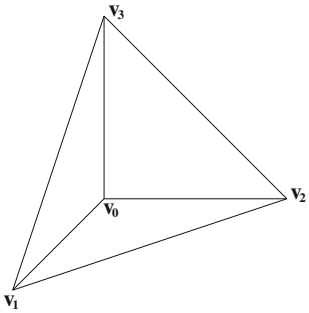


Fig. 2 Alfeld split A_2



n-dimensional Clough–Tocher split, different from the Alfeld split, see [9] and references therein. Each subsimplex in the Alfeld split is a convex hull of a facet of T^n and v_0. We index the subsimplices of the split as follows: The simplex $\sigma_i := [v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}]$ is the unique n-simplex opposite v_i. The common facet of the simplices σ_i and σ_j, $i < j$, will be denoted $\tau_{i,j}$. We may assume that v_0 is the origin, $v_1 = -\sum_{i=1}^{n} e_i$, and $v_{i+1} = e_i$ for $i = 1, \ldots, n$, where e_i is the standard basis vector in $\mathbb{R}^n$. It is immediate to check that for $1 \le i \le n$ the facet $\tau_{1,i+1}$ lies on the hyperplane $x_i = 0$. For a pair (i, j), where $1 \le i \le n-1$ and $i+1 \le j \le n$, the facet $\tau_{i+1,j+1}$ lies on the hyperplane $x_i - x_j = 0$.

In this section, we show that the space of splines of a given polynomial degree d and smoothness r on the Alfeld split is isomorphic to the space of splines over a different partition of T^n that we call the *Alfeld pyramid*. In the rest of the paper, we compute the dimension of the spline space on the Alfeld pyramid split using the Bernstein-Bézier methods and induction on the spatial dimension n. Given the Alfeld split A_n described above, the associated Alfeld pyramid $\widehat{A}_n$ is the partition of T^n into n simplices $\{\widehat{\sigma}_i\}_{i=2}^{n+1}$ and one non-convex polytope $\widehat{\sigma}_1$. For $i = 2, \ldots, n+1$, the simplex $\widehat{\sigma}_i$ has the vertices

$$\{v_0, u_1, v_2, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}\}, \quad \text{where } u_1 := -\frac{v_1}{n}, \tag{1}$$

and the polytope $\widehat{\sigma}_1$ is $T^n \backslash (\mathbb{R}^+)^n$. Figure 3 shows the Alfeld pyramid split in the two-dimensional case. Denoting by $\widehat{\tau}_{i,j}$ the $(n-1)$-simplex which is a common

Fig. 3 Alfeld pyramid $\hat{A}_2$

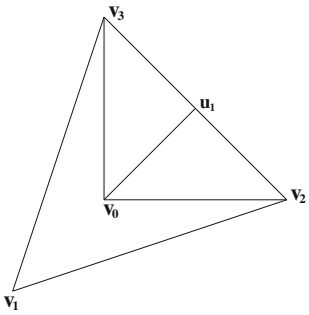
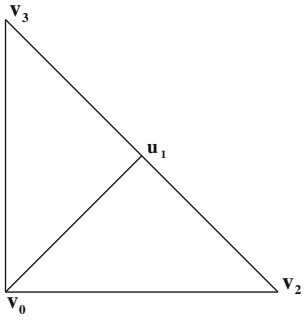


Fig. 4 Pyramid P_2



facet of $\widehat{\sigma}_i$ and $\widehat{\sigma}_j$, we note that for each pair (i, j), the facets $\tau_{i,j}$ and $\widehat{\tau}_{i,j}$ lie on the same hyperplane. This is the key property connecting the Alfeld split A_n and the Alfeld pyramid $\widehat{A}_n$. We denote by P_n the collection of simplices $\widehat{\sigma}_i, i = 2, \ldots, n+1$. This collection is a subset of the Alfeld pyramid, see Fig. 4 for the two-dimensional case.

Theorem 1 *For all $n \geq 2$, and for all $d, r \geq 0$, the spline spaces $S_d^r(A_n)$ and $S_d^r(\widehat{A}_n)$ are isomorphic. In particular,*

$$\dim S_d^r(A_n) = \dim S_d^r(\hat{A}_n).$$

Proof Following [4] or [8], we treat the spline module $\mathcal{R}^r(A_n)$ as the projection onto the first $n+1$ coordinates of the syzygy module of the system of column vectors in the following matrix:

$$\begin{bmatrix}
\delta_{(1,2),1} & \ldots & \delta_{(1,2),n+1} & x_1^{r+1} & 0 & \ldots & \ldots & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\delta_{(1,n+1),1} & \ldots & \delta_{(1,n+1),n+1} & 0 & \ldots & x_n^{r+1} & 0 & \ldots & 0 \\
\delta_{(2,3),1} & \ldots & \delta_{(2,3),n+1} & 0 & \ldots & 0 & (x_1 - x_2)^{r+1} & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & & \vdots & \ddots & \ddots & \vdots \\
\delta_{(n,n+1),1} & \ldots & \delta_{(n,n+1),n+1} & 0 & \ldots & 0 & \ldots & 0 & (x_n - x_{n+1})^{r+1}
\end{bmatrix},$$

where for $1 \le i < j \le n+1$ and $1 \le k \le n+1$,

$$\delta_{(i,j),k} = \begin{cases} 0, & \text{if } k \notin \{i, j\}, \\ (-1)^{k+j}, & \text{if } k = i, \\ (-1)^{i+k+1}, & \text{if } k = j. \end{cases}$$

It remains to note that the matrix associated with the Alfeld pyramid $\widehat{A}_n$ is exactly the same as the one above. Thus the modules $\mathcal{R}^r(A_n)$ and $\mathcal{R}^r(\widehat{A}_n)$ are equal. Since

$$S_d^r(A_n) \cong \{(p_1, \ldots, p_{n+1}) \in \mathcal{R}^r(A_n) |\ \deg(p_i) \le d \ \ \forall i = 1, \ldots, n+1\},$$

and

$$S_d^r(\widehat{A}_n) \cong \{(p_1, \ldots, p_{n+1}) \in \mathcal{R}^r(\widehat{A}_n) |\ \deg(p_i) \le d \ \ \forall i = 1, \ldots, n+1\},$$

the result follows. □

3 A Determining Set for $S_d^1(A_n)$

We recall that given any simplex T^n in $\mathbb{R}^n$, every polynomial p of degree $\le d$ can be written uniquely in the form

$$p = \sum_{i_1+\cdots+i_{n+1}=d} c_{i_1 \ldots i_{n+1}} B^d_{i_1 \ldots i_{n+1}}, \tag{2}$$

where $B^d_{i_1 \ldots i_{n+1}}$ are the Bernstein basis polynomials associated with T^n. As usual, we call the $c_{i_1 \ldots i_{n+1}}$ the B-coefficients of p, and define the associated domain point as

$$\xi^d_{i_1 \ldots i_{n+1}} := (i_1 v_1 + \cdots + i_{n+1} v_{n+1})/d, \quad i_1 + \cdots + i_{n+1} = d. \tag{3}$$

The point $\xi_{i_1 \ldots i_{n+1}}$ is at distance l from the face $[v_1, \ldots, v_k]$ if $i_1 + \cdots + i_k \ge d - l$. A ring $R_l(v_0)$ of radius l around v_0 is the set of domain points at distance l from v_0. The disk $D_l(v_0)$ is the union of rings of radius $\le l$ around v_0. Distances, rings, and disks associated with other faces of T^n are defined similarly. Given Δ, every spline $s \in S_d^0(\Delta)$ can be associated with the set of B-coefficients of its polynomial pieces, and with the set $\mathcal{D}_{d,\Delta}$ of the domain points corresponding to those coefficients.

We begin this section with two simple combinatorial facts. The proofs are based on the tools from the Bernstein-Bézier analysis in order to facilitate the transition to the domain point count in the subsequent theorems.

Lemma 1 *For positive integers n and m, let*

$$\mathcal{I}_m^n := \{(i_1, \ldots, i_{n+1}) \in \mathbb{Z}^{n+1} \mid i_j \geq 0,\ \forall\, j \in \{1, \ldots, n+1\},\ \sum_{j=1}^{n+1} i_j = m\},$$

$$\mathcal{M}_m^n := \{(i_1, \ldots, i_{n+1}) \in \mathcal{I}_m^n \mid \exists\ \textit{a unique}\ j \in \{1, \ldots, n+1\}\ \textit{with}\ i_j = 0\}.$$

Then $|\mathcal{M}_m^n| = (n+1)\binom{m-1}{n-1}$.

Proof Fix $j \in \{1, \ldots, n+1\}$ and set $i_j = 0$. Then $|\mathcal{M}_m^n| = (n+1)|\mathcal{J}_m^n|$, where

$$\mathcal{J}_m^n := \{(i_1, \ldots, i_n) \in \mathbb{Z}^n \mid \quad \forall\, j \in \{1, \ldots, n\},\ \ i_j > 0, \quad i_1 + \cdots + i_n = m\}.$$

In the Bernstein-Bézier analysis, $|\mathcal{J}_m^n|$ is the number of the domain points of a polynomial of degree $\leq m$ in $(n-1)$ variables that are strictly interior to the $(n-1)$-simplex. This number is $\binom{m-1}{n-1}$. Thus, $|\mathcal{M}_m^n| = (n+1)\binom{m-1}{n-1}$. □

Lemma 2 *Let* $\mathcal{I}_m^n$ *be as in Lemma* 1. *Suppose*

$$\mathcal{N}_m^n := \{(i_1, \ldots, i_{n+1}) \in \mathcal{I}_m^n \mid\ \exists\ j \in \{1, \ldots, n+1\}\ \textit{such that}\ i_j = 0\}.$$

Then $|\mathcal{N}_m^n| = \binom{m+n}{n} - \binom{m-1}{n}$.

Proof In the Bernstein-Bézier analysis, this is the number of the domain points of a polynomial of degree $\leq m$ in n variables that are on the boundary of the n-simplex. The easiest way to compute it is to subtract the number of the domain points that are strictly interior to the n-simplex from the total number of the domain points of a polynomial of degree $\leq m$ in n variables in the n-simplex. □

Consider a spline $s \in S_d^1(A_n)$. The set of all B-coefficients associated with this spline is the union of $(n+1)$ sets of B-coefficients associated with the polynomials $s|_{\sigma_i}$, $i = 1, \ldots, n+1$, on each subsimplex. Accordingly, the set of all the domain points associated with s is the union of the domain points for each of the $s|_{\sigma_i}$ in σ_i. One of the key ideas in the argument below is to organize the domain points as

$$\mathcal{D} := \bigcup_{m=0}^{d} R_m(v_0) = \bigcup_{m=0}^{d} \xi_I^m, \quad I \in \mathcal{N}_m^n,$$

where each ξ_I^m is as in (3), $R_m(v_0)$ is the ring of radius m around v_0, and $\mathcal{N}_m^n$ is as in Lemma 2. Indeed, for any $0 \leq m \leq d$, the ring $R_m(v_0)$ of radius m around v_0 is exactly the set of domain points on the boundary of an n-simplex $T_m^n := [mv_1/d, \ldots, mv_{n+1}/d]$. Note that these domain points are the boundary domain points associated with a single polynomial of degree m defined on the simplex T_m^n. This notation establishes a one-to-one correspondence between each domain point and a pair (m, I), where $m \in \{0, \ldots, d\}$ and $I \in \mathcal{N}_m^n$.

We need two basic facts from the Bernstein-Bézier analysis.

Lemma 3 *Let $s \in S_d^1(A_n)$, and let $\mathcal{I}_m^n$ be as in Lemma* 1. *Suppose*

$$\mathcal{F}_m^n := \{(i_1, \ldots, i_{n+1}) \in \mathcal{I}_m^n | \ \exists\, j, k \in \{1, \ldots, n+1\} \text{ such that } j \neq k \text{ and } i_j = i_k = 0\}.$$

Then for each $0 \leq m < d$, the coefficient $c_I^m \in R_m(v_0)$, where $I \in \mathcal{F}_m^n$, can be determined as a linear combination $\mathcal{L}$ of the following $n+1$ coefficients located on $R_{m+1}(v_0)$

$$c^m_{i_1,\ldots,i_{n+1}} = \mathcal{L}(c^{m+1}_{i_1+1,\ldots,i_{n+1}}, c^{m+1}_{i_1,i_2+1,\ldots,i_{n+1}}, \ldots, c^{m+1}_{i_1,\ldots,i_{n+1}+1}). \tag{4}$$

Proof This is a rewrite of the usual smoothness conditions across interior faces of A_n. Indeed, without loss of generality assume $i_1 = i_2 = \cdots = i_k = 0$, $k \geq 2$. Then ξ_I^m lies on the interior face $F_k := [v_0, v_{k+1}, \ldots, v_{n+1}]$ shared by k subsimplices in A_n of the form

$$\sigma_j = [v_0, v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k, v_{k+1}, \ldots, v_{n+1}], \quad j = 1, \ldots, k.$$

Then, we apply C^1 smoothness conditions across F_k to any two such subsimplices. Each smoothness functional combines c_I^m with the $n+1$ coefficients associated with the domain points on T_{m+1}^n that are located at distance one from ξ_I^m in a linear equation and yields (4). □

The proof of the next result can be found in Theorem 6.3 of [10].

Lemma 4 *Let $s \in S_d^1(A_n)$. Then $s \in C^n(v_0)$.*

Combining Lemma 4 and Theorem 1 we obtain the following:

Lemma 5 *Let $s \in S_d^1(\hat{A}_n)$. Then $s \in C^n(v_0)$.*

We note that in both lemmas above, $C^n(v_0)$ is understood in the sense of equality of all partial derivatives of order up to n at v_0. We are now ready to construct a determining set for $S_d^1(A_n)$ as in Definition 5.12 of [7], that is, a subset $\mathcal{G}$ of $\mathcal{D}_{d,\Delta}$ such that if $s \in S_d^1(A_n)$ and the B-coefficients corresponding to all domain points in $\mathcal{G}$ vanish, then s vanishes as well. Note that at this point, we do not prove that this determining set is minimal.

Theorem 2 *Let $s \in S_d^1(A_n)$. Then*

$$DS_d^1(A_n) := \{\xi_I^d, \ \ I \in \mathcal{N}_d^n\} \cup \{\xi_I^m, \ \ j \in \{n, \ldots, d-1\}, \ \ I \in \mathcal{M}_m^n\},$$

is a determining set and $|DS_d^1(A_n)| = \binom{d+n}{n} + n\binom{d-1}{n}$. This set consists of all domain points on the exterior of T^n, along with the domain points strictly interior to each boundary facet of every simplex T_m^n for $m = d-1, \ldots, n$.

Proof It suffices to show that if all coefficients of s corresponding to $DS_d^1(A_n)$ are set to zero, then $s \equiv 0$. We start with setting to zero all the coefficients associated with $\{\xi_I^d, \ \ I \in \mathcal{N}_d^n\}$, or, equivalently, on $R_d(v_0)$. From Lemma 2 we obtain

$$|\{\xi_I^d, \ \ I \in \mathcal{N}_d^n\}| = \binom{d+n}{n} - \binom{d-1}{n}.$$

Next we move to $R_{d-1}(v_0)$. All indices I on this ring have two properties: the number of elements in I is $d-1$, and at least one entry in I is zero. According to Lemma 3, each coefficient corresponding to I with two or more zero entries vanishes. Thus, we only need to set to zero the coefficients corresponding to I with precisely one zero entry. We repeat this process for each m between $d-1$ and n. That is, once $R_{m+1}(v_0)$ is populated with zeros, we move to $R_m(v_0)$, where all indices I have two properties: the number of elements in I is m, and at least one entry in I is zero. According to Lemma 3, each coefficient corresponding to I with two or more zero entries vanishes. Thus, we only need to set to zero the coefficients corresponding to I with precisely one zero entry. From Lemma 1 we obtain

$$|\{\xi_I^m, \ \ I \in \mathcal{M}_m^n\}| = (n+1)\binom{m-1}{n-1}, \ \ m \in \{d-1, d-2 \ldots, n\}.$$

When we populate $R_n(v_0)$ with zeros, we note that by Lemma 4, the disk $D_n(v_0)$ can be considered as a single simplex since the coefficients of s corresponding to this disk form a polynomial .of degree n. We note that from Lemma 2, the total number of domain points on $R_n(v_0)$ is $\binom{2n}{n}$ which is precisely the dimension of polynomials degree $\leq n$ in n variables. Moreover, this polynomial of degree n vanishes on all $n+1$ faces of the simplex T_n^n, and thus it is a zero polynomial. Therefore, all coefficients of s in $D_n(v_0)$ vanish. We now do the final count

$$\begin{aligned} |DS_d^1(A_n)| &= \binom{d+n}{n} - \binom{d-1}{n} + (n+1)\sum_{m=n}^{d-1}\binom{m-1}{n-1} \\ &= \binom{d+n}{n} - \binom{d-1}{n} + (n+1)\binom{d-1}{n} = \binom{d+n}{n} + n\binom{d-1}{n}. \end{aligned}$$

The proof is complete. We note that if the dimension of $S_d^1(A_n)$ is known to be $|DS_d^1(A_n)|$, then $DS_d^1(A_n)$ would be a minimal determining set. □

4 The Main Result

In this section, we compute the dimension of C^1-continuous splines defined over the Alfeld pyramid $\hat{A}_n$ in $\mathbb{R}^n$. By Theorem 1 this is equal to the dimension of C^1-continuous splines over the Alfeld split A_n of a single simplex in $\mathbb{R}^n$. We note that

from Theorem 9.3 in [7] it follows that $\dim S_d^1(A_2) = \binom{d+2}{2} + 2\binom{d-1}{2}$. In Remark 6, we illustrate the idea of our proof for $n = 2$ since this is the only case with clear visual illustration.

Theorem 3 *For all integers $d \geq 0$ and $n \geq 1$,*

$$\dim S_d^1(A_n) = \binom{d+n}{n} + n\binom{d-1}{n}.$$

Proof We use induction on n. Since A_1 is the split of a line segment into two subsegments, it is immediate that $\dim S_d^1(A_1) = 2d$.

For $n \geq 2$, in view of Theorem 1 we can consider $S_d^1(\hat{A}_n)$ instead of $S_d^1(A_n)$. The dimension of $S_d^1(\hat{A}_n)$ is equal to the dimension of polynomials of degree $\leq d$ in n variables $\binom{d+n}{n}$ plus the dimension of

$$S_0 := \{s \in S_d^1(\hat{A}_n) |\ s \equiv 0 \text{ everywhere outside of the pyramid } P_n\}.$$

We treat S_0 as a subspace of $S_d^1(P_n)$. The plan is to use the induction hypothesis to compute the dimension of $S_d^1(P_n)$ and then subtract the number of domain points associated with vanishing B-coefficients due to the condition $s \equiv 0$ outside of P_n. We recall that the pyramid P_n is the split of the simplex $[v_0, v_2, \ldots, v_{n+1}]$ into n subsimplices with the split point $u_1 := -v_1/n$, as in (1). The domain points inside P_n are located on the union of rings $R_i(v_0)$, $i = 0, \ldots, d$. These rings lie on parallel $(n-1)$-simplices $T_i^{n-1} := [iv_2/d, \ldots, iv_{n+1}/d]$. Each simplex T_i^{n-1} is partitioned as $(n-1)$-dimensional Alfeld split A_{n-1}^i by the point iu_1/d. Therefore, the domain points in the pyramid P_n can be considered as the domain points for the Alfeld splits A_{n-1}^i of T_i^{n-1}. Moreover, since all T_i^{n-1} are parallel in $\mathbb{R}^n$, all C^1 smoothness conditions across interior faces of P_n are those for the $(n-1)$-dimensional Alfeld split. Thus, using the induction hypothesis on A_{n-1}^i , we obtain

$$\dim S_d^1(P_n) = \sum_{i=0}^{d} \dim S_i^1(A_{n-1}^i) = \sum_{i=0}^{d} \left[\binom{i+n-1}{n-1} + (n-1)\binom{i-1}{n-1} \right].$$

Moreover, the induction hypothesis along with Theorem 2 provides *minimal* determining sets $DS_i^1(A_{n-1}^i)$. In order to find the dimension of S_0, we need to know the number N_i of points in $DS_i^1(A_{n-1}^i)$ that have associated vanishing B-coefficients after joining the zero function outside of P_n. Then

$$\dim S_0 = \sum_{i=0}^{d} \dim (S_i^1(A_{n-1}^i) - N_i). \tag{5}$$

We now compute N_i. Due to the supersmoothness result of Lemma 5, any $s \in S_0$ is $C^n(v_0)$. Thus,

$$N_i = \binom{i+n-1}{n-1} = \dim S_i^1(A_{n-1}^i) \qquad \text{for} \quad 0 \le i \le n.$$

For each $i > n$, the C^1-smoothness conditions across the boundary of P_n affect the coefficients associated with the domain points located on the boundary and one layer inside of P_n. More precisely, they are located in the rings $R_i(iu_1/d)$ and $R_{i-1}(iu_1/d)$. They form a subset of $DS_i^1(A_{n-1}^i)$. Lemma 1, Lemma 2, and Theorem 2 provide the complete description and the number of such domain points:

$$DS_i^1(A_{n-1}^i) \cap R_i(iu_1/d) = \{\xi_I^i, \ \ I \in \mathcal{N}_i^{n-1}\},$$
$$DS_i^1(A_{n-1}^i) \cap R_{i-1}(iu_1/d) = \{\xi_I^{i-1}, \ \ I \in \mathcal{M}_{i-1}^{n-1}\},$$

and

$$N_i = \binom{i+n-1}{n-1} - \binom{i-1}{n-1} + n\binom{i-2}{n-2}. \tag{6}$$

Substituting (6) into (5) we obtain

$$\dim S_0 = n \sum_{i=n+1}^{d} \left[\binom{i-1}{n-1} - \binom{i-2}{n-2} \right] = n \sum_{i=n+1}^{d} \binom{i-2}{n-1}$$
$$= n \sum_{i=0}^{d-n-1} \binom{i+n-1}{i} = n\binom{d-1}{n}.$$

Finally,

$$\dim S_d^1(A_n) = \dim S_d^1(\hat{A}_n) = \binom{d+n}{n} + \dim S_0 = \binom{d+n}{n} + n\binom{d-1}{n}.$$

The proof is now complete. □

5 Remarks

Remark 1 Theorem 2 combined with Theorem 3 provides a minimal determining set. This set can be used directly to construct C^1-continuous macro-elements based on the Alfeld split of a simplex. However, the polynomial degree d of such macro-elements is at least $2^{n-1} + 1$. Thus for $n \ge 3$, without additional supersmoothness conditions, C^1-continuous macro-elements on the Alfeld split of a simplex have excessive number of free parameters and are hard to implement. For the case $n = 3$, additional smoothness is introduced in [1].

Remark 2 The work on finding dimensions of spline spaces $S_d^r(A_n)$ for higher values of r is in progress. The main difficulty is that the analog of Lemma 4 for $r > 1$ is not known. The lower bound for the supersmoothness at v_0 can be found in [10], but this bound is not exact. The supersmoothness at the split point is one of the main ingredients of the current proof of Theorem 3.

Remark 3 The conjecture on the dimension of $S_d^r(A_n)$ for all values of n, r, and d can be found in [6]. Our result in Theorem 3 proves this conjecture for $r = 1$, for all values of n and d.

Remark 4 There has been a considerable amount of work done with bivariate and trivariate macro-elements based on the Alfeld splits of a triangle and a tetrahedron, respectively. Such macro-element spaces can have dimension different from our result due to additional conditions imposed on them, see [1, 7] and references therein.

Remark 5 The minimal determining sets of Theorem 2 for $n = 2$ and $n = 3$ can be checked directly using P. Alfeld's software available on http://www.math.utah.edu/~pa. The software computes dimension of spline spaces for fixed values of d as well.

Remark 6 In this remark we illustrate the idea of the proof of Theorem 3 for $n = 2$. We consider $S_d^1(\hat{A}_2)$ instead of $S_d^1(A_2)$, and refer to Figs. 2 and 3 to observe that

$$\dim S_d^1(\hat{A}_2) = \binom{d+2}{2} + \dim S_0, \qquad \text{where}$$

$$S_0 := \{s \in S_d^1(\hat{A}_2) |\ \ s \equiv 0 \text{ everywhere outside of } [v_0, v_2, v_3]\}.$$

The split of the triangle $[v_0, v_2, v_3]$ into two subtriangles, $[v_0, v_2, u_1]$ and $[v_0, v_3, u_1]$, forms the pyramid P_2. The domain points inside P_2 are located on the parallel line segments $T_i^1 := [iv_2/d, iv_3/d]$ partitioned into two subsegments $[iv_2/d, iu_1/d]$ and $[iv_3/d, iu_1/d]$ forming the Alfeld splits A_1^i. In Fig. 5, there are five segments T_i^1 split in half. Since all T_i^1 are parallel in $\mathbb{R}^2$, all C^1-smoothness conditions across $[v_0, u_1]$ are those for A_1^i. Each minimal determining set $DS_i^1(A_1^i)$, $i = 1, \ldots, d$, is formed by all domain points on $[iv_2/d, iv_3/d]$ except iu_1/d. The minimal determining set for $DS_i^1(A_1^0)$ is just v_0. In order to find the dimension of S_0, we need to know N_i the number of points in $DS_i^1(A_1^i)$ that have associated vanishing coefficients after joining the zero function outside of $[v_0, v_2, v_3]$. Then

$$\dim S_0 = \sum_{i=0}^{d} \dim (S_i^1(A_1^i) - N_i).$$

Due to supersmoothness two at v_0, the B-coefficients associated with the domain points in $D_2(v_0)$ marked as black dots in Fig. 5 vanish. Thus $N_0 = 1$, $N_1 = 2$, and $N_2 = 4$. For each $i > 2$, the C^1-smoothness conditions across $[v_0, v_2]$, and $[v_0, v_3]$ affect the coefficients associated with the domain points on $[v_0, v_2] \cup [v_0, v_3]$, and

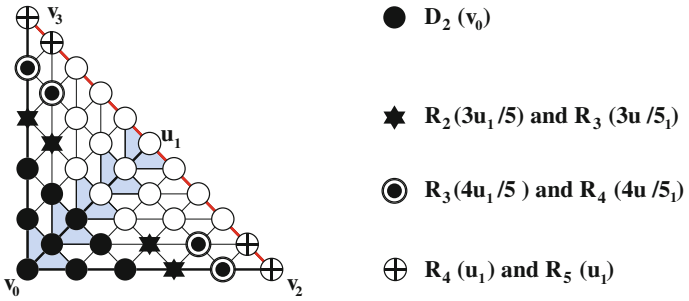


Fig. 5 Domain points in P_2 for $S_5^1(\hat{A}_2)$

one layer inside of P_2. For example, in Fig. 5, the coefficients associated with the stars, the dots in circles, and the crosses all vanish due to C^1 smoothness conditions across $[v_0, v_2]$ and $[v_0, v_3]$. Therefore, $N_i = 4$ for $i > 2$, and

$$\dim S_d^1(\hat{A}_2) = \binom{d+2}{2} + \sum_{i=3}^{d}(2i-4) = \binom{d+2}{2} + 2\binom{d-1}{2}.$$

Acknowledgments The first author was partially supported by the NSF grant DMS-0901315.

References

1. Alfeld, P.: A trivariate Clough-Tocher scheme for tetrahedral data. CAGD **1**, 169–181 (1984)
2. Alfeld, P., Schumaker, L.L., Sirvent, M.: On dimension and existence of local bases for multivariate spline spaces. J. Approx. Theory **70**, 243–264 (1992)
3. Alfeld, p., Sirvent, M.: A recursion formula for the dimension of superspline spaces of smoothness r and degree $d > r2^k$. In: Schempp W., Zeller K. (eds) Approximation Theory V, Proceedings of the Oberwolfach Meeting, pp. 1–8. Birkhäuser (1989)
4. Billera, L.J., Rose, L.L.: A dimension series for multivariate splines. Discrete Comput. Geom. **6**, 107–128 (1991)
5. Billera, L.: Homology of smooth splines: generic triangulations and a conjecture of Strang. Trans. AMS **310**, 325–340 (1988)
6. Foucart, S., Sorokina, T.: Generating dimension formulas for multivariate splines. Albanian J. Math. **7**, 25–35 (2013)
7. Lai, M.-J., Schumaker, L.L.: Spline Functions on Triangulations. Cambridge University Press, Cambridge (2007)
8. Schenck, H., Stilman, M.: Local cohomology of bivariate splines. J. Pure Appl. Algebra **117–118**, 535–548 (1997)
9. Sorokina, T.: A C^1 multivariate Clough-Tocher interpolant. Constr. Approximation **29**(1), 41–59 (2009)
10. Sorokina, T.: Intrinsic supersmoothness of multivariate splines. Numer. Math. **116**, 421–434 (2010)

On Convergence of Singular Integral Operators with Radial Kernels

Sevilay Kırcı Serenbay, Özge Dalmanoğlu and Ertan İbikli

Abstract In this paper, we prove the pointwise convergence of the operator $L(f;x,y;\lambda)$ to the function $f(x_0,y_0)$, as $(x,y;\lambda)$ tends to $(x_0,y_0;\lambda_0)$ by the three parameter family of singular integral operators in $L_1(Q_1)$, where Q_1 is a closed, semi-closed, or open rectangular region $<-a,a>\times<-b,b>$. Here, the kernel function is radial and we take the point(x_0,y_0) as a μ-generalized Lebesgue point of f.

Keywords Singular operators · Radial kernel · Lebesgue point · Pointwise convergence

1 Introduction and Preliminaries

In papers [1] and [6], Gadjiev and Taberski studied the pointwise convergence of integrable functions in $L_1(-\pi,\pi)$ space by a two parameter family of convolution type singular integral operators of the form

$$U(f;x,\lambda)=\int_{-\pi}^{\pi} f(t)K(t-x,\lambda)dt,\ x\in(-\pi,\pi). \tag{1}$$

Here, the kernel function $K(t,\lambda)$ is defined for all t and $\lambda\in\Lambda$ (where Λ is a given set of numbers with accumulation point λ_0), 2π-periodic, even, and measurable

S. K. Serenbay (✉) · Ö. Dalmanoğlu
Faculty of Education, Department of Mathematics Education, Başkent University, Ankara, Turkey
e-mail: sevilaykirci@gmail.com

Ö. Dalmanoğlu
e-mail: odalmanoglu@baskent.edu.tr

E. İbikli
Faculty of Science, Department of Mathematics, Ankara University, Tandoğan, Ankara, Turkey
e-mail: ibikli@ankara.edu.tr

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_17,

with respect to t at each $\lambda \in \Lambda$. The pointwise convergence of the operator (1) was investigated at the point x_0, when x_0 is a continuous point, Lebesgue point, or a generalized Lebesgue point of the function f in $L_1(-\pi, \pi)$.

In [3] Karsli improved the results of Gadjiev and Taberski by considering the singular integral operator of the form

$$T(f; x, \lambda) = \int_a^b f(t)K(t - x, \lambda)dt \quad x \in (a, b). \tag{2}$$

Here, f belongs to the function space $L_1(a, b)$ and the kernel function $K(t, \lambda)$ of $T(f; x, \lambda)$ does not have to be 2π-periodic, positive, or even. So, Karsli extended the results found in [1] and [6]. By taking x_0 a $\mu-$generalized Lebesgue point of f, he showed the pointwise convergence of $T(f; x, \lambda)$ to $f(x_0)$ as (x, λ) tends to (x_0, λ_0).

In papers [5] and [7], the pointwise convergence of integrable functions in $L_1(P)$ was investigated by a three-parameter family of convolution type singular integral operators of the form

$$U(f; x, y; \lambda) = \int_P f(s, t)K(s - x, t - x; \lambda)dsdt \quad (x, y) \in P. \tag{3}$$

(Here, P denotes the region $[-a, a] \times [-b, b]$ and $[-\pi, \pi] \times [-\pi, \pi]$, respectively.)

In [8] Yilmaz et al. investigated the pointwise convergence of the integral operator $L(f, x, y; \lambda)$ to $f(x_{0,} y_0, \lambda_0)$ in the space $L_1(D)$ (space of functions 2π periodic in each variable separately and Lebesgue integrable in the square $D =< -\pi, \pi > \times < -\pi, \pi >$ where $< -\pi, \pi > \times < -\pi, \pi >$ is an arbitrary closed, semi-closed, or open square region), by the three-parameter family of integral operators with radial kernel of the form

$$L(f; x, y; \lambda) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(s, t)K\left(\sqrt{(s - x)^2 + (t - y)^2}; \lambda\right)dsdt \quad (x, y) \in D. \tag{4}$$

Here, (x_0, y_0) is taken as a generalized Lebesgue point of the function $f(s, t)$.

In this work, we have studied the pointwise convergence of the integral operator

$$L^*(f; x, y; \lambda) = \int_{-b}^{b}\int_{-a}^{a} f(s, t)K\left(\sqrt{(s - x)^2 + (t - y)^2}; \lambda\right)dsdt \tag{5}$$

to the function $f(x_{0,} y_0)$ in $L_1(Q_1)$ (space of Lebesque integrable functions in Q_1) by the three parameter family of singular integral operators with radial kernel. Here,

(x_0, y_0) is a $\mu-$generalized Lebesgue point of the function $f(s,t)$, $\lambda \in \Lambda \subset \mathbb{R}$ and the region is extended to a rectangle $Q_1 =< -a, a > \times < -b, b >$ ($< -a, a >$ is an arbitrary interval in $\mathbb{R}$ such that $[-a, a]$, $[-a, a)$, $(-a, a]$, or $(-a, a)$).

Now, we will give some definitions and lemmas that will be used in the next section.

Definition 1 A function $\Psi \in L_1(Q_1)$ $\left(Q_1 \subset \mathbb{R}^2\right)$ is said to be radial, if there exists a function $\kappa\left(\sqrt{s^2+t^2}\right)$, defined on $0 \leq \sqrt{s^2+t^2} < \infty$, such that $\Psi(s,t) = \kappa\left(\sqrt{s^2+t^2}\right)$ a.e. [4].

Definition 2 (*Class A*)
We take a family $\kappa = \left(K\left(\sqrt{s^2+t^2};\lambda\right)\right)_{\lambda\in\Lambda}$ of functions $K\left(\sqrt{s^2+t^2};\lambda\right) : \mathbb{R}^2 \times \Lambda \to \mathbb{R}$. We will say that the function $K\left(\sqrt{s^2+t^2};\lambda\right)$ belongs to class A, if the following conditions are satisfied:

(a) As a function of (s,t), $K\left(\sqrt{s^2+t^2};\lambda\right)$ is defined on $\mathbb{R}^2$ and integrable for each fixed $\lambda \in \Lambda$ (Λ is a given set of numbers with accumulation point λ_0).

(b) $\lim\limits_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right) dsdt = 1.$

(c) There exists a $\delta_0 > 0$ such that the function $K\left(\sqrt{s^2+t^2};\lambda\right)$ takes its maximum value at δ_0 in the region $\mathbb{R}^2 \setminus \left\{(s,t) : \sqrt{s^2+t^2} \leq \delta_0\right\}$ for each $\lambda \in \Lambda$.

(d) $\lim\limits_{\lambda\to\lambda_0} \int\int_{\mathbb{R}^2\setminus\{(s,t):\sqrt{s^2+t^2}\leq\delta\}} K\left(\sqrt{s^2+t^2};\lambda\right) dsdt = 0$ for every $\delta > 0$.

(e) $\lim\limits_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} \sup\limits_{\delta\leq\sqrt{(s-x)^2+(t-y)^2}} K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right) dsdt = 0$ for every $\delta > 0$.

In order to prove our main result, we shall also need the following Lemmas and remarks:

Lemma 1 *Let $1 \leq p < \infty$. If the kernel $K\left(\sqrt{s^2+t^2};\lambda\right)$ belongs to class A, then $L^*(f;x,y;\lambda)$ defines a continuous transformation over $L_p(Q_1)$.*

Proof One can find the proof of similar Lemma in [9].

Lemma 2 [2] *If $g(x,y)$ is continuous over the rectangle $M : (a \leq x \leq b; c \leq y \leq d)$ and $\alpha(x,y)$ is of bounded variation on M, then $g \in RS(\alpha)$.*

Lemma 3 [2] *Assume that $g \in RS(\alpha)$ on M and α is of bounded variation on M. Then,*

$$\left|\int_a^b\int_c^d g(x,y)d_xd_y\alpha(x,y)\right| \leq \sup_{(x,y)\in M}|g(x,y)| \cdot \bigvee_M(\alpha). \tag{6}$$

Remark 1 We denote $\tilde{f} \in L_1(R^2)$ as

$$\tilde{f}(s,t) = \begin{cases} f(s,t), & (s,t) \in Q_1; \\ 0 \quad , & (s,t) \notin Q_1. \end{cases} \tag{7}$$

2 Main Result

We now give our main result with the following theorem:

Theorem 1 *Suppose that the kernel function* $K\left(\sqrt{s^2+t^2};\lambda\right)$ *belongs to class A. Let* (x_0, y_0) *be a* $\mu-$*generalized Lebesgue point of the function* $f(x,y) \in L_1(Q_1)$, *i.e., the condition*

$$\lim_{(h,r)\to(0,0)} \frac{1}{\mu_1(h)\,\mu_2(r)} \int_0^h \int_0^r |f(s+x_0, t+y_0) - f(x_0,y_0)|\,dsdt = 0 \tag{8}$$

is satisfied, where $\mu_1(s)$ *and* $\mu_2(t)$ *are defined on* $< -a, a >$ *and* $< -b, b >$, *respectively.* $\mu_1(s)$ *and* $\mu_2(t)$ *are also increasing, absolutely continuous, and* $\mu_1(0) = \mu_2(0) = 0$.

If $(x, y; \lambda)$ *tends to* $(x_0, y_0; \lambda_0)$ *on any set* Z *on which the functions*

$$\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2 + (y_0-\delta-y)^2};\lambda\right) \left|\mu_1'(x_0-s)\right| ds + 2K\left(|y_0-\delta-y|;\lambda\right)\mu_1(|x_0-x|) \tag{9}$$

$$\int_{y_0-\delta}^{y_0} K\left(\sqrt{(x_0-\delta-x)^2 + (t-y)^2};\lambda\right) \left|\mu_2'(y_0-t)\right| dt + 2K\left(|x_0-\delta-x|;\lambda\right)\mu_2(|y_0-y|) \tag{10}$$

and

$$\mu_1(x_0-s)\ \ \mu_2(y_0-t) \tag{11}$$

are bounded, then

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} L^*(f;x,y;\lambda) = f(x_0,y_0)\,. \tag{12}$$

Proof Suppose that $(x_0, y_0) \in Q_1$ and

$$x_0 + \delta_1 < a, \quad x_0 - \delta_1 > -a, \quad 0 < x_0 - x < \delta_1/2, \tag{13}$$
$$y_0 + \delta_2 < b, \quad y_0 - \delta_2 > -b, \quad 0 < y_0 - y < \delta_2/2, \tag{14}$$

where $0 < \max\{\delta_1, \delta_2\} < \delta_0$. From Remark 1, we can write

$$\begin{aligned}
&\left|L^*(f; x, y; \lambda) - f(x_0, y_0)\right| \\
&= \left| \int_{-b}^{b} \int_{-a}^{a} f(s,t)\, K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt - f(x_0, y_0) \right| \\
&= \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{f}(x,y)\, K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt - f(x_0, y_0) \right| \\
&= \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{f}(s,t)\, K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt - f(x_0, y_0) \right. \\
&\quad + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_0, y_0)\, K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt \\
&\quad \left. - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_0, y_0)\, K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt \right|.
\end{aligned}$$

$$\begin{aligned}
&\left|L^*(f; x, y; \lambda) - f(x_0, y_0)\right| \\
&\le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left|\tilde{f}(s,t) - f(x_0, y_0)\right| K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt \\
&\quad + |f(x_0, y_0)| \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K\left(\sqrt{(s-x)^2 + (t-y)^2}; \lambda\right) dsdt - 1 \right| \\
&= I(x, y; \lambda) + J(x, y; \lambda).
\end{aligned}$$

From condition (b), one can easily obtain that

$$\lim_{(x,y;\lambda) \to (x_0, y_0; \lambda_0)} J(x, y; \lambda) = 0. \tag{15}$$

Now, let us investigate $I(x, y; \lambda)$. We shall divide the region into two parts and examine $I(x, y; \lambda)$ on these two regions.

$$I(x,y;\lambda)=\int\int_{Q_1}|f(s,t)-f(x_0,y_0)|\,K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$+\int\int_{\mathbb{R}^2\setminus Q_1}\left|\tilde{f}(s,t)-f(x_0,y_0)\right|K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$=I_1(x,y;\lambda)+I_2(x,y;\lambda).$$

If we consider $I_2(x,y;\lambda)$, we have

$$I_2(x,y;\lambda)=\int\int_{\mathbb{R}^2\setminus Q_1}\left|\tilde{f}(s,t)-f(x_0,y_0)\right|K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$=|f(x_0,y_0)|\int\int_{\mathbb{R}^2\setminus Q_1}K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$\le|f(x_0,y_0)|\int\int_{\mathbb{R}^2\setminus Q}K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt,$$

where $Q=\left\{(s,t):(s-x_0)^2+(t-y_0)^2\le\delta^2,(x_0,y_0)\in Q_1\right\}$. Now from condition (d), we can write

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)}I_2(x,y;\lambda)=0. \tag{16}$$

Now, we take $I_1(x,y;\lambda)$ into account. We shall again divide the region into two parts.

$$I_1(x,y;\lambda)=\int\int_{Q_1\setminus Q}|f(s,t)-f(x_0,y_0)|\,K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$+\int\int_{Q}|f(s,t)-f(x_0,y_0)|\,K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$=I_{11}(x,y;\lambda)+I_{12}(x,y;\lambda).$$

Our aim is to show that $I_1(x,y;\lambda)$, and thereby $I_{11}(x,y;\lambda)$ and $I_{12}(x,y;\lambda)$, tends to zero as $(x,y;\lambda)\to(x_0,y_0;\lambda_0)$.

First, we consider $I_{11}(x,y;\lambda)$. From (e), we have the following inequalities:

$$I_{11}(x,y;\lambda)\le\int_{Q_1\setminus Q}|f(s,t)|\,K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$
$$+|f(x_0,y_0)|\int\int_{Q_1\setminus Q}K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$

$$\leq \sup_{\delta \leq \sqrt{(s-x)^2+(t-y)^2}} K\left(\sqrt{(s-x)^2+(t-y)^2}; \lambda\right) \times \left(\int_{Q_1} |f(s,t)|\, dsdt + |f(x_0,y_0)| \int_{Q_1} dsdt\right)$$
$$= \sup_{\delta \leq \sqrt{(s-x)^2+(t-y)^2}} \left(\|f\|_{L_1(Q_1)} + 4ab\,|f(x_0,y_0)|\right).$$

Taking the limit of both sides as $(x, y; \lambda) \to (x_0, y_0; \lambda_0)$, we get

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} I_{11}(x, y; \lambda) = 0. \tag{17}$$

Now, let us consider the second integral $I_{12}(x, y; \lambda)$.

$$I_{12}(x, y; \lambda) = \int_Q |f(s,t) - f(x_0,y_0)|\, K\left(\sqrt{(s-x)^2+(t-y)^2}; \lambda\right) dsdt$$
$$\leq \int_{y_0-\delta}^{y_0+\delta} \int_{x_0-\delta}^{x_0+\delta} |f(s,t) - f(x_0,y_0)|\, K\left(\sqrt{(s-x)^2+(t-y)^2}; \lambda\right) dsdt$$
$$= \left(\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0} + \int_{y_0-\delta}^{y_0}\int_{x_0}^{x_0+\delta} + \int_{y_0}^{y_0+\delta}\int_{x_0-\delta}^{x_0} + \int_{y_0}^{y_0+\delta}\int_{x_0}^{x_0+\delta}\right)$$
$$\times |f(s,t) - f(x_0,y_0)|\, K\left(\sqrt{(s-x)^2+(t-y)^2}; \lambda\right) dsdt$$
$$= I_{121}(x, y; \lambda) + I_{122}(x, y; \lambda) + I_{123}(x, y; \lambda) + I_{124}(x, y; \lambda).$$

Since

$$I_{12}(x, y; \lambda) \leq I_{121}(x, y; \lambda) + I_{122}(x, y; \lambda) + I_{123}(x, y; \lambda) + I_{124}(x, y; \lambda) \tag{18}$$

we need to show that the terms on the right hand side of the above inequality tend to zero as $(x, y; \lambda) \to (x_0, y_0; \lambda_0)$.

Now, let us consider the first integral $I_{121}(x, y; \lambda)$. Remember that, from (8) for every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\int_{y_0-h}^{y_0} \int_{x_0-r}^{x_0} |f(s,t) - f(x_0,y_0)|\, dsdt \leq \varepsilon \mu_1(h)\, \mu_2(r) \tag{19}$$

for all $0 < h, r \leq \delta$.

Let us define a new function

$$F(s,t)=\int_{t}^{y_0}\int_{s}^{x_0}|f(u,v)-f(x_0,y_0)|\,dudv. \tag{20}$$

From (19) we can write

$$F(s,t)\le \varepsilon\mu_1(x_0-s)\,\mu_2(y_0-t). \tag{21}$$

Now, we start to evaluate the integral $I_{121}(x,y;\lambda)$. We can write (see [7])

$$\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0}|f(s,t)-f(x_0,y_0)|\,K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dsdt$$

$$=(S)\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0}K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dF(s,t),$$

where (S) denotes the Riemann-Stieltjes integral.

Applying two-dimensional integration by parts to the above Riemann–Stieltjes integral (see [7]) we get,

$$\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0}K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)dF(s,t)$$

$$=\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0}F(s,t)\,dK\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)$$

$$+\int_{x_0-\delta}^{x_0}F(s,y_0-\delta)\,dK\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)$$

$$+\int_{y_0-\delta}^{y_0}F(x_0-\delta,t)\,dK\left(\sqrt{(x_0-\delta-x)^2+(t-y)^2};\lambda\right)$$

$$+F(x_0-\delta,y_0-\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right).$$

From (21) we have,

$$\begin{aligned}
&\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right) dF(s,t)\\
&\le \varepsilon \int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0} \mu_1(x_0-s)\,\mu_2(y_0-t)\,dK\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)\\
&\quad+\varepsilon\int_{x_0-\delta}^{x_0} \mu_1(x_0-s)\,\mu_2(\delta)\,dK\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\\
&\quad+\varepsilon\int_{y_0-\delta}^{y_0} \mu_1(\delta)\,\mu_2(y_0-t)\,dK\left(\sqrt{(x_0-\delta-x)^2+(t-y)^2};\lambda\right)\\
&\quad+\varepsilon\mu_1(\delta)\,\mu_2(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\\
&= i_1+i_2+i_3+i_4. \qquad (22)
\end{aligned}$$

First, we will evaluate the integrals i_2 and i_3.

$$\begin{aligned}
i_2 &= \varepsilon\mu_2(\delta)\int_{x_0-\delta}^{x_0} \mu_1(x_0-s)\,dK\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\\
&= \varepsilon\mu_2(\delta)\left(\mu_1(x_0-s)\,K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\Big|_{x_0-\delta}^{x_0}\right.\\
&\quad\left.+\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-s)\,ds\right)\\
&= \varepsilon\mu_2(\delta)\left(-\mu_1(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right.\\
&\quad\left.+\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-s)\,ds\right)\\
&= -\varepsilon\mu_1(\delta)\,\mu_2(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\\
&\quad+\varepsilon\mu_2(\delta)\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-s)\,ds. \qquad (23)
\end{aligned}$$

Here, we note that if a function f is monotone on $[a,b]$, then

$$\bigvee [f;a,b] = \bigvee_{a}^{b}(f) = |f(b) - f(a)|. \tag{24}$$

Here $\bigvee [f;a,b]$ denotes the total variation of f on $[a,b]$.

According to condition (c) and from (13) and (14), we have the following estimates for the integral on the right hand side of equality (23):

$$\begin{aligned}
&\int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-s)\,ds \\
&= \int_{x_0-\delta-x}^{x_0-x} K\left(\sqrt{(s)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds \\
&= \int_{x_0-\delta-x}^{x_0-x}\left\{\bigvee_{x_0-\delta-x}^{s} K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\right. \\
&\qquad \left. + K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right\}\mu_1'(x_0-x-s)\,ds \\
&= \left(\int_{x_0-\delta-x}^{0} + \int_{0}^{x_0-x}\right)\left\{\bigvee_{x_0-\delta-x}^{s} K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\right\}\mu_1'(x_0-x-s)\,ds \\
&\quad + \int_{x_0-\delta-x}^{x_0-x} K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds \\
&= \int_{x_0-\delta-x}^{0}\left(K\left(\sqrt{(s)^2+(y_0-\delta-y)^2};\lambda\right)\right. \\
&\qquad \left. - K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right)\mu_1'(x_0-x-s)\,ds \\
&\quad + \int_{0}^{x_0-x}\left(\bigvee_{x_0-\delta-x}^{0} + \bigvee_{0}^{s}\right) K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds \\
&\quad + K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)(\mu_1(0)-\mu_1(\delta)).
\end{aligned} \tag{25}$$

For the integral above that includes bounded variation we have the following estimate:

$$\begin{aligned}
&\int_0^{x_0-x}\left(\bigvee_{x_0-\delta-x}^{0}+\bigvee_{0}^{s}\right)K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\quad=\int_0^{x_0-x}\bigvee_{x_0-\delta-x}^{0}K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\qquad+\int_0^{x_0-x}\bigvee_{0}^{s}K\left(\sqrt{(u)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\quad=\int_0^{x_0-x}\left(K\left(|y_0-\delta-y|;\lambda\right)-K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right)\\
&\qquad\times\quad\mu_1'(x_0-x-s)\,ds\\
&\qquad+\int_0^{x_0-x}\left(K\left(|y_0-\delta-y|;\lambda\right)-K\left(\sqrt{(s)^2+(y_0-\delta-y)^2};\lambda\right)\right)\\
&\qquad\times\quad\mu_1'(x_0-x-s)\,ds. \qquad (26)
\end{aligned}$$

Substituting (26) into (25) gives us,

$$\begin{aligned}
&\int_{x_0-\delta}^{x_0}K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-s)\,ds\\
&\quad=\int_{x_0-x-\delta}^{0}K\left(\sqrt{s^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\qquad-\int_0^{x_0-x}K\left(\sqrt{s^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\qquad-\int_{x_0-\delta-x}^{x_0-x}K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\qquad+2\int_0^{x_0-x}K\left(|y_0-\delta-y|;\lambda\right)\mu_1'(x_0-x-s)\,ds\\
&\qquad-\mu_1(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)
\end{aligned}$$

$$= \int_{x_0-x-\delta}^{0} K\left(\sqrt{s^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds$$
$$- \int_{0}^{x_0-x} K\left(\sqrt{s^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds$$
$$- 2K(|y_0-\delta-y|;\lambda)\,\mu_1(x_0-x).$$

Hence we find,

$$\left|\int_{x_0-\delta-x}^{x_0-x} K\left(\sqrt{(s)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(x_0-x-s)\,ds\right|$$
$$\leq \int_{x_0-\delta-x}^{x_0-x} K\left(\sqrt{(s)^2+(y_0-\delta-y)^2};\lambda\right)\mu_1'(|x_0-x-s|)\,ds$$
$$+ 2K(|y_0-\delta-y|;\lambda)\,\mu_1(|x_0-x|)$$
$$= \int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\left|\mu_1'(x_0-s)\right|ds$$
$$+ 2K(|y_0-\delta-y|;\lambda)\,\mu_1(|x_0-x|).$$

Returning back to i_2, we finally obtain

$$|i_2| \leq \varepsilon\mu_1(\delta)\left(\mu_2(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right.$$
$$+ \int_{x_0-\delta}^{x_0} K\left(\sqrt{(s-x)^2+(y_0-\delta-y)^2};\lambda\right)\left|\mu_1'(x_0-s)\right|ds$$
$$\left.+\ 2K(|y_0-\delta-y|;\lambda)\,\mu_1(|x_0-x|)\right).$$

Making similar calculations yields

$$|i_3| \leq \varepsilon\mu_2(\delta)\left(\mu_1(\delta)\,K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\right.$$
$$+ \int_{y_0-\delta}^{y_0} K\left(\sqrt{(x_0-\delta-x)^2+(t-y)^2};\lambda\right)\left|\mu_2'(y_0-s)\right|dt$$
$$\left.+\ 2K(|x_0-\delta-x|;\lambda)\,\mu_2(|y_0-y|)\right).$$

Now from condition (e), (13) and (14), one can easily get

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} \sup_{\delta\le\sqrt{(s-x)^2+(t-y)^2}} K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)=0. \tag{27}$$

Taking conditions (9) and (10) into account we can finally write

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} i_2+i_3+i_4=0. \tag{28}$$

Now, we return back to the integral i_1 in (22). Here, we shall use Lemma 2 in order to get a bound for i_1.

We note that for a function $K\left(\sqrt{s^2+t^2};\lambda\right)$ on $Q_1=<-a,a>\times<-b,b>$ we have

$$\begin{aligned}\bigvee_{Q_1}(K) &= \bigvee(K;<-a,a>,<-b,b>)\\ &= K(a,b)+K(-a,-b)-K(-a,b)-K(a,-b).\end{aligned} \tag{29}$$

So, using Lemma 2 and equality (29) we can write

$$\begin{aligned}|i_1| &= \left|\varepsilon\int_{y_0-\delta}^{y_0}\int_{x_0-\delta}^{x_0}\mu_1(x_0-s)\,\mu_2(y_0-t)\,dK\left(\sqrt{(s-x)^2+(t-y)^2};\lambda\right)\right|\\ &\le \varepsilon\sup\{\mu_1(x_0-s)\,\mu_2(y_0-t)\}\bigvee_{Q}(K)\\ &= \varepsilon\sup\{\mu_1(x_0-s)\,\mu_2(y_0-t)\}\left[K\left(\sqrt{(x_0-x)^2+(y_0-y)^2};\lambda\right)\right.\\ &\quad+K\left(\sqrt{(x_0-\delta-x)^2+(y_0-\delta-y)^2};\lambda\right)\\ &\quad-K\left(\sqrt{(x_0-x)^2+(y_0-\delta-y)^2};\lambda\right)\\ &\quad\left.-K\left(\sqrt{(x_0-\delta-x)^2+(y_0-y)^2};\lambda\right)\right].\end{aligned}$$

Taking condition (e) and property (11) into account we find

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} i_1=0. \tag{30}$$

Consequently, according to (28) and (30) we finally obtain

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda)} I_{121} = 0. \tag{31}$$

Similar to the above procedure the limit of the integrals I_{122}, I_{123}, I_{124} can also be shown to be zero, and this leads to

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} I_{12}\,(x,\,y;\,\lambda) = 0. \tag{32}$$

Hence, from (17) and (32) we get

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} I_{1}\,(x,\,y;\,\lambda) = 0, \tag{33}$$

which implies

$$\lim_{(x,y;\lambda)\to(x_0,y_0;\lambda_0)} L^{*}(f;\,x,\,y;\,\lambda) = f\,(x_0,\,y_0)$$

together with (15), as desired. □

References

1. Gadjiev, A.D.: On the order of convergence of singular integrals depending on two parameters. In: Special Questions of Functional Analysis and its Applications to the Theory of Differential Equations and Functions Theory, pp. 40–44. Baku (1968) (Russian)
2. Jawarneh, Y., Noorani, M.S.M.: Inequalities of Ostrowski and Simpson type for mappings of two variables with bounded variation and applications. TJMM **3**(2), 81–94 (2011)
3. Karsli, H., Ibikli, E.: On convergence of convolution type singular integral operators depending on two parameters. Fasc. Math. **38**, 25–39 (2007)
4. Nessel R.J.: Contributions to the theory saturation for singular integrals in several variables, III, radial kernels. Indag. Math. **29** (Ser.A.), 65–73 (1965)
5. Siudut, S.: A theorem of Romanovski type for double singular integrals. Comment. Math. Prace Mat. **28**, 355–359 (1989)
6. Taberski, R.: Singular integrals depending on two parameters. Roczniki PTM, Seria 1: Prace Mat. **7**, 173–179 (1962)
7. Taberski, R.: On double integrals and Fourier series. Ann. Pol. Math. **15**, 97–1**15** (1964)
8. Yılmaz, M.M., Serenbay, S.K., İbikli, E.: On singular integrals depending on three parameters. App. Math. and Comp. **218**, 1132–1135 (2011)
9. Yılmaz, M.M.: On convergence of singular integrals depending on three parameters with radial kernels. Int. Journal of Math. Analysis **4**(39), 1923–1928 (2010)

Lower Bound on the Dimension of Trivariate Splines on Cells

Jianyun Jimmy Shan

Abstract In [1], Alfeld, Schumaker, and Whiteley determined the generic dimension of the space of C^1 splines of degree $d \geq 8$ on tetrahedral decompositions. In this chapter, we analyze the dimension of C^r, $r = 1, 2$, trivariate splines on cells, which are tetrahedral complexes sharing a single interior vertex. The dimension depends on subtle geometry of the fatpoints corresponding to the configuration of the hyperplanes adjacent to the interior vertex. A key tool is the classification of the relevant fatpoint ideals by Geramita, Harbourne, and Migliore in [2].

Keywords Trivariate splines · Ideals of powers of linear forms · Fat points

1 Introduction

In mathematics, it is often useful to approximate a function f on a region by a "simpler" function. A natural way to do this is to divide the region into simplicies, and then approximate f on each simplex by a polynomial function. A C^r-differentiable piecewise polynomial function on a d-dimensional simplicial complex $\Delta \subset \mathbb{R}^d$ is called a *spline*. The set of splines of degree at most k on Δ form a vector space $C_k^r(\Delta)$. In the case of one-dimensional splines, the dimension and bases for this vector space are completely known. But for higher dimensions, things are more complicated. In the planar case, Alfeld and Schumaker [3] use Bezier-Bernstein techniques to give an explicit formula for the dimension of $C_k^r(\Delta)$ when $k \geq 3r + 1$. In [4], Billera constructed a complex of modules where the spline module $C^r(\Delta)$ appeared as the top homology. Combining this with a vanishing result of Whiteley [5] allowed him to prove a conjecture of Strang [6] on $\dim C_k^1(\Delta)$, for generic complex Δ (that is, complexes where all 2-cells are triangles whose edges are in sufficiently general position).

J. J. Shan (✉)
Department of Mathematics, University of Illinois,
1409 W. Green Street, Urbana, IL 61801, USA
e-mail: shan15@math.uiuc.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_18,

In [7, 8], Schenck and Stillman introduced a chain complex different from that used by Billera, where the top homology also gives the spline module and the lower homologies have nicer properties. Using this, Geramita and Schenck [9] determined the dimension of planar (mixed) splines for sufficiently high degrees. Another interesting aspect of [9] is the use of inverse system between ideals in $\mathbb{R}[x, y]$ generated by powers of homogeneous linear forms and fatpoints in $\mathbb{P}^1$.

In the case of trivariate splines, Alfeld, Schumaker, and Whiteley [1] determined the dimension of C^1 generic tetrahedral splines for degree $d \geq 8$. But for $r > 1$, there is no general formula known. In [10, 11] Alfeld and Schumaker gave upper and lower bounds for $\dim C_k^r(\Delta)$.

It is very natural to first consider some simple tetrahedral complexes, as a first step in understanding splines on general tetrahedral complexes. In this paper, for a tetrahedral complex Δ_v, which consists of several tetrahedra sharing a single interior vertex v, we generalize the approach of Geramita and Schenck [9], and find a lower bound for the $\dim C_k^r(\Delta)$, $r = 1, 2$, see Sect. 6 for a precise statement.

The organization of the paper is as follows. In Sect. 2, we define the spline complex $\mathscr{R}/\mathscr{J}$ and show that $H_1(\mathscr{R}/\mathscr{J}) = H_0(\mathscr{R}/\mathscr{J}) = 0$ and $H_2(\mathscr{R}/\mathscr{J})$ is Artinian (vanishes for high degrees). In Sect. 3, we analyze the dimension of each component of the spline complex, except the last one. In Sect. 4, we review the inverse system between ideals of powers of linear forms and ideals of fatpoints, and the algorithm to compute the dimension of ideals of fatpoints. In Sect. 5, we compute the last component of $\mathscr{R}/\mathscr{J}$ for some examples of tetrahedral complexes. In Sect. 6, we state our main results, compare our bounds with the bounds in the literature, and end with some remarks.

2 Spline Complexes

Let $R = \mathbb{R}[x, y, z]$ be fixed throughout this paper. Our tetrahedral complex Δ_v, which we call a *Cell*, consists of several tetrahedra sharing a single interior vertex v. Following Schenck [7], we define the spline complex $C^r(\Delta_v)$ for any $r \geq 0$.

In general, for a tetrahedral complex Δ, $C^r(\Delta)$ is not a graded module over R and it is convenient to have a graded module to compute the dimension of splines for each degree. Denote by $\hat{\Delta}$ the simplicial complex obtained by embedding the simplicial complex $\Delta \subset \mathbb{R}^3$ in the plane $\{w = 1\} \subset \mathbb{R}^4$ and forming the cone with the origin. Then the set of splines (of all degrees) on $\hat{\Delta}$ is a graded module $C^r(\hat{\Delta})$ over a polynomial ring $S = \mathbb{R}[x, y, z, w]$. We denote its kth graded component by $C^r(\hat{\Delta})_k$. As a vector space, it is isomorphic to the space $C_k^r(\Delta)$ of splines on Δ of degree at most k.

Since there is a single interior vertex v for our tetrahedral complex Δ_v, we can put the vertex v at the origin $O = (0, 0, 0) \in \mathbb{R}^3$, so every linear form defining a hyperplane passing through v will be homogeneous. Thus we do not need to do the above cone construction, $C^r(\Delta)$ is still a graded module over R and $C^r(\Delta)_k$ will be the vector space of splines of smoothness r of degree exactly k.

Let $\Delta = \Delta_v$ in the rest of the paper, unless otherwise stated. Fix an integer $r \geq 0$. Define a complex of ideals of $\mathscr{J}$ on Δ by

$$\begin{aligned} J(\sigma) &= 0, && \text{for } \sigma \in \Delta_3, \\ J(\tau) &= \langle {l_\tau}^{r+1} \rangle, && \text{for } \tau \in \Delta_2^0, \\ J(e) &= \langle {l_\tau}^{r+1} \rangle_{e \in \tau} && \text{for } e \in \Delta_1^0, \\ J(v) &= \langle {l_\tau}^{r+1} \rangle_{v \in \tau} && \text{for } v \in \Delta_0^0. \end{aligned}$$

Here Δ_i^0 are the i-dimensional interior faces of Δ, and we consider all the tetrahedra Δ_3 as interior. l_τ is the homogeneous linear form in R defining the affine hull of τ. We denote h_e and h_v as the number of distinct hyperplanes (not triangular faces) incident to e and v, respectively. Then $J(e)$ is an ideal generated by h_e powers of linear forms, and similarly $J(v)$ is generated by h_v powers of linear forms.

We also define the constant complex $\mathscr{R}$ on Δ by $\mathscr{R}(\sigma) = R$ for each face $\sigma \in \Delta$ with the boundary map ∂_i to be the usual simplicial boundary map. We get the following quotient complex $\mathscr{R}/\mathscr{J}$:

$$0 \to \sum_{\sigma \in \Delta_3} R \xrightarrow{\partial_3} \sum_{\tau \in \Delta_2^0} R/J(\tau) \xrightarrow{\partial_2} \sum_{e \in \Delta_1^0} R/J(e) \xrightarrow{\partial_1} R/J(v) \to 0. \tag{1}$$

Lemma 1 *$H_1(\mathscr{R}/\mathscr{J}) = H_0(\mathscr{R}/\mathscr{J}) = 0$, and $H_2(\mathscr{R}/\mathscr{J})$ is Artinian.*

Proof If we form the cone $\hat{\Delta}$, and define the constant complex $\mathscr{S}$ on Δ by $\mathscr{S}(\Delta) = S$ for each face $\sigma \in \Delta$, we get the quotient complex $\mathscr{S}/\mathscr{J}$, see [7]:

$$0 \to \sum_{\sigma \in \Delta_3} S \xrightarrow{\partial_3} \sum_{\tau \in \Delta_2^0} S/J(\tau) \xrightarrow{\partial_2} \sum_{e \in \Delta_1^0} S/J(e) \xrightarrow{\partial_1} S/J(v) \to 0.$$

Since

$$\mathscr{S}/\mathscr{J} = \mathscr{R}/\mathscr{J} \otimes_R R[w],$$

we have

$$\begin{aligned} H_1(\mathscr{S}/\mathscr{J}) &= H_1(\mathscr{R}/\mathscr{J}) \otimes_R R[w], \\ H_2(\mathscr{S}/\mathscr{J}) &= H_2(\mathscr{R}/\mathscr{J}) \otimes_R R[w]. \end{aligned}$$

By Lemma 3.1 in [7], $\dim H_2(\mathscr{S}/\mathscr{J}) \leq 1$, so we have

$$\dim H_2(\mathscr{R}/\mathscr{J}) \leq 0.$$

Similarly, $\dim H_1(\mathscr{S}/\mathscr{J}) \leq 0$ implies that

$$H_1(\mathscr{R}/\mathscr{J}) = H_0(\mathscr{R}/\mathscr{J}) = 0.$$

This completes the proof.

3 Dimension of Graded Components of the Modules

It is well known that

$$\dim R_k = \binom{k+2}{2}. \tag{2}$$

Since $J(\tau)$ is a principal ideal generated by an element of degree $r+1$, we also have

$$\dim(R/J(\tau))_k = \binom{k+2}{2} - \binom{k-r+1}{2}. \tag{3}$$

3.1 The Case $r = 1$

To compute $\dim(R/J(e))_k$, we use the minimal free resolution of the ideal $J(e)$.

Lemma 2 *The minimal free resolution of $J(e)$ is given by*

$$\begin{aligned} &0 \to R(-4) \to R(-2)^2 \to R \to R/J(e) \to 0, \quad &&\text{if } h_e = 2,\\ &0 \to R(-3)^2 \to R(-2)^3 \to R \to R/J(e) \to 0, \quad &&\text{if } h_e \geq 3. \end{aligned}$$

So we get

$$\dim(R/J(e))_k = \begin{cases} \binom{k+2}{2} - 2\binom{k}{2} + \binom{k-2}{2}, & \text{if } h_e = 2,\\ \binom{k+2}{2} - 3\binom{k}{2} + 2\binom{k-1}{2}, & \text{if } h_e \geq 3. \end{cases} \tag{4}$$

Proof If $h_e = 2$, then $J(e)$ is a complete intersection, generated by two quadratics. If $h_e \geq 3$, then $J(e) = \langle l_1^2, l_1 l_2, l_2^2 \rangle$, and the result follows.

Similarly, we can analyze the ideal $J(v)$, which is generated by squares of the linear forms which define the hyperplanes passing through v. Since the dimension of quadratic forms in R is 6, we only need to consider the case $h_v \leq 6$. If $h_v \geq 6$, then $J(v) = \langle x^2, y^2, z^2, xy, xz, yz \rangle$, so

$$\dim(R/J(v))_k = \begin{cases} 1, & \text{if } \ k = 0,\\ 3, & \text{if } \ k = 1,\\ 0, & \text{if } \ k \geq 2. \end{cases} \tag{5}$$

This is actually the case of Clough-Tocher in Example 5.

At the other extreme, if $h_v = 3$, then

$$J(v) = \langle x^2, y^2, z^2 \rangle,$$

and therefore,

k	0	1	2	3	≥ 4
$\dim(R/J(v))_k$	1	3	3	1	0

We are thus left with the case $h_v = 4$, or 5. If $h_v = 4$, suppose the four hyperplanes passing through v are defined by l_1, l_2, l_3, l_4, so the ideal $J(v) = \langle l_1^2, l_2^2, l_3^2, l_4^2 \rangle$. After a change of variables,

$$J(v) = \langle x^2, y^2, z^2, l^2 \rangle,$$

for some linear form l in x, y, z. This is an example of an almost complete intersection, whose Hilbert series are given by Iarrobino (Lemma C of [12]), giving

$$\dim(R/J(v))_k = \begin{cases} 1, & \text{if } k = 0, \\ 3, & \text{if } k = 1, \\ 2, & \text{if } k = 2, \\ 0, & \text{if } k \geq 3. \end{cases} \tag{6}$$

For $h_v = 5$, there are more variations, depending on the five linear forms defining the hyperplanes passing through v. After a change of variables, we may assume the linear forms are given by $x, y, z, l_1(x, y, z), l_2(x, y, z)$. If the linear forms l_1, l_2 only involve two variables, say x, y, (see Example 3), then

$$J(v) = \langle x^2, y^2, xy, z^2 \rangle,$$

and the $\dim(R/J(v))_k$ is the same as in (6).

In the other cases, we have not been able to analyze the ideal $J(v)$, though we can still compute a Grobner basis and find the dimension as given by

$$\dim(R/J(v))_k = \begin{cases} 1, & \text{if } k = 0, \\ 3, & \text{if } k = 1, \\ 1, & \text{if } k = 2, \\ 0, & \text{if } k \geq 3. \end{cases} \tag{7}$$

We can also get the above formulas of $\dim(R/J(v))_k$ using fatpoints as in Sect. 4.

3.2 The Case $r = 2$

Lemma 3 *The minimal free resolution of $J(e)$ is given by*

$$\begin{aligned} &0 \to R(-6) \to R(-3)^2 \to R \to R/J(e) \to 0, && \text{if } h_e = 2, \\ &0 \to R(-4) \oplus R(-5) \to R(-3)^3 \to R \to R/J(e) \to 0, && \text{if } h_e = 3, \\ &0 \to R(-4)^3 \to R(-3)^4 \to R \to R/J(e) \to 0, && \text{if } h_e \geq 4. \end{aligned}$$

So we get

$$\dim(R/J(e))_k = \begin{cases} \binom{k+2}{2} - 2\binom{k-1}{2} + \binom{k-4}{2}, & \text{if } h_e = 2, \\ \binom{k+2}{2} - 3\binom{k-1}{2} + \binom{k-2}{2} + \binom{k-3}{2}, & \text{if } h_e = 3, \\ \binom{k+2}{2} - 4\binom{k-1}{2} + 3\binom{k-2}{2}, & \text{if } h_e \geq 4. \end{cases} \quad (8)$$

Proof Notice that the ideal $J(e)$ is of codimension 2 in R, so we can apply Hilbert-Burch Theorem [13]. There are 3 cases:

Case 1: $h_e = 2$. This is similar to the case $r = 1$, but $J(e)$ is a complete intersection of two cubics.

Case 2: $h_e = 3$. Suppose the linear forms are given by l_1, l_2 and $l_3 = al_1 + bl_2$. Then it is not hard to see the linear syzygy of l_1^3, l_2^3, l_3^3 is given by

$$-a^3(al_1 + 2bl_2)l_1^3 + b^3(2al_1 + bl_2)l_2^3 + (al_1 - bl_2)l_3^3 = 0,$$

and the quadratic syzygy is given by

$$(a^3l_2^2)l_1^3 + (2a^2bl_1^2 + 2ab^2l_1l_2 + b^3l_2^2)l_2^3 + (-l_2^3)l_3^3 = 0.$$

Then the minimal free resolution of $J(e)$ is given by

$$0 \to R(-4) \oplus R(-5) \xrightarrow{\varphi} R(-3)^3 \xrightarrow{\langle l_1^3, l_2^3, l_3^3 \rangle} R \to R/J(e) \to 0,$$

where

$$\varphi = \begin{bmatrix} -a^3(al_1 + 2bl_2) & a^3l_2^2 \\ b^3(2al_1 + bl_2) & 2a^2bl_1^2 + 2ab^2l_1l_2 + b^3l_2^2 \\ al_1 - bl_2 & -l_2^3 \end{bmatrix}.$$

Case 3: $h_e \geq 4$. Suppose the hyperplanes incident to e are given by $l_1, l_2, \ldots, l_s$, where $l_i = a_il_1 + b_il_2$ for $i \geq 3$. Then it is easy to see the ideal $J(e) = \langle l_1^3, l_1^2l_2, l_1l_2^2, l_2^3 \rangle$, so the minimal free resolution of $J(e)$ is given by

$$0 \to R(-4)^3 \xrightarrow{\psi} R(-3)^4 \xrightarrow{\langle l_1^3, l_1^2l_2, l_1l_2^2, l_2^3 \rangle} R \to R/J(e) \to 0,$$

where

$$\psi = \begin{bmatrix} -l_2 & 0 & 0 \\ l_1 & -l_2 & 0 \\ 0 & l_1 & -l_2 \\ 0 & 0 & l_1 \end{bmatrix}.$$

This completes the proof. □

As in the case above, though the number of hyperplanes passing through v may be big, the dimension of $R/J(v)$ only depends on the ideal $J(v)$. Since the dimension of cubic forms in R is 10, we only need to consider the case $h_v \leq 10$.

Example 1 If $h_v \geq 10$, then $J(v) = \langle x, y, z\rangle^3$ and

k	0	1	2	≥3
$\dim(R/J(v))_k$	1	3	6	0

At the other extreme, if $h_v = 3$, then $J(v) = \langle x^3, y^3, z^3\rangle$, so

k	0	1	2	3	4	5	6	≥7
$\dim(R/J(v))_k$	1	3	6	7	6	3	1	0

We are thus left to consider the possibilities for $h_v \in \{4, 5, .., 9\}$. We use the inverse system dictionary to translate this into a question about the Hilbert function of h_v fatpoints on $\mathbb{P}^2$. Interestingly, there are two distinct cases.

Case 1: $h_v \in \{4, \ldots, 8\}$. In this case, we can give a complete answer to the dimension of $(R/J(v))_k$ for each degree k.

Case 2: $h_v = 9$. There are two subcases depending on whether the cone of numerically effective classes of divisors on the surface obtained by blowup $\mathbb{P}^2$ at the 9 points is finitely generated or not. If the cone is finitely generated, then Harbourne's algorithm, which we will give below (Sect. 4.2), still works and enables us to compute the Hilbert function of fatpoints, thus $\dim(R/J(v))_k$, for each k. However, if the cone is not finitely generated, it is a famous open problem in algebraic geometry (see Miranda's survey article [14]), and therefore the same difficulty to compute $\dim(R/J(v))_k$. However, in any specific case, the dimension may be calculated using Macaulay2 [15].

4 Review of Inverse System and Fatpoints on $\mathbb{P}^2$

4.1 Inverse System

In [16], Emsalem and Iarrobino proved there is a close connection between ideals generated by powers of linear forms and ideals of fatpoints. We use their results in the special case of ideals generated by powers of linear forms in 3 variables and

ideals of fatpoints in $\mathbb{P}^2$, see [9, 17] for more details. Let $p_1, \ldots, p_n \in \mathbb{P}^2$ be a set of distinct points,

$$p_i = [p_{i1} : p_{i2} : p_{i3}],$$
$$I(p_i) = \wp_i \subseteq R' = k[x', y', z'].$$

A fat point ideal is an ideal of the form

$$F = \bigcap_{i=1}^{n} \wp_i^{\alpha_i+1} \subset R'. \tag{9}$$

We define

$$L_{p_i} = p_{i1}x + p_{i2}y + p_{i3}z \in R, \text{for} 1 \leq i \leq n. \tag{10}$$

Define an action of R' on R by partial differentiation:

$$p(x', y', z') \cdot q(x, y, z) = p(\partial/\partial x, \partial/\partial y, \partial/\partial z)q(x, y, z). \tag{11}$$

Since F is a submodule of R', it acts on R. The set of elements annihilated by the action of F is denoted by F^{-1}.

Theorem 1 (Emsalem and Iarrobino [16]) *Let F be an ideal of fatpoints*

$$F = \bigcap_{i=1}^{n} \wp_i^{\alpha_i+1}.$$

Then

$$(F^{-1})_j = \begin{cases} R_j, & \text{for } \ j \leq \max\{\alpha_i\}, \\ L_{p_1}^{j-\alpha_1} R_{\alpha_1} + \cdots, L_{p_n}^{j-\alpha_n} R_{\alpha_n}, & \text{for } \ j \geq \max\{\alpha_i + 1\}. \end{cases} \tag{12}$$

and

$$\dim_k (F^{-1})_j = \dim_k (R/F)_j.$$

Corollary 1 *In the case $r = 1, 2$, let*

$$F = \wp_1^{j-r} \cap \cdots \cap \wp_n^{j-r}$$

be an ideal of fatpoints on $\mathbb{P}^2$. Then $(F^{-1})_j = \langle L_{p_1}^{r+1}, \ldots, L_{p_n}^{r+1} \rangle_j$, and

$$\dim(R/J(v))_j = \begin{cases} \binom{j+2}{2} & \text{for } \ 0 \leq j \leq r, \\ \dim F_j & \text{for } \ j \geq r+1. \end{cases} \tag{13}$$

Therefore, to obtain the dimension of $(R/J(v))_k$, for each k, it is necessary to consider a corresponding ideal of fatpoints on $\mathbb{P}^2$.

4.2 Blowup of Points in $\mathbb{P}^2$

Here we will use some facts about rational surfaces obtained by blowup of n points $p_1, \ldots, p_n$ on $\mathbb{P}^2$, see Hartshorne [18]. We follow Harbourne [2] and only state what is needed in this paper.

There is a well-known correspondence between the graded pieces of an ideal of fat points $F \subset R$ and the global sections of a line bundle on the surface X which is the blowup of $\mathbb{P}^2$ at the points. Let E_i be the class of the exceptional divisor over the point p_i, and L the pullback of a line on $\mathbb{P}^2$. For the fatpoint ideal F in Corollary 1, define

$$D_j = jL - (j-r)(E_1 + \cdots + E_n). \tag{14}$$

Then $\dim J_j = h^0(D_j)$, and thus we have

$$\dim(R/J(v))_j = \begin{cases} \binom{j+2}{2}, & \text{for } \ 0 \le j \le r, \\ h^0(D_j), & \text{for } \ j \ge r+1. \end{cases} \tag{15}$$

Remark 1 This equation tells us that $\dim(R/J(v))_j$ only depends on the divisor D_j, which only depends on the configuration of the fatpoints, and thus only depends on the geometry of the hyperplanes passing through v. See Sect. 5 for examples.

On X, the divisor class group $Cl(X)$ is a free abelian group with basis $L, E_1, \ldots, E_n$ which has the intersection product

$$L^2 = -E_i^2 = 1, \qquad L.E_i = E_j.E_i = 0, \qquad \text{for } \ j \ne i. \tag{16}$$

The canonical class of X is

$$K_X = -3L + E_1 + \cdots + E_n.$$

We also define

$$A_n = (n-2)L - K_X.$$

A prime divisor is the class of a reduced irreducible curve on X, and an effective divisor is a nonnegative integer combination of prime divisors. We denote the set of effective divisors by $EFF(X)$. A divisor whose intersection product with every effective divisor is ≥ 0 is called *numerically effective*(nef). We define $Neg(X)$ as the classes of prime divisors C with $C^2 < 0$. In [2] Proposition 3.1 and 4.1, $Neg(X)$ is explicitly determined, which is the main point for the following algorithm of

Geramita, Harbourne, and Migliore to compute $h^0(F)$ for any divisor F on X. To determine $Neg(X)$, we first define a few classes of divisors on X.

1. $\mathscr{B}_r = \{E_1, \ldots, E_r\}$;
2. $\mathscr{L}_r = \{L - E_{i_1} - \cdots - E_{i_j} | 2 \le j, 0 < i_1 < \cdots < i_j \le r\}$;
3. $\mathscr{Q}_r = \{2L - E_{i_1} - \cdots - E_{i_j} | 5 \le j \le r\}$;
4. $\mathscr{C}_r = \{3L - 2E_{i_1} - E_{i_2} - \cdots - E_{i_j} | 7 \le j \le 8, j \le r\}$;
5. $\mathscr{M}_8 = \{4L - 2E_{i_1} - 2E_{i_2} - 2E_{i_3} - E_{i_4} - \cdots - E_{i_8}, 5L - 2E_{i_1} - 2E_{i_2} - \cdots - 2E_{i_6} - E_{i_7} - E_{i_8}, 6L - 2E_{i_1} - 2E_{i_2} - \cdots - 2E_{i_8}\}$.

Let $\mathscr{N}_r = \mathscr{B}_r \cup \mathscr{L}_r \cup \mathscr{Q}_r \cup \mathscr{C}_r \cup \mathscr{M}_8$. Let X be obtained by blowing up $2 \le r \le 8$ distinct points of $\mathbb{P}^2$. Then

$$Neg(X) \subset \mathscr{N}_r.$$

and

$$Neg(X) = neg(X) \cup \{C \in \mathscr{N}_r | C^2 = -1, C \cdot D \ge 0, \text{for all } D \in neg(X)\},$$

where $neg(X)$ is the subset of $Neg(X)$ of classes of those C with $C^2 = C \cdot C < -1$.

Remark 2 In any given case, we can list the five classes of divisors on X, and $Neg(X)$ is the union of the classes C which has $C^2 < -1$ and the classes C' which has $C'^2 = -1$ and $C' \cdot D \ge 0$ for all $D \in neg(X)$. The classes in $\mathscr{L}_r$ are the pullback of a line passing through the points $p_{i_1}, \ldots, p_{i_j}$ if they are on a line; similarly, the classes in $\mathscr{Q}_r$ are the pullback of a conic passing through the points $p_{i_1}, \ldots, p_{i_j}$ if they are on a conic; and so on. The computation of C^2 and $C \cdot D$ just uses the intersection product, see Eq. (16). See also examples in Sect. 5.

Once we have determined $Neg(X)$, we can use the following algorithm due to Geramita et al. [2] to compute $h^0(F)$ for any class F on X.

Algorithm:
Start with $H = F, N = 0$.
If $H.C < 0$ for some $C \in Neg(X)$, replace H by $H - C$ and replace N by $N + C$.
Eventually either $H.A_n < 0$ or $H.C \ge 0$ for all $C \in Neg(X)$.
In the first case, F is not effective, and $h^0(F) = 0$.
In the latter case, H is nef and effective, and we have a Zariski decomposition

$$F = H + N,$$

with

$$h^0(F) = h^0(H) = (H^2 - H.K_X)/2 + 1.$$

Remark 3 The above algorithm is based on Bezout considerations. See Miranda [14] for an elementary exposition.

5 Examples of Fatpoint Computation

In this section, we will apply the above algorithm to compute $\dim(R/J(v))_j$, depending on the number of hyperplanes h_v passing through v, where $h_v \in \{4, 5, 6, 7, 8\}$. We mainly consider the case $r = 2$ and indicate the similar computation for $r = 1$ in remarks. We demonstrate the computation with examples of tetrahedral complexes constructed from the standard octahedron Δ by perturbing a vertex to get different numbers of hyperplanes passing through O. A key point is to determine $Neg(X)$ in each case, where X as above, is the blowup of $\mathbb{P}^2$ at the fatpoints corresponding to the linear forms defining h_v hyperplanes. For concreteness, we give the coordinates of the vertices of Δ as $O = (0, 0, 0)$, $P_1 = (10, 0, 0)$, $P_2 = (0, 10, 0)$, $P_3 = (-10, 0, 0)$, $P_4 = (0, -10, 0)$, $P_5 = (0, 0, 10)$, $P_6 = (0, 0, -10)$. As said in the Remark 1, the result does not depend on the actual coordinates.

Example 2 (*4 hyperplanes*) By perturbing one vertex along one of the edges, we get an example with 4 hyperplanes. For example, move P_1 along the edge P_1P_2 to get $P_1' = (7, 3, 0)$. Then there are 3 hyperplanes passing through the interior edge OP_5 with defining equations and the corresponding 3 points in $\mathbb{P}^2$ as follows.

$$\begin{aligned} l_1 = x &\longleftrightarrow Q_1 = [1 : 0 : 0], \\ l_2 = y &\longleftrightarrow Q_2 = [0 : 1 : 0], \\ l_3 = 3x - 7y &\longleftrightarrow Q_3 = [3 : -7 : 0]. \end{aligned}$$

The points Q_1, Q_2, Q_3 are colinear. The other hyperplane defined by

$$l_4 = z \longleftrightarrow Q_4 = [0 : 0 : 1].$$

Q_4 is not colinear with the other 3 points. So on the surface X, the divisor

$$C_1 = L - E_1 - E_2 - E_3 \in Neg(X),$$

where L is the pullback of a line on $\mathbb{P}^2$ and E_i is the exceptional divisor corresponding to Q_i for $i = \{1, 2, 3, 4\}$. In fact,

$$Neg(X) = \{C_1, L - E_1 - E_4, L - E_2 - E_4, L - E_3 - E_4, E_1, E_2, E_3, E_4\}.$$

Define D_j as in Eq. (14),

$$D_j = jL - (j - 2)(E_1 + E_2 + E_3 + E_4).$$

Remark 4 In this example, $\mathscr{N}_r = \mathscr{B}_r \cup \mathscr{L}_r$. It seems that we should include $L - E_1 - E_2$, $L - E_1 - E_3$, $L - E_2 - E_3$ in $Neg(X)$. However, these classes are not in $Neg(X)$, because they are not prime. For example,

$$L - E_1 - E_2 = (L - E_1 - E_2 - E_3) + E_3$$

is a sum of two prime divisors. The class $L - E_1 - E_2 - E_3$ is prime because the points Q_1, Q_2, Q_3 are colinear. Moreover, $neg(X) = \emptyset$.

Let's just show that $h^0(D_4) = 4$ as a sample computation, using the intersection product. First,

$$\begin{aligned} D_4 &= 4L - 2(E_1 + E_2 + E_3 + E_4), \\ D_4.C_1 &= 4L^2 + 2E_1^2 + 2E_2^2 + 2E_3^2 \\ &= 4 - 2 - 2 - 2 = -2 < 0. \end{aligned}$$

So we take

$$D_4' = D_4 - C_1 = 3L - E_1 - E_2 - E_3 - 2E_4.$$

It is easy to check that

$$D_4' \cdot C \geq 0, \text{ for any } C \in Neg(X),$$

and therefore D_4' is *nef* and effective. So the Zariski decomposition of D_4 is

$$D_4 = D_4' + C_1.$$

Using the intersection product (16) again, we have

$$\begin{aligned} D_4'^2 &= (3L)^2 + E_1^2 + E_2^2 + E_4^2 + (2E_3)^2 \\ &= 9 - 1 - 1 - 1 - 4 \\ &= 2 \end{aligned}$$

Similarly,

$$\begin{aligned} K_X &= -3L + E_1 + E_2 + E_3 + E_4, \\ D_4'.K_X &= -(3L)^2 - E_1^2 - E_2^2 - E_4^2 - 2E_3^2 \\ &= -9 + 1 + 1 + 1 + 2 = -4. \end{aligned}$$

So we get

$$h^0(D_4) = h^0(D_4') = (D_4'^2 - D_4'.K_X)/2 + 1 = 4.$$

A similar computation shows the Zariski decomposition of D_5 is

$$D_5 = D_5' + 2C_1,$$

where

$$D_5' = 3L - E_1 - E_2 - E_3 - 3E_4,$$

and

$$h^0(D_5) = h^0(D_5') = 1.$$

Summarizing, we have

$$\dim(R/J(v))_j = h^0(D_j) = \begin{cases} 6, & \text{for } j = 3, \\ 4, & \text{for } j = 4, \\ 1, & \text{for } j = 5, \\ 0, & \text{for } j \geq 6. \end{cases} \tag{17}$$

Example 3 (*5 hyperplanes*: Δ_1) By perturbing P_2, P_3 on the plane $z = 0$, there are 4 hyperplanes passing through the interior edge OP_5(or OP_6), so there are 4 corresponding points Q_1, Q_2, Q_3, Q_4 on $\mathbb{P}^2$ which lie on a line l. There is another point $Q_5 = [0 : 0 : 1]$ corresponding to the plane $z = 0$, not lying on l. On the surface X from blowup of the $Q_i's$, as above, E_i corresponds to Q_i, for $1 \leq i \leq 5$, the divisor class

$$C_1 = L - E_1 - E_2 - E_3 - E_4 \in Neg(X).$$

In fact,

$$Neg(X) = \{C_1, L - E_i - E_5, E_i, E_5, i \in \{1, 2, 3, 4\}\}.$$

We also have

$$D_j = jL - (j-2)(E_1 + E_2 + E_3 + E_4 + E_5).$$

We analyze the case $j = 4$ in detail, since it is similar for any j. First,

$$\begin{aligned} D_4 &= 4L - 2(E_1 + E_2 + E_3 + E_4 + E_5), \\ D_4.C_1 &= 4L^2 + 2E_1^2 + 2E_2^2 + 2E_3^2 + 2E_4^2 \\ &= 4 - 2 - 2 - 2 - 2 = -4 < 0, \end{aligned}$$

and so we take

$$D_4' = D_4 - C_1 = 3L - E_1 - E_2 - E_3 - E_4 - 2E_5.$$

Moreover,

$$\begin{aligned} D_4'.C_1 &= 3L^2 + E_1^2 + E_2^2 + E_3^2 + E_4^2 \\ &= 3 - 1 - 1 - 1 - 1 = -1 < 0. \end{aligned}$$

So we subtract C_1 from D_4' to get

$$D_4'' = D_4' - C_1 = 2L - 2E_5.$$

Now, we can check

$$D_4''.C_1 = 2L^2 = 2 > 0.$$

In fact, $D_4''.C \geq 0$ for any $C \in Neg(X)$. Therefore, we have the Zariski Decomposation of D_4 as

$$D_4 = D_4'' + 2C_1.$$

A similar computation will show that,

$$D_4''^2 = 0, \qquad D_4''.K_X = -4.$$

So we get

$$h^0(D_4) = h^0(D_4'') = 3.$$

Summarizing, we have

$$\dim(R/J(v))_j = h^0(D_j) = \begin{cases} 5, & \text{for } j = 3, \\ 3, & \text{for } j = 4, \\ 0, & \text{for } j \geq 5. \end{cases} \tag{18}$$

Remark 5 We have given the formula of $\dim(R/J(v))_k$ in Eq. (6) for the case $r = 1$, by applying a result of Iarrobino [12]. Here we reprove that formula using a similar computation as above. Since $r = 1$, the divisor D_j is given by

$$D_j = jL - (j-1)(E_1 + \cdots + E_5).$$

By Corollary 1, we just need to compute $\dim(R/J(v))_j$, or equivalently $h^0(D_j)$ for $j \geq 2$. For $j = 2$, $D_2 = 2L - (E_1 + \cdots + E_5)$. Since

$$D_2 \cdot C_1 = 2L^2 + E_1^2 + \cdots + E_4^2 = 2 - 4 = -2 < 0,$$

we get $D_2' = D_2 - C_1 = L - E_5$, which is effecitve. Since

$$D_2'^2 = 0, \text{ and } D_2' \cdot K_X = -3 + 1 = -2,$$

we get

$$h^0(D_4) = h^0(D_4') = \frac{0 - (-2)}{2} + 1 = 2.$$

The computation for $j > 2$ is completely similar.

Example 4 (*5 hyperplanes*: Δ_2) By perturbing one vertex along the interior of a face, we can get another example of 5 hyperplanes. For example, if we perturb $P_5 = (0, 0, 10)$ to $P_5' = (1, 1, 8)$, then there are 3 hyperplanes passing through the interior edge OP_1(or OP_3) with defining equations and the corresponding 3 points in $\mathbb{P}^2$ as follows:

$$\begin{aligned} l_1 = z &\longleftrightarrow Q_1 = [0:0:1], \\ l_2 = y &\longleftrightarrow Q_2 = [0:1:0], \\ l_3 = 8y - z &\longleftrightarrow Q_3 = [0:8:-1]. \end{aligned}$$

The points Q_1, Q_2, Q_3 are colinear in $\mathbb{P}^2$. Similarly, through the interior edge OP_2 (or OP_4), there are 3 hyperplanes

$$\begin{aligned} l_1 = z &\longleftrightarrow Q_1 = [0:0:1], \\ l_4 = x &\longleftrightarrow Q_4 = [1:0:0], \\ l_5 = 8x - z &\longleftrightarrow Q_5 = [8:0:-1]. \end{aligned}$$

Similarly, Q_1, Q_4, Q_5 are colinear and Q_1 is the intersection of the two lines. So, on the surface X, the two divisors

$$\begin{aligned} C_1 &= L - E_1 - E_2 - E_3, \\ C_2 &= L - E_1 - E_4 - E_5, \end{aligned}$$

are in $Neg(X)$, where E_i is the exceptional divisor corresponding to Q_i for $i = \{1, 2, 3, 4, 5\}$. In this case, $Neg(X)$ given by

$$\{C_1, C_2, L-E_2-E_4, L-E_2-E_5, L-E_3-E_4, L-E_3-E_5, E_i, i = \{1, 2, 3, 4, 5\}\}.$$

We also have

$$D_j = jL - (j-2)(E_1 + E_2 + E_3 + E_4 + E_5).$$

It is easy to check that D_3 is *nef*, and a similar computation shows

$$D_3^2 = 4, \qquad D_3.K_X = -4,$$

and so

$$h^0(D_3) = 5.$$

For

$$D_4 = 4L - 2(E_1 + E_2 + E_3 + E_4 + E_5),$$

we have

$$\begin{aligned} D_4.C_1 &= 4L^2 + 2E_1^2 + 2E_2^2 + 2E_3^2 \\ &= 4 - 2 - 2 - 2 < 0, \end{aligned}$$

and so we take

$$D_4' = D_4 - C_1 = 3L - E_1 - E_2 - E_3 - 2E_4 - 2E_5.$$

Since

$$\begin{aligned} D_4'.C_2 &= 3L^2 + E_1^2 + 2E_4^2 + 2E_5^2 \\ &= 3 - 1 - 2 - 2 < 0, \end{aligned}$$

and so we take

$$D_4'' = D_4' - C_2 = 2L - E_2 - E_3 - E_4 - E_5.$$

It is easy to check D_4'' is *nef*, so we get the Zariski decomposition

$$D_4 = D_4'' + C_1 + C_2.$$

A similar computation will show that,

$$D_4''^2 = 0, \qquad D_4''.K_X = -2.$$

Thus,

$$h^0(D_4) = h^0(D_4') = 2.$$

For $j \geq 5$, D_j is not effective, so $h^0(D_j) = 0$.

Summarizing, we have

$$\dim(R/J(v))_j = h^0(D_j) = \begin{cases} 5, & \text{for } j = 3, \\ 2, & \text{for } j = 4, \\ 0, & \text{for } j \geq 5. \end{cases} \tag{19}$$

Remark 6 Comparing Example 3 and Example 4, the $\dim(R/J(v))_j$ differ at $j = 4$, even though in both examples, $J(v)$ is an ideal generated by 5 powers of linear forms in x, y, z.

Remark 7 In Eq. (7), we have given a formula of $\dim(R/J(v))_k$ in the case $r = 1$. Here we prove that formula using the same computation. For $k = 2$, we consider the divisor $D_2 = 2L - (E_1 + \cdots + E_5)$. Now

$$D_2 \cdot C_1 = 2 - 3 = -1 < 0,$$

so we get $D_2' = D_2 - C_1 = L - E_4 - E_5$. It is easy to check that $D_2' \cdot D \geq 0$, for all $D \in Neg(X)$, so D_2' is *nef*. Since

$$D_2'^2 = -1, \text{ and } D_2' \cdot K_X = -3 + 2 = -1,$$

we get

$$h^0(D_2) = h^0(D_2') = \frac{-1 - (-1)}{2} + 1 = 1.$$

Example 5 (*6 hyperplanes*: *Clough-Tocher*(*CT*)) This tetrahedral complex *CT* is constructed by choosing an interior point *O*, which we put at the origin $(0, 0, 0)$, in the tetrahedron and decomposing the tetrahedron into four tetrahedra. Through each interior edge of *CT*, there are 3 different hyperplanes, each corresponding to a point in $\mathbb{P}^2$. So we have 4 lines in $\mathbb{P}^2$, with each line corresponding to an interior edge of *CT*, and on each line, there are exactly 3 points. Moreover, each point is the intersection of two lines. For example, through the interior edge OP_1, we have the planes OP_1P_2,OP_1P_3 and OP_1P_4, each corresponding to a point, say Q_1,Q_2 and Q_3 in $\mathbb{P}^2$. Similarly, around OP_2,OP_3 and OP_4, we have the following corresponding points:

OP_2	OP_3	OP_4
$OP_2P_1 \longleftrightarrow Q_1$	$OP_3P_1 \longleftrightarrow Q_4$	$OP_4P_1 \longleftrightarrow Q_5$
$OP_2P_3 \longleftrightarrow Q_4$	$OP_3P_2 \longleftrightarrow Q_2$	$OP_4P_2 \longleftrightarrow Q_3$
$OP_2P_4 \longleftrightarrow Q_5$	$OP_3P_4 \longleftrightarrow Q_6$	$OP_4P_3 \longleftrightarrow Q_6$

The configuration of the 6 points on $\mathbb{P}^2$ is type 10 in the Table of [2]. So on the surface *X* obtained from the blowup the six points, we have the following class of divisors in $Neg(X)$:

$$C_1 = L - E_1 - E_2 - E_3, \quad C_2 = L - E_1 - E_4 - E_5,$$
$$C_3 = L - E_2 - E_4 - E_6, \quad C_4 = L - E_3 - E_5 - E_6,$$

with E_i as the exceptional divisor from blowup of Q_i, for $1 \leq i \leq 6$. In this case, D_3 is *nef*, with

$$D_3^2 = 3, \quad D_3.K_X = -3,$$

and so $h^0(D_3) = 4$. As for D_4, the Zariski decomposition is

$$D_4 = 0 + C_1 + C_2 + C_3 + C_4,$$

and so $h^0(D_4) = h^0(0) = 1$. Summarizing, we have

$$\dim(R/J(v))_j = h^0(D_j) = \begin{cases} 4, & \text{for } j = 3, \\ 1, & \text{for } j = 4, \\ 0, & \text{for } j \geq 5. \end{cases} \tag{20}$$

6 Main Result

6.1 Theorems

Now we have computed the dimension for each component of the complex $\mathscr{R}/\mathscr{J}$. Putting the results together, we get our main result.

For a tetrahedral complex $\Delta = \Delta_v$, denote the number of tetrahedra by f_3, the number of 2-dimensional interior faces passing through v by f_2, the number of interior edges with $h_e = 2$, $h_e = 3$, and $h_e \geq 4$,, respectively, by $f_{1,2}$,$f_{1,3}$, and $f_{1,4}$. Recall that h_e is the number of distinct hyperplanes incident to e. Let f_1 be the number of interior edges, so $f_1 = f_{1,2} + f_{1,3} + f_{1,4}$.

Theorem 2 *The dimension of $C^1(\Delta)_k$, the vector space of splines of smoothness $r = 1$ of degree exactly k, is given by*

$$\dim C^1(\Delta)_k = h_{2,k} + C_k,$$

where

$$h_{2,k} = \dim H_2(\mathscr{R}/\mathscr{J})_k,$$

$$\begin{aligned} C_k = & f_3\binom{k+2}{2} - f_2\left[\binom{k+2}{2} - \binom{k}{2}\right] \\ & + f_{1,2}\left[\binom{k+2}{2} - 2\binom{k}{2} + \binom{k-2}{2}\right] \\ & + (f_{1,3} + f_{1,4})\left[\binom{k+2}{2} - 3\binom{k}{2} + 2\binom{k-1}{2}\right] \\ & - \dim(R/J(v))_k, \end{aligned}$$

and $\dim(R/J(v))_k$ *is given by* (5)*,* (6) *and* (7)*, and can be explicitly computed using the method of* Sect. 4.

Theorem 3 *The dimension of $C^2(\Delta)_k$, the vector space of splines of smoothness $r = 2$ of degree exactly k, is given by*

$$\dim C^2(\Delta)_k = h_{2,k} + D_k,$$

where

$$h_{2,k} = \dim H_2(\mathscr{R}/\mathscr{J})_k,$$

$$\begin{aligned} D_k =& f_3\binom{k+2}{2} - f_2\left[\binom{k+2}{2} - \binom{k-1}{2}\right] \\ &+ f_{1,2}\left[\binom{k+2}{2} - 2\binom{k-1}{2} + \binom{k-4}{2}\right] \\ &+ f_{1,3}\left[\binom{k+2}{2} - 3\binom{k-1}{2} + \binom{k-2}{2} + \binom{k-3}{2}\right] \\ &+ f_{1,4}\left[\binom{k+2}{2} - 4\binom{k-1}{2} + 3\binom{k-2}{2}\right] \\ &- \dim(R/J(v))_k, \end{aligned}$$

and $\dim(R/J(v))_k$ *can be explicitly computed using the method of* Sect. 4.

In the above theorems,

$$\binom{a}{2} = 0, \text{ if } a < 2.$$

Proof (of Theorems 2, 3) The Euler characteristic equation applied to the complex $\mathscr{R}/\mathscr{J}$ is

$$\chi(H(\mathscr{R}/\mathscr{J})) = \chi(\mathscr{R}/\mathscr{J}).$$

Since $C^2(\Delta) \simeq H_3(\mathscr{R}/\mathscr{J})$, this implies that

$$\dim C^2(\Delta)_k = \dim \sum_{i=0}^{3} (-1)^i \bigoplus_{\beta \in \Delta^0_{3-i}} (R/J(\beta))_k + \dim \sum_{i=0}^{2} (-1)^i H_{2-i}(\mathscr{R}/\mathscr{J})_k$$

By Eqs. (4), (8), we get $\dim(R/J(\tau))_k$ and $\dim(R/J(e))_k$. By Lemma 1, $H_1(\mathscr{R}/\mathscr{J}) = H_0(\mathscr{R}/\mathscr{J}) = 0$. Also $H_2(\mathscr{R}/\mathscr{J})$ is Artinian, so its kth graded component vanishes when $k \gg 0$. This completes the proof. □

Corollary 2 $\dim C^1(\Delta)_k \geq C_k$, *and* $\dim C^2(\Delta)_k \geq D_k$.

Remark 8 The two complexes $\mathscr{R}/\mathscr{J}$ for $r = 1$ and $r = 2$ are different, and so are the modules $H_2(\mathscr{R}/\mathscr{J})$.

Corollary 3 *The dimension of* $C^1_d(\Delta)$, *the vector space of splines of smoothness* $r = 1$ *of degree at most* d, *is bounded below as*

$$\dim C_d^1(\Delta) \geq (f_3 - f_2 + f_1)\binom{d+3}{3} + (f_2 - 2f_{1,2} - 3f_{1,3} - 3f_{1,4})\binom{d+1}{3} + 2(f_{1,3} + f_{1,4})\binom{d}{3} + f_{1,2}\binom{d-1}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k. \quad (21)$$

For $d \geq 4$, the inequality simplifies to

$$\dim C_d^1(\Delta) \geq \frac{f_3}{6}d^3 + (f_3 - f_2)d^2 + \left(\frac{11}{6}f_3 - 2f_2 + 3f_1 + f_{1,2}\right)d + (f_3 - f_2 + f_{1,3} + f_{1,4}) - \sum_{k=0}^{d} \dim(R/J(v))_k. \quad (22)$$

Corollary 4 *The dimension of $C_d^2(\Delta)$, the vector space of splines of smoothness $r = 2$ of degree at most d, is bounded below as*

$$\dim C_d^2(\Delta) \geq (f_3 - f_2 + f_1)\binom{d+3}{3} + (f_2 - 2f_{1,2} - 3f_{1,3} - 4f_{1,4})\binom{d}{3} + (f_{1,3} + 3f_{1,4})\binom{d-1}{3} + f_{1,3}\binom{d-2}{3} + f_{1,2}\binom{d-3}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k. \quad (23)$$

For $d \geq 6$, the inequality simplifies to

$$\dim C_d^2(\Delta) \geq \frac{f_3}{6}d^3 + (f_3 - \frac{3}{2}f_2)d^2 + (\frac{11}{6}f_3 - \frac{3}{2}f_2 + 6f_1 + 3f_{1,2} + f_{1,3})d + (f_3 - f_2 - 9f_{1,2} - 4f_{1,3} - 2f_{1,4}) - \sum_{k=0}^{d} \dim(R/J(v))_k. \quad (24)$$

For the extremal cases of exactly 3 or ≥ 10 hyperplanes, we work out $\dim(R/J(v))_k$ in Example 1. Here we put our results on the above examples of 4, 5 or 6 hyperplanes in one place for the readers' convenience. We do not claim these are all the cases of 4, 5 or 6 hyperplanes. Our point is to illustrate the computation of $\dim(R/J(\tau))_k$ by the algorithm. All the remaining cases are similar but more complicated.

Proposition 1 *In the case $r = 2$, $\dim(R/J(v))_k$ for the following cases are given by*

k	0	1	2	3	4	5	≥ 6
4 hyperplanes	1	3	6	6	4	1	0
5 hyperplanes(Δ_1)	1	3	6	5	3	0	0
5 hyperplanes(Δ_2)	1	3	6	5	2	0	0
Clough-Tocher	1	3	6	4	1	0	0

6.2 Comparison and Examples

In the case $r = 1$, Alfeld, Schumaker, and Whiteley also give a lower bound on $\dim C_d^1(\Delta)$ in [1, Theorem 54].

$$\dim C_d^1(\Delta) \geq \frac{d(d-1)(d-5)}{6} T + 3(d-1)V_I + d(d-1)V_B + 1 + 5d - 2d^2, \text{ for } d \geq 3,$$

where T, V_B, V_I are the number of tetrahedra, boundary vertices, and interior vertices, respectively.

In the setting of our paper, $V_I = 1$. Using the relation $V_B = 2f_3 - f_2 + 2$, their bound is given by

$$\dim C_d^1(\Delta) \geq \frac{f_3}{6} d^3 + (f_3 - f_2)d^2 + \left(-\frac{7}{6} f_3 + f_2 + 6\right) d - 2. \tag{25}$$

Comparing our bound in (22) with their bound, the difference is

$$f_{1,2}d + (f_3 - f_2 + f_{1,3} + f_{1,4}) + 2 - \sum_{k=0}^{d} \dim(R/J(v))_k. \tag{26}$$

It is clear that our bound is better if $f_{1,2} > 0$. If $f_{1,2} = 0$, the difference is only

$$-\sum_{k=2}^{d} \dim(R/J(v))_k.$$

For a tetrahedral partition Δ of a simply connected polygonal region $D \subset \mathbb{R}^3$ and $d > r$, Lau [19] proved that a lower bound of $C_d^r(\Delta)$ is given by

$$\begin{aligned} \dim C_d^r(\Delta) \geq & \binom{d+3}{3} + f_2 \binom{d-r+2}{3} \\ & - f_1 \left[\binom{d+3}{3} - \binom{r+3}{3} - (d-r)\binom{r+2}{2} \right] + \delta, \end{aligned} \tag{27}$$

where

$$\delta = \sum_{k=1}^{f_1} \sum_{l=1}^{d-r} \sum_{j=1}^{l} (r+1+j-je_{k*})_+,$$

and e_{k*} is the number of interior faces attached to the interior edge e_k ($k = 1, 2, \ldots, f_1$) which lie on different planes. Here, $(x)_+ = x$, if $x > 0$. Otherwise, $(x)_+ = 0$. The leading term of his formula is only

$$\frac{1-f_1+f_2}{6}d^3 = \frac{f_3-1}{6}d^3,$$

and thus is weaker than our bounds, especially when d is large.

Example 6 For the example of 4 hyperplanes from Example 6 in Sect. 5 of Chap. 1, we have

$$f_3 = 8, f_2 = 12, f_{1,2} = 4, f_{1,3} = 2, f_{1,4} = 0.$$

For the two interior edges OP_5 and OP_6, there are three hyperplanes passing through each edge. For the other four edges OP_1, OP_2, OP_3, OP_4, only two hyperplanes passing through each edge. The formula above gives the following lower bound for $\dim C_d^1(\Delta)$ and

d	0	1	2	3
Bound	1	4	12	30

$$\dim C_d^1(\Delta) \geq 4/3d^3 - 4d^2 + 38/3d - 8, \text{ for } d \geq 4.$$

In this case, the bound is actually exact, because a computer calculation shows that $H_2(\mathscr{R}/\mathscr{J}) = 0$.

In the case $r = 2$, we get the lower bound for $\dim C_d^2(\Delta)$ as and

d	0	1	2	3	4
Bound	1	4	10	22	44

$$\dim C_d^2(\Delta) \geq 4/3d^3 - 10d^2 + 140/3d - 69, \text{ for } d \geq 6.$$

In this case, the bound is actually also exact, because a computer calculation shows that $H_2(\mathscr{R}/\mathscr{J}) = 0$.

Example 7 For the Clough-Tocher cell, we have

$$f_3 = 4, f_2 = 6, f_{1,2} = f_{1,4} = 0, f_{1,3} = 4.$$

The lower bound for $\dim C_d^2(\Delta)$ is given by

$$\dim C_d^2(\Delta) \geq 2\binom{d+3}{3} - 6\binom{d}{3} + 4\binom{d-1}{3} + 4\binom{d-2}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k.$$

The right hand side is equal to 1, 4, 10, 20, 35 for $d = 0, 1, 2, 3, 4$, and

$$2/3d^3 - 5d^2 + 79/3d - 33, \text{ for } d \geq 5.$$

For example, when $d = 17$, this gives 2245, which agrees with the computation in [20].

Similarly, the lower bound for $\dim C_d^1(\Delta)$ is given by

$$\dim C_d^1(\Delta) \geq 2\binom{d+3}{3} - 6\binom{d+1}{3} + 8\binom{d}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k,$$

where $\dim(R/J(v))_k$ is computed in Eq. (5).

Example 8 For the example of 5 hyperplanes from Example 3 in Sect. 5 of Chap. 1, we have

$$f_3 = 8, f_2 = 12, f_{1,2} = 4, f_{1,3} = 0, f_{1,4} = 2.$$

In this case, we have the following bounds

$$\dim C_d^1(\Delta) \geq 2\binom{d+3}{3} - 2\binom{d+1}{3} + 4\binom{d}{3} + 4\binom{d-1}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k,$$

where $\dim(R/J(v))_k$ is given by Eq. (6).

The lower bound for $\dim C_d^2(\Delta)$ is given by

$$\dim C_d^2(\Delta) \geq 2\binom{d+3}{3} - 4\binom{d}{3} + 6\binom{d-1}{3} + 4\binom{d-3}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k,$$

where $\dim(R/J(v))_k$is given by Eq. (18).

For the example of 5 hyperplanes from Example 4 in Sect. 5 of Chap. 1, we have

$$f_3 = 8, f_2 = 12, f_{1,2} = 2, f_{1,3} = 4, f_{1,4} = 0.$$

In this case, we have the following bounds

$$\dim C_d^1(\Delta) \geq 2\binom{d+3}{3} - 4\binom{d+1}{3} + 8\binom{d}{3} + 2\binom{d-1}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k,$$

where $\dim(R/J(v))_k$ is given by (7);

$$\dim C_d^2(\Delta) \geq 2\binom{d+3}{3} - 4\binom{d}{3} + 4\binom{d-1}{3} + 4\binom{d-2}{3} + 2\binom{d-3}{3} - \sum_{k=0}^{d} \dim(R/J(v))_k.$$

where $\dim(R/J(v))_k$ is given by (19).

Remark 9 Using Macaulay2, we found that $C^2(\Delta)$ is a free module over R for the above examples of 4 hyperplanes and Clough-Tocher. By Schenck's Theorem in [7], this implies $H_2(\mathscr{R}/\mathscr{J}) = 0$ and $\dim C^2(\Delta)_k = C_k$, so our bound in Corollary 2 is tight. Corollary 2agrees with Macaulay2's output. This strongly supports our theorem.

Remark 10 To compute the homology $H_2(\mathscr{R}/\mathscr{J})$ is complicated; our calculations were performed with the Macaulay2 software package [15].

Remark 11 For any given tetrahedral complex Δ_v, we can find the configuration of the fatpoints corresponding to the hyperplanes passing through v. The classification of all configurations of fatpoints up to 8 points is given in [2], though some configurations do not correspond to a tetrahedral complex Δ_v.

Acknowledgments Thanks to Hal Schenck for suggesting this problem and useful talks. Thanks to Michael DiPasquale for careful reading of the paper, helpful discussions, and suggestions. Thanks to Tatyana Sorokina for inviting me to give a presentation at San Antonio and also helpful comments and suggestions regarding the extension of our main theorem to the case where the order of smoothness may be different. Also thanks to the organizers of the 14th International Conference in Approximation Theory, San Antonio TX for financial support. I also appreciate the referees' careful reading of this paper and valuable suggestions.

References

1. Alfeld, P., Schumaker, L., Whiteley, W.: The generic dimension of the space of C^1 splines of degree $d \geq 8$ on tetrahedral decompositions. SIAM J. Numer. Anal. **30**, 889–920 (1993)
2. Geramita, A.V., Harbourne, B., Migliore, J.: Classifying Hilbert functions of fat point subschemes in $\mathbb{P}^2$. Collect. Math. **60**(2), 159–192 (2009)
3. Alfeld, P., Schumaker, L.: On the dimension of bivariate spline spaces of smoothness r and degree $d = 3r + 1$. Numer. Math. **57**, 651–661 (1990)
4. Billera, L.: Homology of smooth splines: generic triangulations and a conjecture of Strang. Trans. AMS **310**, 325340 (1988)
5. Whiteley, W.: A matrix for splines. In: "Progress in Approximation Theory". Academic Press, Boston (1991)
6. Strang, G.: Piecewise polynomials and the finite element method. Bull. Amer. Math. Soc. **79**, 1128–1137 (1973)
7. Schenck, H.: A spectral sequence for splines. Adv. Appl. Math. **19**, 183–199 (1997)
8. Schenck, H., Stillman, M.: Local cohomology of bivariate splines. J. Pure Appl. Algebra **117**(118), 535–548 (1997)

9. Geramita, A., Schenck, H.: Fat points, inverse systems, and piecewise polynomial functions. J. Algebra **204**(1), 116–128 (1998)
10. Alfeld, P.: Upper and lower bounds on the dimension of multivariate spline spaces. SIAM J. Numer. Anal. **33**(2), 571–588 (1996)
11. Alfeld, P., Schumaker, L.: Bounds on the dimensions of trivariate spline spaces. Adv. Comput. Math. **29**(4), 315–335 (2008)
12. Iarrobino, A.: Inverse system of a symbolic power III: thin algebras and fat points. Composit. Math. **108**, 319–356 (1997)
13. Eisenbud, D.: Commutative Algebra with a View Towards Algebraic Geometry. Springer, Berlin (1995)
14. Miranda, R.: Linear systems of plane curves. Notices Amer. Math. Soc. **46**(2), 192–201 (1999)
15. Grayson, D.R., Stillman, M.E.: Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/
16. Emsalem, J., Iarrobino, A.: Inverse system of a symbolic power. I. J. Algebra **174**(3), 1080–1090 (1995)
17. Geramita, A.V.: Inverse systems of fat points: Waring's problem, secant varieties of Veronese varieties and parameter spaces for Gorenstein ideals. The Curves Semin Queen's **10**, 2–114 (1996)
18. HartShorne, R.: Algebraic Geometry. Springer, Berlin (1977)
19. Lau, W.: A lower bound for the dimension of trivariate spline spaces. Constr. Approx. **23**(1), 23–31 (2006)
20. Alfeld, P., Schumaker, L., Sirvent, M.: On dimension and existence of local bases for multivariate spline spaces. J. Approx. Theory **70**(2), 243–264 (1992)

One Characterization of Lagrange Projectors

Boris Shekhtman

Abstract Introduced by G. Birkhoff and popularized by C. de Boor, ideal projectors are an elegant generalization of Hermite interpolation projectors to the multivariate setting. An important class of ideal projectors comprises Lagrange interpolation projectors. In this article, we give a characterization of Lagrange projectors in terms of their "restriction property."

Keywords Ideal projector · Lagrange projector · Restriction property

1 Introduction

In this article, the symbol $\mathbb{k}$ will stand for the field $\mathbb{C}$ of complex numbers or the field $\mathbb{R}$ of real numbers. The symbol $\mathbb{k}[\mathbf{x}] = \mathbb{k}[x_1, \ldots, x_d]$ denotes the algebra of polynomials in d variables with coefficients in the field $\mathbb{k}$.

Definition 1 A linear idempotent operator $P : \mathbb{k}[\mathbf{x}] \to \mathbb{k}[\mathbf{x}]$ is called an ideal projector if $\ker P$ is an ideal in $\mathbb{k}[\mathbf{x}]$.

Ideal projectors were defined by Birkhoff in [1] and further studied by de Boor (cf. [4]), Sauer (cf. [8]) as well as the author (cf. [9]).

An important class of ideal projectors is a class of (finite-dimensional) Lagrange interpolation projectors, that is the projectors $P_{\mathscr{Z}}$ with finite-dimensional range that interpolate at a set $\mathscr{Z} := \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ of $N = \dim(\text{ran } P_{\mathscr{Z}})$ distinct points in $\mathbb{k}^d$. Sauer and Xu [7] showed that for any subspace $G_0 \subset \text{ran } P_{\mathscr{Z}}$ one can find a subset $\mathscr{Z}_0 \subset \mathscr{Z}$ of $\dim G_0$ points such that the problem of interpolation from the space G_0

B. Shekhtman (✉)
Department of Mathematics and Statistics, University of South Florida, 4202 E. Fowler Ave., CMC 114, Tampa, FL 33620, USA
e-mail: shekhtma@usf.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_19,

at the points $\mathscr{Z}_0$ is well posed, i.e., there exists a Lagrange interpolation projector $P_{\mathscr{Z}_0}$ onto G_0 that interpolates at the points $\mathscr{Z}_0$. Since

$$\ker P_{\mathscr{Z}} = \{f \in \Bbbk[\mathbf{x}] : f(\mathbf{z}) = 0 \quad \text{, for all } \mathbf{z} \in \mathscr{Z}\},$$

the above fact can be reformulated as the existence of a Lagrange projector $P_{\mathscr{Z}_0}$ onto G_0 such that $\ker P_{\mathscr{Z}_0} \supset \ker P$.

Definition 2 We say an ideal projector P has the restriction property if for every subspace $G_0 \subset \text{ran } P$ there exists an ideal projector P_0 onto G_0 such that $\ker P_0 \supset \ker P$. Equivalently for every $G_0 \subset \text{ran } P$ there exists an ideal $J_0 \supset \ker P$ such that

$$J_0 \oplus G_0 = \Bbbk[\mathbf{x}] = \ker P \oplus \text{ran } P. \tag{1}$$

In this article, we will show that this property indeed characterizes Lagrange projectors within the class of all (finite-dimensional or infinite-dimensional) ideal projectors:

Theorem 1 *An ideal projector P on $\Bbbk[\mathbf{x}]$ is a Lagrange interpolation projector if and only if P has the restriction property.*

2 Preliminaries

In this section, we recall some needed notations and rudimentary facts from algebraic geometry readily available in [2].

With every ideal $J \subset \Bbbk[\mathbf{x}]$ we associate its affine variety:

$$\mathscr{Z}(J) := \{\mathbf{z} \in \Bbbk^d : f(\mathbf{z}) = 0 \text{ for all } f \in J\}.$$

The ideal $J \subset \Bbbk[\mathbf{x}]$ is called zero-dimensional if $\dim(\Bbbk[\mathbf{x}]/J) < \infty$. It is well known (cf. [2, Proposition 8, p. 235], [3]) that the ideal $J \subset \mathbb{C}[\mathbf{x}]$ is zero-dimensional if and only if the variety $\mathscr{Z}(J)$ is finite. Moreover $\#\mathscr{Z}(J) \leq \dim(\mathbb{C}[\mathbf{x}]/J)$. Hence the ideal projector P on $\mathbb{C}[\mathbf{x}]$ is Lagrange if and only if

$$\#\mathscr{Z}(\ker P) = \dim(\Bbbk[\mathbf{x}]/\ker P) = \dim(\text{ran } P),$$

i.e., if and only if $\ker P$ is a zero-dimensional radical ideal.

An affine variety $\mathscr{Z}$ is called irreducible if it cannot be written as a union of two of its proper subvarieties. In particular, every 1-point set $\{\mathbf{z}\} \subset \Bbbk^d$ is an irreducible variety.

By the Hilbert basis theorem (cf. [2, Theorem 2, p. 204]), every affine variety $\mathscr{Z}$ can be written as a finite union of irreducible varieties:

$$\mathscr{Z} = \cup_{j=1}^{k} \mathscr{Z}_j.$$

Such a decomposition is called irredundant if $i \neq j$ implies that $\mathscr{Z}_i$ is not a subvariety of $\mathscr{Z}_j$.

We will need the following lemma that must be well known, but I could not find a reference to it in the literature:

Lemma 1 *An ideal $J \subset \mathbb{C}[\mathbf{x}]$ is zero-dimensional if and only if the set $\mathscr{Z}(J)\backslash\{\mathbf{z}_0\}$ is an affine variety for every $\mathbf{z}_0 \in \mathscr{Z}(J)$.*

Proof If J is zero-dimensional, then $\mathscr{Z}(J)$ is finite. Hence $\mathscr{Z}(J)\backslash\{\mathbf{z}_0\}$ is finite and thus an affine variety. Conversely, assume that J is not zero-dimensional and let

$$\mathscr{Z}(J) = \mathscr{Z}_1 \cup \cdots \cup \mathscr{Z}_m \tag{2}$$

be its irredundant decomposition. At least one of $\mathscr{Z}_j$'s, say $\mathscr{Z}_m$, has infinitely many points. Pick a point $\mathbf{z}_0 \in \mathscr{Z}_m$. Since the decomposition is irredundant, $\{\mathbf{z}_0\} \neq \mathscr{Z}_j$ for any j. If $\mathscr{Z}(J)\backslash\{\mathbf{z}_0\}$ is an affine variety, it has an irredundant decomposition $\mathscr{Z}(J)\backslash\{\mathbf{z}_0\} = \tilde{\mathscr{Z}}_1 \cup \cdots \cup \tilde{\mathscr{Z}}_k$. Hence $\mathscr{Z}(J) = \tilde{\mathscr{Z}}_1 \cup \cdots \cup \tilde{\mathscr{Z}}_k \cup \{\mathbf{z}_0\}$ is an irredundant decomposition of $\mathscr{Z}(J)$ different from (4). This contradicts the uniqueness of irredundant decomposition. □

Theorem 2 [2, Theorem 4, p. 203]. *An affine variety $\mathscr{Z}$ has a unique irredundant decomposition as a finite union of irreducible subvarieties. In particular, any finite affine variety is irreducible iff it is a 1-point set.*

A word about duality: Let $\Bbbk[\![x_1, \ldots, x_d]\!]$ denote the space of formal power series in d variables with coefficients in $\Bbbk$. Via Macaulay duality (inverse systems), the dual space $(\Bbbk[x_1, \ldots, x_d])'$ is identified with $\Bbbk[\![x_1, \ldots, x_d]\!]$ as follows: With every element $\lambda \in \Bbbk[\![x_1, \ldots, x_d]\!]$ we associate the differential operator $\lambda(D) \in \Bbbk[D_1, \ldots, D_d]$ obtained by formally replacing monomials in λ with the appropriate powers of operators D_j that are partial derivatives with respect to x_j. Now, for every $\lambda \in \Bbbk[\![x_1, \ldots, x_d]\!]$ we define the functional $\widetilde{\lambda} \in (\Bbbk[x_1, \ldots, x_d])'$ by

$$\widetilde{\lambda}(f) := (\lambda(D)f)(0) \text{ for every } f \in \Bbbk[x_1, \ldots, x_d].$$

It is well known (cf. [3, 9] and [6]) that the map $\lambda \longmapsto \tilde{\lambda}$ (defined above) is a skew-linear isomorphism between $\Bbbk[\![x_1, \ldots, x_d]\!]$ and $(\Bbbk[x_1, \ldots, x_d])'$. From this point on, we will identify the functional with the corresponding power series and drop the tilde if there is no danger of confusion. In particular if $\mathbf{z} \in \Bbbk^d$, the exponential function

$$e_{\mathbf{z}} : \Bbbk^d \to \Bbbk, \quad e_{\mathbf{z}}(\mathbf{x}) := e^{\mathbf{z}\cdot\mathbf{x}}$$

is identified with its power series and, as such, with the functional

$$\widetilde{e}_{\mathbf{z}}(f) = \sum_{\alpha} \frac{1}{\alpha!} \mathbf{z}^{\alpha} \left(D^{\alpha} f\right)(0) = f(\mathbf{z}) = \delta_{\mathbf{z}}(f), \quad \forall f \in \Bbbk[x_1, \ldots, x_d]. \tag{3}$$

For a subspace $J \subset \Bbbk[x_1, \ldots, x_d]$ we define

$$J^\perp := \left\{\lambda \in \Bbbk[\![x_1, \ldots, x_d]\!] : \lambda(f) = 0 \text{ for all } f \in J\right\}.$$

A subspace $E \subset \Bbbk[\![x_1, \ldots, x_d]\!]$ is called D-invariant if $D_j\lambda \in E$ for every $\lambda \in E$ and every $j = 1, \ldots, d$.

Theorem 3 (cf. [3, 5] and [6]). *A subspace $J \subset \Bbbk[x_1, \ldots, x_d]$ is an ideal if and only if $J^\perp$ is D-invariant. If $J \subset \mathbb{C}[\mathbf{x}]$ is a zero-dimensional ideal, then*

$$J^\perp = \oplus_{\mathbf{z}\in\mathscr{Z}(J)}\delta_{\mathbf{z}} \circ M_{\mathbf{z}}(D), \tag{4}$$

where $M_{\mathbf{z}} \subset \mathbb{C}[\mathbf{x}]$ is a D-invariant subspace of polynomials.

We will now describe a simple process of "complexification" of ideals in $\mathbb{R}[\mathbf{x}]$. If J is an ideal in $\mathbb{R}[\mathbf{x}]$ then, by Hilbert's basis theorem, there exist finitely many polynomials $f_1, \ldots, f_m$ that generate the ideal J:

$$J = \langle f_1, \ldots, f_m \rangle.$$

The polynomials $f_1, \ldots, f_m$ are also polynomials in $\mathbb{C}[x_1, \ldots, x_d]$ and, as such, generate an ideal, say $\hat{J}$, in $\mathbb{C}[x_1, \ldots, x_d]$.

Proposition 1 *The ideal $\hat{J} \subset \mathbb{C}[x_1, \ldots, x_d]$ does not depend on the choice of generators $f_1, \ldots, f_m$ for the ideal $J \subset \mathbb{R}[x_1, \ldots, x_d]$.*

Proof Let $\{g_1, \ldots, g_s\} \subset \mathbb{R}[x_1, \ldots, x_d]$ be another set of generators of the ideal $J \subset \mathbb{R}[x_1, \ldots, x_d]$. Then there exist polynomials $p_{j,k} \in \mathbb{R}[x_1, \ldots, x_d]$ such that $g_j = \sum_k p_{j,k} f_k$ and for every sequence of complex polynomials $h_j \in \mathbb{C}[x_1, \ldots, x_d]$ we have

$$\sum_j h_j g_j = \sum_k \left(\sum_j h_j p_{j,k} \right) f_k \in \hat{J},$$

which proves the proposition. □

Definition 1 The ideal $\hat{J}$ is denoted by $J^{\mathbb{C}}$ and is called the complexification of J.

3 Proof of the Main Result

Proposition 1 *Let G be a subspace in $\mathbb{R}[\mathbf{x}]$ and let J be an ideal in $\mathbb{R}[\mathbf{x}]$ that complements G and such that for every subspace $H \subset G$ with $\dim G/H = 1$ there exists an ideal $K \subset \mathbb{R}[\mathbf{x}]$ that complements H and contains J. Then $\mathscr{Z}\left(J^{\mathbb{C}}\right) = \mathscr{Z}(J) \subset \mathbb{R}^d$.*

Proof Suppose not. Let $\mathbf{z} = (z_1, \ldots, z_d) \in \mathscr{Z}\left(J^{\mathbb{C}}\right)$ be such that $\mathbf{z} \neq \bar{\mathbf{z}}$. Then $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_d) \in \mathscr{Z}\left(J^{\mathbb{C}}\right)$ since every real polynomial in $\mathbb{R}[x_1, \ldots, x_d]$ that

vanishes on $\mathbf{z}$ also vanishes on $\bar{\mathbf{z}}$. Hence $e_{\mathbf{z}}$ and $e_{\bar{\mathbf{z}}}$ annihilate the ideal $J^{\mathbb{C}}$. So, as formal power series with real coefficients,

$$C_{\mathbf{z}}(\mathbf{x}) := e^{\mathrm{Re}\,(\mathbf{z})\cdot\mathbf{x}} \cos((\mathrm{Im}\ \mathbf{z})\cdot\mathbf{x}) \text{ and } S_{\mathbf{z}}(\mathbf{x}) := e^{\mathrm{Re}\,(\mathbf{z})\cdot\mathbf{x}} \sin((\mathrm{Im}\ \mathbf{z})\cdot\mathbf{x})$$

annihilate $J^{\mathbb{C}}$ and, in particular, annihilate J as functionals on $\mathbb{R}[\mathbf{x}]$. Since G complements J, no linear combination of $C_{\mathbf{z}}$ and $S_{\mathbf{z}}$ annihilates G. Hence, in particular,

$$H := \{g \in G : C_{\mathbf{z}}(g) = 0\}$$

is a subspace of G of codimension 1 and $S_{\mathbf{z}} \notin H^{\perp}$. Let $K \supset J$ be an ideal that complements H. Then $\dim K/J = 1$. Since K complements H and $C_{\mathbf{z}}$ annihilates H hence $C_{\mathbf{z}} \notin K^{\perp}$; but $K^{\perp}$ is D-invariant, thus $S_{\mathbf{z}} \notin K^{\perp}$. This means that $C_{\mathbf{z}}$ and $S_{\mathbf{z}}$ are two linearly independent linear functionals over K that annihilate J, and therefore $\dim K/J \geq 2$. □

Theorem 1 *Let J be an ideal in $\mathbb{k}[\mathbf{x}]$ and G be any subspace of $\mathbb{k}[\mathbf{x}]$ that complements J. Then, the following are equivalent:*

(i) $J = \{f \in \mathbb{k}[\mathbf{x}] : f(\mathbf{z}) = 0$ *for all* $\mathbf{z} \in \mathscr{Z}\}$ *for some finite set* $\mathscr{Z} \subset \mathbb{k}^d$.
(ii) *For every subspace G_0 of G of codimension 1 in G, there exists an ideal J_0 complementing G_0 and containing J.*

Proof

(i)⇒(ii): (i) implies that $\delta_{\mathscr{Z}}$ is an injective map from G onto $\mathbb{k}^{\mathscr{Z}}$. Hence the matrix $\delta_{\mathscr{Z}} V$ is invertible for any basis V for G. In particular, with $V =: [V_0, v]$ any basis for G for which V_0 is a basis for G_0, there exists $\mathscr{Z}_0 \subset \mathscr{Z}$ for which $\delta_{\mathscr{Z}_0} V_0$ is invertible, hence $J_0 := \ker \delta_{\mathscr{Z}_0}$ complements G_0 and contains J.

(ii)⇒(i): For any $\mathbf{z} \in \mathscr{Z}(J)$, let

$$G_{\mathbf{z}} := \ker \delta_{\mathbf{z}} \cap G.$$

By Proposition 3.1, $\mathscr{Z}(J) \subset \mathbb{k}^d$, hence, $\dim G/G_{\mathbf{z}} \leq 1$. On the other hand, since G complements J, $G_{\mathbf{z}}$ cannot be all of G, hence

$$\dim G/G_{\mathbf{z}} = 1. \tag{5}$$

By assumption, there exists an ideal $J_{\mathbf{z}}$ that contains J and complements $G_{\mathbf{z}}$. Hence $\mathscr{Z}(J_{\mathbf{z}}) \subset \mathscr{Z}(J)$. In fact, since $J_{\mathbf{z}}$ complements $G_{\mathbf{z}}$, $\mathbf{z} \notin \mathscr{Z}(J_{\mathbf{z}})$, hence $\mathscr{Z}(J_{\mathbf{z}}) \subset \mathscr{Z}(J)\backslash\{\mathbf{z}\}$. More than that:

$$\mathscr{Z}(J_{\mathbf{z}}) = \mathscr{Z}(J)\backslash\{\mathbf{z}\}, \tag{6}$$

as we now prove.

If there were a $\mathbf{z}_1 \notin \mathscr{Z}(J)\backslash\{\mathbf{z}\}$ not in $\mathscr{Z}(J_{\mathbf{z}})$, then, with $\mathbf{z}_0 := \mathbf{z}$, $J_{\mathbf{z}}$ would contain polynomials f_i with $f_i(\mathbf{z}_i) = 1$, $i = 0, 1$. But then, choosing $p_i \in \mathbb{k}[\mathbf{x}]$

such that $p_i(\mathbf{z}_k) = \delta_{i,k}$, we would have $p_i f_i(\mathbf{z}_k) = \delta_{i,k}$ and $p_i f_i \in J_{\mathbf{z}}$. Since G complements J by assumption, we would have $p_i f_i = g_i + h_i$ with $g_i \in G$ and $h_i \in J$, hence

$$g_i(\mathbf{z}_k) = \delta_{i,k}. \tag{7}$$

Assume that for some $a_i \in \mathbb{k}$ we have $\sum_{i=0}^{1} a_i (g_i + G_{\mathbf{z}}) = 0$, i.e., $g := a_0 g_0 + a_1 g_1 \in G_{\mathbf{z}}$. Since $g = \sum_{i=0}^{1} a_i (p_i f_i - h_i) \in J_{\mathbf{z}}$ and $J_{\mathbf{z}}$ complements $G_{\mathbf{z}}$, it would follow that $g = 0$ and, by (7), $a_0 = a_1 = 0$. This would show that $G/G_{\mathbf{z}}$ contains a linearly independent pair of elements, hence $\dim G/G_{\mathbf{z}} \geq 2$ contradicting (5).

Since we now know that (6) holds for arbitrary $\mathbf{z} \in \mathscr{Z}(J)$, we know by Lemma 2.1 that J is 0-dimensional, and therefore

$$J^{\perp} = \oplus_{\mathbf{z} \in \mathscr{Z}(J)} \delta_{\mathbf{z}} \circ M_{\mathbf{z}}(D)$$

for some D-invariant polynomial subspaces $M_{\mathbf{z}}$. Continuing, for arbitrary $\mathbf{z} \in \mathscr{Z}(J)$, with the ideal $J_{\mathbf{z}}$, we know that $J_{\mathbf{z}}$ is 0-dimensional, and hence we must have

$$\oplus_{\mathbf{y} \in \mathscr{Z}(J)} \delta_{\mathbf{y}} \circ N_{\mathbf{y}}(D) = J_{\mathbf{z}}^{\perp} \subset J^{\perp} = \oplus_{\mathbf{y} \in \mathscr{Z}(J)} \delta_{\mathbf{y}} \circ M_{\mathbf{y}}(D)$$

with $N_{\mathbf{y}} \subset M_{\mathbf{y}}$ for all $\mathbf{y}$. But then, $\dim G_{\mathbf{z}} = \dim J_{\mathbf{z}}^{\perp} \leq \dim J^{\perp} - \dim M_{\mathbf{z}} = \dim G - \dim M_{\mathbf{z}}$. Therefore, by (5), $\dim M_{\mathbf{z}} = 1$ for all $\mathbf{z} \in \mathscr{Z}(J)$. This proves (i). □

Acknowledgments I would like to thank Carl de Boor for his many contributions to this paper; both in substance and style.

References

1. Birkhoff, G.: The algebra of multivariate interpolation. In: Constructive Approaches to Mathematical Models (Proc. Conf. in honor of Duffin, R.J., Pittsburgh, Pa., 1978), pp. 345–363. Academic Press, New York (1979)
2. Cox, D., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms. Undergraduate Texts in Mathematics, 3rd edn. Springer, New York (2007)
3. de Boor, C., Ron, A.: On polynomial ideals of finite codimension with applications to box spline theory. J. Math. Anal. Appl. **158**(1), 168–193 (1991). http://dx.doi.org/10.1016/0022-247X(91), 90275–5
4. de Boor, C.: Ideal interpolation. In: Approximation Theory XI: Gatlinburg 2004, Mod. Methods Math., pp. 59–91. Nashboro Press, Brentwood, TN (2005)

5. Macaulay, F.S.: The Algebraic Theory of Modular Systems. Cambridge Mathematical Library. Revised reprint of the 1916 original, with an introduction by Paul Roberts. Cambridge University Press, Cambridge (1994)
6. Möller, H.M.: Hermite interpolation in several variables using ideal-theoretic methods. In: Constructive Theory of Functions of Several Variables (Proc. Conf., Math. Res. Inst., Oberwolfach, 1976), pp. 155–163. Lecture Notes in Math., vol. 571. Springer, Berlin (1977)
7. Sauer, T., Xu, Y.: On multivariate Lagrange interpolation. Math. Comp. **64**(211), 1147–1170 (1995). http://dx.doi.org/10.2307/2153487
8. Sauer, T.: Polynomial interpolation in several variables: lattices, differences, and ideals. In: Topics in Multivariate Approximation and Interpolation, Stud. Comput. Math., vol. 12, pp. 191–230. Elsevier, Amsterdam (2006). http://dx.doi.org/10.1016/S1570-579X(06), 80009–1
9. Shekhtman, B.: A taste of ideal interpolation. J. Concr. Appl. Math. **8**(1), 125–149 (2010)

Minimal Versus Orthogonal Projections onto Hyperplanes in ℓ_1^n and ℓ_∞^n

Boris Shekhtman and Lesław Skrzypek

Abstract In this paper, we explore the relation between the minimal and the orthogonal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n.

Keywords Minimal projection · Orthogonal projection · Hyperplanes

1 Introduction

Suppose X is a Banach space and V a (closed) subspace of X. A projection from X onto V is a linear continuous operator $P : X \to V$ having the property that $Pv = v$ for all $v \in V$. The set of all projections from X onto V is denoted by $P(X, V)$. The set $P(X, V)$ can be empty. If it is not empty then V is said to be complemented.

The *relative projection constant* of V with respect to X is defined by

$$\lambda(V, X) = \inf\{\|P\|, P \in P(X, V)\}. \tag{1}$$

A projection whose norm is equal to this constant is called a minimal projection. It is useful to know whether minimal projections exist, how they are characterized, whether they are unique, and how they are calculated (see [1–8]). The existence of minimal projections onto finite dimensional spaces may be deduced using the compactness argument (see [9] for details).

Note that for any (closed) subspace V of a Hilbert space $\mathscr{H}$ we have $\lambda(V, \mathscr{H}) = 1$ and the orthogonal projection onto V is the unique minimal projection. In general,

B. Shekhtman (✉) · L. Skrzypek
Department of Mathematics and Statistics, University of South Florida,
4202 E. Fowler Avenue, CMC 114, Tampa, FL 33620-5700, USA
e-mail: shekhtma@usf.edu

L. Skrzypek
e-mail: skrzypek@usf.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_20,

a given subspace will not be the range of a projection of norm 1, and minimal projections are very difficult to discover even if their existence is known. In case of the $X = \ell_p^n$, the orthogonal projection seems like a good candidate for a minimal or near minimal projection. Here, by an orthogonal projection we mean the orthogonal projection P from ℓ_2^n onto $V \subset \ell_2^n$ viewed as a projection on ℓ_p^n. This idea is not new and has been previously explored [10–13].

In this paper, we investigate some of the relations between minimal and orthogonal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n. For this reason, we introduce the notation

$$\lambda_O(V, \ell_p^n) = ||P_O||, \tag{2}$$

where P_O is the orthogonal projection from ℓ_p^n onto V.

For any n dimensional space V and any n codimensional space W of X we have, respectively, (see [7, 14])

$$\lambda(V, X) \leq \sqrt{n} \tag{3}$$

and

$$\lambda(W, X) \leq 1 + \sqrt{n}. \tag{4}$$

Applying (4) to hyperplanes H of X we obtain

$$\lambda(H, X) \leq 2. \tag{5}$$

This estimate is sharp, as a consequence of the Daugavet Theorem, for any hyperplane H in $C[0, 1]$

$$\lambda(H, C[0, 1]) = 2. \tag{6}$$

For finite dimesional spaces X, we can improve the estimate in (5) even further (see [15])

$$\lambda(H, X) \leq 2 - \frac{2}{\dim X}. \tag{7}$$

For further results regarding projections onto hyperplanes, see [1, 6, 8, 16–19].

The natural question to ask is how large the quantities $\lambda(H, \ell_p^n)$ and $\lambda_O(H, \ell_p^n)$ can be for hyperplanes $H \subset \ell_p^n$ and what is the relation between them?

For $p = 1, \infty$, (see Theorem 1 and Theorem 2), the maximum of $\lambda(H, \ell_p^n)$ reaches the upper bound in the estimate (7) and this maximum is attained when H is the kernel of the functional $(1, 1, \ldots, 1)$ in ℓ_q^n, i.e.,

$$H = \ker 1 = \left\{ (x_1, \ldots, x_n) \in \ell_p^n : \sum_{i=1}^{n} x_i = 0 \right\}. \tag{8}$$

One can easily see that the minimal projection onto $\ker(1, 1, \ldots 1)$ is in fact the orthogonal projection, i.e.,

$$\lambda\Big(\ker 1, \ell_p^n\Big) = \lambda_O\Big(\ker 1, \ell_p^n\Big). \tag{9}$$

As a result, one may think (as the authors did originally) that the maximal norm of all orthogonal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n is also attained for $\ker 1$. We show a rather intriguing result (see Theorem 3) that the maximum of the norms of orthogonal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n is attained for a hyperplane given by functional f which has all but one of its coordinates equal. The result also shows that

$$\max\left\{\lambda_O(H, \ell_p^n) : H \subset \ell_p^n\right\} \tag{10}$$

can be arbitrary large contrasting the estimate (5).

Recall that the full description of the uniqueness of minimal projection from ℓ_∞^n or ℓ_1^n onto any hyperplane H has been obtained in [1] and [6] (for easier proofs, see [20]).

Finally, note that, for $p \in (1, \infty)$ and $p \neq 2$, it is still an open question whether the maximum of $\lambda(H, \ell_p^n)$ is attained when $H = \ker 1$.

2 Results

Every hyperplane $H \subset \ell_p^n$ is a kernel of a linear functional $f = (f_1, \ldots, f_n) \in S(\ell_q^n)$. The lemma below will allow us to assume, without loss of generality, that $f_i \geq 0$ for every $i = 1, \ldots, n$.

Lemma 1 *Observe that*

$$\lambda\Big(\ker f, \ell_p^n\Big) = \lambda\Big(\ker |f|, \ell_p^n\Big). \tag{11}$$

and

$$\lambda_O\Big(\ker f, \ell_p^n\Big) = \lambda_O\Big(\ker |f|, \ell_p^n\Big). \tag{12}$$

Proof Since ℓ_p^n is a symmetric space there is an isometry $I : \ell_p^n \to \ell_p^n$ such that $I(\ker f) = I(\ker |f|)$. □

The next two theorems may be a part of the mathematical folklore (see [1] and [6]), yet we could not find explicit formulations nor the proofs for these theorems in the literature. Thus, we formulate and prove them for the sake of completeness.

Theorem 1 *The maximal norm of minimal projections onto hyperplanes in ℓ_∞^n equals $2 - \frac{2}{n}$. That is,*

$$\max_{f \in S(\ell_1^n)} \lambda\left(\ker f, \ell_\infty^n\right) = 2 - \frac{2}{n} \tag{13}$$

where the equality is attained only for f which has all of its coordinates equal.

Proof Consider $f \in S(\ell_1^n)$. Without loss of generality, we can assume $f_i > 0$ and $||f||_\infty < 1/2$ (otherwise the norm of minimal projection is 1). By [20] we have

$$\lambda\left(\ker f, \ell_\infty^n\right) = 1 + \frac{2}{\sum_{i=1}^n \frac{2f_i}{1-2f_i}}. \tag{14}$$

Observe that the function $f(x) = \frac{2x}{1-2x}$ is convex on $(0, \frac{1}{2})$. By the classical Jensen inequality, we get

$$\frac{\sum_{i=1}^n \frac{2f_i}{1-2f_i}}{n} \geq \frac{2\left(\frac{\sum_{i=1}^n f_i}{n}\right)}{1-2\left(\frac{\sum_{i=1}^n f_i}{n}\right)} = \frac{\frac{2}{n}}{1-\frac{2}{n}} = \frac{2}{n-2}, \tag{15}$$

and equality holds only when all of f_i are equal. As a result,

$$\lambda(\ker f, \ell_\infty^n) = 1 + \frac{2}{\sum_{i=1}^n \frac{2f_i}{1-2f_i}} \leq 1 + \frac{2}{\frac{n-2}{2n}} = 2 - \frac{2}{n}, \tag{16}$$

and equality holds only when all of f_i are equal. □

Theorem 2 *The maximal norm of minimal projections onto hyperplanes in ℓ_1^n equals $2 - \frac{2}{n}$. That is,*

$$\max_{f \in S(\ell_\infty^n)} \lambda\left(\ker f, \ell_1^n\right) = 2 - \frac{2}{n} \tag{17}$$

where the equality is attained only for f which has all of its coordinates equal.

Proof Consider $f \in S(\ell_1^n)$. Without loss of generality we can assume $f_i > 0$ and that P attains its norm on all extreme points of $S(\ell_1^n)$. By [20] we have

$$\lambda\left(\ker f, \ell_\infty^n\right) = 1 + \frac{2}{\frac{\left(\sum_{i=1}^n f_i\right)\left(\sum_{i=1}^n f_i^{-1}\right)}{n-2} - n}. \tag{18}$$

By the classical inequality between Harmonic and Arithmetic mean, we get

$$\frac{n}{\sum_{i=1}^n f_i^{-1}} \leq \frac{\sum_{i=1}^n f_i}{n} \tag{19}$$

and equality holds only when all of f_i's are equal. As a result,

$$\lambda\left(\ker f, \ell_1^n\right) = 1 + \frac{2}{\frac{\left(\sum_{i=1}^n f_i\right)\left(\sum_{i=1}^n f_i^{-1}\right)}{n-2} - n} \leq 1 + \frac{2}{\frac{n^2}{n-2} - n} = 2 - \frac{2}{n}, \tag{20}$$

and equality holds only when all of f_i's are equal. □

By the above theorems, the maximal norm of all minimal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n is attained for $\ker(1, 1, \ldots, 1)$. But the minimal projection onto $\ker(1, 1, \ldots, 1)$ is orthogonal and is unique. As a result, one may think that the maximal norm of all orthogonal projections onto hyperplanes in ℓ_1^n and ℓ_∞^n is also attained for $\ker(1, 1, \ldots, 1)$. However, this expectation is not true as we obtained the following result:

Theorem 3 *For $p = 1$ and $p = \infty$, the maximal norm of orthogonal projections onto hyperplanes in ℓ_p^n equals $\frac{1+\sqrt{n}}{2}$. That is,*

$$\max_{f \in S(\ell_2^n)} \lambda_O \left(\ker f, \ell_p^n \right) = \frac{1 + \sqrt{n}}{2}. \tag{21}$$

where the equality is attained only for $f \in S(\ell_2^n)$ which is a permutation of the following:

$$f_1 = \sqrt{\frac{\sqrt{n} - 1}{2\sqrt{n}}}, \tag{22}$$
$$f_2 = \cdots = f_n = \sqrt{\frac{1}{2\left(n - \sqrt{n}\right)}}.$$

Proof Fix a hyperplane H in ℓ_1^n. We can assume that H is given by $\ker f$ where $f = (f_1, \ldots, f_n) \in S(\ell_2^n)$ and by Lemma 1 we can assume that $f_i \geq 0$. The orthogonal projection onto H is given by

$$P_f = Id - f \otimes f. \tag{23}$$

Easy calculations yeld $P_f(e_k) = e_k - f_k \cdot f$ and

$$||P_f(e_k)||_1 = 1 - f_k^2 + f_k \left(\sum_{i \neq k} f_i \right). \tag{24}$$

Using the classical inequality between arithmetic and quadratic mean and the fact that $\sum_{i=1}^n f_i^2 = 1$ we get

$$\sum_{i \neq k} f_i \leq \sqrt{(n-1) \sum_{i \neq k} f_i^2} = \sqrt{n-1}\sqrt{1 - f_k^2}. \tag{25}$$

As a result,

$$||P_f(e_k)||_1 \leq 1 - f_k^2 + \sqrt{n-1} f_k \sqrt{1 - f_k^2}$$

$$\leq \max_{0\leq x\leq 1}\left\{1-x^2+\sqrt{n-1}\cdot x\sqrt{1-x^2}\right\}. \tag{26}$$

Using the fact that the set of extreme points of $S(\ell_1^n)$ equals $\{\pm e_1,\ldots,\pm e_n\}$ from the above estimate we get

$$||P_f||_1 = \max_{k=1,\ldots,n} ||P_f(e_k)||_1 \leq \max_{0\leq x\leq 1}\left\{1-x^2+\sqrt{n-1}\cdot x\sqrt{1-x^2}\right\}. \tag{27}$$

Set

$$g(x) = 1-x^2+\sqrt{n-1}\cdot x\sqrt{1-x^2}. \tag{28}$$

Putting $x=\sin(\alpha)$ for $\alpha\in[0,\pi/2]$ we get

$$\begin{aligned} g(x) &= \cos^2(\alpha)+\sqrt{n-1}\sin(\alpha)\cos(\alpha)\\ &= \frac{1+\cos(2\alpha)+\sqrt{n-1}\sin(2\alpha)}{2}\\ &= \frac{1+\sqrt{n}\sin(2\alpha+\theta)}{2} \leq \frac{1+\sqrt{n}}{2}, \end{aligned} \tag{29}$$

where $\theta=\arctan(\frac{1}{\sqrt{n-1}})$. The equality is attained only when

$$2\alpha+\theta=\frac{\pi}{2}, \tag{30}$$

which, after standard computations, yields

$$x^2=\sin^2(\alpha)=\frac{1}{2}-\frac{\cos(2\alpha)}{2}=\frac{1}{2}-\frac{\cos(\frac{\pi}{2}-\theta)}{2}=\frac{1}{2}-\frac{\sin(\theta)}{2}. \tag{31}$$

Using the fact that $\sin(\arctan t)=\frac{t}{\sqrt{1+t^2}}$ we can easily obtain

$$x^2=\frac{1}{2}-\frac{\sin\left(\arctan\left(\frac{1}{\sqrt{n-1}}\right)\right)}{2}=\frac{1}{2}-\frac{\frac{1}{\sqrt{n-1}}}{2\sqrt{1+\frac{1}{n-1}}}=\frac{\sqrt{n}-1}{2\sqrt{n}}. \tag{32}$$

As a result,

$$||P_f||_1 \leq \frac{1+\sqrt{n}}{2} \tag{33}$$

and the equality is attained only when one of the coordinates of f equals $\sqrt{\frac{\sqrt{n}-1}{2\sqrt{n}}}$ and the remaining coordinates equals $\sqrt{\frac{1}{2(n-\sqrt{n})}}$.

The result for ℓ_∞^n follows easily from a standard duality argument (ℓ_∞^n and ℓ_1^n are dual spaces to each other) after noting that if P is an orthogonal projection then P^* (P^* being a dual operator to P) is also an orthogonal projection. □

References

1. Blatter, J., Cheney, E.W.: Minimal projections on hyperplanes in sequence spaces. Ann. Mat. Pura Appl. **101**(4), 215–227 (1974). MR0358179 (50 #10644)
2. Chalmers, B.L., Lewicki, G.: Symmetric spaces with maximal projection constants. J. Funct. Anal. **200**(1), 1–22 (2003). MR1974085 (2004b:46009)
3. Chalmers, B.L., Lewicki, G.: A proof of the Grünbaum conjecture. Studia Math. **200**(2), 103–129 (2010). MR2725896 (2011j:46004)
4. Chalmers, B.L., Metcalf, F.T.: The determination of minimal projections and extensions in l^1. Trans. Amer. Math. Soc. **329**(1), 289–305 (1992). MR1034660 (92e:41017)
5. Grünbaum, B.: Projection constants. Trans. Amer. Math. Soc. **95**, 451–465 (1960). MR0114110 (22 #4937)
6. Odyniec, W., Lewicki, G.: Minimal projections in Banach spaces, Lecture Notes in Mathematics. Problems of Existence and Uniqueness and Their Application, vol. 1449. Springer, Berlin (1990). MR1079547 (92a:41021)
7. Wojtaszczyk, P.: Banach Spaces for Analysts, Cambridge Studies in Advanced Mathematics, vol. 25. Cambridge University Press, Cambridge (1991). MR1144277 (93d:46001)
8. Lewicki, G., Prophet, M.: Codimension-one minimal projections onto Haar subspaces. J. Approx. Theory **127**, 198–206 (2004). MR2058158 (2005h:41070)
9. Isbell, J.R., Semadeni, Z.: Projection constants and spaces of continuous functions. Trans. Amer. Math. Soc. **107**, 38–48 (1963). MR0146649 (26 #4169)
10. Chalmers, B.L., Shekhtman, B.: On minimal and almost locally minimal and orthogonal minimal projections. In: Trends in Approximation Theory (Nashville, TN, 2000), Innovations in Applied Mathematics, pp. 49–52. Vanderbilt University Press, Nashville (2001). MR1937998
11. Chalmers, B.L., Shekhtman, B.: On spaces admitting minimal projections which are orthogonal. In: Approximation Theory X (St. Louis, MO, 2001), Innovations in Applied Mathematics, pp. 113–116. Vanderbilt University Press, Nashville (2002). MR1924853 (2003f:41045)
12. Shekhtman, B.: Some examples concerning projection constants, approximation theory, spline functions and applications, (Maratea, 1991). NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **356**, 471–476 (1992). MR1165993 (93f:41038)
13. Zippin, M.: Orthogonal almost locally minimal projections on ℓ_1^n. Israel J. Math. **115**, 253–268 (2000). MR1749681 (2001h:46022)
14. Kadec', M.I., Snobar, M.G.: Certain functionals on the Minkowski compactum. Mat. Zametki **10**, 453–457 (1971). MR0291770 (45 #861)
15. Bohnenblust, F.: Convex regions and projections in Minkowski spaces. Ann. of Math. **39**, 301–308 (1938)
16. Franchetti, C.: Projections onto hyperplanes in Banach spaces. J. Approx. Theory **38**(4), 319–333 (1983). MR711458 (84h:46023)
17. Franchetti, C.: The norm of the minimal projection onto hyperplanes in $L^p[0, 1]$ and the radial constant. Boll. Un. Mat. Ital. B **7**(4), 803–821 (1990). MR1086705 (92e:46029)
18. Rolewicz, S.: On projections on subspaces of codimension one. Studia Math. **96**(1), 17–19 (1990). MR1055074 (91i:46017)
19. Skrzypek, L.: On the L_p norm of the Rademacher projection and related inequalities. Proc. Amer. Math. Soc. **137**(8), 2661–2669 (2009). MR2497479 (2010d:41043)
20. Skrzypek, L.: Chalmers-Metcalf operator and uniqueness of minimal projections in ℓ_∞^n and ℓ_1^n spaces. In: Neamtu, M., Schumaker L. (eds.) Springer Proceedings in Mathematics, Approximation Theory XIII: San Antonio 2010, vol. 13, pp. 331–344 (2010)

On Hermite Interpolation by Splines with Continuous Third Derivatives

Vesselin Vatchev

Abstract For a C^3-smooth function, we consider a convolution-based method for constructing a C^3 spline interpolant that agrees with the function and its first and second derivatives at the points of interpolation. In the case of equidistant nodes $x_j = \frac{j}{n}$, $j = 0, \ldots, n$ the error of interpolation on $[0, 1]$ is proven to be of order n^{-3} which is one less than the order of the natural spline interpolation at the same points, n^{-4}. Applications are discussed.

Keywords Spline interpolation · Hermite interpolation · Order of approximation

1 Introduction

In this paper for a function $f \in C^3[a, b]$, where $C^k[a, b]$ is the space of all k times continuously differentiable functions on $[a, b]$ we construct a spline function $S \in C^3[a, b]$ such that $S^{(k)}(a) = f^{(k)}(a), S^{(k)}(b) = f^{(k)}(b), k = 0, 1, 2, j = 0, \ldots, n$. By using that procedure, we construct a C^3 Hermite spline interpolant with equidistant interpolating nodes $x_j = a + j\frac{b-a}{n}$, $j = 0, \ldots, n$. In the case when only $S(x_j) = f(x_j)$ is required, we have Lagrange interpolation which is a well-studied problem. Most of the spline methods that construct Lagrange interpolants use cubic splines due to their simplicity to construct and the fact that they are C^2 functions. The problem of interpolating higher derivatives of f by C^2 quintic and higher order splines was considered in [2], and the references within; another treatment of the problem can be found in [1]. The error of approximation is measured by $O(n^{-r})$ where r is an integer related to the smoothness of the spline function. In the case of

V. Vatchev (✉)
University of Texas at Brownsville, One West University Boulevard,
Brownsville, TX 78520, USA
e-mail: vesselin.vatchev@utb.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*,
Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_21,

n equidistant nodes, the natural cubic spline interpolation is of order $O(n^{-4})$, i.e., the error decreases as constant multiple of n^{-4}.

In [4], we studied Hermite interpolation by splines at arbitrary nodes. The method used there can be summarized as first constructing a rough sketch of the interpolant and then refining it by a smoothing procedure. In this work, restricting the interpolating nodes to the case of equally spaced points allows us to use standard fast computational techniques. The resulting splines belong to C^3, are of varying degree (up to sixth), and provide $O(n^{k-3})$ order of approximation to $f^{(k)}$, $k = 0, 1, 2$.

The algorithm is introduced and the error estimate established in Sect. 2. The case of equidistant interpolation points is discussed in Sect. 3.

2 Hermite Spline Interpolation

The cardinal B-spline of order 1 for some real α is defined as the characteristic function $B_{1,\alpha}(x) = \chi_{[-\alpha,\alpha]}(x)$. The cardinal spline of order $k+1 > 0$ is defined recursively by the convolution $B_{k+1,\alpha}(x) = B_{k,\alpha} * B_{1,\alpha}(x) = \int B_{k,\alpha}(x - v)B_{1,\alpha}(v)\,dv$. The spline $B_{k+1,\alpha} \in C^{k-2}$ for $k > 1$ and is zero outside the interval $(-(k+1)\alpha, (k+1)\alpha)$. The central differences of f at x are defined recursively by $\Delta^1_\alpha f(x) = f(x+\alpha) - f(x-\alpha)$, $\Delta^{k+1}_\alpha f(x) = \Delta^k_\alpha f(x+\alpha) - \Delta^k_\alpha f(x-\alpha)$.

The result from the next lemma is an extension of the results about convolution with B-splines of any order presented in [4].

Theorem 1 *For a function* $f \in C^3[a - \frac{b-a}{2}, b + \frac{b-a}{2}]$ *and* $\alpha \le \frac{b-a}{8}$ *let*

$$T_\xi(x) = \frac{f''(\xi)}{2}(x-\xi)^2 + f'(\xi)(x-\xi) + f(\xi) - \frac{2}{3}f''(\xi)\alpha^2,$$
$$R_\alpha(x) = T_a(x)\chi_{[a-\frac{b-a}{2}, a+\frac{b-a}{2}]}(x) + T_b(x)\chi_{(b-\frac{b-a}{2}, b+\frac{b-a}{2}]}(x),$$

and

$$S_\alpha(x) = \frac{1}{(2\alpha)^4} B_{4,\alpha} * R_\alpha(x).$$

Then $S_\alpha \in C^3[a,b]$, $S^{(k)}_\alpha(a) = f^{(k)}(a)$, $S^{(k)}_\alpha(b) = f^{(k)}(b)$ *for* $r = 0, 1, 2$ *and* $S'''_\alpha(a) = S'''_\alpha(b) = 0$. *Furthermore,*

$$\|f^{(k)} - S^{(k)}_\alpha\|_{C[a,b]} \le C(b-a)^{3-k},$$

where C *depends only on* f.

Proof Since α is fixed throughout the proof we drop it from the notation. First, we compute the derivatives of S for $x \in [a,b]$. Since $S(x) = \frac{1}{(2\alpha)^4}\int_{-\alpha}^{\alpha} R * B_3(x+v)\,dv$ and $R * B_3(x+v)$ is continuous then it follows that

$$S'(x) = \frac{1}{(2\alpha)^4}(R * B_3(x+\alpha) - R * B_3(x-\alpha))$$
$$= \frac{1}{(2\alpha)^4}\int_{-\alpha}^{\alpha}(R(x+\alpha+v) - R(x-\alpha+v)) * B_2(v)\,dv$$
$$= \frac{1}{(2\alpha)^4}\int_{-\alpha}^{\alpha}\Delta^1 R(x+v) * B_2(v)\,dv.$$

Similarly, we get that

$$S''(x) = \frac{1}{(2\alpha)^4}\int_{-\alpha}^{\alpha}\Delta^2 R(x+v) * B_1(v)\,dv,$$

and since the integrand function is continuous we can differentiate once again and obtain the formula for the third derivative

$$S'''(x) = \frac{1}{(2\alpha)^4}\int_{-\alpha}^{\alpha}\Delta^3 R(x+v)\,dv. \tag{1}$$

In order to estimate S''' on $[a, b]$ we need to estimate $\Delta^3 R(x) = R(x+3\alpha) - 3R(x+\alpha) + 3R(x-\alpha) - R(x-3\alpha)$ on $x \in [a-\alpha, b+\alpha]$. From the Taylor expansion of f about y such that $a - \frac{b-a}{2} \le y \le \frac{a+b}{2}$ we have that

$$R(y) = \frac{f''(a)}{2}((y-a)^2 - \frac{4}{3}\alpha^2) + f'(a)(y-a) + f(a)$$
$$= f(y) - \frac{2}{3}f''(a)\alpha^2 - \frac{f'''(\xi(y))}{6}(y-a)^3,$$

for some $\xi(y)$ between a and y. Similarly, for $\frac{a+b}{2} < y \le b + \frac{b-a}{2}$ we have that

$$R(y) = f(y) - \frac{2}{3}f''(b)\alpha^2 - \frac{f'''(\xi(y))}{6}(y-b)^3,$$

for some $\xi(y)$ between b and x. By using the above identities in $\Delta^3 R(x)$ we get that $\Delta^3 R(x) = \Delta^3 f(x) + \frac{2r\alpha^2}{3}(f''(b) - f''(a)) - \frac{M(x)}{6}$, where

$$|M(x)| \le (|f'''(\xi(x+3\alpha))| + 3|f'''(\xi(x+\alpha))|)(b-a)^3$$
$$+ (3|f'''(\xi(x-\alpha))| + |f'''(\xi(x-3\alpha))|)(b-a)^3,$$

and hence $\|M\|_{C[a,b]} \le 8\|f'''\|_{C[a-\frac{b-a}{2}, b+\frac{b-a}{2}]}(b-a)^3$. The parameter r is such that

$$r=0 \quad \text{if} \quad x+v-3\alpha > \frac{a+b}{2}, \text{or} \quad x+v+3\alpha < \frac{a+b}{2},$$
$$r=-1 \quad \text{if} \quad x+v-3\alpha < \frac{a+b}{2} < x+v-\alpha, \text{or} \quad x+v+\alpha < \frac{a+b}{2} < x+v+3\alpha,$$
$$r=2 \quad \text{if} \quad x+v-\alpha < \frac{a+b}{2} < x+v+\alpha.$$

Since $\|\Delta^3 f\|_{C[a,b]} < C\|f'''\|_{C[a-\frac{b-a}{2},b+\frac{b-a}{2}]}(b-a)^3$, for a constant C that depends only on f and $f''(b)-f''(a) = f'''(\xi)(b-a)$ for some $\xi \in (a,b)$ and $\alpha < b-a$ then by substituting into (1) and using the triangle inequality we get that $\|S'''\|_{C[a,b]} < C\|f'''\|_{C[a-\frac{b-a}{2},b+\frac{b-a}{2}]}$. At the point a, we have that $[a-4\alpha, a+4\alpha] \subset [a-\frac{b-a}{2}, a+\frac{b-a}{2}]$ and hence $S^{(k)}(a)$ depends only on T_a. Direct computations, similar [4], show that $S^{(k)}(a) = f^{(k)}(a)$ for $k = 0, 1, 2$ and is 0 for $k = 3$. From the continuity of the integral, we can include the middle point, $\frac{a+b}{2}$ to the interval corresponding to T_b in the definition of R, i.e., $\tilde{R}(x) = T_a(x)\chi_{[a-\frac{b-a}{2},a+\frac{b-a}{2})}(x) + T_b(x)\chi_{[b-\frac{b-a}{2},b+\frac{b-a}{2}]}(x)$ and repeat the above consideration with a replaced by b. Finally, by construction $S^{(k)}(a) = f^{(k)}(a)$ for $k = 0, 1, 2$ and since both have continuous third derivatives from the Taylor formula for $f - S$ it follows that $|f^{(k)}(x) - S^{(k)}(x)| \le C|f'''(\xi) - S'''(\xi)|(b-a)^{3-k} \le C(b-a)^{3-k}$ where the constants depend only on f. □

Remark 1 If a function $f \in C^3[a,b]$ and has one-sided derivatives $f^{(k)}(a+) = \lim_{\varepsilon\to 0,\varepsilon>0} f^{(k)}(a+\varepsilon)$ and $f^{(k)}(b-) = \lim_{\varepsilon\to 0,\varepsilon>0} f^{(k)}(b-\varepsilon)$ then f can be extended on the left of a by $Q_a(x) = \frac{f'''(a+)}{6}(x-a)^3 + \frac{f''(a+)}{2}(x-a)^2 + f'(a+)(x-a) + f(a+)$ and on the right of b by $Q_b(x) = \frac{f'''(b-)}{6}(x-b)^3 + \frac{f''(b-)}{2}(x-b)^2 + f'(b-)(x-b) + f(b-)$. In that way the function

$$F = Q_a\chi_{(-\infty,a)} + f\chi_{[a,b]} + Q_b\chi_{(-\infty,a)} \in C^3(-\infty,\infty),$$

$F^{(k)}(a) = f^{(k)}(a)$, $F^{(k)}(b) = f^{(k)}(b)$, $k = 0, 1, 2, 3$ and $F'''(x) = f'''(a)$ for $x \le a$ and $F'''(x) = f'''(b)$ for $x \ge b$. In Theorem 1, the function F can be used instead of f and then the norm of f''' can be restricted only on $[a,b]$.

In the next section, we consider applications of the above construction in the case of equidistant interpolation points and discuss improving the interpolation on subintervals.

3 The Case of Equidistant Points

For a function $f \in C^3[0,1]$, we consider Hermite interpolation at the points $x_j = \frac{j}{n}$, $j = 0, \dots, n$, and $\alpha = \frac{1}{8n}$. Since α is a constant throughout the section we omit it from the notation. To each of the interpolation points x_j, we assign the polynomials T_ξ, defined in Theorem 1, and denote $T_j = T_{x_j}$,

$$T_j(x) = f(x_j) + f'(x_j)(x - x_j) + \frac{f''(x_j)}{2}\left((x - x_j)^2 - \frac{4}{3}\frac{1}{(8n)^2}\right).$$

Then we define R on the extended interval $\left[-\frac{1}{2n}, \frac{2n+1}{2n}\right]$ as

$$R(x) = \sum_{j=0}^{n} T_j(x)\chi_{I_j}(x),$$

where $I_j = \left(\frac{2j-1}{2n}, \frac{2j+1}{2n}\right]$, $j = 0, \ldots, n$ and refine it by convolution

$$S(x) = R * B_4(x). \tag{2}$$

From (2) it follows that S, restricted to the domain of interpolation $[0, 1]$, is a spline with nodes $\frac{j}{4n}$, $j = 0, \ldots, 4n$ and on each of the intervals between the nodes is an algebraic polynomial of degree 6. The following theorem holds true:

Theorem 2 *For $f \in C^3[0, 1]$ the spline S belongs to $C^3[0, 1]$ and for $k = 0, 1, 2$ satisfies the Hermite interpolation conditions $S^{(k)}(x_j) = f^{(k)}(x_j)$, $j = 0, \ldots, n$ and the error estimate*

$$\|f^{(k)} - S^{(k)}\|_{C[0,1]} \le Cn^{k-3},$$

where the constant C depends only on f.

Proof Since α is constant for any interval I_j then the proof is a corollary of Theorem 1 and Remark 1.

Since in the above $\alpha = \frac{1}{8n}$ is constant on the whole interval we can use the Fast Fourier Transform(FFT) for constructing the spline. In that way, we keep the complexity of computations of order $O(n)$ but increase the degree of the spline. In order to use FFT effectively we extend S periodically. Let $x_{-1} = -\frac{1}{n}$ and set $S^{(k)}(x_{-1}) = f^{(k)}(1)$, $k = 0, 1, 2, 3$. In [4], we constructed C^3 splines on $[-\frac{1}{n}, 0]$ that interpolate the Hermite conditions at the end points. The technique described above and Remark 1 also can be used for constructing an extension which third derivative is less than or equal to $\|f'''\|_{C[0,1]}$. The resulting function R is periodic with period $1 + \frac{1}{n}$ and belongs to C^3. The substitution $t = \frac{2\pi n}{n+1}(x + \frac{1}{n})$ transforms the problem to a Hermite interpolation for periodic functions on the interval $[0, 2\pi]$. The Fourier transform of f is $\hat{f}(\xi) = \int_0^{2\pi} f(x)e^{-i\xi x}\, dx$. It is well known, see [3], that $\hat{\chi}(\xi) = \frac{\sin\alpha\xi}{\alpha\xi}$ and hence $\hat{B}_\alpha(\xi) = \left(\frac{\sin\alpha\xi}{\alpha\xi}\right)^4$. For practical applications, FFT can be used for computing the convolution (the error estimates will increase by a factor of $\log n$).

The Hermite interpolation splines discussed above can be used to improve the approximation on subintervals. If after initial interpolation at $x_j = \frac{j}{n}$, $j = 0, 1, \ldots, n$ a better approximation is needed on certain subinterval, say $[x_k, x_{k+1}]$, $k < n - 1$ we can iterate the construction from Theorem 2 on that interval by decreasing

the estimate for the approximation error and preserving the C^3 smoothness of the interpolant.

Corollary 1 *Let $f \in C^3[0,1]$, S_0 be the spline from Theorem 2 that interpolates f, f', f'' at the points x_j, $j = 0,\dots,n$, and S_1 be the spline from Theorem 2 that interpolates f, f', f'' at the points $z_j = x_k + \frac{j}{nm}$, $j = 0,\dots,m$. Then the spline $S = S_1\chi_{[0,x_k]\cup[x_{k+1},1]} + S_2\chi_{[x_k,x_{k+1}]}$ interpolates f, f', f'' at the points $x_j \cup z_i$, $j = 0,\dots,n, i = 1,\dots,m-1$, $S \in C^3[0,1]$, and $\|f^{(r)} - S_2^{(r)}\|_{C[x_k,x_{k+1}]} < C(nm)^{r-3}$, $r = 0,1,2$ where C depends only on f.*

Proof The interpolation and the error estimate follow from Theorem 2. The third derivative at each of the interpolation points is 0 as it was shown in Theorem 1 and hence the spline $S \in C^3[0,1]$. □

References

1. Balabdaoui, F., Wellner, J.: Conjecture of error boundedness in a new Hermite interpolation problem via splines of odd-degree, Technical Report 480. University of Washington, Department of Statistics (2005)
2. Heß, W., Schmidt, J.: Positive quadratic, monotone quintic C^2-spline interpolation in one and two dimensions. J. Comput. Appl. Math. **55**(1), 51–67 (1994)
3. Katznelson, Y.: An Introduction to Harmonic Analysis, 2nd edn. Dover Publications Inc, New York (1976)
4. Vatchev, V.: An inverse of the running average operator for algebraic polynomials and its applications to shape preserving spline interpolation. Jaen J. Approx. **4**(1), 61–71 (2012)

Best Polynomial Approximation on the Unit Sphere and the Unit Ball

Yuan Xu

Abstract This is a survey on best polynomial approximation on the unit sphere and the unit ball. The central problem is to describe the approximation behavior of a function by polynomials via smoothness of the function. A major effort is to identify a correct gadget that characterizes smoothness of functions, either a modulus of smoothness or a K-functional, both of which are often equivalent. We concentrate on characterization of best approximations, given in terms of direct and converse theorems, and report several moduli of smoothness and K-functionals, including recent results that give a fairly satisfactory characterization of best approximation by polynomials for functions in L^p spaces, the space of continuous functions, and Sobolev spaces.

Keywords Best polynomial approximation · Unit sphere · Unit ball · Modulus of smoothness · K-functional

1 Introduction

One of the central problems in approximation theory is to characterize the error of approximation of a function by the smoothness of the function. In this paper, we make a short survey of best approximation by polynomials on the unit sphere $\mathbb{S}^{d-1}$ and the unit ball $\mathbb{B}^d$ in $\mathbb{R}^d$ with

$$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\} \quad \text{and} \quad \mathbb{B}^d = \{x : \|x\| \le 1\},$$

where $\|x\|$ denotes the Euclidean norm of x. To get a sense of the main problem and its solution, let us consider first $\mathbb{S}^1$ and $\mathbb{B}^1$.

Y. Xu (✉)
Department of Mathematics, University of Oregon, Eugene, OR 97403, USA
e-mail: yuan@uoregon.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_22,

If we parameterize $\mathbb{S}^1$ by $(\cos\theta, \sin\theta)$ with $\theta \in [0, 2\pi)$ and identify a function f defined on $\mathbb{S}^1$ with the 2π periodic function $g(\theta) = f(\cos\theta, \sin\theta)$, then polynomials on $\mathbb{S}^1$ are precisely trigonometric polynomials, so that polynomial approximation of functions on the circle $\mathbb{S}^1$ is the same as trigonometric approximation of 2π-periodic functions. Let $\mathscr{T}_n$ denote the space of trigonometric polynomials of degree at most n, $\mathscr{T}_n := \{a_0 + \sum_{k=1}^n a_k \cos k\theta + b_k \sin k\theta : a_k, b_k \in \mathbb{R}\}$. Let $\|\cdot\|_p$ denote the $L^p(\mathbb{S}^1)$ norm of 2π-periodic functions on $[0, 2\pi)$ if $1 \le p < \infty$, and the uniform norm of $C(\mathbb{S}^1)$ if $p = \infty$. For $f \in L^p(\mathbb{S}^1)$ if $1 \le p < \infty$, or $f \in C(\mathbb{S}^1)$ if $p = \infty$, define

$$E_n(f)_p := \inf_{t_n \in \mathscr{T}_n} \|f - t_n\|_p,$$

the error of best approximation by trigonometric polynomials. The convergence behavior of $E_n(f)_p$ is usually characterized by a modulus of smoothness. For $f \in L^p(\mathbb{S}^1)$ if $1 \le p < \infty$ or $f \in C(\mathbb{S}^1)$ if $p = \infty$, $r = 1, 2, \ldots$ and $t > 0$, the modulus of smoothness defined by the forward difference is

$$\omega_r(f; t)_p := \sup_{|\theta| \le t} \left\| \overrightarrow{\Delta}_\theta^r f \right\|_p, \qquad 1 \le p \le \infty,$$

where $\overrightarrow{\Delta}_h f(x) := f(x + h) - f(x)$ and $\overrightarrow{\Delta}_h^r := \overrightarrow{\Delta}_h^{r-1} \overrightarrow{\Delta}_h$. The characterization of best approximation on $\mathbb{S}^1$ is classical (cf. [11, 26]).

Theorem 1.1 *For $f \in L^p(\mathbb{S}^1)$ if $1 \le p < \infty$ or $f \in C(\mathbb{S}^1)$ if $p = \infty$,*

$$E_n(f)_p \le c\, \omega_r \left(f; n^{-1}\right)_p, \quad 1 \le p \le \infty, \quad n = 1, 2, \ldots. \tag{1}$$

On the other hand,

$$\omega_r(f; n^{-1})_p \le c\, n^{-r} \sum_{k=1}^n k^{r-1} E_{k-1}(f)_p, \qquad 1 \le p \le \infty. \tag{2}$$

The theorem contains two parts. The direct inequality (1) is called the Jackson estimate, its proof requires constructing a trigonometric polynomial that is close to the best approximation. The weak converse inequality (2) is called the Bernstein estimate as its proof relies on the Bernstein inequality. Throughout this paper, we let c, c_1, c_2 denote constants independent of f and n. Their values may differ at different times.

Another important gadget, often easier to use in theoretical studies, is the K-functional defined by

$$K_r(f, t)_p := \inf_{g \in W_p^r} \left\{ \|f - g\|_p + t^r \|g^{(r)}\|_p \right\},$$

where W_p^r denotes the Sobolev space of functions whose derivatives up to r-th order are all in $L^p(\mathbb{S}^1)$. The modulus of smoothness $\omega_r(f,t)_p$ and the K-function $K_r(f,t)_p$ are known to be equivalent: for some constants $c_2 > c_1 > 0$, independent of f and t,

$$c_1 K_r(f,t)_p \le \omega_r(f,t)_p \le c_2 K_r(f,t)_p. \tag{3}$$

All characterizations of best approximation, either on the sphere $\mathbb{S}^{d-1}$ or on the ball $\mathbb{B}^d$, encountered in this paper follow along the same line: we need to define an appropriate modulus of smoothness and use it to establish direct and weak converse inequalities; and we can often define a K-functional that is equivalent to the modulus of smoothness.

Convention In most cases, our direct and weak converse estimates are of the same form as those in (1) and (2). In those cases, we simply state that the direct and weak converse theorems hold and will not state them explicitly.

We now turn our attention to approximation by polynomials on the interval $\mathbb{B}^1 := [-1,1]$. Let Π_n denote the space of polynomials of degree n and let $\|\cdot\|_p$ also denote the L^p norm of functions on $[-1,1]$ as in the case of $\mathbb{S}^1$. For $f \in L^p(\mathbb{B}^1)$, $1 \le p < \infty$, or $f \in C(\mathbb{B}^1)$ for $p = \infty$, define

$$E_n(f)_p := \inf_{t_n \in \Pi_n} \|f - p_n\|_p, \qquad 1 \le p \le \infty.$$

The difficulty in characterizing $E_n(f)_p$ lies in the difference between approximation behavior at the interior and at the boundary of $\mathbb{B}^1$. It is well known that polynomial approximation on $\mathbb{B}^1$ displays a better convergence behavior at points close to the boundary than at points in the interior. A modulus of smoothness that is strong enough for both direct and converse estimates should catch this boundary behavior.

There are several successful definitions of modulus of smoothness in the literature. The most satisfactory one is that of Ditzian and Totik in [15]. For $r \in \mathbb{N}$ and $h > 0$, let $\widehat{\triangle}_h^r$ denote the central difference of increment h, defined by

$$\widehat{\triangle}_h f(x) = f(x + \tfrac{h}{2}) - f(x - \tfrac{h}{2}) \quad \text{and} \quad \widehat{\triangle}_h^r = \widehat{\triangle}_h^{r-1} \triangle, \quad r = 2, 3, \ldots. \tag{4}$$

Let $\varphi(x) := \sqrt{1-x^2}$. For $r = 1, 2 \ldots$, and $1 \le p \le \infty$, the Ditzian-Totik moduli of smoothness are defined by

$$\omega_\varphi^r(f,t)_p := \sup_{0 < h \le t} \left\| \widehat{\triangle}_{h\varphi}^r f \right\|_{L^p[-1,1]}, \tag{5}$$

where $\widehat{\triangle}_{h\varphi(x)}^r f(x) = 0$ if $x \pm rh\varphi(x)/2 \notin [-1, 1]$. Both direct theorem and weak converse theorem for $E_n(f)_p$ hold for this modulus of smoothness. Furthermore, the K-functional that is equivalent to this modulus of smoothness is defined by, for $t > 0$ and $r = 1, 2 \ldots$,

$$K_{r,\varphi}(f,t)_p := \inf_{g\in C^r[-1,1]} \left\{ \|f-g\|_p + t^r \|\varphi^r g^{(r)}\|_p \right\}. \tag{6}$$

In the rest of this paper, we discuss characterization of the best approximation on the sphere $\mathbb{S}^{d-1}$ and on the ball $\mathbb{B}^d$. The problem for higher dimension is much harder. For example, functions on $\mathbb{S}^{d-1}$ are no longer periodic, and there are interactions between variables for functions on $\mathbb{S}^{d-1}$ and $\mathbb{B}^d$.

The paper is organized as follows. The characterization of best approximation on the sphere is discussed in the next section, and the characterization on the ball is given in Sect. 3. In Sect. 4 we discuss recent results on Sobolev approximation on the ball, which are useful for spectral methods for numerical solution of partial differential equations. The paper ends with a problem on characterizing best polynomial approximation of functions in Sobolev spaces.

2 Approximation on the Unit Sphere

We start with necessary definitions on polynomial spaces and differential operators.

2.1 Spherical Harmonics and Spherical Polynomials

For $\mathbb{S}^{d-1}$ with $d \ge 3$, spherical harmonics play the role of trigonometric functions for the unit circle. There are many books on spherical harmonic—we follow [10]. Let $\mathscr{P}_n^d$ denote the space of real homogeneous polynomials of degree n and let Π_n^d denote the space of real polynomials of degree at most n. It is known that

$$\dim \mathscr{P}_n^d = \binom{n+d-1}{n} \quad \text{and} \quad \dim \Pi_n^d = \binom{n+d}{n}.$$

Let $\Delta := \partial_1^2 + \cdots + \partial_d^2$ denote the usual Laplace operator. A polynomial $P \in \Pi_n^d$ is called harmonic if $\Delta P = 0$. For $n = 0, 1, 2, \ldots$ let $\mathscr{H}_n^d := \left\{P \in \mathscr{P}_n^d : \Delta P = 0\right\}$ be the linear space of real harmonic polynomials that are homogeneous of degree n. Spherical harmonics are the restrictions of elements in $\mathscr{H}_n^d$ on the unit sphere. It is known that

$$a_n^d := \dim \mathscr{H}_n^d = \dim \mathscr{P}_n^d - \dim \mathscr{P}_{n-2}^d.$$

Let $\Pi_n^d(\mathbb{S}^{d-1})$ denote the space of polynomials restricted on $\mathbb{S}^{d-1}$. Then

$$\Pi_n^d(\mathbb{S}^{d-1}) = \bigoplus_{0\le j\le n/2} \mathscr{H}_{n-2j}^d\Big|_{\mathbb{S}^{d-1}} \quad \text{and} \quad \dim \Pi_n^d(\mathbb{S}^{d-1}) = \dim \mathscr{P}_n^d + \dim \mathscr{P}_{n-1}^d.$$

For $x \in \mathbb{R}^d$, write $x = r\xi$, $r \ge 0$, $\xi \in \mathbb{S}^{d-1}$. The Laplace operator can be written as

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{d-1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\Delta_0,$$

where Δ_0 is a differential operator on ξ, called the Laplace-Beltrami operator; see [10, Sect. 1.4]. The spherical harmonics are eigenfunctions of Δ_0. More precisely,

$$\Delta_0 Y(\xi) = -n(n+d-2)Y(\xi), \qquad Y \in \mathscr{H}_n^d.$$

The spherical harmonics are orthogonal polynomials on the sphere. Let $d\sigma$ be the surface measure, and ω_{d-1} be the surface area of $\mathbb{S}^{d-1}$. For $f, g \in L^1(\mathbb{S}^{d-1})$, define

$$\langle f, g\rangle_{\mathbb{S}^{d-1}} := \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} f(\xi) g(\xi) d\sigma(\xi).$$

If $Y_n \in \mathscr{H}_n^d$ for $n = 0, 1, \ldots$, then $\langle Y_n, Y_m\rangle_{\mathbb{S}^{d-1}} = 0$ if $n \ne m$. A basis $\{Y_\nu^n : 1 \le \nu \le a_n^d\}$ of $\mathscr{H}_n^d$ is called orthonormal if $\langle Y_\nu, Y_\mu\rangle_{\mathbb{S}^{d-1}} = \delta_{\nu,\mu}$. In terms of an orthonormal basis, the reproducing kernel $Z_{n,d}(\cdot,\cdot)$ of $\mathscr{H}_n^d$ can be written as $Z_{n,d}(x, y) = \sum_{1 \le \nu \le a_n^d} Y_\nu(x) Y_\nu(y)$, and the addition formula for the spherical harmonics states that

$$Z_{n,d}(x, y) = \frac{n+\lambda}{\lambda} C_n^\lambda(\langle x, y\rangle), \qquad \lambda = \frac{d-2}{2}, \tag{7}$$

where C_n^λ is the Gegenbauer polynomial of one variable. If $f \in L^2(\mathbb{S}^{d-1})$, then the Fourier orthogonal expansion of f can be written as

$$f = \sum_{n=0}^{\infty} \operatorname{proj}_n f, \qquad \operatorname{proj}_n : L^2(\mathbb{S}^{d-1}) \mapsto \mathscr{H}_n^d,$$

where the projection operator proj_n can be written as an integral

$$\operatorname{proj}_n f(x) = \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} f(y) Z_{n,d}(x, y) d\sigma(y).$$

For $f \in L^p(\mathbb{S}^{d-1})$, $1 \le p < \infty$, or $f \in C(\mathbb{S}^{d-1})$ if $p = \infty$, the error of best approximation by polynomials of degree at most n on $\mathbb{S}^{d-1}$ is defined by

$$E_n(f)_p := \inf_{P \in \Pi_n(\mathbb{S}^{d-1})} \|f - P\|_p, \qquad 1 \le p \le \infty,$$

where the norm $\|\cdot\|_p$ denotes the usual L^p norm on the sphere and $\|\cdot\|_\infty$ denotes the uniform norm on the sphere. Our goal is to characterize this quantity in terms of

some modulus of smoothness. The direct theorem of such a characterization requires a polynomial that is close to the least polynomial that approximates f. For $p = 2$, the n-th polynomial of best approximation is the partial sum,

$$S_n f = \sum_{k=0}^{n} \operatorname{proj}_k f,$$

of the Fourier orthogonal expansion, as the standard Hilbert space theory shows. For $p \neq 2$, a polynomial of near best approximation can be given in terms of a cut-off function, which is a C^∞-function η on $[0, \infty)$ such that $\eta(t) = 1$ for $0 \le \eta(t) \le 1$ and $\eta(t) = 0$ for $t \ge 2$. If η is such a function, define

$$S_{n,\eta} f(x) := \sum_{k=0}^{\infty} \eta\left(\frac{k}{n}\right) \operatorname{proj}_k f(x). \tag{8}$$

Since η is supported on $[0, 2]$, the summation in $S_{n,\eta} f$ can be terminated at $k = 2n - 1$, so that $S_{n,\eta} f$ is a polynomial of degree at most $2n - 1$.

Theorem 2.1 *Let $f \in L^p(\mathbb{S}^{d-1})$ if $1 \le p < \infty$ and $f \in C(\mathbb{S}^{d-1})$ if $p = \infty$. Then*

(1) $S_{n,\eta} f \in \Pi_n(\mathbb{S}^{d-1})$ and $S_{n,\eta} f = f$ for $f \in \Pi_n^d(\mathbb{S}^{d-1})$.
(2) For $n \in \mathbb{N}$, $\|S_{n,\eta} f\|_p \le c \|f\|_p$.
(3) For $n \in \mathbb{N}$, there is a constant $c > 0$, independent of f, such that

$$\|f - S_{n,\eta} f\|_p \le (1 + c) E_n(f)_p.$$

This near-best approximation was used for approximation on the sphere already in [18] and it has become a standard tool by now. For further information, including a sharp estimate of its kernel function, see [10].

2.2 First Modulus of Smoothness and K-functional

The first modulus of smoothness is defined in terms of spherical means.

Definition 2.2 *For $0 \le \theta \le \pi$ and $f \in L^1(\mathbb{S}^{d-1})$, define the spherical means*

$$T_\theta f(x) := \frac{1}{\omega_{d-1}} \int_{\mathbb{S}_x^\perp} f(x \cos\theta + u \sin\theta) d\sigma(u),$$

where $\mathbb{S}_x^\perp := \{y \in \mathbb{S}^{d-1} : \langle x, y \rangle = 0\}$. For $f \in L^p(\mathbb{S}^{d-1})$, $1 \le p < \infty$, or $C(\mathbb{S}^{d-1})$, $p = \infty$, and $r > 0$, define

$$\omega_r^*(f,t)_p := \sup_{|\theta| \le t} \|(I - T_\theta)^{r/2} f\|_p, \tag{9}$$

where $(I - T_\theta)^{r/2}$ is defined by its formal infinite series when $r/2$ is not an integer.

The equivalent K-functional of this modulus is defined by

$$K_r^*(f,t)_p := \inf_g \left\{ \|f - g\|_p + t^r \left\| (-\Delta_0)^{r/2} g \right\|_p \right\}, \tag{10}$$

where Δ_0 is the Laplace-Beltrami operator on the sphere and the infimum is taken over all g for which $(-\Delta_0)^{r/2} g \in L^p(\mathbb{S}^{d-1})$.

This modulus of smoothness was first defined and studied in [4, 23].

Theorem 2.3 *For $1 \le p \le \infty$, the modulus of smoothness $\omega_r^*(f,t)_p$ can be used to establish both direct and weak converse theorems, and it is equivalent to $K_r^*(f,t)_p$.*

The direct and the weak converse theorems were established in various stages by several authors (see [4, 17, 21–23, 27] for further references), before it was finally established in full generality by Rustamov [25]. A complete proof is given in [27] and a simplified proof can be found in [10].

The spherical means T_θ are multiplier operators of Fourier orthogonal series, i.e.,

$$\operatorname{proj}_n T_\theta f = \frac{C_n^\lambda(\cos\theta)}{C_n^\lambda(1)} \operatorname{proj}_n f, \qquad \lambda = \frac{d-2}{2}, \quad n = 0, 1, 2, \ldots. \tag{11}$$

This fact plays an essential role in studying this modulus of smoothness.

It should be mentioned that this multiplier approach can be extended to weighted approximation on the sphere, in which $d\sigma$ is replaced by $h_\kappa^2 d\sigma$, where h_κ is a function invariant under a reflection group. The simplest of such weight function is of the form

$$h_\kappa(x) = \prod_{i=1}^d |x_i|^{\kappa_i}, \qquad \kappa_i \ge 0, \quad x \in \mathbb{S}^{d-1},$$

when the group is $\mathbb{Z}_2^d$. Such weight functions were first considered by Dunkl associated with Dunkl operators. An extensive theory of harmonic analysis for orthogonal expansions with respect to $h_\kappa^2(x) d\sigma$ has been developed (cf. [8, 16]), in parallel with the classical theory for spherical harmonic expansions. The weighted best approximation in $L^p(h_\kappa^2; \mathbb{S}^{d-1})$ norm was studied in [29], where analogs of the modulus of smoothness $\omega_r^*(f,t)_p$ and K-functional $K_r^*(f,t)_p$ are defined with $\|\cdot\|_p$ replaced by the norm of $L^p(h_\kappa^2; \mathbb{S}^{d-1})$ for h_κ invariant under a reflection group, and a complete analog of Theorem 2.3 was established.

The advantages of the moduli of smoothness $\omega_r^*(f,t)_p$ are that they are well-defined for all $r > 0$ and they have a relatively simple structure through multipliers. These moduli, however, are difficult to compute even for simple functions.

2.3 Second Modulus of Smoothness and K-functional

The second modulus of smoothness on the sphere is defined through rotations on the sphere. Let $SO(d)$ denote the group of orthogonal matrix of determinant 1. For $Q \in SO(d)$, let $T(Q)f(x) := f(Q^{-1}x)$. For $t > 0$, define

$$O_t := \left\{ Q \in SO(d) : \max_{x \in \mathbb{S}^{d-1}} \mathrm{d}(x, Qx) \le t \right\},$$

where $\mathrm{d}(x, y) := \arccos \langle x, y \rangle$ is the geodesic distance on $\mathbb{S}^{d-1}$.

Definition 2.4 *For $f \in L^p(\mathbb{S}^{d-1})$, $1 \le p < \infty$, or $C(\mathbb{S}^{d-1})$, $p = \infty$, and $r > 0$, define*

$$\widetilde{\omega}_r(f, t)_p := \sup_{Q \in O_t} \|\triangle_Q^r f\|_p, \qquad \text{where} \quad \triangle_Q^r := (I - T_Q)^r. \tag{12}$$

For $r = 1$ and $p = 1$, this modulus of smoothness was introduced and used in [5] and further studied in [19]. For studying best approximation on the sphere, these moduli were introduced and investigated by Ditzian in [12] and he defined them for more general spaces, including $L^p(\mathbb{S}^{d-1})$ for $p > 0$.

Theorem 2.5 *The modulus of smoothness $\widetilde{\omega}_r(f, t)_p$ can be used to establish both direct and weak converse theorems for $1 \le p \le \infty$, and it is equivalent to the K-functional $K_r^*(f, t)_p$ for $1 < p < \infty$, but the equivalence fails if $p = 1$ or $p = \infty$.*

The direct and weak converse theorems were established in [13] and [12], respectively. The equivalence of $\widetilde{\omega}_r(f; t)_p$ and $K_r^*(f, t)_p$ for $1 < p < \infty$ was proved in [7], and the failure of the equivalence for $p = 1$ and ∞ was shown in [14].

The equivalence passes to the moduli of smoothness and shows in particular that $\widetilde{\omega}_r(f; t)_p$ is equivalent to the first modulus of smoothness $\omega_r^*(f; t)_p$ for $1 < p < \infty$ but not for $p = 1$ and $p = \infty$.

One advantage of the second moduli of smoothness $\widetilde{\omega}_r(f; t)_p$ is that they are independent of the choice of coordinates. These moduli, however, are also difficult to compute even for fairly simple functions.

2.4 Third Modulus of Smoothness and K-functional

The third modulus of smoothness on the sphere is defined in terms of moduli of smoothness of one variable on multiple circles. For $1 \le i, j \le d$, we let $\triangle_{i,j,t}^r$ be the r-rh forward difference acting on the angle of the polar coordinates on the (x_i, x_j) plane. For instance, take $(i, j) = (1, 2)$ as an example,

$$\triangle^r_{1,2,\theta} f(x) = \overrightarrow{\triangle}^r_\theta f\,(x_1\cos(\cdot) - x_2\sin(\cdot), x_1\sin(\cdot) + x_2\cos(\cdot), x_3, \ldots, x_d).$$

Notice that if $(x_i, x_j) = s_{i,j}(\cos\theta_{i,j}, \sin\theta_{i,j})$ then

$$(x_1\cos\theta - x_2\sin\theta, x_1\sin\theta + x_2\cos\theta) = s_{i,j}\cos(\theta_{i,j} + \theta),$$

so that $\triangle^r_{1,2,\theta} f(x)$ can be regarded as a difference on the circle of the (x_i, x_j) plane.

Definition 2.6 *For $r = 1, 2, \ldots$, $t > 0$, and $f \in L^p(\mathbb{S}^{d-1})$, $1 \le p < \infty$, or $f \in C(\mathbb{S}^{d-1})$ for $p = \infty$, define*

$$\omega_r(f,t)_p := \max_{1\le i<j\le d} \sup_{|\theta|\le t} \left\| \triangle^r_{i,j,\theta} f \right\|_p. \tag{13}$$

The equivalent K-functional is defined using the angular derivative

$$D_{i,j} := x_i\partial_j - x_j\partial_i = \frac{\partial}{\partial\theta_{i,j}}, \quad 1 \le i \ne j \le d$$

where $\theta_{i,j}$ is the angle of polar coordinates in (x_i, x_j)-plane defined as above. For $r \in \mathbb{N}_0$ and $t > 0$, the K-functional is defined by

$$K_r(f,t)_p := \inf_g \left\{ \|f - g\|_p + t^r \max_{1\le i<j\le d} \|D^r_{i,j} g\|_p \right\}, \tag{14}$$

where g is taken over all $g \in L^p(\mathbb{S}^{d-1})$ for which $D^r_{i,j} g \in L^p(\mathbb{S}^{d-1})$ for all $1 \le i, j \le d$.

Theorem 2.7 *The modulus of smoothness $\omega_r(f,t)_p$ can be used to establish both direct and weak converse theorems, and is equivalent to $K_r(f,t)_p$ for $1 \le p \le \infty$.*

These moduli and K-functionals were introduced in [8], where the above theorem was proved. Furthermore, it was also shown that

$$K_r(f, n^{-1})_p \sim \|f - S_{n,\eta} f\|_p + n^{-r} \max_{1\le i<j\le d} \|D^r_{i,j} S_{n,\eta} f\|_p,$$

where $S_{n,\eta}$ is the polynomial defined in (8).

For comparison with the other two moduli of smoothness, it was proved in [8] that for $r = 1, 2, \ldots$ and $1 \le p \le \infty$,

$$\omega_r(f,t)_p \le \widetilde{\omega}_r(f,t)_p, \qquad 0 < t < 1.$$

Furthermore, for $1 < p < \infty$, the two moduli of smoothness are equivalent if $r = 1$ or $r = 2$. Thus, the direct theorem with $\omega_r(f,t)_p$ is at least not weaker than the one with either one of the other two moduli of smoothness. Furthermore, all three

moduli are equivalent if $1 < p < \infty$ and $r = 1$ or 2. It remains an open problem if $\omega_r(f,t)_p$ is equivalent to the other two moduli of smoothness for $1 < p < \infty$ and $r \ge 3$ or for $p = 1$ and $p = \infty$.

The angular derivatives are related to the Laplace-Beltrami operator by

$$\Delta_0 = \sum_{1 \le i < j \le d} D_{i,j}^2.$$

Since the K-functional $K_r^*(f,t)_p$ is defined in terms of Δ_0 and the K-function $K_r(f,t)$ is defined in terms of $D_{i,j}$, it indicates that $K_r(f,t)_p$ may be stronger than $K_r^*(f,t)_p$ if we believe that the parts encode more information than the whole.

The main advantage of the modulus of smoothness $\omega_r(f,t)_p$ lies in the fact that it is defined in terms of moduli of smoothness of one variable, which allows us to tap into the well-established theory of trigonometric approximation of one variable, and it also means that $\omega_r(f,t)_p$ can be computed relatively easily (see [8] for examples).

One interesting phenomenon observed from the computational example is that the best approximation on $\mathbb{S}^{d-1}$ for $d \ge 3$ displays a boundary behavior rather like approximation by polynomials on $[-1,1]$. This is not all that surprising on second thought, but it does put $d = 2$ in approximation on $\mathbb{S}^{d-1}$ apart from $d \ge 3$.

3 Approximation on the Unit Ball

On the unit ball, we often work with weighted approximation with a fairly general weight function. We shall restrict our discussion to the classical weight function

$$w_\mu(x) := (1 - \|x\|^2)^{\mu - 1/2}, \qquad \mu > -1/2, \quad x \in \mathbb{B}^d,$$

for which the most has been done. We start with an account of orthogonal structure.

3.1 Orthogonal Structure on the Unit Ball

For the weight function W_μ, we consider the space $L^p(w_\mu, \mathbb{B}^d)$ for $1 \le p < \infty$ or $C(\mathbb{B}^d)$ when $p = \infty$. The norm of the space $L^p(w_\mu, \mathbb{B}^d)$ will be denoted by $\|f\|_{\mu,p}$, taken with the measure $w_\mu(x)dx$. The inner product of $L^2(w_\mu, \mathbb{B}^d)$ is defined by

$$\langle f, g \rangle_{\mu,\,p} := b_\mu \int_{\mathbb{B}^d} f(x) g(x) w_\mu(x) dx,$$

where b_μ is the normalization constant of w_μ such that $\langle 1, 1 \rangle_{\mu,\,p} = 1$. Let $\mathcal{V}_n^d(w_\mu)$ denote the space of polynomials of degree n that are orthogonal to polynomials in

Π_{n-1}^d with respect to the inner product $\langle \cdot, \cdot \rangle_{\mu,p}$. It is known that $\dim \mathcal{V}_n^d(w_\mu) = \binom{n+d-1}{n}$. The orthogonal polynomials in $\mathcal{V}_n^d(w_\mu)$ are eigenfunctions of a second-order differential operator: for $g \in \mathcal{V}_n^d(w_\mu)$,

$$\mathcal{D}_\mu g := \big(\Delta - \langle x, \nabla \rangle^2 - (2\mu + d - 1)\langle x, \nabla \rangle\big) g = -n(n + 2\mu + d - 1)g. \quad (15)$$

For $\nu \in \mathbb{N}_0^d$ with $|\nu| = n$, let P_ν^n denote an orthogonal polynomial in $\mathcal{V}_n^d(w_\mu)$. If $\{P_\nu^n : |\nu| = n\}$ is an orthonormal basis of $\mathcal{V}_n^d$, then the reproducing kernel $P_n(w_\mu; \cdot, \cdot)$ of $\mathcal{V}_n^d(w_\mu)$ can be written as $P_n(w_\mu; x, y) = \sum_{|\nu|=n} P_\nu^n(x) P_\nu^n(y)$. This kernel satisfies a closed-form formula [28] that will be given later in this section. Let $L^2(w_\mu, \mathbb{B}^d)$, then the Fourier orthogonal expansion of f can be written as

$$f = \sum_{n=0}^\infty \operatorname{proj}_n^\mu f, \qquad \operatorname{proj}_n^\mu : L^2(w_\mu, \mathbb{B}^d) \mapsto \mathcal{V}_n^d(w_\mu),$$

where the projection operator proj_n can be written as an integral

$$\operatorname{proj}_n^\mu f(x) = b_\mu \int_{\mathbb{B}^d} f(y) P_n(w_\mu; x, y) w_\mu(y) dy.$$

For $f \in L^p(w_\mu, \mathbb{B}^d)$, $1 \le p < \infty$, or $f \in C(\mathbb{B}^d)$ if $p = \infty$, the error of best approximation by polynomials of degree at most n is defined by

$$E_n(f)_{\mu,\,p} := \inf_{P \in \Pi_n^d} \|f - P\|_{\mu,\,p}, \qquad 1 \le p \le \infty.$$

The direct theorem for $E_n(f)_{\mu,p}$ is also established with the help of a polynomial that is a near-best approximation to f. For $p = 2$, the best polynomial of degree n is again the partial sum, $S_n^\mu f := \sum_{k=0}^n \operatorname{proj}_k^\mu f$, of the Fourier orthogonal expansion, whereas for $p \ne 2$ we can choose the polynomial as

$$S_{n,\eta}^\mu f(x) := \sum_{k=0}^\infty \eta\left(\frac{k}{n}\right) \operatorname{proj}_k^\mu f(x), \quad (16)$$

where η is a cut-off function as in (8). The analog of Theorem 2.1 holds for $S_{n,\eta}^\mu$ and $\|\cdot\|_{\mu,\,p}$ norm.

If μ is an integer or a half-integer, then the orthogonal structure of $L^2(w_\mu, \mathbb{B}^d)$ is closely related to the orthogonal structure on the unit sphere, which allows us to deduce many properties for analysis on the unit ball from the corresponding results on the unit sphere. The connection is based on the following identity: if d and m are positive integers, then for any $f \in L(\mathbb{S}^{d+m-1})$,

$$\int\limits_{\mathbb{S}^{d+m-1}} f(y) d\sigma_{d+m} = \int\limits_{\mathbb{B}^d} (1-\|x\|^2)^{\frac{m-2}{2}} \left[\int\limits_{\mathbb{S}^{m-1}} f\left(x, \sqrt{1-\|x\|^2}\xi\right) d\sigma_m(\xi) \right] dx.$$

This relation allows us to relate the space $\mathcal{V}_n^d(w_\mu)$ with $\mu = \frac{m-1}{2}$ directly to a subspace of $\mathcal{H}_n^{d+m}$, which leads to a relation between the reproducing kernels.

For $\mu = \frac{m-1}{2}$, the reproducing kernel $P_n(w_\mu; \cdot, \cdot)$ satisfies, for $m > 1$,

$$P_n(w_\mu; x, y) = \frac{1}{\omega_m} \int\limits_{\mathbb{S}^{m-1}} Z_{n, d+m}\left((x, x'), (y, \sqrt{1-\|y\|^2}\xi)\right) d\sigma_m(\xi),$$

where $(x, x') \in \mathbb{S}^{d+m-1}$ with $x \in \mathbb{B}^d$ and $x' = \|x'\|\xi \in \mathbb{B}^m$ with $\xi \in \mathbb{S}^{m-1}$, and it satisfies, for $m = 1$ and $y_{d+1} = \sqrt{1-\|y\|^2}$,

$$P_n(w_0; x, y) = \frac{1}{2}\left[Z_{n, d+m}\big((x, x'), (y, y_{d+1})\big) + Z_{n,d+m}\big((x, x'), (y, -y_{d+1})\big)\right].$$

Using the identity (7), we can then obtain a closed-form formula for $P_n(w_\mu; \cdot, \cdot)$, which turns out to hold for all real $\mu > -1/2$.

3.2 First Modulus of Smoothness and K-functional

The first modulus of smoothness on the unit ball is an analog of $\omega_r^*(f, t)_p$ on the sphere, defined in the translation operator T_θ^μ. Let I denote the identity matrix and

$$A(x) := (1-\|x\|^2)I + x^T x, \quad x = (x_1, \ldots, x_d) \in \mathbb{B}^d.$$

For W_μ on $\mathbb{B}^d$, the generalized translation operator is given by

$$T_\theta^\mu f(x) = b_\mu (1-\|x\|^2)^{\frac{d-1}{2}} \int\limits_{\Omega} f\left(\cos\theta x + \sin\theta \sqrt{1-\|x\|^2}\, u\right) \left(1 - uA(x)u^T\right)^{\mu - 1} du,$$

where Ω is the ellipsoid $\Omega = \{u : uA(x)u^T \le 1\}$ in $\mathbb{R}^d$.

Definition 3.1 *Let $f \in L^p(W_\mu, \mathbb{B}^d)$ if $1 \le p < \infty$, and $f \in C(\mathbb{B}^d)$ if $p = \infty$. For $r = 1, 2, \ldots,$ and $t > 0$, define*

$$\omega_r^*(f, t)_{\mu, p} := \sup_{|\theta| \le t} \|\triangle_{\theta, \mu}^r f\|_{p, \kappa}, \qquad \triangle_{\theta, \mu}^r f := \left(I - T_\theta^\mu\right)^{r/2} f.$$

The equivalent K-functional is defined via the differential operator $\mathscr{D}_\mu$ in (15),

$$K_r^*(f, t)_{\mu, p} := \inf_g \left\{ \|f - g\|_{\mu, p} + t^r \|\mathscr{D}_\mu^r g\|_{\mu, p} \right\},$$

where g is taken over all $g \in L^p(W_\mu, \mathbb{B}^d)$ for which $\mathscr{D}_\mu^r g \in L_p(W_\mu, \mathbb{B}^d)$.

Theorem 3.2 *For $1 \le p \le \infty$, the modulus of smoothness $\omega_r^*(f,t)_{\mu,p}$ can be used to establish both direct and weak converse theorems, and it is equivalent to $K_r^*(f,t)_{\mu,p}$.*

These moduli of smoothness and K-functionals were defined in [29] and Theorem 3.2 was also proved there. The integral formula of $T_\theta^\mu f$ was found in [30]. In fact, these results were established for more general weight functions of $h_\kappa^2 w_\mu$ with h_κ being a reflection invariant function. The operator T_θ^μ is a multiplier operator and satisfies

$$\operatorname{proj}_n^\mu \left(T_\theta^\mu f\right) = \frac{C_n^{\lambda_\mu}(\cos\theta)}{C_n^{\lambda_\mu}(1)} \operatorname{proj}_n^\mu f, \qquad \lambda_\mu = \mu + \frac{d-1}{2}, \quad n = 0, 1, \ldots,$$

which is an analog of (11). The proof of Theorem 3.2 can be carried out following the proof of Theorem 2.3.

The advantage of the moduli of smoothness $\omega_r^*(f,t)$ are that they are well-defined for all $r > 0$ and their connection to multipliers, just like the first moduli of smoothness on the sphere. These moduli, however, are difficult to compute even for simple functions.

3.3 Second Modulus of Smoothness and K-functional

The second modulus of smoothness is inherited from the third moduli of smoothness on the sphere. With a slight abuse of notation, we write $w_\mu(x) := (1 - \|x\|^2)^{\mu - \frac{1}{2}}$ for either the weight function on $\mathbb{B}^d$ or that on $\mathbb{B}^{d+1}$, and write $\triangle_{i,j,\theta}^r$ for either the difference operator on $\mathbb{R}^d$ or that on $\mathbb{R}^{d+1}$. This should not cause any confusion from the context. We denote by $\widetilde{f}$ the extension of f defined by

$$\widetilde{f}(x, x_{d+1}) = f(x), \qquad (x, x_{d+1}) \in \mathbb{B}^{d+1}, \quad x \in \mathbb{B}^d.$$

Definition 3.3 *Let $\mu = \frac{m-1}{2}$, $f \in L^p(w_\mu, \mathbb{B}^d)$ if $1 \le p < \infty$ and $f \in C(\mathbb{B}^d)$ if $p = \infty$. For $r = 1, 2 \ldots$, and $t > 0$, define*

$$\omega_r(f,t)_{p,\mu} := \sup_{|\theta| \le t} \left\{ \max_{1 \le i < j \le d} \|\triangle_{i,j,\theta}^r f\|_{L^p(\mathbb{B}^d, W_\mu)}, \right.$$
$$\left. \max_{1 \le i \le d} \|\triangle_{i,d+1,\theta}^r \widetilde{f}\|_{L^p(\mathbb{B}^{d+1}, W_{\mu-1/2})} \right\},$$

where for $m = 1$, $\|\triangle_{i,d+1,\theta}^r \widetilde{f}\|_{L^p(\mathbb{B}^{d+1}, W_{\mu-1/2})}$ is replaced by $\|\triangle_{i,d+1,\theta}^r \widetilde{f}\|_{L^p(\mathbb{S}^d)}$.

The equivalent K-functional is defined in terms of the angular derivatives $D_{i,j}$, and is defined for all $\mu \ge 0$ by

$$K_r(f,t)_{p,\mu} := \inf_{g\in C^r(\mathbb{B}^d)} \Big\{ \|f-g\|_{L^p(W_\mu;\mathbb{B}^d)} + t^r \max_{1\le i<j\le d} \|D^r_{i,j} g\|_{L^p(W_\mu;\mathbb{B}^d)} + t^r \max_{1\le i\le d} \|D^r_{i,d+1}\widetilde{g}\|_{L^p(W_{\mu-1/2};\mathbb{B}^{d+1})} \Big\},$$

where if $\mu = 0$, then $\|D^r_{i,d+1}\widetilde{g}\|_{L^p(W_{\mu-1/2};\mathbb{B}^{d+1})}$ is replaced by $\|D^r_{i,d+1}\widetilde{g}\|_{L^p(\mathbb{S}^d)}$.

Theorem 3.4 *Let $\mu = \frac{m-1}{2}$. For $1 \le p \le \infty$, the modulus of smoothness $\omega_r(f,t)_{\mu,p}$ can be used to establish both direct and weak converse theorems, and is equivalent to $K_r(f,t)_{\mu,p}$.*

The moduli of smoothness $\omega_r(f,t)_{p,\mu}$ and the K-functionals $K_r(f,t)_{p,\mu}$ were introduced in [8] and Theorem 3.4 was proved there. The proof relies heavily on the correspondence between $L^p(w_\mu, \mathbb{B}^d)$ and $L^p(\mathbb{S}^{d+m-1})$. In the definition of $\omega_r(f,t)_{\mu,p}$, the term that involves the difference of $\widetilde{f}$ may look strange but it is necessary, since $\triangle^r_{i,j,\theta}$ are differences in the spherical coordinates.

For comparison with the first modulus of smoothness $\omega^*_r(f,t)_{\mu,p}$, we only have that for $1 < p < \infty$, $r = 1, 2, \ldots$ and $0 < t < 1$,

$$\omega_r(f,t)_{p,\mu} \le c\omega^*_r(f,t)_{p,\mu}.$$

In all other cases, equivalences are open problems. Furthermore, the main results are established only for $\mu = \frac{m-1}{2}$, but they should hold for all $\mu \ge 0$ and perhaps even $\mu > -1/2$, which, however, requires a different proof from that of [8].

One interesting corollary is that, for $d = 1$, $\omega_r(f,t)_{\mu,p}$ defines a modulus of smoothness on $\mathbb{B}^1 = [-1,1]$ that is previously unknown. For $\mu = \frac{m-1}{2}$, this modulus is given by, for $f \in L^p(w_\mu, [-1,1])$,

$$\omega_r(f,t)_{p,mu} := \sup_{|\theta|\le t} \left(c_\mu \int_{\mathbb{B}^2} \left|\triangle^r_\theta f(x_1\cos(\cdot) + x_2\sin(\cdot))\right|^p w_{\mu-\frac{1}{2}}(x)dx \right)^{1/p}.$$

One advantage of the moduli of smoothness is that they can be relatively easily computed. Indeed, they can be computed just like the second modulus of smoothness on the sphere; see [8] for several examples.

3.4 Third Modulus of Smoothness and K-functional

The third modulus of smoothness on the unit ball is similar to $\omega_r(f,t)_{p,\mu}$, but with the term that involves the difference of $\widetilde{f}$ replaced by another term that resembles the difference in the Ditzian–Totik modulus of smoothness. To avoid the complication of the weight function, we state this modulus of smoothness only for $\mu = 1/2$ for which $w_\mu(x) = 1$. In this section, we write $\|\cdot\|_p := \|\cdot\|_{1/2,p}$.

Let e_i be the i-th coordinate vector of $\mathbb{R}^d$ and let $\widehat{\Delta}^r_{he_i}$ be the r-th central difference in the direction of e_i. More precisely,

$$\widehat{\Delta}_{he_i} f(x) := f(x + he_i) - f(x - he_i), \qquad \widehat{\Delta}^{r+1}_{he_i} f(x) = \widehat{\Delta}_{he_i} \widehat{\Delta}^r_{he_i} f(x).$$

As in the case of $[-1, 1]$, we assume that $\widehat{\Delta}^r_{he_i}$ is zero if either of the points $x \pm r\frac{h}{2} e_i$ does not belong to $\mathbb{B}^d$.

Definition 3.5 *Let $f \in L^p(\mathbb{B}^d)$ if $1 \le p < \infty$ and $f \in C(\mathbb{B}^d)$ if $p = \infty$. For $r = 1, 2, \ldots$ and $t > 0$,*

$$\omega^r_\varphi(f, t)_p := \sup_{0 < |h| \le t} \left\{ \max_{1 \le i < j \le d} \|\triangle^r_{i,j,h} f\|_p, \max_{1 \le i \le d} \|\widehat{\Delta}^r_{h\varphi e_i} f\|_p \right\}.$$

With $\varphi(x) := \sqrt{1 - \|x\|^2}$, the equivalent K-functional is defined by

$$K_{r,\varphi}(f, t)_p := \inf_{g \in W^r_p(\mathbb{B}^d)} \left\{ \|f - g\|_p + t^r \max_{1 \le i < j \le d} \|D^r_{i,j} g\|_p + t^r \max_{1 \le i \le d} \|\varphi^r \partial^r_i g\|_p \right\}.$$

Theorem 3.6 *For $1 \le p \le \infty$, the modulus of smoothness $\omega^r_\varphi(f, t)_{\mu,p}$ can be used to establish both direct and weak converse theorems, where the direct estimate takes the form*

$$E_n(f)_p \le c\, \omega^r_\varphi(f, n^{-1})_p + n^{-r} \|f\|_p$$

in which the additional term $n^{-r}\|f\|_p$ can be dropped when $r = 1$, and it is equivalent to $K_{r,\varphi}(f, t)$ in the sense that

$$c^{-1} \omega^r_\varphi(f, t)_p \le K_{r,\varphi}(f, t)_p \le c\, \omega^r_\varphi(f, t)_p + c\, t^r \|f\|_p,$$

where the term $t^r \|f\|_p$ on the right-hand side can be dropped when $r = 1$.

These moduli of smoothness and K-functionals were also defined in [8], and Theorem 3.6 was proved there. For $d = 1$, they agree with the Ditzian–Totik moduli of smoothness and K-functionals. The K-functional $K_{r,\varphi}(f, t)_{\mu,p}$ can be defined by replacing $\|\cdot\|_p$ with $\|\cdot\|_{\mu,p}$ in the definition of $K_{r,\varphi}(f, t)_p$, which were used to prove direct and weak converse theorems for $E_n(f)_{\mu,p}$ in terms of the K-functionals in [8].

For comparison with the second K-functional $K_r(f, t)_{\mu,p}$, which is only defined for $\mu = \frac{m-1}{2}$, $m = 1, 2, \ldots$, we know that for $1 \le p \le \infty$,

$$K_{1,\varphi}(f, t)_{\mu,p} \sim K_1(f, t)_{\mu,p}$$

and, for $r > 1$, there is a $t_r > 0$ such that

$$K_r(f, t)_{\mu,p} \le c\, K_{r,\varphi}(f, t)_{\mu,p} + c\, t^r \|f\|_{\mu,p}, \quad 0 < t < t_r,$$

where we need to assume that r is odd if $p = infty$. We can also state the result for comparison of the moduli of smoothness $\omega_{r,\varphi}(f,t)_p$ and $\omega_r(f,t)_{1/2,p}$ accordingly. The other direction of the equivalence for $r = 2, 3, \ldots$ remains open.

The advantages of the modulus of smoothness $\omega^r_\varphi(f,t)_p$ and the K-functional $\omega^r_\varphi(f,t)_p$ are that they are more intuitive, as direct extensions of the Ditizian–Totik modulus of smoothness and K-functional, and that the modulus of smoothness is relatively easy to compute.

4 Approximation in the Sobolev Space on the Unit Ball

For $r = 1, 2, \ldots$ we consider the Sobolev space $W^p_r(\mathbb{B}^d)$ with the norm defined by

$$\|f\|_{W^p_r(\mathbb{B}^d)} = \Big(\sum_{|\alpha| \le r} \|\partial^\alpha f\|_p \Big)^{1/p}.$$

The direct theorem given in terms of the K-functional yields immediately an estimate of $E_n(f)_p$ for functions in the Sobolev space. In the spectral method for solving partial differential equations, we often want estimates for the errors of derivative approximation as well. In this section, we again let $\|\cdot\|_p = \|\cdot\|_{1/2,p}$.

Approximation in Sobolev space requires estimates of derivatives. One such result was proved in [9], which includes the following estimates:

$$\|D^r_{i,j}(f - S^\mu_n f)\|_{p,\mu} \le cE_n(D^r_{i,j} f)_{p,\mu}, \quad 1 \le i < j \le d,$$

and a similar estimate that involves $D_{i,d+1}\widetilde{f}$. However, what we need is an estimate that involves only derivatives ∂^α instead of $D^r_{i,j}$. In this regard, the following result can be established:

Proposition 4.1 *If $f \in W^s_p(\mathbb{B}^d)$ for $1 \le p < \infty$, or $f \in C^s(\mathbb{B}^d)$ for $p = \infty$, then for $|\alpha| = s$,*

$$\big\|\phi^{|\alpha|/p}(\partial^\alpha f - \partial^\alpha S_{n,\eta} f)\big\|_p \le cE_{n-|\alpha|}(\partial^\alpha f)_p \le cn^{-s}\|f\|_{W^s_p(\mathbb{B}^d)}, \tag{17}$$

where $S_{n,\eta} f = S^{1/2}_{n,\eta} f$ is the near-best approximation defined in (16).

The estimate (17) in the proposition, however, is still weaker than what is needed in the spectral method, which requires an estimate similar to (17) but without the term $[\phi(x)]^{|\alpha|/p} = (1 - \|x\|^2)^{|\alpha|/p}$. It turns out that the near-best approximation $S_{n,\eta}$ is inadequate for obtaining such an estimate. What we need is the orthogonal structure of the Sobolev space $W^r_2(\mathbb{B}^d)$.

The orthogonal structure of $W^r_2(\mathbb{B}^d)$ was studied first in [32] for the case $r = 1$, and in [24, 31] for the case $r = 2$, and in [20] for general r. The inner product of $W^r_2(\mathbb{B}^d)$ is defined by

$$\langle f, g\rangle_{-s} := \langle \nabla^s f, \nabla^s g\rangle_{\mathbb{B}^d} + \sum_{k=0}^{\lceil \frac{s}{2}\rceil - 1} \langle \Delta^k f, \Delta^k g\rangle_{\mathbb{S}^{d-1}}.$$

Let $\mathcal{V}_n^d(w_{-s})$ denote the space of polynomials of degree n that are orthogonal to polynomials in Π_{n-1}^d with respect to the inner product $\langle\cdot,\cdot\rangle_{-s}$. Then $\mathcal{V}_n^d(w_{-1})$ satisfies a decomposition

$$\mathcal{V}_n^d(w_{-1}) = (1 - \|x\|^2)\mathcal{V}_{n-2}^d(w_1) \oplus \mathcal{H}_n^d,$$

where $\mathcal{H}_n^d$ is the space of spherical harmonics of degree n, and $\mathcal{V}_n^d(w_{-2})$ satisfies a decomposition

$$\mathcal{V}_n^d(w_{-2}) = (1 - \|x\|^2)^2\mathcal{V}_{n-4}^d(w_2) \oplus (1 - \|x\|^2)\mathcal{H}_{n-2}^d \oplus \mathcal{H}_n^d.$$

For each of these two cases, an orthonormal basis can be given in terms of the Jacobi polynomials and spherical harmonics, and the basis resembles the basis of $\mathcal{V}_n^d(w_\mu)$ for $\mu = -1$ and $\mu = -2$, which is why we adopt the notation $\mathcal{V}_n^d(w_{-s})$. The pattern of orthogonal decomposition, however, breaks down for $r > 2$. Nevertheless, an orthonormal basis can still be defined for $\mathcal{V}_n^d(w_{-s})$, which allows us to define an analog of the near-best polynomial $S_{n,\eta}^{-s} f$. The result for approximation in the Sobolev space is as follows:

Theorem 4.2 *Let $r, s = 1, 2, \ldots$ and $r \ge s$. If $f \in W_p^r(\mathbb{B}^d)$ with $r \ge s$ and $1 < p < \infty$. Then, for $n \ge s$,*

$$\|f - S_{n,\eta}^{-s} f\|_{W_p^k(\mathbb{B}^d)} \le c n^{-r+k} \|f\|_{W_p^r(\mathbb{B}^d)}, \quad k = 0, 1, \ldots, s,$$

where $S_{n,\eta}^{-s} f$ can be replaced by $S_n^{-s} f$ if $p = 2$.

This theorem is established in [20], which contains further refinements of such estimates in Sobolev spaces. The proof of this theorem, however, requires substantial work and uses a duality argument that requires $1 < p < \infty$.

The estimate in the theorem can be used to obtain an error estimate for the Galerkin spectral method, which looks for approximate solutions of a partial differential equations that are polynomials written in terms of orthogonal polynomials on the ball and their coefficients are determined by the Galerkin method. We refer to [20] for applications on a Helmholtz equation of second-order and a biharmonic equation of fourth-order on the unit ball. The method can also be applied to Poisson equations considered in [1–3].

These results raise the question of characterizing the best approximation by polynomials in Sobolev spaces, which is closely related to simultaneous approximation traditionally studied in approximation theory. But there are also distinct differences as the above discussion shows. We conclude this paper by formulating this problem in a more precise form.

Let Ω be a domain in $\mathbb{R}^d$ and w be a weight function on Ω. For $s = 1, 2 \ldots$, and $f \in W_p^s(w, \Omega)$. Define

$$E_n(f)_{W_p^s(w,\Omega)} := \inf_{p_n \in \Pi_n^d} \|f - p_n\|_{W_p^s(w,\Omega)}.$$

Problem 4.3 Establish direct and (weak) converse estimates of $E_n(f)_{W_p^s(w,\Omega)}$.

In the case of $\Omega = \mathbb{B}^d$ and $w(x) = 1$, Theorem 4.2 gives a direct estimate of $E_n(f)_{W_p^s(w,\, s\Omega)}$ for $f \in W_p^r(w, \Omega)$ with $r \geq s$. However, the estimate is weaker than what is needed. A direct estimate should imply that $E_n(f)_{W_p^s(w,\,\Omega)}$ goes to zero as $n \to \infty$ whenever $f \in W_p^s(w, \Omega)$. What this calls for is an appropriate K-functional, or a modulus of smoothness, for $f \in W_p^s(w, \Omega)$ that characterizes the best approximation $E_n(f)_{W_p^s(w,\,\Omega)}$.

References

1. Atkinson, K., Chien, D., Hansen, O.: A spectral method for elliptic equations: the Dirichlet problem. Adv. Comput, Math. **33**, 169–189 (2010)
2. Atkinson, K., Chien, D., Hansen, O.: A spectral method for elliptic equations: the Neumann problem. Adv. Comput, Math. **34**, 295–317 (2011)
3. Atkinson, K., Han, W.: Spherical Harmonics and Approximations on the Unit Sphere: An Introduction. Lecture Notes in Mathematics 2044. Springer, Heidelberg (2012)
4. 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)
5. Calderón, A.P., Weiss, G., Zygmund, A.: On the existence of singular integrals. In: Singular Integrals: Proceedings of the Symposium in Pure Mathematics, Chicago, IL, pp. 56–73. Amer. Math. Soc. Providence, RI (1966)
6. Dai, F., Ditzian, Z.: Jackson inequality for Banach spaces on the sphere. Acta Math. Hungar. **118**, 171–195 (2008)
7. Dai, F., Ditzian, Z., Huang, H.W.: Equivalence of measures of smoothness in $L^p(S^{d-1})$, $1 < p < \infty$. Studia Math. **196**, 179–205 (2010)
8. Dai, F., Xu, Y.: Moduli of smoothness and approximation on the unit sphere and the unit ball. Adv. Math. **224**(4), 1233–1310 (2010)
9. Dai, F., Xu, Y.: Polynomial approximation in Sobolev spaces on the unit sphere and the unit ball. J. Approx. Theory **163**, 1400–1418 (2011)
10. Dai, F., Xu, Y.: Approximation Theory and Harmonic Analysis on Spheres and Balls. Springer Monographs in Mathematics. Springer, New York (2013)
11. DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Springer, New York (1993)
12. Ditzian, Z.: A modulus of smoothness on the unit sphere. J. Anal. Math. **79**, 189–200 (1999)
13. Ditzian, Z.: Jackson-type inequality on the sphere. Acta Math. Hungar. **102**, 1–35 (2004)
14. Ditzian, Z.: Optimality of the range for which equivalence between certain measures of smoothness holds. Studia Math. **198**, 271–277 (2010)
15. Ditzian, Z., Totik, V.: Moduli of Smoothness. Springer, New York (1987)
16. Dunkl, C.F., Xu, Y.: Orthogonal Polynomials of Several Variables. In: Encyclopedia of Mathematics and its Applications 81. Cambridge University Press, Cambridge (2001)
17. Kalybin, G.A.: On moduli of smoothness of functions given on the sphere. Soviet Math. Dokl. **35**, 619–622 (1987)

18. Kamzolov, A.I.: The best approximation on the classes of functions $W_p^\alpha(S^n)$ by polynomials in spherical harmonics, Mat. Zametki **32**, 285–293. English transl in Math Notes **32**, 622–628 (1982)
19. Kurtz, D.S., Wheeden, R.L.: Results on weighted norm inequalities for multiplier. Trans. Amer. Math. Soc. **255**, 343–362 (1979)
20. Li, H., Xu, Y.: Spectral approximation on the unit ball, preprint (2013), arXiv:1310.2283
21. Nikolskii, S.M., Lizorkin, P.I.: Approximation of functions on the sphere, Izv. AN SSSR, Ser. Mat. **51**(3) , 635–651 (1987)
22. Nikolskii, S.M., Lizorkin, P.I.: Approximation on the sphere, survey, Translated from the Russian by Jerzy Trzeciak. Banach Center Publ., 22, Approximation and function spaces (Warsaw, 1986), pp. 281–292, PWN, Warsaw (1989)
23. Pawelke, S.: Über Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen. Tôhoku Math. J. **24**, 473–486 (1972)
24. Piñar, M., Xu, Y.: Orthogonal polynomials and partial differential equations on the unit ball. Proc. Amer. Math. Soc. **137**, 2979–2987 (2009)
25. Rustamov, KhP: On the approximation of functions on a sphere, (Russian), Izv. Ross. Akad. Nauk Ser. Mat. 57 (1993), 127–148; translation in Russian Acad. Sci. Izv. Math. **43**(2), 311–329 (1994)
26. Timan, A.F.: Theory of approximation of functions of a real variable, Translated from the Russian by J. Berry. Translation edited and with a preface by J. Cossar. Reprint of the 1963 English translation. Dover Publ. Inc., Mineola, New York (1994)
27. Wang, K.Y., Li, L.Q.: Harmonic Analysis and Approximation on the Unit Sphere. Science Press, Beijing (2000)
28. Xu, Y.: Summability of Fourier orthogonal series for Jacobi weight on a ball in $\mathbb{R}^d$. Trans. Amer. Math. Soc. **351**, 2439–2458 (1999)
29. Xu, Y.: Weighted approximation of functions on the unit sphere. Constr. Approx. **21**, 1–28 (2005)
30. Xu, Y.: Generalized translation operator and approximation in several variables. J. Comp. Appl. Math. **178**, 489–512 (2005)
31. Xu, Y.: A family of Sobolev orthogonal polynomials on the unit ball. J. Approx. Theory **138**, 232–241 (2006)
32. Xu, Y.: Sobolev orthogonal polynomials defined via gradient on the unit ball. J. Approx. Theory **152**, 52–65 (2008)

Support Vector Machines in Reproducing Kernel Hilbert Spaces Versus Banach Spaces

Qi Ye

Abstract In this article, we compare the support vector classifiers in Hilbert spaces versus those in Banach spaces. Recently, we developed a new concept of reproducing kernel Banach spaces (RKBSs). These spaces are a natural generalization of reproducing kernel Hilbert spaces (RKHSs) by extending the reproduction property from inner products to dual bilinear products. Based on the techniques of Fourier transforms, we can construct RKBSs by many well-known positive definite functions, e.g., Matérn functions and Gaussian functions. In addition, we can obtain finite-dimensional solutions of support vector machines defined in infinite-dimensional RKBSs. Finally, the numerical examples provided in this paper show that the solution of support vector machines in a RKBS can be computed and easily coded just as the classical algorithms given in RKHSs.

Keywords Support vector machine · Reproducing kernel Banach space · Positive definite function · Matérn function · Sobolev spline · Gaussian function

1 Introduction

The review paper [9] states that kernel-based approximation methods have become a general mathematical tool in the fields of machine learning and meshfree approximation. One of the popular supervised learning models is the support vector machine for classification and regression analysis. Many of the results in support vector machines have involved reproducing kernel Hilbert spaces (RKHSs). Recently, several papers [3, 10–12, 16–18] have carried out abstract theoretical results of support vector machines in reproducing kernel Banach spaces (RKBSs) while there are few

Q. Ye (✉)
Department of Mathematics, Syracuse University, 215 Carnegie Building, Syracuse, NY 13244, USA
e-mail: qiye@syr.edu

G. E. Fasshauer and L. L. Schumaker (eds.), *Approximation Theory XIV: San Antonio 2013*, Springer Proceedings in Mathematics & Statistics 83, DOI: 10.1007/978-3-319-06404-8_23,

papers that mention practical numerical tests of support vector classifiers in RKBSs. The goal of this article is to show that the solutions of support vector machines in RKBSs also work well for programming.

The classical works of support vector machines focus on how to minimize empirical regularized risks in the RKHS $\mathscr{H}$ by the given training data $\{(\mathbf{x}_k, y_k)\}_{k=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}$, i.e.,

$$\min_{f\in\mathscr{H}} \sum_{j=1}^{N} \frac{1}{N} L\big(\mathbf{x}_j, y_j, f(\mathbf{x}_j)\big) + R\big(\|f\|_{\mathscr{H}}\big),$$

where L is a loss function and R is a regularization function (see [1, 13, 14]). Since Banach spaces have different geometric structures, people have a great interest in the extension of support vector machines from Hilbert spaces to Banach spaces. The recent papers [3, 17] give a new concept of RKBSs. These spaces are the generalization of the reproduction property from inner products to dual bilinear products. Moreover, support vector machines can be well-posed in such a RKBS $\mathscr{B}$, i.e.,

$$\min_{f\in\mathscr{B}} \sum_{j=1}^{N} \frac{1}{N} L\left(\mathbf{x}_j, y_j, f(\mathbf{x}_j)\right) + R\big(\|f\|_{\mathscr{B}}\big).$$

The main ideas of this article are based on the theoretical results of our recent paper [3] that shows how to construct RKBSs by positive definite functions. We find that the solutions of support vector machines in RKBSs are still spanned by finite-dimensional kernel bases so that a global minimizer over the infinite-dimensional space belongs to some known finite-dimensional space. This allows to develop numerical algorithms of support vector classifiers defined in RKBSs. We find that the formulas for the solutions of support vector machines in RKBSs could be different from those in RKHSs when both RKBSs and RKHSs are introduced by the same positive definite functions. For examples, the RKHS $\mathscr{H}_\Phi(\mathbb{R}^2)$ and the RKBS $\mathscr{B}_\Phi^4(\mathbb{R}^2)$ are constructed by the same positive definite function Φ (see Theorem 1). The solutions of the support vector machines in the RKHS $\mathscr{H}_\Phi(\mathbb{R}^2)$ given in Eq. (11) are set up by the reproducing kernel but the solutions of the support vector machines in the RKBS $\mathscr{B}_\Phi^4(\mathbb{R}^2)$ given in Eq. (14) are induced by a different kernel function. This discovery gives a novel formula for learning solutions in Banach spaces which is different from the paper [8].

Finally, we summarize the structure of this article. We mainly focus on a classical binary classification problem described in Sect. 2. Then Sect. 3 primarily reviews the definitions and theorems mentioned in [3]. In Sect. 4, we illustrate the solutions of support vector machines in RKBSs by two examples of Matérn functions and Gaussian functions. For the set of training and testing data given in Sect. 2, we compare the performance of the support vector classifiers defined in RKHSs and RKBSs (see Sect. 5).

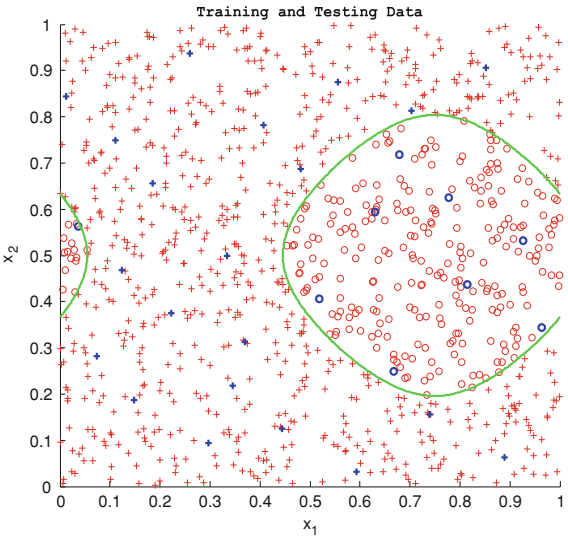


Fig. 1 A classification example in two dimensions. The classes are coded as a binary variable (*cross* $= +1$ and *circle* $= -1$). The *green line* is the original decision boundary defined by $g(\mathbf{x}) = 0$ where $g(\mathbf{x}) := \sin(2\pi x_1)/2 + \cos(2\pi x_2)/2 + 1/3$. The *blue* symbols denote the training data and the *red* symbols denote the testing data

2 Background

We discuss the support vector machines starting from simple binary classification, which is learning a class from its positive and negative samples. Standard binary classification can be presented as follows.

Given the training data consisting of N pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, with $\mathbf{x}_j \in \mathbb{R}^d$ and $y_j \in \{\pm 1\}$, we want to find a nonlinear separable boundary to classify two different classes of the training data. This separable boundary could be represented by a decision function $f : \mathbb{R}^d \to \mathbb{R}$, i.e., $\left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 0\right\}$. In other words, we can create a classification rule $r : \mathbb{R}^d \to \{\pm 1\}$ induced by the decision function f, i.e.,

$$r(\mathbf{x}) := \text{sign}\left(f(\mathbf{x})\right),$$

such that $r(\mathbf{x}_j) = y_j$ or $y_j f(\mathbf{x}_j) > 0$ for all $j = 1, \ldots, N$.

In addition, we will apply another set of testing data $(\tilde{\mathbf{x}}_1, \tilde{y}_1), \ldots, (\tilde{\mathbf{x}}_M, \tilde{y}_M)$ to compute the absolute mean error for the decision function f, i.e.,

$$\text{Error} := \frac{1}{M}\sum_{k=1}^{M} I\left(\tilde{y}_k \neq r(\tilde{\mathbf{x}}_k)\right) = \frac{1}{M}\sum_{k=1}^{M}\left|\frac{1 - \text{sign}\left(\tilde{y}_k f(\tilde{\mathbf{x}}_k)\right)}{2}\right|,$$

where $I\,(a \neq b)$ is 1 if $a \neq b$ and is 0 if $a = b$.

In our numerical experiments, we will use a test function $g : \mathbb{R}^2 \to \mathbb{R}$ to generate the training and testing data, i.e., $y_j := \text{sign}(g(\mathbf{x}_j))$ for $j = 1, \ldots, N$ and $\tilde{y}_k := \text{sign}(g(\tilde{\mathbf{x}}_k))$ for $k = 1, \ldots, M$, given in Fig. 1. In the following sections, we will show

how to solve for the optimal decision functions from support vector machines defined in RKHSs and RKBSs. These training data will be used to obtain the support vector classifiers in RKHSs and RKBSs, and we will compare these two classifications with the help of the testing data.

3 Reproducing Kernel Hilbert and Banach Spaces

For the reader's convenience, we will review the definitions and theorems of the RKHSs and RKBSs mentioned in the books [4, 13, 15] and the papers [3, 16, 17].

First we look at the definitions of RKHSs and RKBSs.

Definition 1 [15, Definition 10.1] Let $\mathscr{H}$ be a Hilbert space consisting of functions $f : \mathbb{R}^d \to \mathbb{R}$. $\mathscr{H}$ is called a *reproducing kernel Hilbert space* (RKHS) and a kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a *reproducing kernel* for $\mathscr{H}$ if

$$(I)\ K(\cdot, \mathbf{y}) \in \mathscr{H} \text{ and } (II)\ f(\mathbf{y}) = (f, K(\cdot, \mathbf{y}))_{\mathscr{H}}, \quad \text{for all } f \in \mathscr{H} \text{ and all } \mathbf{y} \in \mathbb{R}^d,$$

where $(\cdot, \cdot)_{\mathscr{H}}$ is used to denote the inner product of $\mathscr{H}$.

In this article, $\langle \cdot, \cdot \rangle_{\mathscr{B}}$ denotes the dual bilinear product on a Banach space $\mathscr{B}$ and its dual space $\mathscr{B}'$, i.e.,

$$\langle f, G \rangle_{\mathscr{B}} := G(f), \quad \text{for all } G \in \mathscr{B}' \text{ and all } f \in \mathscr{B}.$$

Obviously the dual space of a Hilbert space is equal to itself and the inner product can be viewed as a dual bilinear product of a Hilbert space. Then we can generalize RKHSs to RKBSs by replacing inner products with dual bilinear products.

Definition 2 [3, Definition 3.1] Let $\mathscr{B}$ be a Banach space composed of functions $f : \mathbb{R}^d \to \mathbb{R}$, whose dual space (i.e., the space of continuous linear functionals) $\mathscr{B}'$ is isometrically equivalent to a normed space $\mathscr{F}$ consisting of functions $g : \mathbb{R}^d \to \mathbb{R}$. Note that $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel function.

We call $\mathscr{B}$ a *right-sided reproducing kernel Banach space* (RKBS) and K its *right-sided reproducing kernel* if

$$(I)\ K(\cdot, \mathbf{x}) \in \mathscr{F} \equiv \mathscr{B}' \text{ and } (II)\ f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathscr{B}}, \quad \text{for all } f \in \mathscr{B} \text{ and all } \mathbf{x} \in \mathbb{R}^d.$$

If the Banach space $\mathscr{B}$ reproduces from the other side, i.e.,

$$(III)\ K(\mathbf{y}, \cdot) \in \mathscr{B} \text{ and } (IV)\ \langle K(\mathbf{y}, \cdot), g \rangle_{\mathscr{B}} = g(\mathbf{y}), \quad \text{for all } g \in \mathscr{F} \equiv \mathscr{B}' \text{ and all } \mathbf{y} \in \mathbb{R}^d,$$

then $\mathscr{B}$ is called a *left-sided reproducing kernel Banach space* and K its *left-sided reproducing kernel*.

For two-sided reproduction as above, we say that $\mathscr{B}$ is a *two-sided reproducing kernel Banach space* with the *two-sided reproducing kernel* K.

Remark 1 Actually, RKHSs and RKBSs can be defined on locally compact Hausdorff spaces equipped with finite Borel measures. To simplify the notation and discussion, we only consider the whole space $\mathbb{R}^d$ as the domain in this article.

Comparing Definitions 1 and 2, a RKHS is obviously a special case of a RKBS. More precisely, the reproductions (I) and (III) of RKBSs are a generalization of the reproduction (I) of RKHSs, and the reproductions (II) and (IV) of RKBSs are an extension of the reproduction of (II) for RKHSs.

3.1 Positive Definite Functions

Next we will look at a two-sided RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ driven by a positive definite function Φ given in [3].

Definition 3 [15, Definition 6.1] A continuous even function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is called *positive definite* if, for all $N \in \mathbb{N}$ and all sets of pairwise distinct centers $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$, the quadratic form

$$\sum_{j,k=1}^{N,N} c_j c_k \Phi(\mathbf{x}_j - \mathbf{x}_k) > 0, \quad \text{for all nonzero } \mathbf{c} := (c_1, \ldots, c_N)^T \in \mathbb{R}^N.$$

Definition 3 indicates that the function Φ is positive definite if and only if all its associated matrices $\mathsf{A}_{\Phi,X} := \big(\Phi(\mathbf{x}_j - \mathbf{x}_k)\big)_{j,k=1}^{N,N}$ are positive definite. In the historical terminology, the positive definite function given in Definition 3 may be called a strictly positive definite function. In this article we want to use the same definition of positive definite functions as in the books [4, 15] and the paper [3].

[15, Theorem 6.11] assures that the Fourier transform $\hat{\Phi}$ of the positive definite function $\Phi \in \mathrm{L}_1(\mathbb{R}^d) \cap \mathrm{C}(\mathbb{R}^d)$ is nonnegative and nonvanishing. Therefore, we can apply the Fourier transform $\hat{\Phi}$ to define a normed space

$$\mathscr{B}_\Phi^p(\mathbb{R}^d) := \Big\{ f \in \mathrm{L}_p(\mathbb{R}^d) \cap \mathrm{C}(\mathbb{R}^d) : \text{ the distributional Fourier transform } \hat{f} \text{ of } f \text{ is a measurable function such that } \hat{f}/\hat{\Phi}^{1/q} \in \mathrm{L}_q(\mathbb{R}^d) \Big\}, \quad (1)$$

equipped with the norm

$$\|f\|_{\mathscr{B}_\Phi^p(\mathbb{R}^d)} := \left((2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\big|\hat{f}(\mathbf{x})\big|^q}{\hat{\Phi}(\mathbf{x})} \mathrm{d}\mathbf{x} \right)^{1/q}, \quad (2)$$

where $1 < q \leq 2 \leq p < \infty$ and $p^{-1} + q^{-1} = 1$. In particular, when $f \in \mathrm{L}_1(\mathbb{R}^d)$ then $\hat{f}$ is the L_1-Fourier transform of f, i.e., $\hat{f}(\mathbf{x}) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\mathbf{y}) e^{-i\mathbf{x}^T\mathbf{y}} \mathrm{d}\mathbf{y}$, where i is the imaginary unit, i.e., $i^2 = -1$.

Combining the results of [3, Theorem 4.1 and Corollary 4.2] we can obtain the reproduction properties for the space $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ as follows.

Theorem 1 *Let $1 < q \leq 2 \leq p < \infty$ and $p^{-1} + q^{-1} = 1$. Suppose that $\Phi \in \mathrm{L}_1(\mathbb{R}^d) \cap \mathrm{C}(\mathbb{R}^d)$ is a positive definite function on $\mathbb{R}^d$ and that $\hat{\Phi}^{q-1} \in \mathrm{L}_1(\mathbb{R}^d)$. Then $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ given in Eqs. (1 and 2) is a two-sided reproducing kernel Banach space with the two-sided reproducing kernel*

$$K(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x} - \mathbf{y}), \quad \textit{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

In particular, when $p = 2$ then $\mathscr{B}_\Phi^2(\mathbb{R}^d) = \mathscr{H}_\Phi(\mathbb{R}^d)$ is a reproducing kernel Hilbert space.

Remark 2 The paper [3] discusses the RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ in complex values. As in the statement of [3, Remark 5.2], the restriction of the theorems given in [3] to the reals does not affect their conclusions by [7, Proposition 1.9.3]. In the interest of reducing the complexity, we just regard the real RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ in this article.

3.2 Support Vector Machines

Now we will find the optimal decision function of the binary classification by the support vector machines defined in the RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ based on the given finitely many training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \in \mathbb{R}^d \times \{\pm 1\}$.

Suppose that the regularization function $R : [0, \infty) \to [0, \infty)$ is convex and strictly increasing, and the loss function $L : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to [0, \infty)$ such that $L(\mathbf{x}, y, \cdot)$ is a convex map for any fixed $\mathbf{x} \in \mathbb{R}^d$ and any fixed $y \in \mathbb{R}$. Using the loss function L and the regularization function R, we can construct the minimization problem (support vector machine)

$$\min_{f \in \mathscr{B}_\Phi^p(\mathbb{R}^d)} \frac{1}{N} \sum_{j=1}^{N} L\big(\mathbf{x}_j, y_j, f(\mathbf{x}_j)\big) + R\Big(\|f\|_{\mathscr{B}_\Phi^p(\mathbb{R}^d)}\Big). \tag{3}$$

According to [3, Theorem 4.4] we can obtain finite-dimensional representations of support vector machine solutions of (3).

Theorem 2 *Let $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ for $2 \leq p < \infty$ be defined in Theorem 1. Then the support vector machine (3) has the unique optimal solution*

$$s(\mathbf{x}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{\Phi}(\mathbf{y})^{p-1} \sum_{k=1}^{N} c_k e^{i(\mathbf{x}-\mathbf{x}_k)^T \mathbf{y}} \left| \sum_{l=1}^{N} c_l e^{-i\mathbf{x}_l^T \mathbf{y}} \right|^{p-2} \mathrm{d}\mathbf{y}, \quad \textit{for } \mathbf{x} \in \mathbb{R}^d, \tag{4}$$

where $c_1, \ldots, c_N \in \mathbb{R}$. *Moreover, the norm of* s *has the form*

$$\|s\|_{\mathscr{B}_\Phi^p(\mathbb{R}^d)} = \left(\sum_{k=1}^{N} c_k s(\mathbf{x}_k) \right)^{1/q}. \tag{5}$$

The classical representer theorem [13, Theorem 5.5] provides that the support vector machine solution in the RKHS $\mathscr{H}_\Phi(\mathbb{R}^d)$ has a simple representation in terms of Φ, i.e.,

$$s(\mathbf{x}) := \sum_{k=1}^{N} c_k \Phi(\mathbf{x} - \mathbf{x}_k), \quad \text{for } \mathbf{x} \in \mathbb{R}^d.$$

Should we also be able to obtain a well-computable formula for s similar to the classical ones? To this end, we are going to simplify Eq. (4) when p is an even integer, i.e., $p = 2n$ for $n \in \mathbb{N}$.

The fact that $\hat{\Phi}^{p-1} \in \mathrm{L}_1(\mathbb{R}^d) \cap \mathrm{C}(\mathbb{R}^d)$ guarantees that $\hat{\Phi}^{p-1}$ has the L_1-inverse Fourier transform $\Phi^{*(p-1)}$, i.e., $\hat{\Phi}^{p-1} = \mathscr{F}\left(\Phi^{*(p-1)}\right)$, where $\mathscr{F}$ is the Fourier transform map. Obviously $\Phi^{*1} = \Phi$. Furthermore, it is easy to check that $\Phi^{*(p-1)}$ is even and continuous. Subsequently, we can introduce a typical representer theorem of the RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ using the kernel function $\Phi^{*(p-1)}$ when p is an even number.

Theorem 3 *If* $p = 2n$ *for some* $n \in \mathbb{N}$, *then the support vector machine solution given in Eq.* (4) *can be rewritten as*

$$s(\mathbf{x}) = \sum_{k_1, \ldots, k_{p-1}=1}^{N, \ldots, N} \left(\prod_{j=1}^{p-1} c_{k_j} \right) \Phi^{*(p-1)} \left(\mathbf{x} + \sum_{l=1}^{p-1} (-1)^l \mathbf{x}_{k_l} \right), \quad \textit{for } \mathbf{x} \in \mathbb{R}^d, \tag{6}$$

and its norm has the form

$$\|s\|_{\mathscr{B}_\Phi^p(\mathbb{R}^d)} = \left(\sum_{k_0, k_1, \ldots, k_{p-1}=1}^{N, N, \ldots, N} \left(\prod_{j=0}^{p-1} c_{k_j} \right) \Phi^{*(p-1)} \left(\sum_{l=0}^{p-1} (-1)^l \mathbf{x}_{k_l} \right) \right)^{1/q}, \tag{7}$$

where $c_1, \ldots, c_N \in \mathbb{R}$ *and* $\Phi^{*(p-1)}$ *is the inverse transform of* $\hat{\Phi}^{p-1}$.

Proof Expanding Eq. (4) for $p = 2n$, we can obtain that

$$
\begin{aligned}
s(\mathbf{x}) &= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{\Phi}(\mathbf{y})^{p-1} \sum_{k_1,\ldots,k_{p-1}=1}^{N,\ldots,N} \left(\prod_{j=1}^{p-1} c_{k_j} \right) e^{i\mathbf{y}^T \left(\mathbf{x} + \sum_{l=1}^{p-1} (-1)^l \mathbf{x}_{k_l} \right)} \mathrm{d}\mathbf{y} \\
&= \sum_{k_1,\ldots,k_{p-1}=1}^{N,\ldots,N} \left(\prod_{j=1}^{p-1} c_{k_j} \right) (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathscr{F} \left(\Phi^{*(p-1)} \right) (\mathbf{y}) e^{i\mathbf{y}^T \left(\mathbf{x} + \sum_{l=1}^{p-1} (-1)^l \mathbf{x}_{k_l} \right)} \mathrm{d}\mathbf{y} \\
&= \sum_{k_1,\ldots,k_{p-1}=1}^{N,\ldots,N} \left(\prod_{j=1}^{p-1} c_{k_j} \right) \Phi^{*(p-1)} \left(\mathbf{x} + \sum_{l=1}^{p-1} (-1)^l \mathbf{x}_{k_l} \right), \quad \text{for } \mathbf{x} \in \mathbb{R}^d.
\end{aligned}
$$

Putting the above expansion of s into Eq. (5), we can check that the $\mathscr{B}_\Phi^p(\mathbb{R}^d)$-norm of s can be also written as in Eq. (7). □

Remark 3 Since the Fourier transform of $\Phi^{*(p-1)}$ is equal to $\hat{\Phi}^{p-1}$ which is nonnegative and nonvanishing, the kernel function $\Phi^{*(p-1)}$ is also positive definite, e.g., Matérn functions and Gaussian functions discussed in Sect. 4. When $p = 2$, then Theorem 3 covers the classical results of RKHSs driven by positive definite functions. Actually, the $\mathscr{B}_\Phi^p(\mathbb{R}^d)$-norm of s given in Eq. (5) is computed by its semi-inner product and much more details of the proof are mentioned in the proof of [3, Theorem 4.4] (the representer theorem of $\mathscr{B}_\Phi^p(\mathbb{R}^d)$). Roughly speaking, the $\mathscr{B}_\Phi^p(\mathbb{R}^d)$-norm of s for $p = 2n$ can be seen as the qth root of the generalization of the quadratic form.

The solution s of the support vector machines in the RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$ can be seen as one choice of an optimal decision function for the binary classification. Its optimal separable boundary is given by $\left\{ \mathbf{x} \in \mathbb{R}^d : s(\mathbf{x}) = 0 \right\}$ and its classification rule has the form

$$
r(\mathbf{x}) := \operatorname{sign}(s(\mathbf{x})), \quad \text{for } \mathbf{x} \in \mathbb{R}^d.
$$

4 Examples of Matérn Functions and Gaussian Functions

In this section, we focus on some popular positive definite functions such as Matérn functions and Gaussian functions. We will show that the kernel function $\Phi^{*(p-1)}$ of the Matérn function or the Gaussian function Φ discussed in Theorem 3 is still a Matérn function or a Gaussian function, respectively. This indicates that the support vector machine solutions driven by Matérn functions or Gaussian functions given in Eqs. (6 and 7) can work well for computer programs.

4.1 Matérn Functions (Sobolev Splines)

The Matérn function (Sobolev spline) with the shape parameter $\theta > 0$ and degree $m > d/2$

$$\Phi_{\theta,m}(\mathbf{x}) := \frac{2^{1-m-d/2}}{\pi^{d/2}\Gamma(m)\theta^{2m-d}}(\theta\,\|\mathbf{x}\|_2)^{m-d/2}K_{d/2-m}(\theta\,\|\mathbf{x}\|_2), \quad \text{for } \mathbf{x}\in\mathbb{R}^d,$$

is a positive definite function, where $t \mapsto K_\nu(t)$ is the modified Bessel function of the second kind of order ν and $t \mapsto \Gamma(t)$ is the Gamma function. We can compute its Fourier transform

$$\hat{\Phi}_{\theta,m}(\mathbf{x}) = \left(\theta^2 + \|\mathbf{x}\|_2^2\right)^{-m}, \quad \text{for } \mathbf{x}\in\mathbb{R}^d.$$

(Many more details of the Matérn functions are mentioned in [5, Example 5.7].)

Let $1 < q \le 2 \le p < \infty$ with $p^{-1} + q^{-1} = 1$ such that $mq/p > d/2$. Since $\hat{\Phi}_{\theta,m}^{q-1} \in \mathrm{L}_1(\mathbb{R}^d)$, Theorem 1 assures that $\mathscr{B}^p_{\Phi_{\theta,m}}(\mathbb{R}^d)$ is a two-sided RKBS with a two-sided reproducing kernel $K_{\theta,m}(\mathbf{x},\mathbf{y}) = \Phi_{\theta,m}(\mathbf{x}-\mathbf{y})$.

Finally, we compute $\Phi_{\theta,m}^{*(p-1)}$ when $p = 2n$ for $n \in \mathbb{N}$. Since

$$\hat{\Phi}_{\theta,m}(\mathbf{x})^{p-1} = \left(\theta^2 + \|\mathbf{x}\|_2^2\right)^{-(p-1)m}, \quad \text{for } \mathbf{x}\in\mathbb{R}^d,$$

the inverse Fourier transform of $\hat{\Phi}_{\theta,m}^{p-1}$ has the form

$$\Phi_{\theta,m}^{*(p-1)}(\mathbf{x}) = \Phi_{\theta,(p-1)m}(\mathbf{x}), \quad \text{for } \mathbf{x}\in\mathbb{R}^d.$$

For typical examples,

$$\Phi_{\theta,m}^{*1} = \Phi_{\theta,m} \text{ when } p = 2, \quad \Phi_{\theta,m}^{*3} = \Phi_{\theta,3m} \text{ when } p = 4. \tag{8}$$

4.2 Gaussian Functions

By [4, Example 1 in Section 4.1], the Gaussian function with the shape parameter $\theta > 0$

$$\Phi_\theta(\mathbf{x}) := e^{-\theta^2\|\mathbf{x}\|_2^2}, \quad \text{for } \mathbf{x}\in\mathbb{R}^d,$$

and its Fourier transform has the form

$$\hat{\Phi}_\theta(\mathbf{x}) = 2^{-d/2}\theta^{-d}e^{-2^{-2}\theta^{-2}\|\mathbf{x}\|_2^2}, \quad \text{for } \mathbf{x}\in\mathbb{R}^d.$$

Let $1 < q \le 2 \le p < \infty$ with $p^{-1} + q^{-1} = 1$. Since $\hat{\Phi}_\theta^{q-1} \in \mathrm{L}_1(\mathbb{R}^d)$, Theorem 1 provides that $\mathscr{B}^p_{\Phi_\theta}(\mathbb{R}^d)$ is a two-sided RKBS with a two-sided reproducing kernel $K_\theta(\mathbf{x},\mathbf{y}) = \Phi_\theta(\mathbf{x}-\mathbf{y})$.

Furthermore, we want to obtain $\Phi_\theta^{*(p-1)}$ when $p = 2n$ for $n \in \mathbb{N}$. Since

$$\hat{\Phi}_\theta(\mathbf{x})^{p-1} = 2^{-(p-1)d/2}\theta^{-(p-1)d}e^{-(p-1)2^{-2}\theta^{-2}\|\mathbf{x}\|_2^2}, \quad \text{for } \mathbf{x} \in \mathbb{R}^d,$$

we have

$$\Phi_\theta^{*(p-1)}(\mathbf{x}) = 2^{-(p-2)d/2}(p-1)^{-d/2}\theta^{-(p-2)d}e^{-(p-1)^{-1}\theta^2\|\mathbf{x}\|_2^2}, \quad \text{for } \mathbf{x} \in \mathbb{R}^d.$$

In particular,

$$\Phi_\theta^{*1} = \Phi_\theta \text{ when } p = 2, \quad \Phi_\theta^{*3} = 12^{-d/2}\theta^{-2d}\Phi_{\sqrt{3}\theta/3} \text{ when } p = 4. \tag{9}$$

5 Numerical Tests for $\mathscr{B}_\Phi^2(\mathbb{R}^2)$ and $\mathscr{B}_\Phi^4(\mathbb{R}^2)$

In this section, we will first do some numerical tests of the support vector classifiers in the RKHSs and RKBSs driven by positive definite functions defined on the two-dimensional space $\mathbb{R}^2$.

Let Φ be a Matérn function with $\theta = 22$ and $m = 7/2$ or a Gaussian function with $\theta = 12$ defined on $\mathbb{R}^2$ (see Sect. 4). Here the shape parameter θ is chosen by personal experience. We do not consider how to find the optimal parameters of the kernel functions in this article. By Theorem 1 we can use the positive definite function Φ to set up two kinds of normed spaces $\mathscr{B}_\Phi^2(\mathbb{R}^2)$ and $\mathscr{B}_\Phi^4(\mathbb{R}^2)$. Obviously $\mathscr{B}_\Phi^2(\mathbb{R}^2) = \mathscr{H}_\Phi(\mathbb{R}^2)$ is a Hilbert space but $\mathscr{B}_\Phi^4(\mathbb{R}^2)$ is merely a Banach space.

The training data of binary classes given in Fig. 1 will be applied in the numerical tests of support vector machines in $\mathscr{B}_\Phi^2(\mathbb{R}^2)$ and $\mathscr{B}_\Phi^4(\mathbb{R}^2)$. We compare the separable boundaries and the classification rules induced by the minimizers of regularized empirical risks over RKHSs and RKBSs, respectively. We choose the hinge loss

$$L(\mathbf{x}, y, t) := \max\{0, 1 - yt\}, \quad \text{for } \mathbf{x} \in \mathbb{R}^2, \ y \in \{\pm 1\}, \ t \in \mathbb{R},$$

and the regularization function

$$R_p(z) := \sigma z^{\frac{p}{p-1}}, \quad \text{for } z \in [0, \infty), \quad \text{where } \sigma := 0.1 \text{ and } p = 2 \text{ or } 4,$$

to construct the support vector machines.

The representer theorem in RKHSs [13, Theorem 5.5] provides that the support vector machine in the RKHS $\mathscr{B}_\Phi^2(\mathbb{R}^2) = \mathscr{H}_\Phi(\mathbb{R}^2)$

$$\min_{f \in \mathscr{B}_\Phi^2(\mathbb{R}^2)} \frac{1}{N}\sum_{j=1}^N L\big(\mathbf{x}_j, y_j, f(\mathbf{x}_j)\big) + R_2\left(\|f\|_{\mathscr{B}_\Phi^2(\mathbb{R}^2)}\right), \tag{10}$$

has the unique optimal solution

$$s_2(\mathbf{x}) = \sum_{k=1}^{N} c_k \Phi\left(\mathbf{x} - \mathbf{x}_k\right) = \sum_{k=1}^{N} c_k \Phi^{*1}\left(\mathbf{x} - \mathbf{x}_k\right), \quad \text{for } \mathbf{x} \in \mathbb{R}^2, \tag{11}$$

and its norm has the form

$$\|s_2\|^2_{\mathscr{B}^2_\Phi(\mathbb{R}^2)} = \sum_{j,k=1}^{N,N} c_j c_k \Phi\left(\mathbf{x}_j - \mathbf{x}_k\right) = \sum_{j,k=1}^{N,N} c_j c_k \Phi^{*1}\left(\mathbf{x}_j - \mathbf{x}_k\right), \tag{12}$$

where $c_1, \ldots, c_N \in \mathbb{R}$. According to Theorem 3 the support vector machine in the RKBS $\mathscr{B}^4_\Phi(\mathbb{R}^2)$

$$\min_{f \in \mathscr{B}^4_\Phi(\mathbb{R}^2)} \frac{1}{N} \sum_{j=1}^{N} L\left(\mathbf{x}_j, y_j, f(\mathbf{x}_j)\right) + R_4\left(\|f\|_{\mathscr{B}^4_\Phi(\mathbb{R}^2)}\right), \tag{13}$$

has the unique optimal solution

$$s_4(\mathbf{x}) = \sum_{k_1,k_2,k_3=1}^{N,N,N} c_{k_1} c_{k_2} c_{k_3} \Phi^{*3}\left(\mathbf{x} - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right), \quad \text{for } \mathbf{x} \in \mathbb{R}^2, \tag{14}$$

and its norm can be written as

$$\|s_4\|^{4/3}_{\mathscr{B}^4_\Phi(\mathbb{R}^2)} = \sum_{k_0,k_1,k_2,k_3=1}^{N,N,N,N} c_{k_0} c_{k_1} c_{k_2} c_{k_3} \Phi^{*3}\left(\mathbf{x}_{k_0} - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right), \tag{15}$$

where $c_1, \ldots, c_N \in \mathbb{R}$. Here the kernel functions Φ^{*1} and Φ^{*3} are given in Eq. (8) or (9).

The next step is to solve the coefficients of s_2 and s_4, respectively, with the help of Matlab programs. Let the functions be

$$T_2(\mathbf{b}) := \frac{1}{N} \sum_{j=1}^{N} L\left(\mathbf{x}_j, y_j, \phi_{2,j}(\mathbf{b})\right) + R_2\left(\Gamma_2(\mathbf{b})\right),$$

and

$$T_4(\mathbf{b}) := \frac{1}{N} \sum_{j=1}^{N} L\left(\mathbf{x}_j, y_j, \phi_{4,j}(\mathbf{b})\right) + R_4\left(\Gamma_4(\mathbf{b})\right),$$

where

$$\phi_{2,j}(\mathbf{b}) := \sum_{k=1}^{N} b_k \Phi^{*1}\left(\mathbf{x}_j - \mathbf{x}_k\right), \quad j = 1, \ldots, N,$$

$$\Gamma_2(\mathbf{b}) := \left(\sum_{j,k=1}^{N,N} b_j b_k \Phi^{*1}\left(\mathbf{x}_j - \mathbf{x}_k\right) \right)^{1/2},$$

and

$$\phi_{4,j}(\mathbf{b}) := \sum_{k_1,k_2,k_3=1}^{N,N,N} b_{k_1} b_{k_2} b_{k_3} \Phi^{*3}\left(\mathbf{x}_j - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right), \quad j = 1, \ldots, N,$$

$$\Gamma_4(\mathbf{b}) := \left(\sum_{k_0,k_1,k_2,k_3=1}^{N,N,N,N} b_{k_0} b_{k_1} b_{k_2} b_{k_3} \Phi^{*3}\left(\mathbf{x}_{k_0} - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right) \right)^{3/4},$$

for $\mathbf{b} := (b_1, \cdots, b_N)^T \in \mathbb{R}^N$. We find that $\phi_{2,j}$ are linear functions and $\phi_{4,j}$ are polynomials of degree three. Moreover, the support vector machines (10) and (13) can be transformed into the optimization problems

$$\min_{\mathbf{b}\in\mathbb{R}^N} T_2(\mathbf{b}), \tag{16}$$

and

$$\min_{\mathbf{b}\in\mathbb{R}^N} T_4(\mathbf{b}), \tag{17}$$

respectively. This indicates that the parameters $\mathbf{c} := (c_1, \ldots, c_N)^T$ of s_2 and s_4 are the minimizers of the optimization problems (16) and (17), respectively. In our following numerical tests, the Matlab function "fminunc" will be used to solve the global minimizers of the optimization problems (16) and (17).

Remark 4 If we let

$$\alpha_{k_1,k_2,k_3} := c_{k_1} c_{k_2} c_{k_3}, \quad \text{for } k_1, k_2, k_3 = 1, \ldots, N,$$

then s_4 can be viewed as the linear combination of $\Phi^{*3}\left(\cdot - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right)$. (Here we do not consider the duplicate kernel bases for convenience.) This indicates that we can solve for the coefficients $\alpha := \left(\alpha_{k_1,k_2,k_3}\right)_{k_1,k_2,k_3=1}^{N,N,N}$ of s_4 as the minimizer of the optimization problem

$$\min_{\beta\in\mathbb{R}^{N^3}} \frac{1}{N} \sum_{j=1}^{N} L\left(\mathbf{x}_j, y_j, f_\beta(\mathbf{x}_j)\right) + R_4\left(\left\| f_\beta \right\|_{\mathscr{B}_\Phi^4(\mathbb{R}^2)}\right), \tag{18}$$

where

$$f_\beta(\mathbf{x}) = \sum_{k_1,k_2,k_3=1}^{N,N,N} \beta_{k_1,k_2,k_3} \Phi^{*3}\left(\mathbf{x} - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right),$$

for $\beta := \left(\beta_{k_1,k_2,k_3}\right)_{k_1,k_2,k_3=1}^{N,N,N} \in \mathbb{R}^{N^3}$; hence $s_4 = f_\alpha$. However, in the minimization problem (18) it is difficult to compute the norm of $\|f_\beta\|_{\mathscr{B}_\Phi^4(\mathbb{R}^2)}$ by Eq. (2) for general $\beta \in \mathbb{R}^{N^3}$ because, for any $\beta = \left(\beta_{k_1,k_2,k_3}\right)_{k_1,k_2,k_3=1}^{N,N,N} \in \mathbb{R}^{N^3}$, there may be no sequence $\mathbf{b} = (b_k)_{k=1}^N \in \mathbb{R}^N$ such that $\beta_{k_1,k_2,k_3} = b_{k_1} b_{k_2} b_{k_3}$ for all $k_1, k_2, k_3 = 1, \ldots, N$. Since the norm of s_4 has an explicit form as in Eq. (15), we can add constraint conditions to the optimization problem (18) to obtain another well-computable minimization problem

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^{N^3}} \frac{1}{N} \sum_{j=1}^{N} L\left(\mathbf{x}_j, y_j, f_\beta(\mathbf{x}_j)\right) + R_4\left(\Gamma_4(b_1, \cdots, b_N)\right), \\ &\text{s.t. } \beta_{k_1,k_2,k_3} = b_{k_1} b_{k_2} b_{k_3}, \quad \text{for all } k_1, k_2, k_3 = 1, \ldots, N. \end{aligned} \tag{19}$$

We also find that the optimization problem (19) is another equivalent format of the optimization problem (17).

Firstly we compare the formulas of the support vector machine solutions s_2 and s_4 driven by three different data points. Suppose that the three training data points $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ are noncollinear. The kernel bases of s_2 are given by

$$\Phi^{*1}(\cdot - \mathbf{x}_1), \ \Phi^{*1}(\cdot - \mathbf{x}_2), \ \Phi^{*1}(\cdot - \mathbf{x}_3),$$

(see Fig. 2). This means that the training data points $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ can be viewed as the kernel-based points of s_2, and the kernel bases of s_2 are the shifts of the functions Φ^{*1} by the kernel-basis points (Fig. 2).

The kernel bases of s_4, in contrast, are represented as $\Phi^{*3}\left(\cdot - \mathbf{x}_{k_1} + \mathbf{x}_{k_2} - \mathbf{x}_{k_3}\right)$ for $k_1, k_2, k_3 = 1, 2, 3$. Since the training data points $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ are non-collinear, we can obtain the twelve pairwise distinct data points $\mathbf{z}_j := \mathbf{x}_{k_1} - \mathbf{x}_{k_2} + \mathbf{x}_{k_3}$ that comprise three training data points and another nine data points around the training data points (see Fig. 3). Then the kernel-based-point set $\{\mathbf{z}_1, \ldots, \mathbf{z}_{12}\}$ of s_4 denotes the set $\left\{\mathbf{x}_{k_1} - \mathbf{x}_{k_2} + \mathbf{x}_{k_3} : k_1, k_2, k_3 = 1, 2, 3\right\}$. Using the kernel-based points $\mathbf{z}_j$, the kernel bases of s_4 can be rewritten as

$$\Phi^{*3}(\cdot - \mathbf{z}_1), \ \Phi^{*3}(\cdot - \mathbf{z}_2), \ \ldots, \ \Phi^{*3}(\cdot - \mathbf{z}_{12}).$$

(see Fig. 4). This means that the kernel-based points of s_4 are rearranged and recombined by the training data points, and the kernel bases of s_4 are the shifts of the functions Φ^{*3} by the kernel-basis points.

For the general case of N training data points, the support vector machine solution s_2 has $\mathscr{O}(N)$ kernel bases while the kernel bases of the solution s_4 have $\mathscr{O}(N^3)$

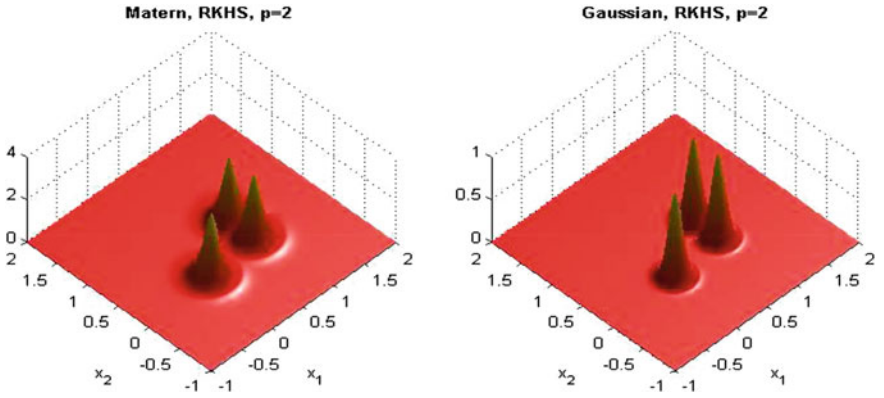


Fig. 2 The kernel bases $\Phi^{*1}(\cdot - \mathbf{x}_1)$, $\Phi^{*1}(\cdot - \mathbf{x}_2)$, $\Phi^{*1}(\cdot - \mathbf{x}_3)$ of s_2

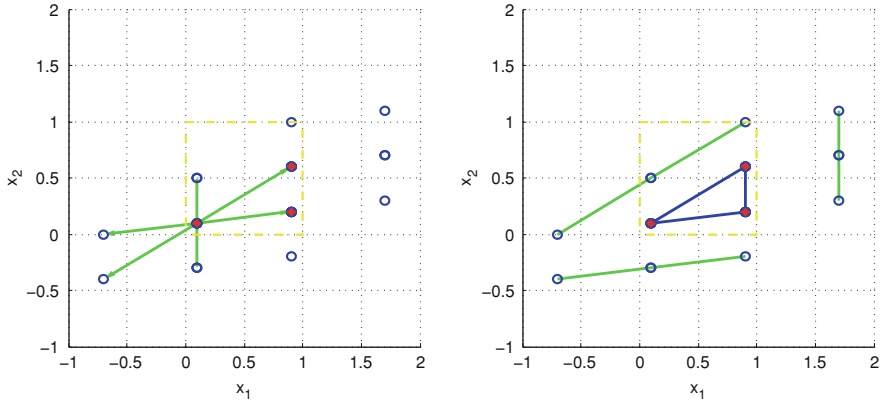


Fig. 3 The kernel-based points $\{\mathbf{z}_1, \ldots, \mathbf{z}_{12}\}$ of s_4 induced by three training data points $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. The *left panel* shows that each training data point $\mathbf{x}_k$ is associated with seven kernel-based points $\mathbf{z}_j$ and the *right panel* shows that the *triangle* driven by the training data points $\mathbf{x}_k$ are equivalent to the *triangles* driven by the kernel-based data points $\mathbf{z}_j$ other than the training data points. The training data points $\mathbf{x}_k$ are marked by *red dots* and the kernel-based points $\mathbf{z}_j$ are marked as the *blue circles*

elements, e.g., Fig. 5. This shows that the computational complexity of s_4 is larger than s_2 because the kernel bases of s_4 are more than s_2. Roughly speaking, the complexity of s_2 is $\mathcal{O}(N)$ and the complexity of s_4 is $\mathcal{O}(N^3)$.

Remark 5 Currently the Matlab function "fminunc" is not an efficient program to solve for the parameters $c_1, \ldots, c_N$ of s_4. The paper [3] shows that the parameters $c_1, \ldots, c_N$ of the solutions of support vector machines in RKBSs can be seen as a fixed point of some function. This means that they can be further solved by a fixed-point iterative algorithm. We will try to develop another algorithm to faster solve for their parameters later.

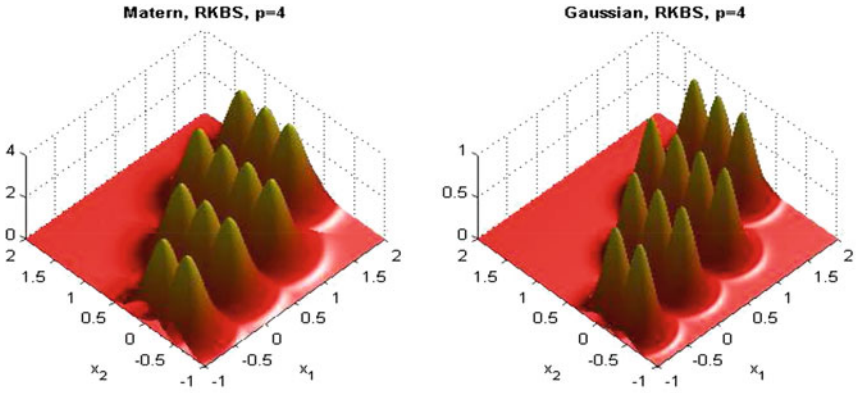


Fig. 4 The kernel bases $\Phi^{*3}(\cdot - \mathbf{z}_1)$, $\Phi^{*3}(\cdot - \mathbf{z}_2)$, ..., $\Phi^{*3}(\cdot - \mathbf{z}_{12})$ of s_4

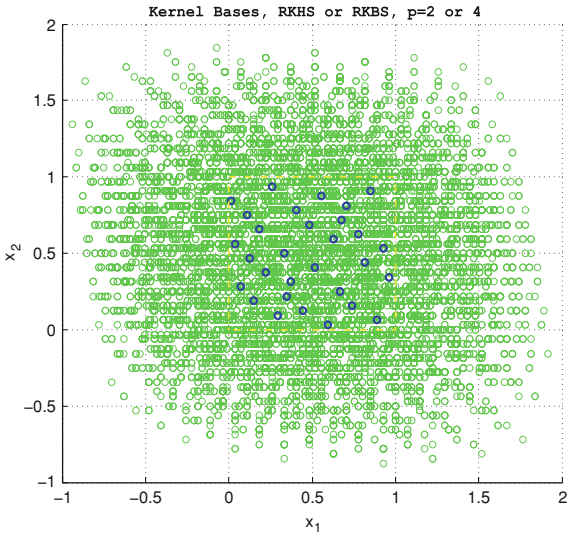


Fig. 5 The kernel-based points of s_4 induced by the 30 training data points, i.e., $N = 30$. The *green circles* mark the kernel-based data points $\mathbf{z}_j$ and the *blue circles* mark the training data points $\mathbf{x}_k$

Next, we do numerical tests of support vector classifiers in RKHSs and RKBSs using the training and testing data given in Fig. 1. The numerical results of Fig. 6 show that the support vector machine solution s_4 has the same learning ability as the solution s_2. Since the smoothness of the kernel bases affects the geometric structures of the separable boundaries, the separable boundaries $s_4(\mathbf{x}) = 0$ are smoother than the separable boundaries $s_2(\mathbf{x}) = 0$. In other words, the 2-norm margin of $s_2(\mathbf{x}) = \pm 1$ is bigger than $s_4(\mathbf{x}) = \pm 1$ because the hinge loss is designed to maximize the 2-norm margin for binary classification and the RKHS $\mathscr{H}_\Phi(\mathbb{R}^2) = \mathscr{B}_\Phi^2(\mathbb{R}^2)$ is associated with the 2-norm margin. Then we guess that the RKBS $\mathscr{B}_\Phi^4(\mathbb{R}^2)$ would be connected to the margin for some other τ-norm for $1 \leq \tau \leq \infty$, and we may need to construct another loss function for the support vector machines in RKBSs. Moreover, we look

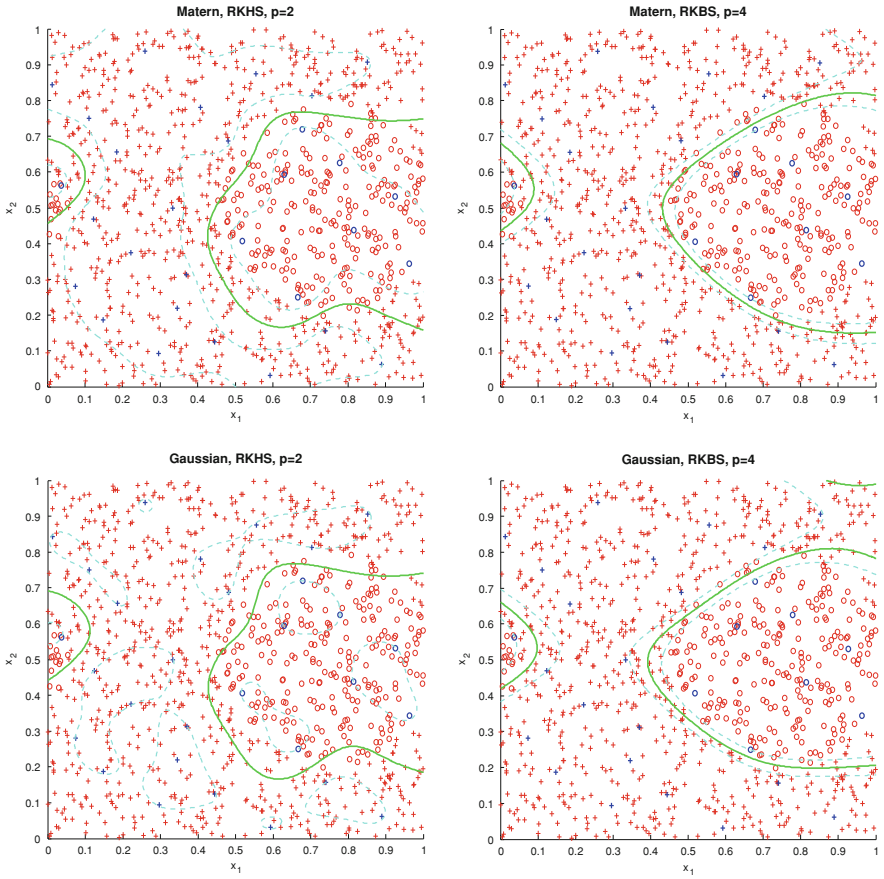


Fig. 6 A classification example in the observation domain $(0, 1)^2$ using the training and testing data given in Fig. 1. The classes are coded as a binary variable (*cross* $= +1$ and *circle* $= -1$). The *top left* and *right panels* show the binary classification for the solutions s_2 and s_4 induced by the Matérn function, respectively. The *bottom left*and *right panels* show the binary classification for the solutions s_2 and s_4 induced by the Gaussian function, respectively. The *blue circles* and crosses denote the training data and the *red circles* and crosses denote the testing data. The decision boundaries are the *green solid lines*, i.e., $s_{2,4}(\mathbf{x}) = 0$. The *cyan broken lines* mark the margins for $s_{2,4}(\mathbf{x}) = \pm 1$

at the absolute mean errors of s_2 and s_4 given in Table 1 and find that the predictions of the support vector solutions s_2 and s_4 are alike.

In addition, we consider another numerical example in the extended observation domain $(-1, 2)^2$ using the same training data in Fig. 1. Let $\Omega_{\text{out}} := (-1, 2)^2 \backslash (0, 1)^2$. In this case, we can still use the same decision function $g(\mathbf{x}) := \sin(2\pi x_1)/2 + \cos(2\pi x_2)/2 + 1/3$ given in Fig. 1 to generate the random uniformly distributed testing data in Ω_{out}. Since g is a periodic function, the testing data have a periodic distribution in $(-1, 2)^2$. But there are no training data in Ω_{out}. Then the support vector machine solutions $s_{2,4}$ discussed in Fig. 7 and Table 2 are the same as in Fig. 6

Table 1 Absolute mean errors of s_2 and s_4 for the random uniformly distributed testing data in the observation domain $(0, 1)^2$

Number of testing data	$M = 250$		$M = 500$		$M = 1000$	
Support vector machines	s_2	s_4	s_2	s_4	s_2	s_4
Error for Matérn function	0.0840	0.0560	0.0920	0.0740	0.0840	0.0700
Error for Gaussian function	0.0880	0.0640	0.0880	0.0580	0.0820	0.0550

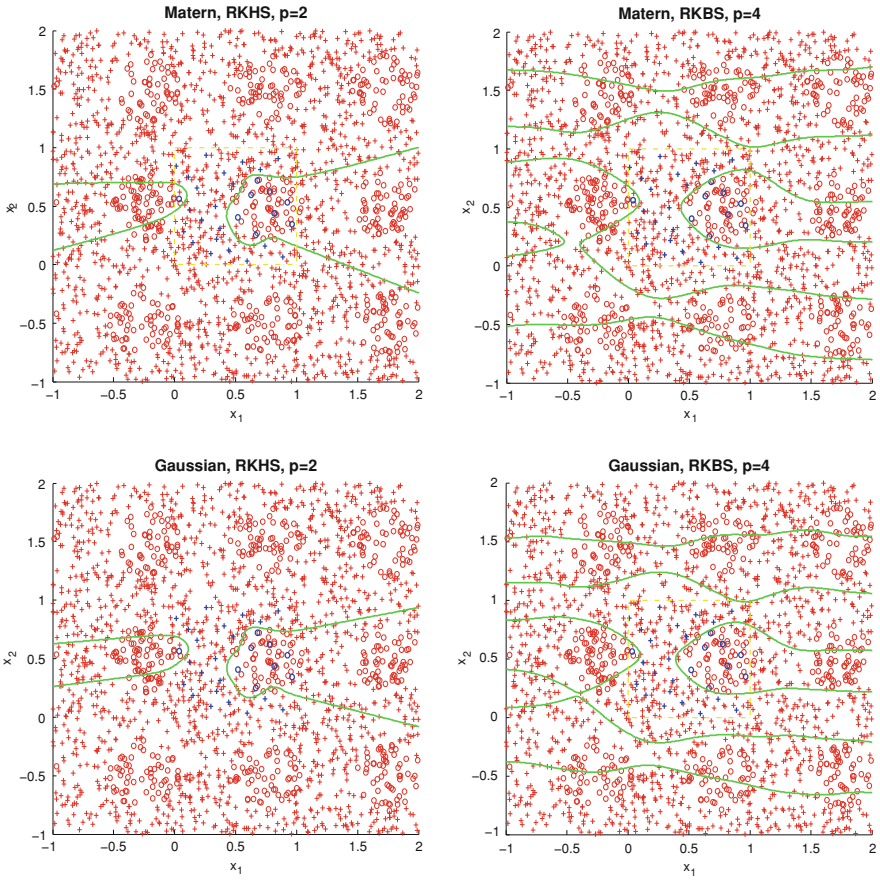


Fig. 7 A classification example in the observation domain $(-1, 2)^2$ using the training data given in Fig. 1. The testing data are introduced by the same original decision function g given in Fig. 1. The classes are coded as a binary variable (*cross* $= +1$ and *circle* $= -1$). The top *left* and *right* panels show the binary classification for the solutions s_2 and s_4 induced by the Matérn function, respectively. The *bottom left* and *right* panels show the binary classification for the solutions s_2 and s_4 induced by the Gaussian function, respectively. The *blue circles* and crosses denote the training data and the *red circles* and crosses denote the testing data. The decision boundaries are the *green solid lines*, i.e., $s_{2,4}(\mathbf{x}) = 0$

Table 2 Absolute mean errors of s_2 and s_4 for the random uniformly distributed testing data in the observation domain $(-1, 2)^2$

Number of testing data	$M = 500$		$M = 1000$		$M = 2000$	
Support vector machines	s_2	s_4	s_2	s_4	s_2	s_4
Error for Matérn function	0.3080	0.2860	0.3050	0.2810	0.2875	0.2815
Error for Gaussian function	0.2960	0.2960	0.2940	0.2750	0.2750	0.2745

and Table 1. However, observing Fig. 7, the classification rules of s_2 and s_4 have big differences in the domain Ω_{out} because s_4 has the kernel bases set up by kernel-based points in Ω_{out} but s_2 does not. We further look at the absolute mean errors of s_2 and s_4 for the additional testing data chosen in Ω_{out}. Obviously the absolute mean errors given in Table 1 are smaller than Table 2 because the support vector machine solutions $s_{2,4}$ do not include any training data information in Ω_{out}. But the absolute mean errors of s_2 and s_4 are slightly different.

In conclusion, we obtain interesting and novel support vector classifiers in RKBSs driven by positive definite functions, but at this point we can not determine whether the methods driven in RKBSs or RKHSs are better. We will try to look at numerical examples of other support vector machines in RKBSs to understand the learning methods in Banach spaces more deeply.

6 Final Remarks

In this article, we compare support vector machines (regularized empirical risks) in Hilbert versus those in Banach spaces. According to the theoretical results in our recent paper [3], we can develop support vector machines in RKBSs induced by positive definite functions, which are to minimize loss functions subject to regularization conditions related to the norms of the RKBSs. Their support vector machine solutions have finite-dimensional representations in terms of the positive definite functions. These formulas can be different from the classical solutions in RKHSs. Our numerical experiments further guarantee that the new support vector classifiers in RKBSs can be well computed and easily coded just as the classical algorithms driven in RKHSs.

In many current works, people try to generalize machine learning over Hilbert spaces to Banach spaces by replacing the regularization related to various norms of Banach spaces. But they still give the same linear kernel-based representations as the Hilbert spaces. Our theorems and algorithms show that the global minimizers of regularized risks over Banach spaces may be written as finite-dimensional kernel-based solutions. It is well known that the hinge loss is associated with RKHSs to maximize the 2-norm margin. The classical loss functions may be not the best choices for support vector machines in Banach spaces. In our next papers, we will try to construct other loss functions for different choices of the RKBSs. Moreover, we hope to approximate the general support vector machines

$$\min_{f\in\mathscr{B}_\Phi^p(\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}} L\left(\mathbf{x}, y, f(\mathbf{x})\right) \mathbb{P}(\mathrm{d}y|\mathbf{x})\mu(\mathrm{d}\mathbf{x}) + R\left(\|f\|_{\mathscr{B}_\Phi^p(\mathbb{R}^d)}\right),$$

where $\mathbb{P}(\cdot|\mathbf{x})$ is a conditional probability distribution defined on $\mathbb{R}$ dependent on $\mathbf{x} \in \mathbb{R}^d$ and μ is a positive measure defined on $\mathbb{R}^d$, by the empirical support vector machines in the RKBS $\mathscr{B}_\Phi^p(\mathbb{R}^d)$.

Acknowledgments The author would like to express his gratitude to the organizing committee of the 14th International Conference on Approximation Theory (AT 14) for the invitation and the travel grant to the AT 14 at San Antonio, Texas.

References

1. Alpaydin, E.: Introduction to Machine Learning. MIT Press, Cambridge (2010)
2. Erickson, J.F., Fasshauer, G.E.: Generalized native spaces. In: Neamtu, M., Schumaker, L.L. (eds.) Approximation Theory XII: San Antonio 2007, pp. 133–142. Nashboro Press, Brentwood (2008)
3. Fasshauer, G.E., Hickernell, F.J., Ye, Q.: Solving support vector machines in reproducing kernel Banach spaces with positive definite functions. Appl. Comput. Harmon. Anal. doi:10.1016/j.acha.2014.03.007
4. Fasshauer, G. E.: Meshfree Approximation Methods with Matlab. World Scientific Publishing Co., Pte. Ltd., Hackensack, NJ (2007)
5. Fasshauer, G.E., Ye, Q.: Reproducing kernels of generalized Sobolev spaces via a Green function approach with distributional operators. Numer. Math. **119**, 585–611 (2011)
6. Fasshauer, G.E., Ye, Q.: Reproducing kernels of Sobolev spaces via a Green kernel approach with differential operators and boundary operators. Adv. Comput. Math. **38**, 891–921 (2013)
7. Megginson, R.E.: An Introduction to Banach Space Theory. Springer, New York (1998)
8. Micchelli, C.A., Pontil, M.: A function representation for learning in Banach spaces. In: Shawe-Taylor, J., Singer, Y. (eds.) Learning Theory, pp. 255–269. Springer, Berlin (2004)
9. Schaback, R., Wendland, H.: Kernel techniques: from machine learning to meshless methods. Acta Numerica. **15**, 543–639 (2006)
10. Song, G., Zhang, H.: Reproducing kenrel Banach spaces with l^1 norm II: error analysis for regularized least square regression. Neural Compt. **23**(10), 2713–2729 (2011)
11. Song, G., Zhang, H., Hickernell, F.J.: Reproducing kernel Banach spaces with the l^1-norm. Appl. Comput. Harmon. Anal. **34**(1), 96–116 (2013)
12. Sriperumbudur, B.K., Fukumizu, K., Lanckriet, G.R.G.: Learning in Hilbert vs. Banach spaces: a measure embedding viewpoint. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P., Pereira, F., Weinberger, K.Q. (eds.) Neural Information Processing Systems, pp. 1773–1781. MIT Press, Cambridge (2011)
13. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
14. Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)
15. Wendland, H.: Scattered Data Approximation. Cambridge University Press, Cambridge (2005)
16. Ye, Q.: Analyzing reproducing kernel approximation methods via a Green function approach. Ph.D. thesis, Illinois Institute of Technology, Chicago (2012)
17. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. J. Mach. Learn. Res. **10**, 2741–2775 (2009)
18. Zhang, H., Zhang, J.: Regularized learning in Banach spaces as an optimization problem: representer theorems. J. Global Optim. **54**(2), 235–250 (2012)