Springer Proceedings in Mathematics & Statistics

Peter Eichelsbacher · Guido Elsner
Holger Kösters · Matthias Löwe
Franz Merkl · Silke Rolles *Editors*

Limit Theorems in Probability, Statistics and Number Theory

Springer

Springer Proceedings in Mathematics and Statistics

Volume 42

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

Springer Proceedings in Mathematics and Statistics

This book series features volumes composed of selected 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.

Peter Eichelsbacher • Guido Elsner
Holger Kösters • Matthias Löwe • Franz Merkl
Silke Rolles
Editors

Limit Theorems in Probability, Statistics and Number Theory

In Honor of Friedrich Götze

Springer

Editors
Peter Eichelsbacher
Mathematics Faculty
Ruhr-University Bochum
Bochum
Germany

Guido Elsner
Mathematics Faculty
Bielefeld University
Bielefeld
Germany

Holger Kösters
Mathematics Faculty
Bielefeld University
Bielefeld
Germany

Matthias Löwe
Institute for Mathematical statistics
University of Münster
Münster
Germany

Franz Merkl
Mathematics Institute
University of München
München
Germany

Silke Rolles
Centre for Mathematics
Technische Universität München
Garching bei München
Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-642-36067-1 ISBN 978-3-642-36068-8 (eBook)
DOI 10.1007/978-3-642-36068-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013936545

Mathematical Subject Classification (2010): 60F05, 62E20, 60-06, 46L54, 60B20, 60E10, 11J83

Printed on acid-free paper

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

Preface

Mathematical developments and breakthroughs have in many of their most shining examples been achieved by researchers who were fluid in the language of more than one field of mathematics and able to translate ideas from one field to another one. An outstanding example, almost a proof, for this claim is Friedrich Götze. Over the past more than 30 years he obtained a series of beautiful results in both probability theory and analytic number theory (among others), and, quite often, a central idea of the proof of these results is to borrow a technique from another field. This volume is dedicated to him. It consists of thirteen papers, the majority of which are based on contributions to a workshop that took place from August 4 to 6, 2011 at Bielefeld University on the occasion of Friedrich's sixtieth birthday. This workshop was supported by CRC 701 "Spectral Structures and Topological Methods in Mathematics".

The scope of the articles in this collection is as broad as Friedrich's interest. He started out as a pure mathematician studying complex geometry and topology for his diploma thesis with Friedrich Hirzebruch in Bonn. After that he changed his field of research to statistics to do his Ph.D. with Johann Pfanzagl in Cologne. His first papers analyze the speed of convergence and asymptotic expansions in central limit theorems for various statistics. He soon became a worldwide acknowledged expert for giving best-known or even optimal rates of convergence in limit theorems. In the 1990s he broadened his spectrum by a series of articles on the geometry of numbers. The choice of this particular subject was not a coincidence. Already Friedrich's analysis of the convergence rates in limit theorems for quadratic forms led to questions in analytic number theory, in particular to the investigation of the number of lattice points in an ellipsoid. These questions were first tackled by Hardy and Littlewood, on the one hand, and Landau, on the other, in the 1920s, but it was not until a series of fundamental papers by Friedrich and his coauthors that these questions were ultimately solved with the help of probabilistic methods.

Apart from the two subjects mentioned above Friedrich has always been open for new trends in probability, statistics, and many other fields. Over the past 2 decades he made major contributions to the theory of random matrices and free probability, the theory of resampling techniques, and log-Sobolev inequalities, among others.

Many of Friedrich's results were derived in collaboration with colleagues and friends, many of whom presented talks on the occasion of his sixtieth birthday and also contributed to this volume. In total, it collects 13 papers (all of which have been peer-reviewed) by researchers in fields in which Friedrich has become famous for his contributions, namely number theory, probability, statistics and combinatorics, and the theory of random matrices. Many of these papers have been stimulated by his work, either by the choice of subject or by his techniques. The articles are prefixed by an interview by Willem van Zwet, which illuminates Friedrich's achievements in the context of his personal experiences.

We thus hope to be able to shed some light on Friedrich's preeminent scientific work.

Münster, Germany Matthias Löwe

Contents

A Conversation with Friedrich Götze

Willem R. van Zwet

Friedrich Götze
Photo kindly provided by Friedrich Götze (2011)

Abstract Friedrich Götze has made signal contrbutions to mathematical statistics, probability theory and related areas in mathematics. He also rendered many other important services to the profession. His 60th birthday provides and excellent opportunity for a conversation about his career and his views on various matters.

1 Early Days: A Talented Tinkerer

Interviewer: Friedrich, let us start at the beginning. You were born in 1951 in Hameln and as a boy you must have shown great promise as a scientist. Can you tell us about your scientific activities while you were in school?

W.R. van Zwet (✉)
Department of Mathematics, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: vanzwet@math.leidenuniv.nl

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_1,

F.G.: Well, I grew up in a household where my father was a small grocer, and of course his perspective for my future was taking over his grocery. After finishing elementary school I wanted to go to the gymnasium. My father had something more practical in mind, but the teacher convinced him that the gymnasium would be a better choice. From there on, I did not show much interest in the grocery store, but rather in the libraries of our town. After a while I had all kinds of interests, especially in soldering together radios and some electronics that I was fascinated with at that time.

Interviewer: We have a beautiful picture of you from those days. It appeared in the local newspaper and shows you and a friend with some fantastic looking equipment you put together.

Friedrich and his friend Friedrich Hupe with their computer.
The photo was published in the local newspaper of Hameln: DeWeZet

F.G.: Yes, there was the centennial celebration of the school, and on this occasion each of the students should carry out a project, for instance some chemical or physics experiment. I was fascinated by computers, which were not available for the general public at the time, and I thought I would make a demonstration computer. And because the necessary components were very expensive at that time, I went to some of the so-called scrap-shops of the telecom companies where they were throwing out their switchboards made of electromagnetic relays. Huge numbers, which you could buy for a few Marks, just for the weight of them. I collected them and then we soldered them together to make accumulators and for doing some basic binary addition, and even multiplication which was a great thing. To wire these things up to have something like a main switchboard that would do these things and display the result, took us more than a year and I still remember that it was very difficult to keep all of these wirings in mind. I had a pal who was better at soldering than I was, and he did the soldering I told him to do. So we finally made it and had this thing displayed at the centennial celebration of our school.

Interviewer: It could actually multiply?

F.G.: Yes, it could actually multiply, but it took a while before you could see what had happened. It was very slow.

Interviewer: So we should consider ourselves lucky that you didn't go into electrical engineering or computer science.

F.G.: Of course computer science was something interesting for young people. But what you heard about it was not happening in Europe. It was happening in the United States. The first things we heard of—and were very fond of—where these programmable pocket calculators. But they were extremely expensive and cost about a thousand Marks, which was about half of the salary of an assistant. Also, it all happened in America, with Fairchild and later with Intel. To get into computer science you would have to go to the US.

Interviewer: Let me return for a moment to your father. Was he ultimately convinced that an academic career would suit you better than minding the store?

F.G.: Yes, sooner or later he recognized that the store was not my cup of tea and as my grades in school got better and better, he said okay, if that is your future . . .

2 A Wander-Student

Interviewer: So you went to the gymnasium and obviously finished there, and then went to Göttingen in 1970 to study physics and mathematics. There was probably very little computer soldering going on there, so you must have prepared yourself in some other way.

F.G.: Yes, I'm afraid that my computer would not have got me very far in Göttingen. I already mentioned my interest in libraries that would lend me physics and mathematics books for 2 or 3 weeks. During that time I copied what I thought was important in little notebooks. There were no copying machines in those days, so it all had to be done by hand.

Interviewer: So in fact you constructed your own private library. It is certainly a good way to learn. It takes a lot of time but you never forget these things anymore.

F.G.: That is right. Even though the lectures at Göttingen were more rigorous, I think I was quite well prepared.

Interviewer: So at Göttingen you basically studied physics with a second major in mathematics.

F.G.: With my kind of interests I might have become an electrical engineer, but I was not sure and felt that a general physics and mathematics education could not do any harm. Also computer science was still part of mathematics at that time. Finally, Göttingen was only 80 km from where I lived.

Interviewer: During your 2 years at Göttingen you switched from physics to mathematics as a first major, and passed your pre-diploma exam in mathematics. What had happened?

F.G.: Well, I passed through the usual physics and mathematics curricula. The lectures were all right, but in physics there were also these dreadful experimental sessions. I had expected to see wonderful new laboratory equipment, but instead we merely worked with traditional old instruments dating back to the 1920s. They really belonged in a science museum. This may well be useful to develop your skills with basic measurement devices, to trim them to higher precision and learn physics that way, but it was not very exciting.

Interviewer: I also moved from physics to mathematics after similar experiences. However, that was much earlier and you might have expected these lab classes would have been modernized a bit in the meantime.

F.G.: Well, the 1920s were the great days of Göttingen physics and, as generations of students before us, we were supposed to learn by using the same marvellous old instruments to repeat the experiments of those days, the outcomes of which you should of course know from your courses.

Interviewer: Another thing that turned me off was the way in which mathematics was handled in some physics courses. Our mathematics teachers taught us to be rigorous, but that didn't seem to hold for experimental physicists.

F.G.: I took lectures in quantum mechanics and I started pestering the teacher afterwards asking what the meaning of this measurement process was, because I had heard there was a debate among physicists about what they were describing by this. And then I got this nice reply: "Young man, first try to learn the trade and do your exercises. Leave this type of question to the time when you get a Nobel Prize and then you can do philosophy".

Interviewer: So much for physics and physicists. Anything remarkable about the mathematics courses?

F.G.: Well, there was certainly no lack of rigor there! Our first calculus course was taught by Brieskorn who had his first position as a full professor in Göttingen and later became a well-known geometer. He started his calculus course by teaching logic first, and we were trained so thoroughly that it took us three or four semesters before we could actually write our proofs in normal mathematical style again without using seven or eight quantors.

Interviewer: You were treated the rough way!

F.G.: Yes, but it was good training.

Interviewer: Having survived all of this, you did get your pre-diploma in mathematics at Göttingen in 1972.

F.G.: Yes, I did, even though there was a slight problem. I had completely forgotten that linear algebra was also part of the exam, and I had done nothing to prepare

myself for this. I discovered this shortly before the exam and literally worked day and night to catch up. As a result I overslept on the day of the linear algebra part of the exam. It was scheduled at 9 a.m. and I showed up at noon. Luckily they still allowed me to take the exam.

Interviewer: But then you left Göttingen and went to Bonn to continue your mathematics study. Why?

F.G.: I had a stipend from the Studienstiftung, a foundation supporting kids that did well at school. Among other things they held nice meetings in the semester break, organized by professors in various disciplines who were interested in young people and, of course, were looking for talent. I attended one of those seminars in Alpach in the Austrian Alps that was organized by Professor Hirzebruch from Bonn. He impressed me by the way he could explain essentially complicated matters in a simple way. It is rather common in Germany to change universities after your pre-diploma, so I decided to go to Bonn and study complex geometry and topology with Hirzebruch. I got my diploma under his guidance in 1975.

Hirzebruch had an interesting style. There was the Wednesday afternoon seminar. Everybody in the geometry group was supposed to be there. If you missed a seminar, he would say the next day: "I didn't see you yesterday". Then you knew that you'd better be there next week.

Interviewer: Sounds like Jerzy Neyman. Any other interesting characters in Bonn at the time?

F.G.: Definitely. There was Don Zagier. He was about my age and in those days usually dressed in a formal suit. We thought this a bit strange, but we took into account that he finished high school at age 13, received his master's degree at M.I.T. at age 16, and his Ph.D. with Hirzebruch in Bonn at age 20. He was about to finish his Habilitation when we were attending his lectures. He had clearly been a child prodigy, but without the difficult characteristics that often go with this. I attended his lectures on modular forms and was tremendously impressed by the speed at which he could do calculations on the blackboard. We used to say that he would be the only guy who could go shopping at a shopping centre for 2 weeks of supplies, and by looking at the numbers the cashier pushed, would know the grand total before she did. But he was also very practical and owned one of the first computers to check up on his number theory, and he was also interested in applied matters.

Interviewer: It is interesting to hear you say this. Many years ago I gave a lecture in Bonn on a topic in probability theory which most of the pure mathematicians present clearly considered a waste of their precious time. Afterwards Zagier took me out to a very pleasant dinner where we had a very sensible discussion of some probability problems. So apparently pure and applied mathematicians can get on quite well, but we have to realize that Don and you and I may not be the prototypical pure and applied types!

F.G.: At that time I also met another interesting person. For the work I did for my diploma I had to read an original paper in Russian and took a Russian course

provided by the university. However, like many older papers, this one was written in a verbal and descriptive style for which you need a better command of the language than my elementary course had provided. More important, during the course I met my future wife Irene. After I got my diploma in 1975, we got married in 1976.

Interviewer: After obtaining your diploma you were thinking of getting a Ph.D. with Hirzebruch in Bonn.

F.G.: It was not so clear what to do. Hirzebruch was quite pleased with my diploma thesis. However, it was also becoming clear that the expansion of the university system in Germany was coming to a sudden end. The oil crisis had frightened people, in particular the politicians who were no longer willing to finance further expansion. I was thinking of having a family and it was becoming increasingly doubtful that getting a Ph.D. in geometry would provide a stable basis for a family income. So I thought that maybe something more practical would be better. I looked around for advice and was told that a diploma is fine, but something more applied would be even better for getting a job in industry or an insurance company. Also, if you really want to have a decent career in industry in Germany it is an advantage to have a Ph.D.

Interviewer: It seems to me that you ended up by getting the best of both worlds. You did get a Ph.D. in an applied field and you did not end up in an insurance company. So off to Pfanzagl in Köln because it was close?

F.G.: Not so fast. My wife was studying medicine in Bonn and she thought I could perhaps be a medical doctor. This went as far as her taking me once to a dissection course in anatomy. She thought it quite interesting and very fascinating. But when I came into this large hall where a lot of students were around and people were opening up skulls, my lunch was protesting and I had to leave immediately. I thought this is not my cup of tea. So far for my taking up medical studies.

Interviewer: Yes, this sounds quite drastic. I believe that even for medical students, this is a test of stamina. Fortunately, less extreme forms of medical studies also exist.

F.G.: In view of possibilities in finance and insurance companies, I thought of brushing up the knowledge I acquired in Göttingen in courses on ergodic theory and measure theory of Ulrich Krengel. I was pleasantly surprised that the institute in Göttingen offered me a tutor position to complement my weekly allowance. It also seemed to make sense for my new career plans to renew old acquaintances, so I decided to return to Göttingen for a semester. Of course Krengel was there, together with visitors like Ahlswede, an American named Lee Jones who was the life of the party, and someone from Fribourg teaching rank tests in Hájek's style.

Interviewer: Was that André Antille, by any chance?

F.G.: You are right: it was. Everyone was in a good mood, we used to sit around posing problems to each other and I actually learned some statistics. Krengel was of course doing ergodic theory and to me that looked very much like analysis, and

rather than doing that, I might as well have stayed in geometry. I also turned down an offer from Ahlswede to come to Bielefeld with him. I was interested in finding a real statistician and someone told me there was someone named Pfanzagl in Köln, close to Bonn where Irene was studying. So one day I went to Köln, it was late in the evening, I didn't really expect to find Pfanzagl there, I knocked on his office door and said: "Hello. I'm a pure mathematician coming from Bonn, I have had some lectures in probability and I want to study statistics". This must have been something that he didn't expect and he was looking at me a little doubtfully. I handed him all my certificates which didn't look too bad, I would say. He was quite nice and willing to give me a try, offered me a tutor position, and asked me to give some lectures and seminars, which I did, and so I stayed there.

Interviewer: How long did you stay there until your Ph.D.?

F.G.: That must have been from 1976 to 1978 when I did my Ph.D. there. Of course I learned a lot of things and I was also involved in Pfanzagl's projects, correcting and checking manuscripts, helping him out in seminars, etc. The work for my Ph.D. thesis I did more or less on my own. Since Edgeworth expansions were a hot topic at the time, I wrote about expansions in Banach spaces.

I consider myself fortunate that I learned statistics from Pfanzagl. He got his Ph.D. with Hlawka in Vienna in number theory, worked in the statistical office of Austria, obtained a chair in the social sciences faculty and then moved to mathematics. But he was not content to investigate the mathematical properties of standard statistical procedures and recipes. He first wanted to discuss how appropriate such a procedure was for a given application. He loved to do mathematics, but never ignored his early history in applied statistics.

Interviewer: I have the same experience with my Ph.D. advisor Jan Hemelrijk, and of course, this is what statistics is all about! Unfortunately, many of today's mathematical statisticians lack experience with genuine applications, and as a result have little feeling for the validity of their models and procedures in practice.

F.G.: Pfanzagl had organized his life in an interesting way. He was at the university 2 days a week, when he had his appointments, taught his courses, attended seminars, etc. The remainder of the week he spent at his very nice home in the countryside on the other side of the Rhine, and during the semester breaks he was in Vienna with his family. Of course as students we also liked this 2-day workweek, because it gave us lots of time for research. If I would join my wife who had to be at the clinic at 7 a.m., it was a quiet time for me too, because if you appear at a mathematics department at 7 a.m., there is nobody there.

Interviewer: Friedrich, I think we have reached the end of your days as a student. You said earlier in this interview that it is quite normal in Germany to study at two universities. Now that we have seen that you moved from Göttingen to Bonn, then back to Göttingen again and finally to Köln, you won't mind that we call this section of the interview "A wander-student".

3 To the USA and Back

Interviewer: After getting your Ph.D. in 1978 you remained with Pfanzagl in Köln until you finished your Habilitation in 1983. However, in fact you spent quite a bit of time elsewhere.

F.G.: In November 1977 I attended my first meeting in Oberwolfach. It was on asymptotic methods of statistics organized by Pfanzagl and Witting. The meeting was crucial for my further development because I met you and Peter Bickel, which was the beginning of a long friendship and collaboration in various matters. At one point Peter asked: "Why don't you come and visit us in Berkeley?". Irene and I wanted to see the US anyhow, but it had to be cheap. So in 1978 we flew Icelandic Airways and made a detour through Mexico where my wallet containing my credit card was stolen in the underground of Mexico City. Via Atlanta we flew to San Francisco. By that time our money was almost gone and we went to the Bank of America where they advised us to let our bank send them a cheque for us. We followed their advice but the cheque never showed up.

Interviewer: Well, you couldn't have known that the European and American banking systems together are unable to send ten cents across the ocean successfully.

F.G.: At that time I was still very shy and didn't want to involve Peter who would probably have helped us immediately. Finally we got the money wired in some way. We went back via Princeton were we had friends, so apart from this one mishap, our first visit to the US was very pleasant.

The next year I went back to Berkeley as a visiting assistant professor for the academic year 1980–1981. When I told Pfanzagl about this plan, he looked a bit dubious and obviously thought he'd never see me again. In Berkeley momentous things had just happened. Neyman had died, and so had Kiefer who had just come to Berkeley. When I started teaching I found this a bit different from what I was used too. In a course for engineers you obviously have to follow the book, but I sometimes tried to explain a few things on a slightly more advanced level. That didn't sit well with the students, so Betty Scott, who was the department chair at the time, told me "Young man. You have to realize that we are teaching American students. You are not in Germany".

A fabulous occurrence during my stay was the Joint Statistical Meeting of ASA and IMS in Las Vegas. Participants could get a suite in the MGM Grand hotel for very much reduced prices, because the owners figured that statisticians who always talk about coin tossing, would be fanatic gamblers. Their gambling losses would easily make up for the reduced prices. When hardly anyone turned up at the gambling tables, they obviously felt cheated.

When my year as a visitor in Berkeley ended, I was wondering whether I should perhaps stay in the United States. However, at that time there were not very many positions at my level and my stay at Berkeley had spoiled me for the rest of the US. So I decided to go back to Köln and continue my career in Germany by getting my Habilitation, which I did with Pfanzagl in 1983.

4 Bielefeld

Interviewer: After your Habilitation in November 1983 you were appointed as an associate professor at Bielefeld in March 1984. They were obviously eager for you to finish your Habilitation.

F.G.: Yes, shortly before my Habilitation I was invited to apply for this position. So I went there and gave a talk on asymptotic statistics. I don't think that the pure mathematicians had the slightest idea of what I was talking about, until I mentioned that a variant of the Cramér-von Mises statistic was associated with a nice theta function. You could almost hear a sigh of relief from the audience: at least this person has heard of theta functions! So in early 1984 I was offered the position and accepted.

You may not remember, but earlier I had applied for a job in Amsterdam. Piet Groeneboom was at the CWI in Amsterdam at the time, but he was visiting Seattle and had not applied for the job. However, while the selection procedure was underway, Piet suddenly showed up and declared an interest in the position. At that time I was still considered very theoretical and no match for Piet, who was appointed.

Interviewer: So if Piet hadn't suddenly shown up, you might have been in Amsterdam.

F.G.: Yes, I was actually thinking that maybe I should learn a bit of Dutch and bought a dictionary. Irene and I both went to Amsterdam and I gave a talk there. But this idea didn't last very long.

And then I accepted the offer from Bielefeld and Irene said: "Great! Now this uncertainty over what is going to happen is past, and we go to Bielefeld". But once we came here and she saw the university building, she looked at me and said "You don't intend to stay here very long, do you?".

Interviewer: Well, admittedly Bielefeld University is not exactly a shining example of modern architecture. But there is some truth in what the Russians used to say about the architecture of Stalin's time: "It is best to be inside these buildings, where you can't see the outside".

F.G.: Yes, and when I learned a bit more about Bielefeld and the department, I was quite pleased and felt that I had landed at a quite good place.

They didn't have these barriers between the different kinds of mathematics that they had at many of the classical universities: an institute of pure mathematics here, one for applied mathematics there, and perhaps one for stochastics somewhere else, with each group tending its own little garden. In Bielefeld they had just one mathematics department. That didn't mean that they didn't have problems with allocation of funds or the hiring of new professors, but they knew they had to talk to each other and reach a compromise.

Interviewer: Even at the same institute people often didn't talk to each other. When I was appointed in Leiden, some of my colleagues thought it was ridiculous to have a professor of statistics. What stabilized the situation was the common interest. In my case people began to realize that in 20 years they might not have any students if they only taught pure mathematics. I'm not saying that everybody liked everybody else, but they could work together.

F.G.: Another pleasant thing about Bielefeld University was that you were appointed as a professor of mathematics, rather than geometry or statistics. That meant that you could change your field of research to another area of mathematics as long as you taught your courses. It was also possible to take a double teaching load in 1 year and be completely free for doing research the next year. Finally it was a lively and scientifically excellent department.

Interviewer: I understand that the philosophy of Bielefeld was certainly favourable for applied mathematics. Please tell us something about the concept of 'Mathematization' that was so typical for Bielefeld.

F.G.: The task of founding Bielefeld University in the nineteen-seventies was carried out under the leadership of Schelsky, a single person rather than a committee. Mathematics was one of the founding departments of the university because Schelsky had the idea of 'Mathematisierung' in mind. He wanted to change the social sciences and all other fields where people worked qualitatively into quantitative sciences. For that he needed all kinds of mathematics, especially applied mathematics. But not only a mathematics department, but also some bridging institutes such as the one for mathematics and econometrics. Such an institute was not permanent and would have to be reviewed every 8 years.

Schelsky was a conservative sociologist—I don't know whether this species still exists—and at that time universities were going wild. Especially the newly founded ones like Bielefeld were very progressive, including some of the staff. Schelsky found this a little bit too much, and left.

Interviewer: I suppose the mathematics staff also contained its share of progressives.

F.G.: We had a number of people in the staff who were joining the student demonstrations. There is one story of a young, energetic and leftist professor who organized a large student demonstration and marched with the students. He had his bicycle with him and after a while he looked at his watch and said: "Oh, I'm sorry. You march on. I have something important to do".

We had a very progressive institution here where all the members including secretaries and non-scientific staff had considerable voting power in all affairs of the university. This didn't make life any easier.

Interviewer: We have had the same in the Netherlands. It wasn't too bad because in the end people were usually willing to listen.

F.G.: Yes, but an older colleague told me that once in a while they had to strike a deal. They would tell the students: "We really want to appoint so-and-so as a

professor, whether you like it or not. But the next appointment you can make." Of course this would never happen, even though there was a signed agreement which is presumably well hidden by now.

Interviewer: I always have to explain to people why I stayed at the same university throughout my career. So let me put the question to you too. Why are you still here?

F.G.: In 1987 I was offered a full professorship at Kaiserslautern which I turned down. In 1989 I turned down a full professorship at the T.U. Berlin. The day Irene and I spent in Berlin to discuss the matter happened to be November 9, 1989. In the afternoon, with a small child in a stroller, we went to the wall to see people sitting on the wall and dismantling it. Then in 1990 I accepted a full professorship in Bielefeld. Finally in 2003 I turned down a full professorship at the Humboldt University in Berlin. A complicating factor at that time was that now that our children were growing up, Irene had just been licensed in Bielefeld as a doctor in residence, which allowed her to resume her medical career. It was highly unlikely that she could be licensed in Berlin, so moving there would have been a major sacrifice on her part.

Interviewer: From what you just told me, I think there is more to it than that. You explained that in the department in Bielefeld there were no barriers between different kinds of mathematics. People were appointed as professors of mathematics rather than geometry or statistics and they could change their field if they felt like it. Well, what could be a better place for someone who had difficulty deciding whether to get a Ph.D. in geometry or statistics and actually worked in number theory later? It seems to me that Bielefeld and you were made for each other, even though the Bielefeld architecture is best forgotten.

F.G.: You may have a point there.

Interviewer: And after 1989 there were the Sonderforschungsbereiche (SFB's) in mathematics that made Bielefeld such an attractive place to be. People tell me that you played an important role there, as a person who had the interests of all of mathematics at heart.

F.G.: The first SFB in Bielefeld started in 1989 and was devoted to "Discrete structures in mathematics". It was a broad collaborative effort to combine discrete methods used in combinatorics, information theory, but also (numerical linear) algebra, number theory, topology and arithmetic algebraic geometry. Many of the younger people in the department profited from the increased possibilities of communicating with colleagues and visitors from other fields within this SFB. Personally, I started out working on asymptotic approximations in mathematical statistics, but slowly moved in the direction of more discrete objects in stochastic algorithms and number theory related to this.

However, some of the senior members of the department who enjoyed a great reputation in a particular area of the SFB were more interested in seeing their leadership for the whole project acknowledged than in investing time and resources into collaborations with others. I served as chairman of the SFB for a number of years, and it was not always easy to balance the various views and keep the peace.

After the end of this SFB in 2000 there was an interim phase when the Bielefeld department went through a considerable generation change. My own interests in between pure and applied mathematics found a home in a subsequent smaller collaborative grant from the DFG, followed in 2005 by the current SFB on "Spectral structures and topological methods in mathematics". The explicitly stated aim of this SFB is to study developments connecting these rather diverging classical areas of mathematical research. Thanks to the efforts of all of the senior and junior colleagues who joined me on this adventurous road through largely uncharted terrain, this collaborative program has been quite successful in two 4-year periods so far. In the last 3 years the department has also been successful in hiring new younger staff members, who are eager to accompany us on the path we have chosen. Perhaps mathematicians are becoming more adventurous.

Interviewer: Well Friedrich, it looks like you are in Bielefeld to stay.

5 Oberwolfach

Interviewer: Friedrich, you mentioned earlier that we first met in Oberwolfach in 1977 and I would almost say: "Where else?". We have both spent a significant time of our lives there, but you have also been involved in its organization. So let's talk about the Mathematisches Forschungsinstitut Oberwolfach.

F.G.: I was appointed to the Beirat around 1990 and stayed there for nearly a decade. Then I became a member of the Executive Board of Oberwolfach, which I still am today. So I can speak about Oberwolfach during the last 20 years.

The present institute was built in the late sixties and the seventies. In a very courageous move, director Barner started the building procedure. Before the funding contract with the VW foundation had actually appeared in writing, he already ordered the construction companies to start and the bulldozers arrived. Well that was the way you did business in those days. You could count on oral agreements without fear of sudden budget cuts. So in 1968 they first constructed the building with the rooms for participants, kitchen, dining room, wine cellar and office, together with the bungalows. Then in the seventies they tore down the old villa and replaced it with the new building with lecture rooms, library, etc. Of course this was a great improvement, but at the time many people were sad to see the villa go. The institute had started there three decades earlier, and it had become a symbol for Oberwolfach.

The next 20 years passed without major problems. The State of Baden-Württemberg had no financial problems and Barner's relationship with the State administration was excellent. However, there was one inconsistency in the financial set-up. The VW Foundation paid for erecting the buildings, but not for their maintenance, and the State paid for the operational cost, also excluding maintenance. So Barner agreed informally with the authorities to save a bit of the operational cost to pay for maintenance. Again, such agreements were quite common in those days.

However, times changed. In the nineties the financial situation of Baden-Württemberg deteriorated, budget cuts became necessary, and of course Oberwolfach suffered. Perhaps even more important was the generation change in the State government. The old guard with whom Barner had such excellent relations went out, and a new generation arrived that was more interested in budget cutting than in longstanding relations. They insisted on new State rules of accounting superseding the previous arrangement that allowed Oberwolfach to save funds for maintenance.

Interviewer: Of course this is the risk of being funded by a single organization. I seem to remember that repeated efforts were made to be funded by the Federal Government too. Wasn't there a list of institutes funded by both the State and the Federal Government, the so-called blue list. Did Oberwolfach ever get on this list?

F.G.: Barner's successor Matthias Kreck (1994–2002) started to increase the annual budget to a level necessary for running Oberwolfach by obtaining support from individuals, industry and the European Union. Efforts were also made to get Oberwolfach on the blue list. The role of the blue list itself had changed with time. It was now called the Leibniz Foundation and also served the needs of institutes in the former DDR. To get on the list was a very difficult project that advanced only slowly, but we suddenly got help from an unexpected quarter, namely President Rau of the Federal Republic of Germany. On a visit to Denmark, during a dinner at a meeting on science and technology, Rau was seated next to our friend and colleague Ole Barndorff-Nielsen. Ole was quite indignant about the way Oberwolfach was treated by the German authorities and raised the topic with Rau. Rau listened carefully and on his return to Germany he sent what he had heard down the bureaucratic channels, and lo and behold, we received a call from the Federal Ministry of Science and Technological Development inquiring whether Oberwolfach needed something. Of course this was only the beginning of a long bureaucratic process, because Oberwolfach is not a standard research institute, but in the end Oberwolfach joined the Leibniz Foundation on January 1, 2005, and the foundation is now rather proud of us. So Oberwolfach is now supported by the Local as well as the Federal Government, and for a start, all buildings received a major overhaul to bring them up to date.

Interviewer: I'm really happy to hear this. Congratulations!

F.G.: There were also important changes in the scientific program of Oberwolfach. There used to be annually returning meetings with very broad topics and organized by the same small group of people. This doesn't happen anymore.

After Kreck left, we were fortunate to find an excellent successor in Gerd Martin Greuel. Greuel was able to increase the amount of additional grant money from various sources considerably. He was also responsible for making Oberwolfach a center for mathematical documents of various kinds. This was in line with the application to the Leibniz Foundation where Oberwolfach was presented not only as a meeting place, but also as a keeper of records, including those of its own history because many important mathematical results were first presented at Oberwolfach. Starting with the excellent library which has many books that universities cannot afford because of budget constraints, Greuel added electronic

records of the Oberwolfach meetings where one can find what was discussed at any given meeting and who were present.

Interviewer: Yes, I have just taken a look at this on internet and I'm impressed. It is a real service to the profession. You can also find out how often you have attended Oberwolfach meetings yourself, which could become a new mathematical game like comparing Erdös numbers. Let me add that I like Greuel very much.

F.G.: Everybody does, but he has turned 65 and will be retiring when we have found a successor. We are hopeful to find a good person.

6 Relations with Eastern Europe

Interviewer: For many years you and I have both been heavily involved in establishing relations with fellow scientists in Eastern Europe and the former Soviet Union. It all started when a few major scientists like Hájek and Révész showed up in places like Berkeley and Oberwolfach and we got to know each other. Next the European Meetings of Statisticians came to be held in Eastern Europe with some regularity: first in Budapest in 1972, next in Prague, Varna, East Berlin and Wroclaw. At the same time Western participation in existing locally organized meetings in Prague, Budapest and Vilnius increased sharply. In 1975 the Bernoulli Society was founded partly to build bridges between Eastern Europe and the West and in 1986 the World Meeting in Tashkent proved its success. At that time you had attended quite a few of these meetings.

F.G.: Yes, it was often the only way to meet people from Eastern Europe and the Soviet Union, because they would be allowed to attend meetings in the West only rarely. After the collapse of the communist regimes this was still difficult for financial reasons. The Tashkent World Meeting was the first statistics meeting in the Soviet Union attended by a huge number of participants from all over the world, in particular from the United States. For many of them, used to Cold War rhetoric, it was an eye-opener to see the circumstances in the Soviet Union for the first time and notice the natural anarchy that was present everywhere and made even getting there an exciting adventure.

Interviewer: These meetings were always held as far away from Moscow as possible to avoid official interference. In that respect Vilnius was good, but Tashkent was even better!

F.G.: At that time we didn't have many possibilities to invite colleagues from Eastern Europe and the Soviet Union to visit us and take part in research projects. That changed after the collapse of the Soviet Union. First the Soros Foundation provided funds for this and then the European Union started a program named INTAS for scientific collaboration with the former Soviet Union. Living conditions of scientists in the former Soviet Union were desperate: salaries were not sufficient

to cover basic needs, if they were paid at all. Many scientists were looking for opportunities for leaving the country and finding a position in the West. The INTAS program was aimed at joint research projects of participants in the former Soviet Union and in the European Union. The former participants would receive a supplement of their salaries, whereas the latter should apply for and administer the research program, which turned out to consist of writing an endless series of reports and solving problems with the Brussels bureaucracy and Russian banks. Most of the people running INTAS had no previous involvement with the former Soviet Union, which made it difficult for them to understand what goes on in the minds of the many different peoples who made up the Soviet Union. We also learned quickly that calling Brussels bureaucrats before 3 p.m. was useless because they were apparently having a good lunch. The banks turned out to be a more difficult problem. The cost of transferring money to participants was astronomical without any guarantee that the money would actually arrive. I understand that in some cases the money was actually brought to Moscow by a messenger who handed it out at the airport to people showing a passport with the right name.

Before we had even heard of the existence of INTAS, you and I both received a request from friends in Moscow to apply for an INTAS program for collaboration with a large group of really excellent probabilists and statisticians in the former Soviet Union. We did and proposed a program that needed a large number of special skills that were well represented in the group of participants. The program immediately got funded in the first round in 1993 and was extended for another period each time it ran out, until INTAS stopped operations in 2006.

Interviewer: Until my secretary and I both retired in 1999 we did the administration of the program in Leiden. My secretary used to refer to this job as the INTAS disaster, or words of similar meaning. After that the job went to you in Bielefeld, so it is clear that we truly shared the load.

One final remark: There was no provision for travel in the INTAS grants because the INTAS philosophy was clearly that everybody should stay where they were. No mass emigration to the West! However, there was no rule against consultations between different groups in the program, so of course quite a few people spent quite some time in Bielefeld and Leiden.

F.G.: After INTAS stopped, the Deutsche Forschungsgemeinschaft (D.F.G.) provided new programs for joint research with scientists in the former Soviet Union, which allowed them to spend time in Bielefeld. Other visitors were supported by the Humboldt Foundation and by the Sonderforschungsbereiche that we have had in Bielefeld.

Interviewer: Yes, there used to be an entire corridor in the institute in Bielefeld that was jokingly referred to as Moscow Boulevard.

Let me raise a further point in this connection. I have the impression that in probability and statistics we have had excellent relations with our colleagues in the former Soviet Union much earlier than in any other branch of mathematics. This may well be due to the fact that we always made contact on a personal rather than

an institutional basis. When the idea of the World Meeting in Tashkent came up, the Bernoulli Society, with much help from the members of its Russian branch, discussed the matter at length with all of their Russian friends, until almost everyone was confident it was a great idea that might be possible. Then after much lobbying the Soviet Academy of Science decided not to oppose it and it went through. If the Bernoulli Society had chosen the institutional way and wrote a letter to the Academy, they would certainly have received a negative reply, if any at all.

F.G.: Yes, definitely. The relations we had with the former Soviets were based on the fact that we attended all of these meetings together, got to know each other at a very early stage and developed a mutual trust. I have never noticed this in other branches of mathematics to this extent.

We should also not discount the role of the various nationalities in the Soviet Union. To organize a scientific meeting in Vilnius with a large international participation would be strongly supported by the local government, because it would show that Lithuania was not merely a part of the Soviet Union but also internationally recognized.

Interviewer: You are absolutely right.

7 German Reunification

Interviewer: Under the communist regimes, scientists basically had the same problems everywhere in Eastern Europe and the Soviet Union. Still East Germany was a somewhat special case. How did the German reunification affect East German science and scientists?

F.G.: The situation in East Germany was indeed very special. Two neighboring countries with the same language, the Eastern part watching Western television and seeing the wonderful world of Western luxury, while they couldn't buy bananas in the supermarket. After 1989 the original idea was that the two countries would remain separate for the time being, each with its own currency and a cheap-labor part in East Germany, and then slowly evolve into a single country. But it quickly became clear to the politicians that with the heavily guarded border gone, there was nothing to stop people moving to the more prosperous part. So they had to act fast and the East German parliament decided to join West Germany. It was not unification of two states and of two political systems, but of East Germany becoming part of the West German republic, with all of the legal and bureaucratic consequences that this implied. Of course this was easy for the West Germans because they wouldn't have to change anything.

But this created a problem for the sciences, because in East Germany, like in most Eastern European countries, scientific research was organized through the Academy of Sciences, and the East German academy employed about 30,000 people under the heading of scientific socialist production. Presumably these people did scientific research to enhance socialist production. This was not a happy idea, because the

people employed by the academy were not content to do purely applied work for industry and mostly did their own thing, whereas industry wasn't pleased to be told to let these people interfere with production, socialist or not! Most of this would have to go, because there was no way to employ 30,000 people at Western salaries, as would be a legal obligation for all civil servants.

A similar problem existed at the universities where people below the rank of professor also had permanent positions and would have to be paid Western salaries if they remained. Finally there were people with Party or Stasi affiliations which should not be retained. All of these problems would have to be solved before everybody became civil servants under West German law.

As a first step all members of the scientific staff of the Academy and the universities were dismissed. It was decided to follow the West German model where research is performed at the universities and at a limited number of specialized institutions with a far smaller staff than the former Academy institutes. Every qualified person could apply for one of the available positions, which meant that after being dismissed one could apply for one's previous position or any other. At the same time scientists from outside East Germany could also apply. At the universities there were honors committees of East German members as well as external hiring committees to reappoint people. This created uncomfortable situations where people like you and me had to review senior East German colleagues competing for their own jobs with young applicants from West Germany.

Interviewer: Yes, that's what I really found shocking.

F.G.: Another question was how to fund the specialized research institutions. Max Planck didn't want them. Fraunhofer didn't view these institutes as helpful for applied research with industry. So the only place for these institutes was the blue list of institutes financed jointly by the local State and the Federal Government that I mentioned earlier when speaking about Oberwolfach. For this new role the blue list was renamed Leibniz Foundation, which certainly sounds more dignified. When discussing which of the institutes should go to Leibniz I heard the representatives of the Federal Government make a promise that I never heard before or after. They told us to decide which institutes were really of a very high level, and whatever you find good, we'll pay for. It was amazing!

Interviewer: And did they actually keep their promise?

F.G.: Yes. Well look at the West German deficit at that time.

Of course this gave many people a very bad time. They had permanent positions and never expected to have to look for another job. But there was no other way. It all had to be done in a few months, which was a hectic time for all of these committees too. I was mainly involved with the Weierstrass Institute which was cut down from 200 to about 80 people.

So this was the reunification process, but then it was argued that this would have been a unique time for a real unification, in the sense of also cutting out some of the fat in the West German system. But as you can imagine with these time constraints other people argued convincingly against this.

Interviewer: Should things have been done differently or was there really no alternative?

F.G.: Theoretically we could have gone to a new structure together. But the West German institutions would not have been in a great hurry to submit plans and would certainly have fought this.

8 Visiting Committees

Interviewer: Friedrich, we have both spent a considerable amount of time on visiting committees that show up at mathematics departments of other universities and report what we think of them. If this opinion is not fit to be printed, we typically confuse the reader with a mountain of generalities.

F.G.: In Germany the visiting committee is a fairly recent phenomenon that was introduced when the golden years of expansion were over and serious cutbacks started in the eighties. Before that time all universities were supposed to have been created equal and it was blasphemy to try and rank them.

Interviewer: I think that this was true in most other countries too, and if it was done earlier, it was merely an intellectual exercise without serious consequences.

F.G.: By the end of the eighties the data collected by the visiting committees on research and teaching began to play a role in the allocation of funds.

Interviewer: Still the effect has generally been pretty minimal. The mathematics department at Leiden is usually declared to be the best in the country, but this never brought us a penny. It does make us more popular with the president of the university, though.

F.G.: Yes that is true in general. But at some smaller universities founded during the time of university expansion the staff was roughly of the same age so there was massive retirement 20–25 years later. In such a case it is probably not a good idea to let these senior people decide on the direction the department should take, so outside advice can be very useful.

Interviewer: Sure, such cases exist, but in most universities they are rare.

F.G.: Yes, but outside advice is becoming more and more common. A generation change that I just mentioned is one thing, but it now happens regularly. There is this excellence competition between universities, which forces them to choose main areas of research. Anytime a number of positions in a department are open for reappointment and a change of direction might be possible, the new constitution gives the university the option to make such strategic decisions its own responsibility. Of course they are also a bit at a loss what to do, so they call for outside advice.

Interviewer: I can imagine that you are concerned about this as it sounds really rather extreme to me. Does this also exist in this extreme form in other countries? Of course the president and rector are formally responsible for everything that goes on in a university and there is much talk of departments choosing main research areas, but in practice the department still decides whom to appoint in the Netherlands. I suppose that this outside advice also leads to a new bureaucracy to handle this advice and act on it?

F.G.: Exactly! The office of the president or the rector is acquiring a whole new group of people who organize these reviews, help in formulating long-term policies etc. So far these kinds of jobs seem to be more attractive to people with a background in the humanities rather the natural sciences.

Interviewer: It sounds gruesome. I can see only one positive side to this, which is that using visiting committees is certainly better than basing decisions on numbers of publications and citations. Bureaucrats usually prefer these numbers assuming that they provide "hard" evidence in contrast to peer review which is considered "soft".

9 Scientific Interests

Interviewer: During this interview we have repeatedly seen that your interest in mathematics doesn't stop with mathematical statistics and probability theory. During the symposium held on the occasion of your sixtieth birthday, Professor Hirzebruch who guided the work for your diploma in geometry, gave a lively account of this work. Let us now talk about your recent work in number theory.

F.G.: Willem, as you well know I started out in mathematical statistics with an interest in asymptotic expansions for the distributions of nonlinear statistical procedures like goodness-of-fit procedures, such as the Cramér-von Mises test. These may be viewed as expansions for the probabilities of ellipsoids in the Central Limit Theorem (CLT) in function spaces. The methods I developed for obtaining such expansions seemed to be interesting for a group of the Kolmogorov / Linnik school in probability in Russia working on these questions since the sixties. In the first SFB (1989–2000) in Bielefeld I originally worked on statistical problems in Markov chain Monte-Carlo, image restoration, as well as resampling methods, time-series and stochastic processes. But a number of researchers applied for and received Humboldt-fellowships to work with me on Berry-Esseen type bounds and asymptotic expansions for quadratic forms, U-statistics and Student statistics. These were Bentkus, Bloznelis, Rachkauskas, Tikhomirov, Zalesky and Zaitsev. The Humboldt-Foundation also helped to finance via Humboldt-Prizes the collaboration with Rabi Bhattacharya, David Mason and you.

Of all of the remaining open probabilistic questions concerning the rate of convergence in the high and infinite-dimensional CLT for regions defined by

quadratic forms, I felt that one was particularly important. It was raised in the seminal work by C.G. Esseen (1945) who proved that the error in the CLT for balls in dimension d is $\mathcal{O}(n^{-d/(d+1)})$. He noted that for sums of random vectors taking values in a lattice, his result is the equivalent in probability of classical results in analytic number theory by Landau and his students in the 1920s. They proved asymptotic rate bounds for the difference between the number of points of the standard lattice in ellipsoids of fixed shape blown up by a large radius factor and their corresponding Lebesgue-volume.

Interviewer: How did you approach these problems in number theory?

F.G.: In order to find the optimal rate in he CLT I started to study the old papers of Landau, Hardy and Littlewood and related papers by Weyl on this topic. First of all, I found out how rewarding it is to go back to the original sources concerning a problem. There you see the full force of the original arguments, whereas later publications often deal with refined versions of combinations of several methods of often undisclosed origin, which makes understanding the basic ideas and the further development much harder. It was very interesting to see a variety of methods that were either similar to or different from the ones I had used for the probabilistic questions. After intensive work I found ways to combine stochastic ideas with those of the classical analytic number theory establishing in this way a firm link of both worlds, where distributions on lattices turned out to provide the worst cases to be dealt with in the CLT.

I finally succeeded together with V. Bentkus to show the optimal rates of order n^{-1} for a sum of n vectors in the CLT, as well as in corresponding distributional problems in number theory in dimension 9 and larger. The chain of arguments started in number theory, improving Landau's bounds by new ones of optimal order and after that proceeding by representing distributional errors for sums by averages over errors for multinomial distributions on randomly selected lattices.

Interviewer: What was the role of the dimension in this problem. Esseen did not have any restriction in his bounds as far as I remember?

F.G.: It was clear by old results in number theory that dimensions 2–4 were different, hence one could not expect the same rate of convergence $\mathcal{O}(n^{-1})$ in the CLT for this case. But it took nearly a decade to get from dimension 9 to the final result for dimensions 5 and larger.

First this was done for ellipsoids in number theory in 2004, but the transfer to probability needed results for indefinite forms in number theory, which were obtained by means of quantitative equi-distribution results for orbits of 1-parameter (unipotent) subgroups jointly with G. Margulis 5 years later. The final transfer of these methods to the CLT in dimension 5 and larger with rate $\mathcal{O}(n^{-1})$ (without any $\log n$ factors) has been achieved last year jointly with A. Zaitsev. This closes the circle back to the original problem of Esseen.

Interviewer: Friedrich, thank you very much for taking so much of your time to tell us something about your career and the many activities that went with it.

Part I
Number Theory

Distribution of Algebraic Numbers and Metric Theory of Diophantine Approximation

V. Bernik, V. Beresnevich, F. Götze, and O. Kukso

Abstract In this paper we give an overview of recent results regarding close conjugate algebraic numbers, the number of integral polynomials with small discriminant and pairs of polynomials with small resultants.

Keywords Diophantine approximation • approximation by algebraic numbers • discriminant • resultant • polynomial root separation

2010 *Mathematics Subject Classification.* 11J83, 11J13, 11K60, 11K55

1 Introduction

Throughout the paper μA stands for the Lebesgue measure of a measurable set $A \subset \mathbb{R}$ and dim B denotes the Hausdorff dimension of B. Given $\psi : \mathbb{N} \to (0+\infty)$, let $\mathcal{L}(\psi)$ denote the set of $x \in \mathbb{R}$ such that

$$\left| x - \frac{p}{q} \right| < \frac{\psi(q)}{q} \tag{1}$$

has infinitely many solutions $(p, q) \in \mathbb{Z} \times \mathbb{N}$. We begin by recalling two classical results in metric theory of Diophantine approximation.

V. Bernik · O. Kukso
Institute of Mathematics, Academy of Sciences of Belarus, Minsk, Belarus

V. Beresnevich (✉)
Department of Mathematics, University of York, Heslington, York, England

F. Götze
Faculty of Mathematics, Bielefeld University, Bielefeld, Germany

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_2,

Khintchine's theorem [35]. *Let $\psi : \mathbb{N} \to (0, +\infty)$ be monotonic and I be an interval in $\mathbb{R}$. Then*

$$\mu(I \cap \mathcal{L}(\psi)) = \begin{cases} 0, & \text{if } \sum_{q=1}^{\infty} \psi(q) < \infty, \\ \mu(I), & \text{if } \sum_{q=1}^{\infty} \psi(q) = \infty. \end{cases} \tag{2}$$

Jarník–Besicovitch theorem [25, 34]. *Let $v > 1$ and for $q \in \mathbb{N}$ let $\psi_v(q) = q^{-v}$. Then*

$$\dim \mathcal{L}(\psi_v) = \frac{2}{v+1}.$$

The condition that ψ is monotonic can be omitted from the convergence case of Khintchine's theorem, though it is vital in the case of divergence—see [12, 33, 42] for a further discussion. By the turn of the millennium the above theorems were generalised in various directions. One important direction of research has been Diophantine approximation by algebraic numbers and/or integral polynomials, which has eventually grown into an area of number theory known as Diophantine approximation on manifolds.

Given a polynomial $P = a_n x^n + \cdots + a_1 x + a_0 \in \mathbb{Z}[x]$, the number $H = H(P) = \max_{0 \le i \le n} |a_j|$ will be called the (naive) height of P. Given $n \in \mathbb{N}$ and an approximation function $\Psi : \mathbb{N} \to (0, +\infty)$, let $\mathcal{L}_n(\Psi)$ be the set of $x \in \mathbb{R}$ such that

$$|P(x)| < \Psi(H(P)) \tag{3}$$

for infinitely many $P \in \mathbb{Z}[x] \setminus \{0\}$ with $\deg P \le n$. Note that $\mathcal{L}_1(\Psi)$ is essentially the same as the set $\mathcal{L}(\Psi)$ introduced above. Thus, the following statement represents an analogue of Khintchine's theorem for the case of polynomials.

Theorem 1. *Let $n \in \mathbb{N}$ and $\Psi : \mathbb{N} \to (0, +\infty)$ be monotonic. Then for any interval I*

$$\mu(I \cap \mathcal{L}_n(\Psi)) = \begin{cases} 0, & \text{if } \sum_{h=1}^{\infty} h^{n-1}\Psi(h) < \infty, \\ \mu(I), & \text{if } \sum_{h=1}^{\infty} h^{n-1}\Psi(h) = \infty. \end{cases} \tag{4}$$

The case of convergence of Theorem 1 was proved in [17], the case of divergence was proved in [4]. The condition that Ψ is monotonic can be omitted from the case of convergence as shown in [6]. Theorem 1 was generalised to the case of approximation in the fields of complex and p-adic numbers [9, 19], to simultaneous approximations in $\mathbb{R}\times\mathbb{C}\times\mathbb{Q}_p$ [22,26] and to various other settings. When $\Psi = \Psi_w$ is given by $\Psi_w(q) = q^{-w}$ Theorem 1 reduces to a famous problem of Mahler [37, 41] solved by Sprindžuk. The versions of Theorem 1 for monic polynomials were established in [27, 40]. For the more general case of Diophantine approximation on manifolds see, for example, [5, 7, 10, 15, 18, 20, 36, 42].

The more delicate Jarník-Besicovitch theorem was also generalised to the case of polynomials and reads as follows.

Theorem 2. *Let $w > n$ and $\Psi_w(q) = q^{-w}$. Then*

$$\dim \mathcal{L}_n(\Psi_w) = \frac{n+1}{w+1}. \tag{5}$$

The lower bound $\dim \mathcal{L}_n(\Psi_w) \geq \frac{n+1}{w+1}$ was obtained by Baker and Schmidt [2] who also conjectured (5). The conjecture was proved in full generality in [16]. It is worth noting that the generalised Baker-Schmidt problem for manifolds remains an open challenging problem in dimensions $n \geq 3$; the case of $n = 2$ was settled by R.C. Baker [3], see also [1, 8] and [7, 11, 43] for its analogue for simultaneous rational approximations.

The various techniques used to prove Theorems 1 and 2 make a substantial use of the properties of discriminants and resultants of polynomials and to some extent the distribution of algebraic numbers. The main substance of this paper will be to overview some relevant recent developments and techniques in this area.

2 Distribution of Discriminants of Integral Polynomials

The discriminant of a polynomial is a vital characteristic that crops up in various problems of number theory. For example, they play an important role in Diophantine equations, Diophantine approximation and algebraic number theory [41].

Let

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

be a polynomial of degree n and $\alpha_1, \alpha_2, \ldots, \alpha_n$ be its roots. By definition, the discriminant of P is given by

$$D(P) = a_n^{2n-2} \prod_{1 \leq i < j \leq n} (\alpha_i - \alpha_j)^2. \tag{6}$$

The following matrix formula for $D(P)$ is well known:

$$D(P) = (-1)^{n(n-1)/2} \begin{vmatrix} 1 & a_{n-1} & a_{n-2} & \ldots & a_0 & 0 & \ldots \\ 0 & a_n & a_{n-1} & \ldots & a_1 & a_0 & \ldots \\ & & & \ldots & & & \\ 0 & \ldots & 0 & a_n & \ldots & a_1 & a_0 \\ n & (n-1)a_{n-1} & (n-2)a_{n-2} & \ldots & 0 & 0 & \ldots \\ 0 & na_n & (n-1)a_{n-1} & \ldots & a_1 & 0 & \ldots \\ & & & \ldots & & & \\ 0 & \ldots & \ldots & 0 & na_n & \ldots & a_1 \end{vmatrix}.$$

Thus, the discriminant is an integer polynomial of the coefficients of P. Consequently, whenever P has rational integer coefficients the discriminant $D(P)$ is also an integer. Furthermore,

$$|D(P)| \geqslant 1 \quad \text{for any } P \in \mathbb{Z}[x] \text{ with } \deg P \geq 1 \text{ and } D(P) \neq 0. \tag{7}$$

Clearly, by (6), $D(P) \neq 0$ if and only if P has no multiple roots.

Fix $n \in \mathbb{N}$. Let $Q > Q_0(n)$, where $Q_0(n)$ is a sufficiently large number. Let $\mathcal{P}_n(Q)$ denote the set of non-zero polynomials $P \in \mathbb{Z}[x]$ with $\deg P \leq n$ and $H(P) \leq Q$. Throughout c_j, $j = 0, 1, \ldots$ will stand for positive constants depending on n only. When it is not essential for calculations we will denote these constants as $c(n)$. Also we will use the Vinogradov symbols: $A \ll B$ meaning that $A \leq c(n)B$. The expression $A \asymp B$ will mean $B \ll A \ll B$. Finally $\#S$ means the cardinality of a finite set S. In what follows we consider polynomials such that

$$c_1 Q < H(P) \leq Q, \ 0 < c_1 < 1. \tag{8}$$

Using the matrix representation for $D(P)$ one readily verifies that $|D(P)| < c(n)Q^{2n-2}$ for $P \in \mathcal{P}_n(Q)$. Thus, by (7), we have that

$$1 \leq |D(P)| < c(n)Q^{2n-2} \tag{9}$$

for polynomials $P \in \mathcal{P}_n(Q)$ with no multiple roots. Further, it is easily verified that

$$\#\mathcal{P}_n(Q) < 2^{2n+2}Q^{n+1}.$$

The latter together with (9) shows that $[1, c(n)Q^{2n-2}]$ contains intervals of length $c(n)Q^{n-3}$ that are not hit by the values of $D(P)$ for any $P \in \mathcal{P}_n(Q)$ whatsoever. For $n \geqslant 4$ these intervals can be arbitrarily large. Thus, the discriminants $D(P)$ are rather sparse in the interval $[1, c(n)Q^{2n-2}]$.

In order to understand the distribution of the values of $D(P)$ as P varies within $\mathcal{P}_n(Q)$, for each given $v \geq 0$ we introduce the following subclass of $\mathcal{P}_n(Q)$:

$$\mathcal{P}_n(Q, v) = \{P \in \mathcal{P}_n(Q) \ : \ |D(P)| < Q^{2n-2-2v} \text{ and (8) holds}\}. \tag{10}$$

These subclasses are of course dependant on the choice of c_1, but for the moment let us think of c_1 as a fixed constant.

We initially discuss some simple techniques utilizing the theory of continued fractions that enable one to obtain non-trivial lower bounds for $\#\mathcal{P}_n(Q, v)$ in terms of Q and v.

The first observation concerns shifts of the variable x by integers. More precisely, if $m \in \mathbb{Z}$ then $D(P(x)) = D(P(x - m))$. The height of $P(x - m)$ changes as m varies. It is a simple matter to see that imposing (8) on $P(x - m)$ restricts m to at most c_2 values. Furthermore, (8) ensures that polynomials of relatively small height cannot be in $\mathcal{P}_n(Q, v)$.

By (6), the fact that P belongs to $\mathcal{P}_n(Q, v)$ with $v > 0$ implies that P must necessarily have at least two close roots. This gives rise to a natural path to constructing polynomials P in $\mathcal{P}_n(Q, v)$—we have to make sure that they have close roots. We now describe a very special procedure that enables one to do exactly this.

Take n best approximations (convergents) $\frac{p_j}{q_j}$ to the number $\sqrt{2}$ with $k + 1 \le j \le k + n$ for some $k \in \mathbb{N}$. Define the polynomial

$$T_{\sqrt{2}}(x) = \prod_{j=k+1}^{k+n} (q_j x - p_j)$$

of degree n. Clearly the above mentioned best approximations to $\sqrt{2}$ are the roots of T. Also note that the height of T is $\ll a_n$, where

$$a_n = \prod_{k+1 \le i \le k+n} q_i .$$

From the theory of continued fractions we know that $q_j \le 3q_{j-1}$. Thus $a_n \le c(n)q_{k+1}^n$. On the other hand, we also know that for $i < j$

$$\left| \frac{p_i}{q_i} - \frac{p_j}{q_j} \right| \le \left| \frac{p_i}{q_i} - \frac{p_{i+1}}{q_{i+1}} \right| = \frac{1}{q_i q_{i+1}} < \frac{1}{q_i^2}.$$

Therefore, we can estimate the following product

$$\Pi_1 = \prod_{k+1 \le i < j \le k+n} \left| \frac{p_i}{q_i} - \frac{p_j}{q_j} \right|^2 \le \prod_{k+1 \le i \le k+n-1} \left(\frac{1}{q_i^2} \right)^{2(k+n-i)} \ll q_{k+1}^{-\sigma},$$

where

$$\sigma = 2 \sum_{i=k+1}^{k+n-1} 2(k + n - i) = 4 \sum_{\ell=1}^{n-1} \ell = 2n(n-1).$$

and see that

$$|D(T_{\sqrt{2}}(x))| \le a_n^{2n(n-1)} \Pi_1 \ll q_{k+1}^{2n(n-1)} q_{k+1}^{-\sigma} = 1.$$

This way we construct a polynomial of degree n with arbitrarily large height and discriminant as small as $c(n)$. However, to get quantitative bounds for $\#\mathcal{P}_n(Q, v)$ more needs to be done. The following lemmas underpin the construction.

Lemma 1. *Let I be an interval, $I \subset \mathbb{R}$, c_3 and c_4 be positive constants such that $\max\{c_3, c_4\} \le 1$. Given a sufficiently large Q, let $\mathcal{L}_{1,Q}(c_3, c_4)$ be the set of $x \in I$ such that the system of inequalities*

$$\begin{cases} |qx - p| < c_3 Q^{-1}, \\ 1 \le q \le c_4 Q, \end{cases} \tag{11}$$

has a solution in coprime $(p, q) \in \mathbb{Z} \times \mathbb{N}$. *Then for* $c_3 c_4 < \lambda$, $0 < \lambda < \frac{1}{3}$, *we have*

$$\mu \mathcal{L}_{1,Q}(c_3, c_4) < 3\lambda |I|.$$

Lemma 2. *Let* $M_n(Q)$ *denote the set of* $x \in I$ *such that the following* n *systems of inequalities*

$$\begin{cases} \frac{Q^{-1}}{3^i (n+1)^i} < |q_i x - p_i| < \frac{Q^{-1}}{3^{i-1}(n+1)^{i-1}} \\ 3^{i-2}(n+1)^{i-2} Q \le q_i \le 3^{i-1}(n+1)^{i-1} Q, \ 1 \le i \le n \end{cases}$$

have solutions in $\{(p_i, q_i)\}_{i=1}^n$. *Then* $\mu M_n(Q) > \frac{|I|}{n+1}$.

Lemma 1 is proved by summing up the measures of intervals given by the first inequality of (11). Lemma 2 is a corollary of Lemma 1 (which should be applied n times) and Minkowski's theorem for convex bodies. See [21] for details.

Now take any point $x_1 \in M_n(Q)$ and define

$$T_1(x) = \prod_{j=1}^{n} (q_j x - p_j) \qquad \text{and} \qquad Q = c(n) \prod_{i=1}^{n} q_i,$$

where (p_j, q_j) arise from Lemma 2. Estimating $|D(T_1)|$ gives

$$|D(T_1)| \ll Q^{2n-2} \prod_{1 \le i < j \le n} \left| \frac{p_i}{q_i} - \frac{p_j}{q_j} \right| \ll 1.$$

We now use the fact that $M_n(Q)$ is a fairly large subset of I to produce other polynomials with this property. For this purpose we choose points $x_2, x_3, \ldots \in M_n(Q)$ that are well separated. As a result we obtain

Theorem 3 ([21]). *For any sufficiently large* Q *there are* $c(n) Q^{\frac{2}{n}}$ *polynomials* $P \in \mathcal{P}_n(Q)$ *such that* $1 \le |D(P)| \le c(n)$.

The above ideas can be generalised to give a similar bound for the number of polynomials $P \in \mathcal{P}_n(Q)$ such that $|D(P)|$ lies in a neighborhood of some K with $c(n) < K < c(n) Q^{2n-2}$.

Theorem 4 ([21]). *For any* θ, $0 \le \theta \le 2n - 2$, *there are at least* $c(n) Q^{2/n}$ *polynomials* $P \in \mathcal{P}_n(Q)$ *with discriminants satisfying the inequalities*

$$c_5 Q^{\theta} < |D(P)| < c_6 Q^{\theta}.$$

We proceed by describing a more sophisticated method from [23] that produces lower bounds for $\#\mathcal{P}_n(Q, v)$. The main result is as follows.

Theorem 5 ([23]). *Let $v \in [0, \frac{1}{2}]$. Then there are at least $c(n)Q^{n+1-2v}$ polynomials $P \in \mathcal{P}_n(Q)$ with discriminants*

$$|D(P)| < Q^{2n-2-2v}. \tag{12}$$

Establishing upper bounds for $\#\mathcal{P}_n(Q, v)$ is likely a more difficult task. We expect that if we impose some reasonable conditions on polynomials P from $\mathcal{P}_n(Q)$ (for example excluding reducible polynomials) then the lower bound given by Theorem 5 would become sharp. We now state this formally as the following

Problem 1. Find reasonable constrains on polynomials P that chop a subclass $\mathcal{P}'_n(Q)$ off $\mathcal{P}_n(Q)$ such that $\#\mathcal{P}'(Q) \asymp \#\mathcal{P}_n(Q)$ and for $v \in [0, \frac{1}{2}]$

$$\#\{P \in \mathcal{P}'_n(Q) : |D(P)| < Q^{2(n-1-v)}\} \asymp Q^{n+1-2v}. \tag{13}$$

Obtaining the estimates of this ilk for a larger range of v is another problem. We wish to note that (13) is false for $\mathcal{P}_n(Q)$—see [32] for precise upper and lower bounds in the case $v < 3/5$ and $n = 3$.

Problem 2. For each n find the function $f_n(v)$, if it exists at all, such that for all sufficiently large Q one has the estimates

$$\#\{P \in \mathcal{P}_n(Q) : |D(P)| < Q^{2(n-1-v)}\} \asymp Q^{n+1-f_n(v)}. \tag{14}$$

It was shown in [32] that $f_3(v) = \frac{5}{3}v$ for $0 \le v \le 3/5$.

2.1 Sketch of the Proof of Theorem 5

Underlying the proof of Theorem 5 is the following result, which essentially plays the role of Lemma 1 in this more general context. In what follows, given an interval $I \subset [-1/2, 1/2]$, let $\mathcal{L}_n(I, Q, v, c_7, c_8)$ be the set of $x \in I$ such that

$$\begin{cases} |P(x)| < c_7 Q^{-n+v}, \\ |P'(x)| < c_8 Q^{1-v} \end{cases} \tag{15}$$

holds for some $P \in \mathcal{P}_n(Q)$.

Theorem 6. *Let Q denote a sufficiently large number, $v \in [0, \frac{1}{2}]$ and let c_7 and c_8 be positive constants such that $c_7 c_8 < n^{-1}2^{-n-12}$. Then*

$$\mu\mathcal{L}_n(I, Q, v, c_7, c_8) < \frac{|I|}{4}.$$

We now explain the role of Theorem 6 in establishing Theorem 5. Suppose that $P \in \mathbb{Z}[x]$, $\deg P \leq n$, $|a_n| > cH$. If $|a_n| \leq cH$ then the polynomial can be transformed into one with a large leading coefficient with the same discriminant—see [41].

By Dirichlet's pigeonhole principle, for any point $x \in I$ and $Q > 1$ the following system

$$\begin{cases} |P(x)| < Q^{-n+v}, \\ |P'(x)| < 8nQ^{1-v} \end{cases} \tag{16}$$

holds for some polynomial $P \in \mathcal{P}_n(Q)$. Let $\gamma = n^{-2}2^{-n-15}$ and $I = [-1/2, 1/2]$. Then, by Theorem 6, the set

$$B_1 = I \setminus \mathcal{L}_n(I, Q, v, 1, \gamma) \cup \mathcal{L}_n(I, Q, v, \gamma, 8n)$$

satisfies $\mu B_1 \geqslant \frac{1}{2}$ for all sufficiently large Q. Hence for any $x_1 \in B_1$ the solution $P \in \mathcal{P}_n(Q)$ to the system (16) must satisfy

$$\begin{cases} \gamma Q^{-n+v} < |P(x_1)| < Q^{-n+v}, \\ \gamma Q^{1-v} < |P'(x_1)| < 8nQ^{1-v}. \end{cases} \tag{17}$$

For all x in the interval $|x - x_1| < Q^{-\frac{2}{3}}$, the Mean Value Theorem gives

$$P'(x) = P'(x_1) + P''(\xi_1)(x - x_1) \text{ for some } \xi_1 \in [x, x_1]. \tag{18}$$

The obvious estimate $|P''(\xi_2)| < n^3 Q$ implies $|P''(\xi_1)(x - x_1)| < n^3 Q^{\frac{1}{3}}$. But $|P'(x_1)| \gg Q^{\frac{1}{2}}$ for $v \leq \frac{1}{2}$ and therefore, by (18) and the second inequality of (17), for sufficiently large Q we have that

$$\tfrac{\gamma}{2} Q^{1-v} < \tfrac{1}{2}|P'(x_1)| < |P'(x)| < 2|P'(x_1)| < 16nQ^{1-v}.$$

There are four possible combinations for signs of $P(x_1)$ and $P'(x_1)$. To illustrate the ideas we consider the case when $P_1(x_1) < 0$ and $P_1'(x_1) > 0$—the others are dealt with in a similar way. Our goal for now is to find a root of P close to x_1. Once again we appeal to the Mean Value Theorem:

$$P(x) = P(x_1) + P'(\xi_2)(x - x_1) \text{ for some } \xi_2 \in [x_1, x]. \tag{19}$$

Write $x = x_1 + \Delta$ and suppose that $\Delta > 2\gamma^{-1} Q^{-n-1+2v}$. If $P(x_1) < P(x_1 + \Delta) < 0$ then the first inequality of (17) implies

$$0 < P(x_1 + \Delta) - P(x_1) < Q^{-n+v}.$$

On the other hand we have

$$|P'(\xi_2)\Delta| > \frac{\gamma}{2} Q^{1-v} 2\gamma^{-1} Q^{-n-1+2v} = Q^{-n+v}.$$

Thus in view of (19) we obtain a contradiction. This means that $P_1(x_1 + \Delta) > 0$ and there is a real root α of the polynomial $P(x)$ between x_1 and $x_1 + \Delta$. Once again using the Mean Value Theorem and the estimates for $P(x)$ and $P'(\alpha)$ we get

$$|x_1 - \alpha| < 2\gamma^{-1} Q^{-n-1+2v} = n2^{n+13} Q^{-n-1+2v}. \tag{20}$$

Note that as well as ensuring that α, a root of P, is close to x_1 inequalities (17) keep α sufficiently away from x_1. We now explain this more formally. Again we consider only one of the four possibilities: $P(x_1) > 0$, $P'(x_1) < 0$. With $x = x_1 + \Delta_1$, by the Mean Value Theorem, we have

$$P(x) = P(x_1) + P'(\xi_3)\Delta_1, \ \xi_3 \in [x_1, x]. \tag{21}$$

If $\Delta_1 < 2^{-4} n^{-1} \gamma Q^{-n-1+2v}$ then in (21) the following holds: $|P(x_1)| > \gamma Q^{-n+v}$ and $|P'(\xi_3)\Delta_1| < \gamma Q^{-n+v}$. It implies that the polynomial $P(x)$ cannot have any root in the interval $[x_1, x_1 + \Delta_1]$ and therefore for any root α, we have

$$n^{-1} 2^{-n-13} Q^{-n-1+2v} < |x - \alpha|.$$

This time let α be the root of P closest to x_1. By the Mean Value Theorem,

$$P'(\alpha) = P'(x_1) + P''(\xi_4)(x_1 - \alpha), \ \xi_4 \in [x, \alpha],$$

the estimate $|P''(\xi)| < n^3 Q$ and (20) for sufficiently large Q we get

$$n^{-1} 2^{-n-13} Q^{1-v} < |P'(\alpha)| < 16n Q^{1-v}.$$

The square of derivative is a factor of the discriminant of P. Taking into account that for $|a_n| \asymp H(P)$ all roots of the polynomial are bounded, see [41]. Then we can estimate the differences $|\alpha_i - \alpha_j|, 2 \le i < j \le n$, by a constant $c(n)$. This way we obtain (12). Since $\mu B_1 \ge 1/2$ and x_1 is an arbitrary point in B_1 we must have $\gg Q^{n+1-2v}$ different α's that arise from (20). Since each polynomial P of degree $\le n$ has at most n roots this gives $\gg Q^{n+1-2v}$ polynomials in $\mathcal{P}_n(Q, v)$ satisfying (12)—see [13] for further details.

2.2 *Sketch of the Proof of Theorem 6*

The purpose of this section is to discuss the key ideas of the proof of Theorem 6 given in [13] as they may be useful in a variety of other tasks. We start by estimating the measure of x such that the system

$$\begin{cases} |P(x)| < c_{11} Q^{-n+v}, \\ Q^{1-v_1} < |P'(x)| < c_{12} Q^{1-v} \end{cases} \tag{22}$$

is solvable for $P \in \mathcal{P}_n(Q)$, where v_1 satisfies $v < v_1 \le 1$ and will be specified later.

We shall see that $P'(x)$ can be replaced with $P'(\alpha)$ in the second inequality of (22), where α denotes the root of P nearest to x. Indeed, using the Mean Value Theorem gives

$$P'(x) = P'(\alpha) + P''(\xi_1)(x-\alpha), \quad \xi_1 \in (\alpha, x).$$

We apply the following inequality for $|x-\alpha|$

$$|x-\alpha| < n\frac{|P(x)|}{|P'(x)|},$$

which was proved in [17, 41]. Then

$$|P'(\alpha)| = |P'(x) - P''(\xi_5)(x-\alpha)|, \ \xi_5 \in (\alpha, x).$$

As

$$|P''(\xi_1)(x-\alpha)| \le n^3 Q c_{11} n Q^{-n-1+v+v_1} = c_{11} n^4 Q^{-n+v+v_1}$$

for sufficiently large Q we obtain

$$\frac{3}{4} Q^{1-v_1} \le \frac{3}{4}|P'(x)| \le |P'(\alpha)| \le \frac{4}{3}|P'(x)| \le \frac{4}{3} c_{12} Q^{1-v}$$

and

$$\frac{3}{4}|P'(\alpha)| \le |P'(x)| \le \frac{4}{3}|P'(\alpha)|.$$

Therefore for sufficiently large Q inequality (22) implies

$$\begin{cases} |P(x)| < c_{11} Q^{-n+v} \\ \frac{3}{4} Q^{1-v_1} < |P'(\alpha)| < \frac{4}{3} c_{12} Q^{1-v} \\ |a_j| \le Q. \end{cases} \tag{23}$$

Let $\mathcal{L}'_n(v)$ denote the set of x, for which system (23) is solvable for $P \in \mathcal{P}_n(Q)$. Now we are able to prove that $\mu\mathcal{L}'_n(v) < \frac{3}{8}|I|$.

Consider the intervals:

$$\sigma_1(P) = \{x : |x-\alpha| < \frac{4}{3} c_{11} n Q^{-n+v} |P'(\alpha)|^{-1}\}$$

and

$$\sigma_2(P) = \{x : |x-\alpha| < c_{13} Q^{-1+\nu} |P'(\alpha)|^{-1}\}.$$

The value of c_{13} will be specified below. Of course, each polynomial P has up to n roots and potentially we have to consider all the different intervals $\sigma_1(P)$ and $\sigma_2(P)$ that correspond to each P. However, this will only affect the constant in the estimates. Thus, without loss of generality we confine ourselves to a single choice of $\sigma_1(P)$ and $\sigma_2(P)$. Obviously

$$|\sigma_1(P)| \le \tfrac{4}{3} c_{11} c_{13}^{-1} n Q^{-n+1} |\sigma_2(P)|. \tag{24}$$

Fix a vector $\overline{b} = (a_n, \dots, a_2)$ of the coefficients of P. The polynomials $P \in \mathcal{P}_n(Q)$ with the same vector $\bar{b}$ form a subclass of $\mathcal{P}_n(Q)$ which will be denoted by $\mathcal{P}(\overline{b})$.

The interval $\sigma_2(P_1)$ with $P_1 \in \mathcal{P}(\overline{b})$ is called *inessential* if there is another interval $\sigma_2(P_2)$ with $P_2 \in \mathcal{P}(\overline{b})$ such that

$$|\sigma_2(P_1) \cap \sigma_2(P_2)| \geqslant \tfrac{1}{2} |\sigma_2(P_1)|.$$

Otherwise for any $P_2 \in \mathcal{P}(\overline{b})$ different from P_1

$$|\sigma_2(P_1) \cap \sigma_2(P_2)| < \tfrac{1}{2} |\sigma_2(P_1)|$$

and the interval $\sigma_2(P_2)$ is called *essential*.

The case of essential intervals. In this case every point $x \in I$ belongs to at most two essential intervals $\sigma_2(P)$. Hence for any vector $\overline{b}$

$$\sum_{\substack{P \in \mathcal{P}(\overline{b}) \\ \sigma_2(P) \text{ is essential}}} |\sigma_2(P)| \le 2|I|. \tag{25}$$

The number of all possible vectors $\overline{b}$ is at most $(2Q+1)^{n-1} < 2^n Q^{n-1}$. Then, by (24) and (25), we obtain

$$\sum_{\overline{b}} \sum_{\substack{P \in \mathcal{P}(\overline{b}) \\ \sigma_2(P) \text{ is essential}}} |\sigma_1(P)| < \tfrac{4}{3} c_{11} c_{13}^{-1} n Q^{-n+1} 2|I| 2^n Q^{n-1} = n 2^{n+2} c_{11} c_{13}^{-1} |I|.$$

Thus for $c_{13} = n 2^{n+5} c_{11}$ the measure will be not larger than $\frac{1}{8}|I|$.

The case of inessential intervals. In this case we need to estimate the values of $|P_j(x)|$, $j = 1, 2$, for $x \in \sigma_2(P_1) \cap \sigma_2(P_2)$. By Taylor's formula,

$$P_j(x) = P_j'(\alpha)(x-\alpha) + \tfrac{1}{2} P_j''(\xi_6)(x-\alpha)^2 \text{ for some } \xi_6 \in (\alpha, x),$$

where α is the root of either P_1 or P_2 as appropriate, and

$$P_j'(x) = P_j'(\alpha) + P_j''(\xi_7)(x-\alpha) \text{ for some } \xi_7 \in (\alpha, x).$$

The second summand is estimated by

$$|P''(\xi_6)(x-\alpha)^2| \le 2n^3 c_{13}^2 Q^{-3+2v+2v_1},$$

while

$$|P'(\alpha)(x-\alpha)| < c_{13} Q^{-1+v}.$$

As $2v_1 < 2 - v$ for an appropriate choice of $v_1 < \frac{3}{4}$ we obtain

$$|P_j(x)| \le \tfrac{7}{6} c_{13} Q^{-1+v}, \ j = 1, 2. \tag{26}$$

Similarly we obtain the following estimate for $P_j'(x)$ when $v_1 \le 2 - 2v$:

$$|P_j'(x)| \le \tfrac{4}{3} c_{12} Q^{1-v}, \ j = 1, 2. \tag{27}$$

Let $K(x) = P_2(x) - P_1(x) \in \mathbb{Z}[x]$. Obviously $K(x)$ is non-zero and has the form $K(x) = b_1 x + b_0$. By (26) and (27), we readily obtain that

$$|b_1 x + b_0| < \tfrac{8}{3} c_{13} Q^{-1+v} \tag{28}$$

and

$$|b_1| = |K'(x)| < \tfrac{8}{3} c_{12} Q^{1-v}. \tag{29}$$

Thus, the union of inessential intervals can be covered by intervals $\Delta(b_1, b_0) \subset I$ given by (28). For fixed b_0 and b_1 the length of $\Delta(b_1, b_0)$ is bounded by $\frac{16}{3} c_{13} Q^{-1+v} b_1^{-1}$. Given that $x \in I$ and (28) is satisfied we conclude that b_0 takes at most $|I||b_1| + 2$ values. Then

$$\sum_{b_0} |\Delta(b_1, b_0)| \le \tfrac{16}{3} c_{13} Q^{-1+v} b_1^{-1} (|I||b_1| + 2) < 6 c_{13} Q^{-1+v} |I|. \tag{30}$$

Using (29) we further obtain that

$$\sum_{b_1} \sum_{b_0} |\Delta(b_1, b_0)| \le 2^5 c_{12} c_{13} Q^{1-v-1+v} |I| = n 2^{n+8} c_{11} c_{12} |I| = \tfrac{1}{8} |I|$$

for we have that $c_{11} c_{12} < n^{-1} 2^{-n-11}$. Finally, combining the estimates for essential and inessential intervals we obtain $\frac{1}{4}|I|$ as an upper bound for their total measure. The case $v \ge v_1$ can be deal with using methods described in [17] and [36].

3 Divisibility of Discriminants by Prime Powers

Let p be a prime number. Throughout μ_p denotes the Haar measure on $\mathbb{Q}_p$ normalized to have $\mu_p(\mathbb{Z}_p) = 1$. In this section we consider the divisibility of the discriminant $D(P)$, $P \in \mathcal{P}_n(Q)$ by prime powers p^l. This natural arithmetical question has usual interpretation in terms of Diophantine approximation in $\mathbb{Q}_p$, the field of p-adic number. Indeed, $p^l | D(P)$ if and only if $|D(P)|_p \le p^{-l}$, where $|\cdot|_p$ stands for the p-adic norm. Thus, the question we outlined above becomes a p-adic analogue of the problems we have considered in the previous section. Naturally, we proceed with the following p-adic analogue of Theorem 5.

Theorem 7 ([24]). *Let $v \in [0, \frac{1}{2}]$. Then there are at least $c(n)Q^{n+1-2v}$ polynomials $P \in \mathcal{P}_n(Q)$ with*

$$|D(P)|_p < Q^{-2v}. \tag{31}$$

The proof of this result relies on the following p-adic version of Theorem 6.

Theorem 8 ([24]). *Let Q denote a sufficiently large number and c_{14} and c_{15} denote constants depending only on n. Also, let K be a disc in $\mathbb{Q}_p$. Assume that $c_{14}c_{15} < 2^{-n-11}p^{-8}$ and $v \in [0, \frac{1}{2}]$. If $M_{n,Q}(c_{14}, c_{15})$ is the set of $w \in K \subset \mathbb{Q}_p$ such that the system of inequalities*

$$\begin{cases} |P(w)|_p < c_{14}Q^{-n-1+v}, \\ |P'(w)|_p < c_{15}Q^{-v} \end{cases}$$

has solutions in polynomials $P \in \mathcal{P}_n(Q)$, then

$$\mu_p(M_{n,Q}(c_{14}, c_{15})) < \tfrac{1}{4}\mu_p(K).$$

The techniques used in the proof of this theorem are essentially the p-adic analogues of those used for establishing Theorem 6 and draw on the estimates obtained in [39]—see [24] for more details. Skipping any explanation of the proof of Theorem 8, we now show how it is used for establishing Theorem 7.

Let K be a disc in $\mathbb{Q}_p$, $M_{n,Q}$ be the same as in Theorem 8, $\gamma = p^{-11}2^{-n-11}$ and

$$B = K \setminus \big(M_{n,Q}(\gamma, p^3) \cup M_{n,Q}(1, \gamma)\big).$$

Then, by Theorem 8, $\mu_p(B) \geqslant \frac{1}{2}\mu_p(K)$. Take any $w_1 \in B$. Then, using Dirichlet's pigeonhole principle we can find a polynomial $P \in \mathcal{P}_n(Q)$ such that $|P(w_1)|_p < Q^{-n-1+v}$ and $|P'(w_1)|_p < p^3Q^{-v}$. Since $w_1 \in B$ we have that

$$\begin{cases} \gamma Q^{-n-1+v} \le |P(w_1)|_p < Q^{-n-1+v}, \\ \gamma Q^{-v} \le |P'(w_1)|_p < p^3Q^{-v}. \end{cases} \tag{32}$$

Let $w \in K_1 = \{w \in \mathbb{Q}_p : |w - w_1|_p < Q^{-\frac{3}{4}}\}$. By Taylor's formula

$$P'(w) = P'(w_1) + \sum_{i=2}^{n} \frac{P^{(i)}(w_1)(w - w_1)^{i-1}}{(i-1)!}.$$

Since

$$|(i-1)!|_p^{-1} |P^{(i)}(w_1)|_p |w - w_1|_p^{i-1} \ll Q^{-\frac{3}{4}},$$

and

$$|P'(w)|_p \geqslant \gamma Q^{-v} \gg Q^{-\frac{1}{2}},$$

for all $w \in K_1$ we obtain that $|P'(w)|_p = |P'(w_1)|_p$. Let α be the closest root of $P(w)$ to the point w_1. Then, using the Mean Value Theorem, we get that $|w_1 - \alpha|_p \leq |P(w_1)|_p |P'(w_1)|_p^{-1}$. By (32),

$$|w_1 - \alpha|_p \leq \gamma^{-1} Q^{-n-1+2v}. \tag{33}$$

To estimate the distance between w_1 and the root of the polynomial we can also apply Hensel's lemma. Since $|P(w_1)|_p < |P'(w_1)|_p^2$ we obtain that the sequence $w_{n+1} = w_n - \frac{P_1(w_n)}{P_1'(w_n)}$ converges to the root α_1 of P that lies in $\mathbb{Q}_p$ and satisfies the inequality

$$|w_1 - \alpha_1|_p \leq |P(w_1)|_p |P'(w_1)|_p^{-2} \leq \gamma^{-2} Q^{-n-1+3v}. \tag{34}$$

Since $0 < \gamma < 1$ and $v > 0$ the right hand side of (33) is less than that of (34). This implies that the root α belongs to the disc with center w_1 of radius less than the radius for the disc defined for the root α_1. By Hensel's lemma, we find that $\alpha_1 \in \mathbb{Q}_p$ but estimate (33) does not guarantee that $\alpha \in \mathbb{Q}_p$. Suppose that $\alpha \neq \alpha_1$ and consider the discriminant of the polynomial $P \in \mathcal{P}_n(Q)$

$$D(P) = a_n^{2n-2} \prod_{1 \leq i < j \leq n} (\alpha_i - \alpha_j)^2. \tag{35}$$

From $|\alpha_j|_p \ll 1$ follows that $|\alpha_i - \alpha_j|_p \ll 1$. The product in (35) contains the difference $(\alpha - \alpha_1)$ for some $i \neq j$. We have $D(P) \in \mathbb{Z}$ and $|D(P)| \ll Q^{2n-2}$. Assume for the moment that $D(P) \neq 0$. Then $|D(P)|_p \geq |D(P)|^{-1} \gg Q^{-2n+2}$. From (33) and (34) we further obtain that

$$|\alpha_1 - \alpha|_p = |\alpha_1 - w_1 + w_1 - \alpha|_p \leq \max\{|w_1 - \alpha_1|_p, |w_1 - \alpha|_p\} \leq \gamma^{-2} Q^{-n-1+3v}.$$

Therefore

$$Q^{-2n+2} \ll |D(P)|_p \ll |\alpha_1 - \alpha|_p^2 < \gamma^{-4} Q^{-2n-2+6v}. \tag{36}$$

For $v \leq \frac{1}{2}$ and $Q > Q_0$ the inequality $Q^{-2n+2} \ll \gamma^{-4} Q^{-2n-2+6v}$ is a contradiction. Hence, $\alpha_1 = \alpha$. Thus, $\alpha \in \mathbb{Q}_p$ and $|w_1 - \alpha|_p$ satisfies condition (33).

In the case when $D(P) = 0$ one has to use the above argument with P replaced by its factor, say $\tilde{P}$. If α and α_1 are conjugate over $\mathbb{Q}$ one can take $\tilde{P}$ to be the minimal polynomials (over $\mathbb{Z}$) of α. Otherwise, $\tilde{P}$ is taken to be the product of the minimal polynomials for α and α_1.

By Taylor's formula,

$$P'(\alpha) = P'(w_1) + \sum_{i=2}^{n}((i-1)!)^{-1}P^{(i)}(w_1)(\alpha - w_1)^{i-1}. \tag{37}$$

Observe that

$$|(i-1)!|_p^{-1}|P^{(i)}(w_1)|_p|\alpha - w_1|_p^{i-1} \ll Q^{-n-1+2v}.$$

Then, by (33), we obtain

$$|P'(\alpha)|_p = |P'(w_1)|_p < p^3 Q^{-v}.$$

Therefore

$$|D(P)|_p = |a_n^2 \prod_{k=2}^{n}(\alpha_1 - \alpha_k)^2|_p \, |a_n^{2n-4} \prod_{2\le i<j\le n}(\alpha_i - \alpha_j)^2|_p \ll |P'(\alpha)|_p^2 \ll Q^{-2v}. \tag{38}$$

Inequality (33) implies that in the neighborhood of the point $w_1 \in B$ there exists a root α of the polynomial P with discriminant satisfying (38).

By (33), w_1 lies in the disc $K(\alpha, c(n)Q^{-n-1+2v})$. Since w_1 is an arbitrary point of B and $\mu_p(B) \ge \frac{1}{2}\mu_p(K)$, we must have $\ge c(n)Q^{n+1-2v}\mu_p(K)$ discs $K(\alpha, c(n)Q^{-n-1+2v})$ to cover B, where α is a root of some $P \in \mathcal{P}_n(Q)$ satisfying (38). Since each polynomial $P \in \mathcal{P}_n(Q)$ has at most n roots we must have $\ge c(n)Q^{n+1-2v}$ different polynomials $P \in \mathcal{P}_n(Q)$ satisfying (38), that is (31).

4 Close Conjugate Algebraic Numbers

Estimating the distance between conjugate algebraic numbers of degree n (in $\mathbb{C}$) has been investigated over the last 50 years. There are various upper and lower bounds found. However, the exact answers are known in the case of degree 2 and 3 only. Define κ_n (respectively κ_n^*) to be the infimum of κ such that the inequality

$$|\alpha_1 - \alpha_2| > H(\alpha_1)^{-\kappa}$$

holds for arbitrary conjugate algebraic numbers (respectively algebraic integers) $\alpha_1 \ne \alpha_2$ of degree n with sufficiently large height $H(\alpha_1)$. Here and elsewhere $H(\alpha)$ denotes the height of an algebraic number α, which is the absolute height of the minimal polynomial of α over $\mathbb{Z}$. Clearly, $\kappa_n^* \le \kappa_n$ for all n.

In 1964 Mahler [37] proved the upper bound $\kappa_n \leq n-1$, which is the best estimate up to date. It can be easily shown that $\kappa_2 = 1$ (see, e.g. [30]). Evertse [31] proved that $\kappa_3 = 2$. In the case of algebraic integers $\kappa_2^* = 0$ and $\kappa_3^* \geq 3/2$. The latter has been proved by Bugeaud and Mignotte [30].

For $n > 3$ estimates for κ_n are less satisfactory. At first Mignotte [38] showed that $\kappa_n, \kappa_n^* \geq n/4$ for all $n \geq 3$. Subsequently Bugeaud and Mignotte [29, 30] proved that

$$\begin{aligned}
&\kappa_n \geq n/2 && \text{when } n \geq 4 \text{ is even},\\
&\kappa_n^* \geq (n-1)/2 && \text{when } n \geq 4 \text{ is even},\\
&\kappa_n \geq (n+2)/4 && \text{when } n \geq 5 \text{ is odd},\\
&\kappa_n^* \geq (n+2)/4 && \text{when } n \geq 5 \text{ is odd}.
\end{aligned}$$

In a recent paper Bugeaud and Dujella [28] have further shown that

$$\kappa_n \geq \frac{n}{2} + \frac{n-2}{4(n-1)}. \tag{39}$$

On taking an alternative route it has been shown in [14] that there are numerous examples of close conjugate algebraic numbers:

Theorem 9 ([13, 14]). *For any $n \geq 2$ we have that*

$$\min\{\kappa_n, \kappa_{n+1}^*\} \geq \frac{n+1}{3}. \tag{40}$$

There are at least $c(n)Q^{\frac{n+1}{3}}$ pairs of conjugate algebraic numbers of degree n (or conjugate algebraic integers of degree $n+1$) α_1 and α_2 such that

$$|\alpha_1 - \alpha_2| \asymp H(\alpha_1)^{-\frac{n+1}{3}}$$

The proof of Theorem 9 is based on solvability of system of Diophantine inequalities for analytic functions [7] on the set of positive density on any interval $J \subset [-\frac{1}{2}, \frac{1}{2}]$. The interval $[-\frac{1}{2}, \frac{1}{2}]$ is taken to simplify the calculation of constants.

As above μ will denote Lebesgue measure in $\mathbb{R}$ while λ will be a non-negative constant. Given an interval $J \subset \mathbb{R}$, $|J|$ will denote the length of J. Also, $B(x, \rho)$ will denote the interval in $\mathbb{R}$ centered at x of radius ρ.

Let $n \geq 2$ be an integer, $\lambda \geq 0$, $0 < \nu < 1$ and $Q > 1$. Let $\mathbb{A}_{n,\nu}(Q, \lambda)$ be the set of algebraic numbers $\alpha_1 \in \mathbb{R}$ of degree n and height $H(\alpha_1)$ satisfying

$$\nu Q \leq H(\alpha_1) \leq \nu^{-1} Q \tag{41}$$

and

$$\nu\, Q^{-\lambda} \leq |\alpha_1 - \alpha_2| \leq \nu^{-1} Q^{-\lambda} \qquad \text{for some } \alpha_2 \in \mathbb{R}, \text{ conjugate to } \alpha_1. \tag{42}$$

Theorem 10. *For any $n \geq 2$ there is a constant $v > 0$ depending on n only with the following property. For any λ satisfying*

$$0 < \lambda \leq \frac{n+1}{3} \tag{43}$$

and any interval $J \subset [-\frac{1}{2}, \frac{1}{2}]$, for all sufficiently large Q

$$\mu\left(\bigcup_{\alpha_1 \in \mathbb{A}_{n,v}(Q,\lambda)} B(\alpha_1, Q^{-n-1+2\lambda}) \cap J\right) \geq \tfrac{3}{4}|J|. \tag{44}$$

Remark. The constant $\frac{3}{4}$ in the right hand side of (44) can be replaced by any positive number < 1.

Corollary 1. *For any $n \geq 2$ there is a positive constant v depending on n only such that for any λ satisfying (43) and any interval $J \subset [-\frac{1}{2}, \frac{1}{2}]$, for all sufficiently large Q*

$$\#\big(\mathbb{A}_{n,v}(Q,\lambda) \cap J\big) \geq \tfrac{1}{8} Q^{n+1-2\lambda}|J|. \tag{45}$$

The deduction of the corollary is rather simple. Indeed, if we have that $B(\alpha_1, Q^{-n-1+2\lambda}) \cap \frac{1}{2}J \neq \varnothing$ then $\alpha_1 \in J$ provided that Q is sufficiently large. Then, using (44) we obtain

$$\#\big(\mathbb{A}_{n,v}(Q,\lambda) \cap J\big) 2Q^{-n-1+2\lambda} \geq$$

$$\geq \mu\left(\bigcup_{\alpha_1 \in \mathbb{A}_{n,v}(Q,\lambda)} B(\alpha_1, Q^{-n-1+2\lambda}) \cap \tfrac{1}{2}J\right) \overset{(44)}{\geq} \tfrac{1}{4}|J|,$$

whence (45) readily follows. Taking the largest possible value of λ gives Theorem 9.

The key element of the proof of Theorem 10 is a far reaching generalisation of the arguments around (17) shown earlier. The appropriate analogue of Theorem 6 is given by Theorem 5.8 from [7]. In order to give a formal statement we first introduce some further notation. In what follows $\xi_0, \ldots, \xi_n \in \mathbb{R}^+$ will be positive real parameters satisfying the following conditions

$$\begin{aligned} &\xi_i \ll 1 \quad \text{when } 0 \leq i \leq m-1, \\ &\xi_i \gg 1 \quad \text{when } m \leq i \leq n, \\ &\xi_0 < \varepsilon, \qquad \xi_n > \varepsilon^{-1} \end{aligned} \tag{46}$$

for some $0 < m \leq n$ and $\varepsilon > 0$, where the constants in the Vinogradov's symbol $\ll$ depend on n only. We will also assume that

$$\prod_{i=0}^{n} \xi_i = 1. \tag{47}$$

Lemma 3. *For every $n \geq 2$ there are positive constants δ_0 and c_0 depending on n only with the following property. For any interval $J \subset [-\frac{1}{2}, \frac{1}{2}]$ there is a sufficiently small $\varepsilon = \varepsilon(n, J) > 0$ such that for any $\xi_0, \dots, \xi_n$ satisfying (46) and (47) there is a measurable set $G_J \subset J$ satisfying*

$$\mu(G_J) \geq \tfrac{3}{4}|J| \tag{48}$$

such that for every $x \in G_J$ there are $n+1$ linearly independent primitive irreducible polynomials $P \in \mathbb{Z}[x]$ of degree exactly n such that

$$\delta_0 \xi_i \leq |P^{(i)}(x)| \leq c_0 \xi_i \qquad \text{for all } i = 0, \dots, n\,. \tag{49}$$

We now reprocess the main steps of the proof of this statement. Let $n \geq 2$ and let $\xi_0, \dots, \xi_n$ be given and satisfy (46) and (47) for some m and ε. Let $J \subset [-\frac{1}{2}, \frac{1}{2}]$ be any interval and $x \in J$. Consider the system of inequalities

$$|P^{(i)}(x)| \leq \xi_i \qquad \text{when } 0 \leq i \leq n\,, \tag{50}$$

where $P(x) = a_n x^n + \cdots + a_1 x + a_0$. Let B_x be the set of $(a_0, \dots, a_n) \in \mathbb{R}^{n+1}$ satisfying (50). Note that B_x is a convex body in $\mathbb{R}^{n+1}$ symmetric about the origin. By (47), the volume of B_x equals $2^{n+1} \prod_{i=1}^{n} i!^{-1}$. Let $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_n$ be the successive minima of B_x. By Minkowski's theorem for successive minima,

$$\frac{2^{n+1}}{(n+1)!} \;\leq\; \lambda_0 \dots \lambda_n \operatorname{vol} B_x \;\leq\; 2^{n+1}.$$

Substituting the value of vol B_x gives $\lambda_0 \dots \lambda_n \;\leq\; \prod_{i=1}^{n} i!$, whence we get that

$$\lambda_n \leq \lambda_0^{-n} \prod_{i=1}^{n} i!. \tag{51}$$

Further we define certain subsets of J that we will 'avoid'. The avoidance will alow us to find the polynomials P of interest as well as to establish the lower bounds in (49). Let $E_\infty(J, \delta_1)$ be the set of $x \in J$ such that $\lambda_0 = \lambda_0(x) \leq \delta_1$, where $\delta_1 < 1$. By the definition of λ_0, there is a non-zero polynomial $P \in \mathbb{Z}[x]$, $\deg P \leq n$ satisfying

$$|P^{(i)}(x)| \leq \delta_1 \xi_i \qquad (0 \leq i \leq n). \tag{52}$$

Applying Lemma 3 from [7] gives

$$\mu(E_\infty(J, \delta_1)) \ll \left(1 + \frac{1}{\delta_J^\alpha} \max\left\{\frac{\delta_1 \xi_0}{\delta_1}, \frac{1}{\xi_n}\right\}^\alpha\right) \delta_1^{\frac{\alpha}{n+1}} |J|,$$

where $\delta_J > 0$ is a constant. By (46), $\max\{\xi_0, \xi_n^{-1}\} < \varepsilon$. Therefore $\mu(E_\infty(J, \delta_1)) \ll \delta_1^{\alpha/(n+1)}|J|$ provided that $\varepsilon < \delta_J$. Then there is a sufficiently small δ_1 depending on n only such that

$$\mu(E_\infty(J, \delta_1)) \le \tfrac{1}{4n+8}|J|. \tag{53}$$

By construction, for any $x \in J \setminus E_\infty(J, \delta_1)$ we have that

$$\lambda_0 \ge \delta_1. \tag{54}$$

Combining (51) and (54) gives

$$\lambda_n \le c_{16} := \delta_1^{-n} \prod_{i=1}^{n} i!, \tag{55}$$

where c_{16} depends on n only. By the definition of λ_n, there are $(n+1)$ linearly independent integer points $\mathbf{a}_j = (a_{0,j}, \dots, a_{n,j})$ $(0 \le j \le n)$ lying in the body $\lambda_n B_x \subset c_{16} B_x$. In other words, the polynomials $P_j(x) = a_{n,j}x^n + \cdots + a_{0,j}$ $(0 \le j \le n)$ satisfy the system of inequalities

$$|P_j^{(i)}(x)| \le c_{16}\xi_i \qquad (0 \le i \le n). \tag{56}$$

Let $A = (a_{i,j})_{0 \le i,j \le n}$ be the integer matrix composed from the integer points $\mathbf{a}_j$ $(0 \le j \le n)$. Since all these points are contained in the body $c_{16}B_x$, we have that $|\det A| \ll \mathrm{vol}(B_x) \ll 1$. That is $|\det A| < c_{17}$ for some constant c_{17} depending on n only. By Bertrand's postulate, choose a prime number p satisfying

$$c_{17} \le p \le 2c_{17}. \tag{57}$$

Therefore, $|\det A| < p$. Since $\mathbf{a}_0, \dots, \mathbf{a}_n$ are linearly independent and integer, $|\det A| \ge 1$. Therefore, $\det A \not\equiv 0 \pmod p$ and the following system

$$A\bar{t} \equiv \bar{b} \pmod p \tag{58}$$

has a unique non-zero integer solution $\bar{t} = {}^t(t_0, \dots, t_n) \in [0, p-1]^{n+1}$, where $\bar{b} := {}^t(0, \dots, 0, 1)$ and ^{t} denotes transposition. For $l = 0, \dots, n$ define $\bar{r}_l = {}^t(1, \dots, 1, 0, \dots, 0) \in \mathbb{Z}^{n+1}$, where the number of zeros is l. Since $\det A \not\equiv 0 \pmod p$, for every $l = 0, \dots, n$ the following system

$$A\bar{\gamma} \equiv -\frac{A\bar{t} - \bar{b}}{p} + \bar{r}_l \pmod p \tag{59}$$

has a unique non-zero integer solution $\bar{\gamma} = \bar{\gamma}_l \in [0, p-1]^{n+1}$. Define $\bar{\eta}_l := \bar{t} + p\bar{\gamma}_l$ $(0 \le l \le n)$. Consider the $(n+1)$ polynomials of the form

$$P_l(x) = a_n x^n + \cdots + a_0 := \sum_{i=0}^{n} \eta_{l,i} P_i(x) \in \mathbb{Z}[x], \tag{60}$$

where $(\eta_{l,0}, \ldots, \eta_{l,n}) = \overline{\eta}_l$. Since $\overline{r}_0, \ldots, \overline{r}_n$ are linearly independent modulo p, the vectors $-\frac{A\overline{t}-\overline{b}}{p} + \overline{r}_l$ $(l = 0, \ldots, n)$ are linearly independent modulo p. Hence, by (59), the vectors $\gamma_0, \ldots, \gamma_n$ are linearly independent modulo p. Hence, $\gamma_0, \ldots, \gamma_n$ are linearly independent over $\mathbb{Z}$. Since these vectors are integer, they are also linearly independent over $\mathbb{R}$. Therefore, the vectors $\overline{\eta}_l := \overline{t} + p\overline{\gamma}_l$ $(0 \le l \le n)$ are linearly independent over $\mathbb{R}$. Hence the polynomials given by (60) are linearly independent and so are non-zero.

Let $\overline{\eta} = \overline{\eta}_l$. Observe that $A\overline{\eta}$ is the column ${}^t(a_0, \ldots, a_n)$ of coefficients of P. By construction, $\overline{\eta} \equiv \overline{t} \pmod{p}$ and therefore $\overline{\eta}$ is also a solution of (58). Then, since $\overline{b} = {}^t(0, \ldots, 0, 1)$ and $A\overline{\eta} \equiv \overline{b} \pmod{p}$, we have that $a_n \not\equiv 0 \pmod{p}$ and $a_i \equiv 0 \pmod{p}$ for $i = 0, \ldots, n-1$. Furthermore, by (59), we have that $A\overline{\eta} \equiv \overline{b} + p\overline{r}_l \pmod{p^2}$. Then, on substituting the values of $\overline{b}$ and $\overline{r}_l$ into this congruence one readily verifies that $a_0 \equiv p \pmod{p^2}$ and so $a_0 \not\equiv 0 \pmod{p^2}$. By Eisenstein's criterion, P is irreducible.

Since both $\overline{t}$ and $\overline{\gamma}_l$ lie in $\in [0, p-1]^{n+1}$ and $\overline{\eta} = \overline{t} + p\overline{\gamma}_l$, it is readily seen that $|\eta_i| \le p^2$ for all i. Therefore, using (56) and (57) we obtain that

$$|P^{(i)}(x)| \le c_0 \xi_i \qquad (0 \le i \le n) \tag{61}$$

with $c_0 = 4(n+1)c_{16}c_{17}^2$. Without loss of generality we may assume that the $(n+1)$ linearly independent polynomials P constructed above are primitive (that is the coefficients of P are coprime) as otherwise the coefficients of P can be divided by their greatest common divisor. Thus, $P \in \mathbb{Z}[x]$ are primitive irreducible polynomials of degree n which satisfy the right hand side of (49). The final part of the proof is aimed at establishing the left hand side of (49).

Let $\delta_0 > 0$ be a sufficiently small parameter depending on n. For every $j = \overline{0, n}$ let $E_j(J, \delta_0)$ be the set of $x \in J$ such that there is a non-zero polynomial $R \in \mathbb{Z}[x]$, $\deg R \le n$ satisfying

$$|R^{(i)}(x)| \le \delta_0^{\delta_{i,j}} c_0^{1-\delta_{i,j}} \xi_i, \tag{62}$$

where $\delta_{i,j}$ equals 1 if $i = j$ and 0 otherwise. Let $\theta_i = \delta_0^{\delta_{i,j}} c_0^{1-\delta_{i,j}} \xi_i$. Then $E_j(J, \delta_0) \subset A_n(J; \theta_0, \ldots, \theta_n)$. In view of (46) and (47), Lemma 3 from [7] is applicable provided that $\varepsilon < \min\{c_0^{-1}, c_0\delta_0\}$. Then, by the same lemma,

$$\mu(E_j(J, \delta_0)) \ll \left(1 + \frac{1}{\delta_J^{\alpha}} \max\left\{\frac{c_0\xi_0}{c_0^n\delta_0}, \frac{1}{\delta_0 c_0 \xi_n}\right\}^{\alpha}\right) (\delta_0 c_0^n)^{1/(n+1)} |J|,$$

where $\delta_J > 0$ is a constant. It is readily seen that the above maximum is $\le \delta_J$ if $\varepsilon < \delta_J \delta_0 c_0$. Then

$$\mu(E_j(J, \delta_0)) \le \tfrac{1}{4n+8} |J| \tag{63}$$

provided that $\varepsilon < \min\{\delta_J \delta_0 c_0, c_0^{-1}, c_0 \delta_0\}$ and $\delta_0 = \delta_0(n)$ is sufficiently small. By construction, for any x in the set G_J defined by

$$G_J := J \setminus \left(\bigcup_{j=0}^{n} E_j(J, \delta_0) \cup E_\infty(J, \delta_1) \right)$$

we must necessarily have that $|P^{(i)}(x)| \geq \delta_0 \xi_i$ for all $i = 0, \dots, n$, where P is the same as in (61). Therefore, the left hand side of (49) holds for all i. Finally, observe that

$$\mu(G_J) \geq |J| - \sum_{i=0}^{n} \mu(E_i(J, \delta_0)) - \mu(E_\infty(J, \delta_1))$$
$$\overset{(53)\,\&\,(63)}{\geq} |J| - (n+2)\,\tfrac{1}{(4n+8)}|J| = \tfrac{3}{4}|J|.$$

The latter verifies (48) and completes the proof.

The following appropriate analogue of Lemma 3 for monic polynomials can be obtained using the techniques of [27].

Lemma 4. *For every $n \geq 2$ there are positive constants δ_0 and c_0 depending on n only with the following property. For any interval $J \subset [-\frac{1}{2}, \frac{1}{2}]$ there is a sufficiently small $\varepsilon = \varepsilon(n, J) > 0$ such that for any positive $\xi_0, \dots, \xi_n$ satisfying (46) and (47) there is a measurable set $G_J \subset J$ satisfying*

$$\mu(G_J) \geq \tfrac{3}{4}|J| \tag{64}$$

such that for every $x \in G_J$ there is an irreducible monic polynomials $P \in \mathbb{Z}[x]$ of degree $n+1$ satisfying (49).

5 On the Distribution of Resultants

In this section we discuss the distribution of the resultant $R(P_1, P_2)$ of polynomials P_1 and P_2 from $\mathcal{P}_n(Q)$. It is well known that

$$R(P_1, P_2) = a_n^m b_m^n \prod_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} (\alpha_i - \beta_j), \tag{65}$$

where $\alpha_1, \dots, \alpha_n$ are the roots of P_1 and $\beta_1, \dots, \beta_m$ are the roots of P_2; a_n and b_m stand for the leading coefficients of P_1 and P_2 respectively, where $n = \deg P_1$ and $m = \deg P_2$. The resultant $R(P_1, P_2)$ equals zero if and only if the polynomials P_1 and P_2 have a common root. Since the resultant can be represented as the

determinant of the Sylvester matrix of the coefficients of P_1 and P_2 it follows that R is integer. Furthermore,

$$|R(P_1, P_2)| \ll Q^{2n} \tag{66}$$

for $P_1, P_2 \in \mathcal{P}_n(Q)$. Akin to the already discussed results for the distribution of determinants we now state their analogue for resultants.

Theorem 11 ([13]). *Let $m \in \mathbb{Z}$ with $0 \le m < n$. Then there exist $\gg Q^{\frac{2(n+1)}{(m+1)(m+2)}}$ pairs of different primitive irreducible polynomials (P_1, P_2) from $\mathcal{P}_n(Q)$ of degree n such that*

$$1 \le |R(P_1, P_2)| \ll Q^{\frac{2(n-m-1)}{m+2}}. \tag{67}$$

Note that the left had side of (67) is obvious since P_1 and P_2 are primitive and irreducible. There are a few interesting corollaries of the above theorem. For $m = 0$ we have at least $c_1 Q^{n+1}$ pairs (P_1, P_2) that satisfy $|R(P_1, P_2)| \ll Q^{n-1}$. For $m = n - 1$ we have at least $c_2 Q^{\frac{2}{n}}$ pairs (P_1, P_2) that satisfy $|R(P_1, P_2)| \le c(n)$.

To introduce the ideas of the proof we first consider the case $m = 0$. By Lemma 3 given in the previous section, for any $x \in G_J$ there are different irreducible polynomials P_1 and P_2 of degree n and height $\ll Q$ such that

$$\begin{gathered} \delta_0 Q^{-n} < |P_i(x_1)| < c_0 Q^{-n},\ i = 1, 2 \\ \delta_0 Q < |P_i'(x_1)| < c_0 Q, \end{gathered} \tag{68}$$

Denote by α_1 the root of P_1 closest to x, and by β_1 the root of P_2 closest to x. Using (68) and the Mean Value Theorem, one can easily find that

$$|x - \alpha_1| \ll Q^{-n-1},\ |x - \beta_1| \ll Q^{-n-1}. \tag{69}$$

By (69), we get $|\alpha_1 - \beta_1| \ll Q^{-n-1}$. This together with (65) gives

$$|R(P_1, P_2)| \ll Q^{n-1}. \tag{70}$$

For a fixed pair of (α_1, β_1) inequalities (69) are satisfied only for a set of x of measure $\ll Q^{-n-1}$. Since $\mu(G_J) \gg |J|$, we must have $\gg Q^{n+1}$ diffract pairs (α_1, β_1) with the above properties. Since each polynomial in $\mathcal{P}_n(Q)$ has at most n root, we must have $\gg Q^{n+1}$ pairs of different irreducible polynomials (P_1, P_2) satisfying (70).

Now let $1 \le m \le n - 1$. Let $v_0, \dots, v_m \ge -1$ and $v_0 + v_1 + \cdots + v_m = n - m$. By Lemma 3, for any $x \in G_J$ there exists a pair of irreducible polynomials P_1, $P_2 \in \mathbb{Z}[x]$ of degree $\le n$ such that for $i = 1, 2$ we have that

$$\begin{gathered} \delta_0 Q^{-v_0} \le |P_i(x)| \le c_0 Q^{-v_0}, \\ \delta_0 Q^{-v_j} \le |P_i^{(j)}(x)| \le c_0 Q^{-v_j},\ 1 \le j \le m, \\ \delta_0 Q \le |P_i^{(j)}(x)| \le c_0 Q, \quad m + 1 \le j. \end{gathered} \tag{71}$$

Let $d_0, d_1, \ldots, d_{m+1}$ be a non-increasing sequence of real numbers such that

$$d_j = v_{j-1} - v_j,\ 1 \le j \le m,\ d_{m+1} = v_m + 1. \tag{72}$$

Order the roots α_i with respect to x as follows:

$$|x - \alpha_1| \le |x - \alpha_2| \le \ldots \le |x - \alpha_n|.$$

We claim that the roots α_j with $1 \le j \le m$ satisfy the following inequalities

$$\begin{aligned} |x - \alpha_j| &\ll Q^{-v_{j-1}+v_j}, \quad (1 \le j \le m-1) \\ |x - \alpha_m| &\ll Q^{-v_m-1}. \end{aligned} \tag{73}$$

The $(j-1)$-th derivative of $P(x) = a_n(x - \alpha_1)\cdots(x - \alpha_n)$ is

$$P^{(j-1)}(x) = (j-1)!a_n\Big(\prod_{i=j}^{n}(x - \alpha_i) + \sum_{i_j}(x - \alpha_{i_1})\cdot\cdots\cdot(x - \alpha_{i_{n-j}})\Big), \tag{74}$$

where the sum $\sum_{i_j}$ involves all summands with factor $(x - \alpha_{i_j})$, where $i_j < j$. If for $i < j$ there is a sufficiently large number $s_1 = c(n)$ such that

$$|x - \alpha_j| < s_1|x - \alpha_i| \tag{75}$$

then (72) implies (73) for $|x - \alpha_j|$. Otherwise, (74) implies

$$|P^{(j-1)}(x)| \gg |x - \alpha_j||P^{(j)}(x)|$$

because in this case the summand $(x_1 - \alpha_j)(x_1 - \alpha_{j+1})\cdots(x_1 - \alpha_n)$ in the above expression for $P^{(j-1)}(x_1)$ dominates all the others. Now choose v_j so that

$$v_0 = (m+1)v_m + m \ \text{ and } \ v_0 = (k+1)v_k - kv_{k+1}\ (1 \le k \le m-1). \tag{76}$$

By the first equation of (76), we get

$$v_m = \frac{v_0 - m}{m+1}. \tag{77}$$

By the other equalities of (76) we have that

$$v_{m-1} = \frac{2v_0 - m + 1}{m+1}, \quad v_k = \frac{(m-k+1)v_0 - k}{m+1}\ (1 \le k \le m-2). \tag{78}$$

Finally, by (77) and (78), we obtain

$$v_{j-1} - v_j = v_m + 1 = \frac{v_0 + 1}{m + 1} \quad (0 \le j \le m). \tag{79}$$

Taking into account the condition

$$v_0 + v_1 + \cdots + v_m = n - m,$$

by (77) and (78), we have

$$v_0 = \frac{2n - m}{m + 2}.$$

Thus roots $\alpha_1, \alpha_2, \ldots, \alpha_m$ of P_1 lie within $\ll Q^{-(v_0+1)(m+1)^{-1}}$ of x. The same is true for the roots $\beta_1, \beta_2, \ldots, \beta_m$ of P_2. Hence

$$T(m) = \prod_{1 \le i,j \le m+1} |\alpha_i - \beta_j| \ll Q^{-\frac{2(n+1)(m+1)}{m+2}}.$$

Consequently

$$|R(P_1, P_2)| \ll Q^{\frac{2(n-m-1)}{m+2}}. \tag{80}$$

It remains to give a lower bound for the number of pairs of (P_1, P_2) constructed above. Once again we use the fact that $\alpha_1, \ldots, \alpha_m, \beta_1, \beta_m$ lie within $\ll Q^{-(v_0+1)(m+1)^{-1}}$ of x. In other worlds x lies in the interval

$$\Delta(P_1, P_2) = \left\{x : |\max_i \{\max\{\alpha_i - x|, |\beta_i - x|\}\} \ll Q^{-(v_0+1)(m+1)^{-1}}\right\}.$$

Since x is an arbitrary point of G_J and $\mu(G_J) \gg |J|$, we must have $\gg Q^{(v_0+1)(m+1)^{-1}}$ different pairs (P_1, P_2) to cover G_J with intervals $\Delta(P_1, P_2)$. Substituting the value of v_0 we conclude that the number of different pairs (P_1, P_2) as above is at least $c(n)Q^{\frac{2(n+1)}{(m+1)(m+2)}}$ as required.

Acknowledgements The authors are very grateful to the anonymous referee for the very useful comments. The authors are grateful to SFB701 for its support and making this collaborative work possible.

References

1. D. Badziahin, Inhomogeneous Diophantine approximation on curves and Hausdorff dimension. Adv. Math. **223**, 329–351 (2010)
2. A. Baker, W.M. Schmidt, Diophantine approximation and Hausdorff dimension. Proc. Lond. Math. Soc. **21**, 1–11 (1970)

3. R.C. Baker, Dirichlet's theorem on Diophantine approximation. Math. Proc. Camb. Phil. Soc. **83**, 37–59 (1978)
4. V. Beresnevich, On approximation of real numbers by real algebraic numbers. Acta Arith. **90**, 97–112 (1999)
5. V. Beresnevich, A Groshev type theorem for convergence on manifolds. Acta Math. Hungar. **94**(1–2), 99–130 (2002)
6. V. Beresnevich, On a theorem of Bernik in the metric theory of Diophantine approximation. Acta Arith. **117**, 71–80 (2005)
7. V. Beresnevich, Rational points near manifolds and metric Diophantine approximation. Ann. Math. (2) **175**(1), 187–235 (2012)
8. V. Beresnevich, V. Bernik, M. Dodson, On the Hausdorff dimension of sets of well-approximable points on nondegenerate curves. Doklady NAN Belarusi **46**(6), 18–20 (2002)
9. V. Beresnevich, V. Bernik, E. Kovalevskaya, On approximation of p-adic numbers by p-adic algebraic numbers. J. Number Theor. **111**, 33–56 (2005)
10. V. Beresnevich, D. Dickinson, S. Velani, Measure theoretic laws for lim sup sets. Mem. Am. Math. Soc. **179**(846), x+91 (2006)
11. V. Beresnevich, D. Dickinson, S. Velani, Diophantine approximation on planar curves and the distribution of rational points. Ann. Math. (2) **166**, 367–426 (2007)
12. V. Beresnevich, V. Bernik, M. Dodson, S. Velani, *Classical Metric Diophantine Approximation Revisited*. Analytic number theory (Cambridge University Press, Cambridge, 2009), pp. 38–61
13. V. Beresnevich, V. Bernik, F. Götze, On distribution of resultants of integral polynomials with bounded degree and height. Doklady NAN Belarusi **54**(5), 21–23 (2010)
14. V. Beresnevich, V. Bernik, F. Götze, The distribution of close conjugate algebraic numbers. Composito Math. **5**, 1165–1179 (2010)
15. V.V. Beresnevich, V.I. Bernik, D. Kleinbock, G.A. Margulis, Metric diophantine approximation: the Khintchine-Groshev theorem for nondegenerate manifolds. Dedicated to Yuri I. Manin on the occasion of his 65th birthday. Mosc. Math. J. **2**(2), 203–225 (2002)
16. V.I. Bernik, An application of Hausdorff dimension in the theory of Diophantine approximation. Acta Arith. **42**, 219–253 (1983) (In Russian). English transl. in Am. Math. Soc. Transl. **140**, 15–44 (1988)
17. V.I. Bernik, The exact order of approximating zero by values of integral polynomials. Acta Arith. **53**(1), 17–28 (1989) (In Russian)
18. V.I. Bernik, M.M. Dodson, *Metric Diophantine Approximation on Manifolds*, vol. 137. Cambridge Tracts in Mathematics (Cambridge University Press, Cambridge, 1999)
19. V.I. Bernik, D.V. Vasil'ev, Diophantine approximations on complex manifolds in the case of convergence. Vestsī Nats. Akad. Navuk Belarusī Ser. Fīz.-Mat. Navuk, **1**, 113–115, 128 (2006) (In Russian)
20. V.I. Bernik, D. Kleinbock, G.A. Margulis, Khintchine-type theorems on manifolds: the convergence case for standard and multiplicative versions. Int. Math. Res. Notices 453–486 (2001)
21. V. Bernik, F. Götze, O. Kukso, Bad-approximable points and distribution of discriminants of the product of linear integer polynomials. Chebyshevskii Sb. **8**(2), 140–147 (2007)
22. V. Bernik, N. Budarina, D. Dickinson, A divergent Khintchine theorem in the real, complex, and p-adic fields. Lith. Math. J. **48**(2), 158–173 (2008)
23. V. Bernik, F. Götze, O. Kukso, Lower bounds for the number of integral polynomials with given order of discriminants. Acta Arith. **133**, 375–390 (2008)
24. V. Bernik, F. Götze, O. Kukso, On the divisibility of the discriminant of an integral polynomial by prime powers. Lith. Math. J. **48**, 380–396 (2008)
25. A.S. Besicovitch, Sets of fractional dimensions (IV): On rational approximation to real numbers. J. Lond. Math. Soc. **9**, 126–131 (1934)
26. N. Budarina, D. Dickinson, V. Bernik, Simultaneous Diophantine approximation in the real, complex and p-adic fields. Math. Proc. Camb. Phil. Soc. **149**(2), 193–216 (2010)
27. Y. Bugeaud, Approximation by algebraic integers and Hausdorff dimension. J. Lond. Math. Soc. **65**, 547–559 (2002)

28. Y. Bugeaud, A. Dujella, Root separation for irreducible integer polynomials, Root separation for irreducible integer polynomials. Bull. Lond. Math. Soc. **43**(6), 1239–1244 (2011)
29. Y. Bugeaud, M. Mignotte, On the distance between roots of integer polynomials. Proc. Edinb. Math. Soc. (2) **47**, 553–556 (2004)
30. Y. Bugeaud, M. Mignotte, Polynomial root separation. Int. J. Number Theor. **6**, 587–602 (2010)
31. J.-H. Evertse, Distances between the conjugates of an algebraic number. Publ. Math. Debrecen **65**, 323–340 (2004)
32. F. Götze, D. Kaliada, O. Kukso, Counting cubic integral polynomials with bounded discriminants. Submitted
33. G. Harman, *Metric Number Theory*. LMS Monographs New Series, vol. 18 (Clarendon Press, Oxford, 1998)
34. V. Jarnik, Diophantische approximationen und Hausdorffsches mass. Mat. Sb. **36**, 371–382 (1929)
35. A.Ya. Khintchine, Einige Sätze über Kettenbrüche, mit Anwendungen auf die Theorie der Diophantischen Approximationen. Math. Ann. **92**, 115–125 (1924)
36. D.Y. Kleinbock, G.A. Margulis, Flows on homogeneous spaces and Diophantine approximation on manifolds. Ann. Math. **148**, 339–360 (1998)
37. K. Mahler, An inequality for the discriminant of a polynomial. Michigan Math. J. **11**, 257–262 (1964)
38. M. Mignotte, Some useful bounds, in *Computer Algebra* (Springer, Vienna, 1983), pp. 259–263
39. A. Mohammadi, A.Salehi Golsefidy, S-arithmetic Khintchine-type theorem. Geom. Funct. Anal. **19**(4), 1147–1170 (2009)
40. N. Shamukova, V. Bernik, Approximation of real numbers by integer algebraic numbers and the Khintchine theorem. Dokl. Nats. Akad. Nauk Belarusi **50**(3), 30–33 (2006)
41. V. Sprindžuk, *Mahler's Problem in the Metric Theory of Numbers*, vol. 25. Translations of Mathematical Monographs (Amer. Math. Soc., Providence, 1969)
42. V.G. Sprindžuk, *Metric Theory of Diophantine Approximations.* Scripta Series in Mathematics (V. H. Winston & Sons, Washington, DC; A Halsted Press Book, Wiley, New York-Toronto, London, 1979), p. 156. Translated from the Russian and edited by Richard A. Silverman. With a foreword by Donald J. Newman
43. R.C. Vaughan, S. Velani, Diophantine approximation on planar curves: the convergence theory. Invent. Math. **166**(1), 103–124 (2006)

Fine-Scale Statistics for the Multidimensional Farey Sequence

Jens Marklof

Dedicated to Friedrich Götze on the occasion of his 60th birthday

Abstract We generalize classical results on the gap distribution (and other fine-scale statistics) for the one-dimensional Farey sequence to arbitrary dimension. This is achieved by exploiting the equidistribution of horospheres in the space of lattices, and the equidistribution of Farey points in a certain subspace of the space of lattices. The argument follows closely the general approach developed by A. Strömbergsson and the author [Ann. Math. 172:1949–2033, 2010].

Keywords Farey sequence • void probability • horosphere

2010 *Mathematics Subject Classification.* 11B57; 37D40

Denote by $\hat{\mathbb{Z}}^{n+1}$ the set of integer vectors in $\mathbb{R}^{n+1}$ with relatively prime coefficients, i.e., $\hat{\mathbb{Z}}^{n+1} = \{\boldsymbol{m} \in \mathbb{Z}^{n+1} \setminus \{\mathbf{0}\} : \gcd(\boldsymbol{m}) = 1\}$. The Farey points of level $Q \in \mathbb{N}$ are defined as the finite set

$$\mathcal{F}_Q = \left\{ \frac{\boldsymbol{p}}{q} \in [0,1)^n : (\boldsymbol{p}, q) \in \hat{\mathbb{Z}}^{n+1},\ 0 < q \le Q \right\}. \tag{1}$$

The number of Farey points of level Q is asymptotically, for large Q,

$$|\mathcal{F}_Q| \sim \sigma_Q := \frac{Q^{n+1}}{(n+1)\,\zeta(n+1)}. \tag{2}$$

J. Marklof (✉)
School of Mathematics, University of Bristol, Bristol BS8 1TW, UK
e-mail: j.marklof@bristol.ac.uk

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_3,

In fact, for any bounded set $\mathcal{D} \subset [0,1)^n$ with boundary of Lebesgue measure zero and non-empty interior,

$$|\mathcal{F}_Q \cap \mathcal{D}| \sim \mathrm{vol}(\mathcal{D})\, \sigma_Q, \tag{3}$$

which means the Farey sequence is uniformly distributed in $[0,1)^n$.

The objective of the present paper is to understand the fine-scale statistical properties of $\mathcal{F}_Q$. To this end, it will be convenient to identify $[0,1)^n$ with the unit torus $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ via the bijection $[0,1)^n \to \mathbb{T}^n$, $\boldsymbol{x} \mapsto \boldsymbol{x}+\mathbb{Z}^n$. We will consider the following two classical statistical measures of randomness of a deterministic point process: Given $k \in \mathbb{Z}_{\geq 0}$ and two test sets $\mathcal{D} \subset \mathbb{T}^n$ and $\mathcal{A} \subset \mathbb{R}^n$, both bounded, with boundary of Lebesgue measure zero and non-empty interior, define

$$P_Q(k,\mathcal{D},\mathcal{A}) = \frac{\mathrm{vol}\{\boldsymbol{x} \in \mathcal{D} : |(\boldsymbol{x} + \sigma_Q^{-1/n}\mathcal{A} + \mathbb{Z}^n) \cap \mathcal{F}_Q| = k\}}{\mathrm{vol}(\mathcal{D})} \tag{4}$$

and

$$P_{0,Q}(k,\mathcal{D},\mathcal{A}) = \frac{|\{\boldsymbol{r} \in \mathcal{F}_Q \cap \mathcal{D} : |(\boldsymbol{r} + \sigma_Q^{-1/n}\mathcal{A} + \mathbb{Z}^n) \cap \mathcal{F}_Q| = k\}|}{|\mathcal{F}_Q \cap \mathcal{D}|}. \tag{5}$$

The scaling of the test set $\mathcal{A}$ by a factor $\sigma_Q^{-1/n}$ ensures that the expectation value

$$\mathbb{E}P_Q(k,\mathcal{D},\mathcal{A}) := \sum_{k=0}^{\infty} k P_Q(k,\mathcal{D},\mathcal{A}) \tag{6}$$

is asymptotic to $\mathrm{vol}(\mathcal{A})$ for large Q. The quantity $P_{0,Q}(0,\mathcal{D},\mathcal{A})$ is the natural higher dimensional generalization of the gap distribution of sequences in one dimension, which, in the case of the Farey sequence for $\mathcal{A} = [0,s]$ and $\mathcal{D} = \mathbb{T}$, was calculated by Hall [6]. $P_Q(0,\mathbb{T},[0,s])$ corresponds in one dimension to the probability that the distance between a random point on $\mathbb{T}$ and the nearest element of the sequence is at least $s/2$. An elementary argument shows that in one dimension the density of this distribution is equal to $P_{0,Q}(0,\mathbb{T},[0,s])$, see e.g. [11, Theorem 2.2] and (36) below. The most comprehensive result in one dimension is due to Boca and Zaharescu [3], who calculate the limiting n-point correlation measures. We refer the reader to the survey article [4] for an overview of the relevant literature.

The results we will discuss here are valid in arbitrary dimension, and will also extend to the distribution in several test sets $\mathcal{A}_1,\dots,\mathcal{A}_s$. To keep the notation simple, we will restrict the discussion to one test set; the proofs are otherwise identical, cf. [15, Sect. 6] for the necessary tools.

It is evident that the distribution of Farey sequences is intimately linked to the distribution of directions of visible lattice points studied in [15, Sect. 2]. The only difference is in the ordering of the sequence of primitive lattice points and the way they are projected: In the Farey case we take all primitive lattice points in a blow-up

of the polytope $\{(\boldsymbol{x}, y) \in (0,1]^{n+1} : x_j \le y\}$, draw a line from each lattice point to the origin and record the intersection of these lines with the hyperplane $\{(\boldsymbol{x}, 1) : \boldsymbol{x} \in \mathbb{R}^n\}$. In the case of directions, we take all points in a fixed cone with arbitrary cross-section projected radially onto the unit sphere. Since the cross section of the cone is arbitrary, this yields (by a standard approximation argument) the statistics of primitive lattice points in the blow-up of any star-shaped domain (with boundary of measure zero), which are projected radially onto a suitably chosen hypersurface of codimension one. The proof of a limit distribution for $P_Q(k, \mathcal{D}, \mathcal{A})$ for Farey fractions is therefore a corollary of the results of [15].

If the points in $\mathcal{F}_Q$ were independent, uniformly distributed random variables in $\mathbb{T}^n$, we would have, almost surely, convergence to the Poisson distribution:

$$\lim_{Q\to\infty} P_Q(k, \mathcal{D}, \mathcal{A}) = \lim_{Q\to\infty} P_{0,Q}(k, \mathcal{D}, \mathcal{A}) = \frac{\mathrm{vol}(\mathcal{A})^k}{k!} \mathrm{e}^{-\mathrm{vol}(\mathcal{A})} \qquad \text{a.s.} \quad (7)$$

The $\mathcal{F}_Q$ are of course not Poisson distributed. But, as we will see, the limit distributions exist, are independent of $\mathcal{D}$, and are given by probability measures on certain spaces of random lattices in $\mathbb{R}^{n+1}$. The reason for this is as follows.

Define the matrices

$$h(\boldsymbol{x}) = \begin{pmatrix} 1_n & {}^t\mathbf{0} \\ -\boldsymbol{x} & 1 \end{pmatrix}, \qquad a(y) = \begin{pmatrix} y^{1/n} 1_n & {}^t\mathbf{0} \\ \mathbf{0} & y^{-1} \end{pmatrix} \quad (8)$$

and the cone

$$\mathfrak{C}(\mathcal{A}) = \{(\boldsymbol{x}, y) \in \mathbb{R}^n \times (0,1] : \boldsymbol{x} \in \sigma_1^{-1/n} y\mathcal{A}\} \subset \mathbb{R}^{n+1}. \quad (9)$$

Then, for any $(\boldsymbol{p}, q) \in \mathbb{R}^{n+1}$,

$$\frac{\boldsymbol{p}}{q} \in \boldsymbol{x} + \sigma_Q^{-1/n}\mathcal{A}, \qquad 0 < q \le Q, \quad (10)$$

if and only if

$$(\boldsymbol{p}, q)h(\boldsymbol{x})a(Q) \in \mathfrak{C}(\mathcal{A}). \quad (11)$$

Thus, if Q is sufficiently large so that $\sigma_Q^{-1/n}\mathcal{A} \subset (0,1]^n$, then

$$|(\boldsymbol{x} + \sigma_Q^{-1/n}\mathcal{A} + \mathbb{Z}^n) \cap \mathcal{F}_Q| = |\hat{\mathbb{Z}}^{n+1} h(\boldsymbol{x})a(Q) \cap \mathfrak{C}(\mathcal{A})|. \quad (12)$$

This observation reduces the question of the distribution of the Farey sequence to a problem in the geometry of numbers. In particular, (4) and (5) can now be expressed as

$$P_Q(k, \mathcal{D}, \mathcal{A}) = \frac{\mathrm{vol}\{\boldsymbol{x} \in \mathcal{D} : |\hat{\mathbb{Z}}^{n+1} h(\boldsymbol{x})a(Q) \cap \mathfrak{C}(\mathcal{A})| = k\}}{\mathrm{vol}(\mathcal{D})} \quad (13)$$

and

$$P_{0,Q}(k,\mathcal{D},\mathcal{A}) = \frac{|\{\boldsymbol{r}\in\mathcal{F}_Q\cap\mathcal{D} : |\hat{\mathbb{Z}}^{n+1}h(\boldsymbol{r})a(Q)\cap\mathfrak{C}(\mathcal{A})| = k\}|}{|\mathcal{F}_Q\cap\mathcal{D}|}. \tag{14}$$

Let $G = \mathrm{SL}(n+1,\mathbb{R})$ and $\Gamma = \mathrm{SL}(n+1,\mathbb{Z})$. The quotient $\Gamma\backslash G$ can be identified with the space of lattices in $\mathbb{R}^{n+1}$ of covolume one. We denote by μ the unique right G-invariant probability measure on $\Gamma\backslash G$. Let furthermore be μ_0 the right G_0-invariant probability measure on $\Gamma_0\backslash G_0$, with $G_0 = \mathrm{SL}(n,\mathbb{R})$ and $\Gamma_0 = \mathrm{SL}(n,\mathbb{Z})$.

Define the subgroups

$$H = \left\{M\in G : (\mathbf{0},1)M = (\mathbf{0},1)\right\} = \left\{\begin{pmatrix} A & \boldsymbol{b}\\ \mathbf{0} & 1\end{pmatrix} : A\in G_0,\ \boldsymbol{b}\in\mathbb{R}^n\right\} \tag{15}$$

and

$$\Gamma_H = \Gamma\cap H = \left\{\begin{pmatrix} \gamma & \boldsymbol{m}\\ \mathbf{0} & 1\end{pmatrix} : \gamma\in\Gamma_0,\ \boldsymbol{m}\in\mathbb{Z}^n\right\}. \tag{16}$$

Note that H and Γ_H are isomorphic to $\mathrm{ASL}(n,\mathbb{R})$ and $\mathrm{ASL}(n,\mathbb{Z})$, respectively. We normalize the Haar measure μ_H of H so that it becomes a probability measure on $\Gamma_H\backslash H$. That is,

$$d\mu_H(M) = d\mu_0(A)\,d\boldsymbol{b}, \qquad M = \begin{pmatrix} A & \boldsymbol{b}\\ \mathbf{0} & 1\end{pmatrix}. \tag{17}$$

The main ingredient in the proofs of the limit theorems for $P_{0,Q}(k,\mathcal{D},\mathcal{A})$ and $P_Q(k,\mathcal{D},\mathcal{A})$ are the following two equidistribution theorems. The first is the classic equidistribution theorem for closed horospheres of large volume (cf. [15, Sect. 5] for background and references), the second the equidistribution of Farey points on closed horospheres [13, Theorem 6]. In the latter, a key observation is that [13, (3.53)]

$$\Gamma h(\boldsymbol{r})a(Q)\in\Gamma\backslash\Gamma Ha(\tfrac{Q}{q}) \simeq \Gamma_H\backslash Ha(\tfrac{Q}{q}). \tag{18}$$

Theorem 1. *For $f : \mathbb{T}^n\times\Gamma\backslash G\to\mathbb{R}$ bounded continuous,*

$$\lim_{Q\to\infty}\int_{\mathbb{T}^n} f\big(\boldsymbol{x},h(\boldsymbol{x})a(Q)\big)\,d\boldsymbol{x} = \int_{\mathbb{T}^n\times\Gamma\backslash G} f(\boldsymbol{x},M)\,d\boldsymbol{x}\,d\mu(M). \tag{19}$$

Theorem 2. *For $f : \mathbb{T}^n\times\Gamma\backslash G\to\mathbb{R}$ bounded continuous,*

$$\lim_{Q\to\infty}\frac{1}{|\mathcal{F}_Q|}\sum_{\boldsymbol{r}\in\mathcal{F}_Q} f\big(\boldsymbol{r},h(\boldsymbol{r})a(Q)\big) = \int_0^1\int_{\mathbb{T}^n\times\Gamma_H\backslash H} f(\boldsymbol{x},Ma(\lambda^{-\frac{1}{n+1}}))\,d\boldsymbol{x}\,d\mu_H(M)\,d\lambda. \tag{20}$$

Both theorems can be derived from the mixing property of the action of the diagonal subgroup $\{a(y)\}_{y\in\mathbb{R}_{>0}}$. The exponential decay of correlations of this

action was exploited by H. Li to calculate explicit rates of convergence [9]. One can furthermore generalize Theorem 2 to general lattices Γ in G and non-closed horospheres [12]. Theorem 2 may also be interpreted as an equidistribution theorem for periodic points of the return map of the horocycle flow (in the case $n = 1$) to the section

$$\Gamma\backslash\Gamma H\{a(y) : y \in \mathbb{R}_{>1}\} \simeq \Gamma_H\backslash H\{a(y) : y \in \mathbb{R}_{>1}\} \tag{21}$$

which is discussed in [1]. The identification of (21) as an embedded submanifold, which is transversal to closed horospheres of large volume, is central to the proof of Theorem 2 in [13].

By standard probabilistic arguments, the statements of both theorems remain valid if f is a characteristic function of a subset $\mathcal{S} \subset \mathbb{T}^n \times \Gamma\backslash G$ whose boundary has measure zero with respect to the limit measure $d\boldsymbol{x}\, d\mu(M)$ or $d\boldsymbol{x}\, d\mu_H(M)\, d\lambda$, respectively. The relevant set in our application is

$$\mathcal{S} = \mathcal{D} \times \{M \in \Gamma\backslash G : |\hat{\mathbb{Z}}^{n+1} M \cap \mathfrak{C}(\mathcal{A})| \geq k\}. \tag{22}$$

The fact that $\mathcal{S}$ has indeed boundary of measure zero with respect to $d\boldsymbol{x}\, d\mu(M)$ is proved in [15, Sect. 6]. We can therefore conclude:

Theorem 3. *Let $k \in \mathbb{Z}_{\geq 0}$, and $\mathcal{D} \subset \mathbb{T}^n$, $\mathcal{A} \subset \mathbb{R}^n$ bounded with boundary of Lebesgue measure zero. Then*

$$\lim_{Q\to\infty} P_Q(k, \mathcal{D}, \mathcal{A}) = P(k, \mathcal{A}) \tag{23}$$

with

$$P(k, \mathcal{A}) = \mu(\{M \in \Gamma\backslash G : |\hat{\mathbb{Z}}^{n+1} M \cap \mathfrak{C}(\mathcal{A})| = k\}), \tag{24}$$

which is independent of the choice of $\mathcal{D}$.

In the second case, we require that the set

$$\{M \in \Gamma_H\backslash H : |\hat{\mathbb{Z}}^{n+1} M \cap \mathfrak{C}_\lambda(\mathcal{A})| \geq k\}, \qquad \mathfrak{C}_\lambda(\mathcal{A}) := \mathfrak{C}(\mathcal{A}) a(\lambda^{\frac{1}{n+1}}), \tag{25}$$

has boundary of measure zero with respect to μ_H, which follows from analogous arguments. With this, we have:

Theorem 4. *Let $k \in \mathbb{Z}_{\geq 0}$, and $\mathcal{D} \subset \mathbb{T}^n$, $\mathcal{A} \subset \mathbb{R}^n$ bounded with boundary of Lebesgue measure zero. Then*

$$\lim_{Q\to\infty} P_{0,Q}(k, \mathcal{D}, \mathcal{A}) = P_0(k, \mathcal{A}) = \int_0^1 p_0(k, \mathfrak{C}_\lambda(\mathcal{A}))\, d\lambda. \tag{26}$$

where

$$p_0(k, \mathfrak{C}) = \mu_H(\{M \in \Gamma_H\backslash H : |\hat{\mathbb{Z}}^{n+1} M \cap \mathfrak{C}| = k\}), \tag{27}$$

which is independent of the choice of $\mathcal{D}$.

In dimension $n \geq 2$, it is difficult to obtain a more explicit description of the limit distributions $P(k, \mathcal{A})$ and $P_0(k, \mathcal{A})$. It is however possible to provide asymptotic estimates for large and small sets $\mathcal{A}$ when $k = 0$ and $k = 1$, see [2, 18] for general results in this direction. The case of fixed $\mathcal{A}$ and large k is discussed in [10].

The geometry of $\Gamma\backslash G$ is significantly simpler in the case $n = 1$. This permits the derivation of explicit formulas for the limit distributions in many instances, cf. [5, 14, 19]. For example, take $\mathcal{A} = (0, s]$, and the cone $\mathfrak{C}_\lambda(\mathcal{A})$ becomes the triangle

$$\Delta_{s,\lambda} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 \leq \tfrac{\pi^2}{3} x_2 \lambda s,\ 0 < x_2 \leq \lambda^{-1/2}\}, \tag{28}$$

where we have used $\sigma_1 = \frac{1}{2\zeta(2)} = \frac{3}{\pi^2}$. Furthermore $\Gamma_H\backslash H$ is simply the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, μ_H is the standard Lebesgue measure. Hence

$$p_0(k, \Delta_{s,\lambda}) = \mathrm{meas}(\{x \in \mathbb{T} : |\{(p, q) \in \hat{\mathbb{Z}}^2 : (p, px + q) \in \Delta_{s,\lambda}\}| = k\}). \tag{29}$$

It is now a geometric exercise to work out the case $k = 0$: With the shorthand $y = \lambda^{1/2}$ and $a = (\frac{\pi^2}{3} s)^{-1}$, we deduce

$$p_0(0, \Delta_{s,\lambda}) = \begin{cases} 1 & \text{if } y \leq a \\ 1 - \frac{1}{y} + \frac{a}{y^2} & \text{if } a < y \leq a(1-y)^{-1} \\ 0 & y > a(1-y)^{-1}. \end{cases} \tag{30}$$

Solving for y, we have in the case $0 < a \leq \frac{1}{4}$

$$p_0(0, \Delta_{s,\lambda}) = \begin{cases} 1 & \text{if } y \in [0, a] \\ 1 - \frac{1}{y} + \frac{a}{y^2} & \text{if } y \in [a, \frac{1}{2} - \sqrt{\frac{1}{4} - a}] \cup [\frac{1}{2} + \sqrt{\frac{1}{4} - a}, 1] \\ 0 & \text{if } y \in [\frac{1}{2} - \sqrt{\frac{1}{4} - a}, \frac{1}{2} + \sqrt{\frac{1}{4} - a}]. \end{cases} \tag{31}$$

For $\frac{1}{4} < a < 1$, we have

$$p_0(0, \Delta_{s,\lambda}) = \begin{cases} 1 & \text{if } y \in [0, a] \\ 1 - \frac{1}{y} + \frac{a}{y^2} & \text{if } y \in [a, 1], \end{cases} \tag{32}$$

and for $a \geq 1$, we have

$$p_0(0, \Delta_{s,\lambda}) = 1, \qquad y \in [0, 1]. \tag{33}$$

The gap distribution $P_0(0, [0, s])$ is now an elementary integral (recall (26)), which yields

$$P_0(0,[0,s]) = \begin{cases} 1 & \text{if } a \in [1,\infty) \\ -1+2a-2a\log a & \text{if } a \in [\frac{1}{4},1] \\ -1+2a+2\sqrt{\frac{1}{4}-a}-4a\log(\frac{1}{2}+\sqrt{\frac{1}{4}-a}) & \text{if } a \in [0,\frac{1}{4}]. \end{cases} \tag{34}$$

which reproduces Hall's distribution [6]. The density of this distribution is

$$-\frac{d}{ds}P_0(0,[0,s]) = \frac{\pi^2}{3}a^2\frac{d}{da}P_0(0,[0,s]) = \begin{cases} 0 & \text{if } a \in [1,\infty) \\ -\frac{2\pi^2}{3}a^2\log a & \text{if } a \in [\frac{1}{4},1] \\ -\frac{4\pi^2}{3}a^2\log(\frac{1}{2}+\sqrt{\frac{1}{4}-a}) & \text{if } a \in [0,\frac{1}{4}], \end{cases} \tag{35}$$

cf. [4, Theorem 2.1]. By [11, Theorem 2.2], we have

$$-\frac{d}{ds}P(0,[0,s]) = P_0(0,[0,s]), \tag{36}$$

and hence formula (34) yields directly the density of the distribution of the distance to the nearest element. Formula (34) was rediscovered in [8, Lemma 2.6].

Theorems 1 and 2 reduce in the case $n = 1$ to classic statements in the theory of automorphic forms, with precise bounds on the rate of convergence. Sarnak [16] proved Theorem 1 for test functions $f \in C_0^\infty$ (infinitely differentiable, compactly supported) that are independent of the first coordinate x, and showed that the optimal rate of convergence holds if and only if the Riemann Hypothesis is true (this phenomenon was first pointed out by Zagier [20]). The reason for the appearance of the Riemann zeros is that the only relevant harmonics in the problem are Eisenstein series $E_{2k}(z,s)$ of even weight $2k$, whose poles are located at the poles of

$$\sum_{q=1}^{\infty}\frac{\varphi(q)}{q^{2s}} = \frac{\zeta(2s-1)}{\zeta(2s)}. \tag{37}$$

where $\varphi(s)$ is Euler's totient function and $\zeta(s)$ the Riemann zeta function.

Under the Riemann Hypothesis, Sarnak's rate is significantly better than what one would expected from square-root cancellations—it is the square-root of that. If the test function f depends on x (we assume again f is C_0^∞), the work of Hejhal [7] and Strömbergsson [17] shows that the convergence rate slows to the square-root of the horocycle length (or worse) as other terms in the harmonics dominate the error coming from of the Riemann zeros. The object replacing the Eisenstein series in this setting is the Poincaré series $P_{m,2k}(z,s)$ of weight $2k$.

The proof Theorem 2 for $n = 1$ on the other hand quickly reduces to estimates of sums of Kloosterman sums. To see this, note first of all that the statement of Theorem 2 is equivalent to: For every bounded continuous function $f : \mathbb{T}\times\Gamma\backslash\mathbb{H} \to \mathbb{R}$ (where $\mathbb{H}$ is the complex upper half plane, on which $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ acts by Möbius transformations) we have

$$\lim_{Q\to\infty}\frac{1}{|\mathcal{F}_Q|}\sum_{q=1}^{Q}\sum_{p\in\mathbb{Z}_q^\times} f\left(\frac{p}{q},\frac{\overline{p}}{q}+\mathrm{i}\frac{Q^2}{q^2}\right)=\int_0^\infty\int_0^1\int_0^1 f(x,u+\mathrm{i}v)\,dx\,du\,\frac{dv}{v^2}. \tag{38}$$

Here $\mathbb{Z}_q^\times$ denotes the multiplicative group of invertible residues mod q, and $\overline{p}$ is the inverse of p mod q. One way of proving (38) is to expand $f\in \mathrm{C}_0^\infty$ in its harmonics (Fourier series in x and u and Mellin transform in v) which leads to Selberg's Kloosterman zeta function

$$Z_{m_1,m_2}(s)=\sum_{q=1}^{\infty}\frac{K(m_1,m_2,q)}{q^{2s}} \tag{39}$$

with the Kloosterman sum

$$K(m_1,m_2,q)=\sum_{p\in\mathbb{Z}_q^\times}\mathrm{e}^{2\pi\mathrm{i}(m_1p+m_2\overline{p})/q}. \tag{40}$$

As in the case of the equidistribution of closed horocycles, where the asymptotics was determined by the poles of the Eisenstein and Poincaré series, the poles of $Z_{m_1,m_2}(s)$ now determine the asymptotics of (38). Note that $Z_{0,m_2}(s)$ are precisely the Fourier coefficients of $E_0(z,s)$ and, as already understood by Selberg, $Z_{m_1,m_2}(s)$ is the m_2-th Fourier coefficient of the Poincaré series $P_{m_1,0}(z,s)$. Hence the appearance of the Riemann hypothesis in the error term of Theorem 2 mirrors exactly the situation in Theorem 1.

Acknowledgements J.M. is supported by a Royal Society Wolfson Research Merit Award, a Leverhulme Trust Research Fellowship and ERC Advanced Grant HFAKT.

References

1. J.S. Athreya, Y. Cheung, A Poincaré section for horocycle flow on the space of lattices. arXiv:1206.6597 (2012)
2. J.S. Athreya, G.A. Margulis, Logarithm laws for unipotent flows I. J. Mod. Dyn. **3**, 359–378 (2009)
3. F.P. Boca, A. Zaharescu, The correlations of Farey fractions. J. Lond. Math. Soc. **72**, 25–39 (2005)
4. F.P. Boca, A. Zaharescu, Farey fractions and two-dimensional tori, in *Noncommutative Geometry and Number Theory*. Aspects Math., vol. E37 (Vieweg, Wiesbaden, 2006), pp. 57–77
5. N.D. Elkies, C.T. McMullen, Gaps in $\sqrt{n}$ mod 1 and ergodic theory. Duke Math. J. **123**, 95–139 (2004)
6. R.R. Hall, A note on Farey series. J. Lond. Math. Soc. **2**, 139–148 (1970)
7. D.A. Hejhal, On the uniform equidistribution of long closed horocycles. Asian J. Math. **4**, 839–853 (2000)
8. P.P. Kargaev, A.A. Zhigljavsky, Asymptotic distribution of the distance function to the Farey points. J. Number Theor. **65**, 130–149 (1997)

9. H. Li, Effective limit distribution of the Frobenius numbers. arXiv:1101.3021
10. J. Marklof, The n-point correlations between values of a linear form. Ergod. Theor. Dyn. Syst. **20**, 1127–1172 (2000)
11. J. Marklof, Distribution modulo one and Ratner's theorem, in *Equidistribution in Number Theory, an Introduction*. NATO Sci. Ser. II Math. Phys. Chem., vol. 237 (Springer, Dordrecht, 2007), pp. 217–244
12. J. Marklof, Horospheres and Farey fractions. Contemp. Math. **532**, 97–106 (2010)
13. J. Marklof, The asymptotic distribution of Frobenius numbers. Invent. Math. **181**, 179–207 (2010)
14. J. Marklof, A. Strömbergsson, Kinetic transport in the two-dimensional periodic Lorentz gas. Nonlinearity **21**, 1413–1422 (2008)
15. J. Marklof, A. Strömbergsson, The distribution of free path lengths in the periodic Lorentz gas and related lattice point problems. Ann. Math. **172**, 1949–2033 (2010)
16. P. Sarnak, Asymptotic behavior of periodic orbits of the horocycle flow and Eisenstein series. Comm. Pure Appl. Math. **34**, 719–739 (1981)
17. A. Strömbergsson, On the uniform equidistribution of long closed horocycles. Duke Math. J. **123**, 507–547 (2004)
18. A. Strömbergsson, On the probability of a random lattice avoiding a large convex set. Proc. Lond. Math. Soc. **103**, 950–1006 (2011)
19. A. Strömbergsson, A. Venkatesh, Small solutions to linear congruences and Hecke equidistribution. Acta Arith. **118**, 41–78 (2005)
20. D. Zagier, Eisenstein series and the Riemann zeta function, in *Automorphic Forms, Representation Theory and Arithmetic* (Bombay, 1979). Tata Inst. Fund. Res. Studies in Math., vol. 10 (Tata Inst. Fundamental Res., Bombay, 1981), pp. 275–301

Part II
Probability Theory

On the Problem of Reversibility of the Entropy Power Inequality

Sergey G. Bobkov and Mokshay M. Madiman

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract As was shown recently by the authors, the entropy power inequality can be reversed for independent summands with sufficiently concave densities, when the distributions of the summands are put in a special position. In this note it is proved that reversibility is impossible over the whole class of convex probability distributions. Related phenomena for identically distributed summands are also discussed.

Keywords Convex measures • Entropy power inequality • Log-concave • Reverse Brunn-Minkowski inequality • Rogers-Shephard inequality

2010 *Mathematics Subject Classification.* 60F05.

S.G. Bobkov (✉)
School of Mathematics, University of Minnesota, Vincent Hall 228, 206 Church St SE, Minneapolis, MN 55455, USA
e-mail: bobkov@math.umn.edu

M.M. Madiman
Department of Mathematical Sciences, University of Delaware, 501 Ewing Hall, Newark, DE 19716, USA
e-mail: mokshay.madiman@yale.edu

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_4,

1 The Reversibility Problem for the Entropy Power Inequality

Given a random vector X in $\mathbb{R}^n$ with density f, introduce the entropy functional (or Shannon's entropy)

$$h(X) = -\int_{\mathbb{R}^n} f(x) \log f(x)\, dx,$$

and the entropy power

$$H(X) = e^{2h(X)/n},$$

provided that the integral exists in the Lebesgue sense. For example, if X is uniformly distributed in a convex body $A \subset \mathbb{R}^n$, we have

$$h(X) = \log |A|, \qquad H(X) = |A|^{2/n},$$

where $|A|$ stands for the n-dimensional volume of A.

The entropy power inequality due to Shannon and Stam indicates that

$$H(X + Y) \geq H(X) + H(Y), \tag{1}$$

for any two independent random vectors X and Y in $\mathbb{R}^n$, for which the entropy is defined ([27, 28], cf. also [14, 15, 29]). This is one of the fundamental results in Information Theory, and it is of large interest to see how sharp (1) is.

The equality here is only achieved, when X and Y have normal distributions with proportional covariance matrices. Note that the right-hand side is unchanged when X and Y are replaced with affine volume-preserving transformation, that is, with random vectors

$$\tilde{X} = T_1(X), \qquad \tilde{Y} = T_2(Y) \qquad (|\det T_1| = |\det T_2| = 1). \tag{2}$$

On the other hand, the entropy power $H(\tilde{X} + \tilde{Y})$ essentially depends on the choice of T_1 and T_2. Hence, it is reasonable to consider a formally improved variant of (1),

$$\inf_{T_1, T_2} H(\tilde{X} + \tilde{Y}) \geq H(X) + H(Y), \tag{3}$$

where the infimum is running over all affine maps $T_1, T_2 : \mathbb{R}^n \to \mathbb{R}^n$ subject to (2). (Note that one of these maps may be taken to be the identity operator.) Now, equality in (3) is achieved, whenever X and Y have normal distributions with arbitrary positive definite covariance matrices.

A natural question arises: When are both the sides of (3) of a similar order? For example, within a given class of probability distributions (of X and Y), one wonders whether or not it is possible to reverse (3) to get

$$\inf_{T_1, T_2} H(\tilde{X} + \tilde{Y}) \leq C(H(X) + H(Y)) \tag{4}$$

with some constant C.

The question is highly non-trivial already for the class of uniform distributions on convex bodies, when it becomes to be equivalent (with a different constant) to the inverse Brunn-Minkowski inequality

$$\inf_{T_1,T_2} \left|\tilde{A} + \tilde{B}\right|^{1/n} \le C\left(|A|^{1/n} + |B|^{1/n}\right). \tag{5}$$

Here $\tilde{A} + \tilde{B} = \{x + y : x \in \tilde{A},\ y \in \tilde{B}\}$ stands for the Minkowski sum of the images $\tilde{A} = T_1(A)$, $\tilde{B} = T_2(B)$ of arbitrary convex bodies A and B in $\mathbb{R}^n$. To recover such an equivalence, one takes for X and Y independent random vectors uniformly distributed in A and B. Although the distribution of $X + Y$ is not uniform in $A + B$, there is a general entropy-volume relation

$$\frac{1}{4}\,|A + B|^{2/n} \le\ H(X + Y) \le\ |A + B|^{2/n},$$

which may also be applied to the images $\tilde{A}$, $\tilde{B}$ and $\tilde{X}$, $\tilde{Y}$ (cf. [3]).

The inverse Brunn-Minkowski inequality (5) is indeed true and represents a deep result in Convex Geometry discovered by V. D. Milman in the mid 1980s (cf. [21–24]). It has connections with high dimensional phenomena, and we refer an interested reader to [1, 12, 16, 17]. The questions concerning possible description of the maps T_1 and T_2 and related isotropic properties of the normalized Gaussian measures are discussed in [6].

Based on (5), and involving Berwald's inequality in the form of C. Borell [9], the inverse entropy power inequality (4) has been established recently [2,3] for the class of all probability distributions having log-concave densities. Involving additionally a general submodularity property of entropy [19], it turned out also possible to consider more general densities of the form

$$f(x) = V(x)^{-\beta}, \qquad x \in \mathbb{R}^n, \tag{6}$$

where V are positive convex functions on $\mathbb{R}^n$ and $\beta \ge n$ is a given parameter. More precisely, the following statement can be found in [3].

Theorem 1.1. *Let X and Y be independent random vectors in $\mathbb{R}^n$ with densities of the form (6) with $\beta \ge 2n + 1$, $\beta \ge \beta_0 n$ ($\beta_0 > 2$). There exist linear volume preserving maps $T_i : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$H\left(\tilde{X} + \tilde{Y}\right) \le\ C_{\beta_0}\,(H(X) + H(Y)), \tag{7}$$

where $\tilde{X} = T_1(X)$, $\tilde{Y} = T_2(Y)$, and where C_{β_0} is a constant, depending on β_0, only.

The question of what maps T_1 and T_2 can be used in Theorem 1.1 is rather interesting, but certainly the maps that put the distributions of X and Y in M-position suffice (see [3] for terminology and discussion). In a more relaxed form, one

needs to have in some sense "similar" positions for both distributions. For example, when considering identically distributed random vectors, there is no need to appeal in Theorem 1.1 to some (not very well understood) affine volume-preserving transformations, since the distributions of X and Y have the same M-ellipsoid. In other words, we have for X and Y drawn independently from the *same* distribution (under the same assumption on form of density as Theorem 1.1) that

$$H(X+Y) \le C_{\beta_0}(H(X)+H(Y)) = 2C_{\beta_0} H(X). \tag{8}$$

Since the distributions of X and $-Y$ also have the same M-ellipsoid, it is also true that

$$H(X-Y) \le C_{\beta_0}(H(X)+H(Y)) = 2C_{\beta_0} H(X). \tag{9}$$

We strengthen this observation by providing a quantitative version with explicit constants below (under, however, a convexity condition on the convolved measure). Moreover, one can give a short and relatively elementary proof of it without appealing to Theorem 1.1.

Theorem 1.2. *Let X and Y be independent identically distributed random vectors in $\mathbb{R}^n$ with finite entropy. Suppose that $X-Y$ has a probability density function of the form (6) with $\beta \ge \max\{n+1, \beta_0 n\}$ for some fixed $\beta_0 > 1$. Then*

$$H(X-Y) \le D_{\beta_0} H(X)$$

and

$$H(X+Y) \le D_{\beta_0}^2 H(X),$$

where $D_{\beta_0} = \exp(\frac{2\beta_0}{\beta_0 - 1})$.

In the special case of X and Y being log-concave, a similar quantitative result was recently obtained by [18] using a different approach.

Let us return to Theorem 1.1 and the class of distributions involved there. For growing β, the families (6) shrink and converge in the limit as $\beta \to +\infty$ to the family of log-concave densities which correspond to the class of log-concave probability measures. Through inequalities of the Brunn-Minkowski-type, the latter class was introduced by A. Prékopa [25], while the general case $\beta \ge n$ was studied by C. Borell [10, 11], cf. also [5, 13]. In [10, 11] it was shown that probability measures μ on $\mathbb{R}^n$ with densities (6) (and only they, once μ is absolutely continuous) satisfy the geometric inequality

$$\mu\big(tA + (1-t)B\big) \ge \big[\, t\mu(A)^\kappa + (1-t)\mu(B)^\kappa \,\big]^{1/\kappa} \tag{10}$$

for all $t \in (0,1)$ and for all Borel measurable sets $A, B \subset \mathbb{R}^n$, with negative power

$$\kappa = -\frac{1}{\beta - n}.$$

Such μ's form the class of so-called κ-concave measures. In this hierarchy the limit case $\beta = n$ corresponds to $\kappa = -\infty$ and describes the largest class of measures on $\mathbb{R}^n$, called *convex*, in which case (10) turns into

$$\mu(tA + (1-t)B) \geq \min\{\mu(A), \mu(B)\}.$$

This inequality is often viewed as the weakest convexity hypothesis about a given measure μ.

One may naturally wonder whether or not it is possible to relax the assumption on the range of β in (7)–(9), or even to remove any convexity hypotheses. In this note we show that this is impossible already for the class of all one-dimensional convex probability distributions. Note that in dimension one there are only two admissible linear transformations, $\tilde{X} = X$ and $\tilde{X} = -X$, so that one just wants to estimate $H(X+Y)$ or $H(X-Y)$ from above in terms of $H(X)$. As a result, the following statement demonstrates that Theorem 1.1 and its particular cases (8)–(9) are false over the full class of convex measures.

Theorem 1.3. *For any constant C, there is a convex probability distribution μ on the real line with a finite entropy, such that*

$$\min\{H(X+Y), H(X-Y)\} \geq C\, H(X),$$

where X and Y are independent random variables, distributed according to μ.

A main reason for $H(X+Y)$ and $H(X-Y)$ to be much larger than $H(X)$ is that the distributions of the sum $X+Y$ and the difference $X-Y$ may lose convexity properties, when the distribution μ of X is not "sufficiently convex". For example, in terms of the convexity parameter κ (instead of β), the hypothesis of Theorem 1.1 is equivalent to

$$\kappa \geq -\frac{1}{(\beta_0 - 1)n} \quad (\beta_0 > 2), \qquad \kappa \geq -\frac{1}{n+1}.$$

That is, for growing dimension n we require that κ be sufficiently close to zero (or the distributions of X and Y should be close to the class of log-concave measures). These conditions ensure that the convolution of μ with the uniform distribution on a proper (specific) ellipsoid remains to be convex, and its convexity parameter can be controled in terms of β_0 (a fact used in the proof of Theorem 1.1). However, even if κ is close to zero, one cannot guarantee that $X+Y$ or $X-Y$ would have convex distributions.

We prove Theorem 1.2 in Sect. 2 and Theorem 1.3 in Sect. 3, and then conclude in Sect. 4 with remarks on the relationship between Theorem 1.3 and recent results about Cramer's characterization of the normal law.

2 A "Difference Measure" Inequality for Convex Measures

Given two convex bodies A and B in $\mathbb{R}^n$, introduce $A - B = \{x - y : x \in A,\ y \in B\}$. In particular, $A - A$ is called the "difference body" of A. Note it is always symmetric about the origin.

The Rogers-Shephard inequality [26] states that, for any convex body $A \subset \mathbb{R}^n$,

$$|A - A| \ \le\ C_{2n}^n\, |A|, \tag{11}$$

where $C_n^k = \frac{n!}{k!(n-k)!}$ denote usual combinatorial coefficients. Observe that putting the Brunn-Minkowski inequality and (11) together immediately yields that

$$2 \le \frac{|A - A|^{\frac{1}{n}}}{|A|^{\frac{1}{n}}} \le \left[C_{2n}^n\right]^{\frac{1}{n}} < 4,$$

which constrains severely the volume radius of the difference body of A relative to that of A itself. In analogy to the Rogers-Shephard inequality, we ask the following question for entropy of convex measures.

Question. *Let X and Y be independent random vectors in $\mathbb{R}^n$, which are identically distributed with density $V^{-\beta}$, with V positive convex, and $\beta \ge n+\gamma$. For what range of $\gamma > 0$ is it true that $H(X - Y) \le C_\gamma H(X)$, for some constant C_γ depending only on γ?*

Theorems 1.2 and 1.3 partially answer this question. To prove the former, we need the following lemma about convex measures, proved in [4].

Lemma 2.1. *Fix $\beta_0 > 1$. Assume a random vector X in $\mathbb{R}^n$ has a density $f = V^{-\beta}$, where V is a positive convex function on the supporting set. If $\beta \ge n + 1$ and $\beta \ge \beta_0 n$, then*

$$\log \|f\|_\infty^{-1} \le h(X) \le c_{\beta_0} n + \log \|f\|_\infty^{-1}, \tag{12}$$

where one can take for the constant $c_{\beta_0} = \frac{\beta_0}{\beta_0 - 1}$.

In other words, for sufficiently convex probability measures, the entropy may be related to the L^∞-norm $\|f\|_\infty = \sup_x f(x)$ of the density f (which is necessarily finite). Observe that the left inequality in (12) is general: It trivially holds without any convexity assumption. On the other hand, the right inequality is an asymptotic version of a result from [4] about extremal role of the multidimensional Pareto distributions.

Now, let f denote the density of the random variable $W = X - Y$ in Theorem 1.2. It is symmetric (even) and thus maximized at zero, by the convexity hypothesis. Hence, by Lemma 2.1,

$$h(W) \le \log \|f\|_\infty^{-1} + c_{\beta_0} n = \log f(0)^{-1} + c_{\beta_0} n.$$

But, if p is the density of X, then $f(0) = \int_{\mathbb{R}^n} p(x)^2\,dx$, and hence

$$\log f(0)^{-1} = -\log \int_{\mathbb{R}^n} p(x) \cdot p(x)\,dx \leq \int_{\mathbb{R}^n} p(x)[-\log p(x)]\,dx$$

by using Jensen's inequality. Combining the above two displays immediately yields the first part of Theorem 1.2.

To obtain the second part, we need the following lemma on the submodularity of the entropy of sums proved in [19].

Lemma 2.2. *Given independent random vectors X, Y, Z in $\mathbb{R}^n$ with absolutely continuous distributions, we have*

$$h(X + Y + Z) + h(Z) \leq h(X + Z) + h(Y + Z),$$

provided that all entropies are well-defined and finite.

Taking X, Y and $-Z$ to be identically distributed, and using the monotonicity of entropy (after adding an independent summand), we obtain

$$h(X + Y) + h(Z) \leq h(X + Y + Z) + h(Z) \leq h(X + Z) + h(Y + Z)$$

and hence

$$h(X + Y) + h(X) \leq 2h(X - Y).$$

Combining this bound with the first part of Theorem 1.2 immediately gives the second part.

It would be more natural to state Theorem 1.2 under a shape condition on the distribution of X rather than on that of $X - Y$, but for this we need to have better understanding of the convexity parameter of the convolution of two κ-concave measures when $\kappa < 0$.

Observe that in the log-concave case of Theorem 1.2 (which is the case of $\beta \to \infty$, but can easily be directly derived in the same way without taking a limit), one can impose only a condition on the distribution of X (rather than that of $X-Y$) since closedness under convolution is guaranteed by the Prékopa-Leindler inequality.

Corollary 2.3. *Let X and Y be independent random vectors in $\mathbb{R}^n$ with log-concave densities. Then*

$$h(X - Y) \leq h(X) + n,$$
$$h(X + Y) \leq h(X) + 2n.$$

In particular, observe that putting the entropy power inequality (1) and Corollary 2.3 together immediately yields that

$$2 \le \frac{H(X-Y)}{H(X)} \le e^2,$$

which constrains severely the entropy power of the "difference measure" of μ relative to that of μ itself.

A result similar to Corollary 2.3 (but with different constants) was recently obtained in [18] using a different approach.

3 Proof of Theorem 1.3

Given a (large) parameter $b > 1$, let a random variable X_b have a truncated Pareto distribution μ, namely, with the density

$$f(x) = \frac{1}{x \log b}\, 1_{\{1<x<b\}}(x).$$

By the construction, μ is supported on a bounded interval $(1, b)$ and is convex.

First we are going to test the inequality

$$H(X_b + Y_b) \le C H(X_b) \tag{13}$$

for growing b, where Y_b is an independent copy of X_b. Note that

$$\begin{aligned} h(X_b) &= \int_1^b f(x)\, \log(x \log b)\, dx \\ &= \log\log b + \frac{1}{\log b} \int_1^b \frac{\log x}{x}\, dx \;=\; \log\log b + \frac{1}{2}\, \log b, \end{aligned}$$

so $H(X_b) = b\, \log^2 b$.

Now, let us compute the convolution of f with itself. The sum $X_b + Y_b$ takes values in the interval $(2, 2b)$. Given $2 < x < 2b$, we have

$$g(x) = (f * f)(x) = \int_{-\infty}^{+\infty} f(x-y) f(y)\, dy = \frac{1}{\log^2 b} \int_\alpha^\beta \frac{dy}{(x-y)y},$$

where the limits of integration are determined to satisfy the constraints $1 < y < b$, $1 < x - y < b$. So,

$$\alpha = \max(1, x-b), \quad \beta = \min(b, x-1),$$

and using $\frac{1}{(x-y)y} = \frac{1}{x}\,(\frac{1}{y} + \frac{1}{x-y})$, we find that

$$g(x) = \frac{1}{x \log^2 b} \left(\log(y) - \log(x-y) \right|_{x=\alpha}^{\beta} = \frac{1}{x \log^2 b} \log \frac{y}{x-y} \Big|_{x=\alpha}^{\beta}$$
$$= \frac{1}{x \log^2 b} \left(\log \frac{\beta}{x-\beta} - \log \frac{\alpha}{x-\alpha} \right).$$

Note that $x - \alpha = x - \max(1, x-b) = \min(b, x-1) = \beta$. Hence,

$$g(x) = \frac{2}{x \log^2 b} \log \frac{\beta}{\alpha} = \frac{2}{x \log^2 b} \log \frac{\min(b, x-1)}{\max(1, x-b)}.$$

Equivalently,

$$g(x) = \frac{2}{x \log^2 b} \log(x-1), \text{ for } 2 < x < b+1,$$
$$g(x) = \frac{2}{x \log^2 b} \log \frac{b}{x-b}, \text{ for } b+1 < x < 2b.$$

Now, on the second interval $b + 1 < x < 2b$, we have

$$g(x) \le \frac{2}{x \log^2 b} \log b = \frac{2}{x \log b} < \frac{2}{(b+1) \log b} < 1,$$

where the last bound holds for $b \ge e$, for example. Similarly, on the first interval $2 < x < b + 1$, using $\log(x-1) < \log b$, we get

$$g(x) \le \frac{2}{x \log b} < \frac{1}{\log b} \le 1.$$

Thus, as soon as $b \ge e$, we have $g \le 1$ on the support interval. From this,

$$h(X_b + Y_b) = \int_2^{2b} g(x) \log(1/g(x))\, dx \ge \int_2^b g(x) \log(1/g(x))\, dx.$$

Next, using on the first interval the bound $g(x) \le \frac{2}{x \log b} \le \frac{1}{x}$, valid for $b \ge e^2$, we get for such values of b that

$$h(X_b + Y_b) \ge \int_2^b g(x) \log x\, dx = \frac{2}{\log^2 b} \int_2^b \frac{\log(x-1)\, \log x}{x}\, dx.$$

To further simplify, we may write $x - 1 \ge \frac{x}{2}$, which gives

$$\int_2^b \frac{\log(x-1)\,\log x}{x}\,dx \geq \int_2^b \frac{\log^2 x}{x}\,dx - \log 2 \int_2^b \frac{\log x}{x}\,dx$$
$$= \frac{1}{3}\left(\log^3 b - \log^3 2\right) - \frac{\log 2}{2}\left(\log^2 b - \log^2 2\right)$$
$$> \frac{1}{3}\log^3 b - \frac{\log 2}{2}\log^2 b.$$

Hence, $h(X_b + Y_b) > \frac{2}{3}\log b - \log 2$, and so

$$H(X_b + Y_b) > \frac{1}{4} b^{4/3} \quad (b \geq e^2).$$

In particular,

$$\frac{H(X_b + Y_b)}{H(X_b)} > \frac{b^{1/3}}{4\log^2 b} \to +\infty, \quad \text{as } b \to +\infty.$$

Hence, the inequality (13) may not hold for large b with any prescribed value of C.

To test the second bound

$$H(X_b - Y_b) \leq C H(X_b), \tag{14}$$

one may use the previous construction. The random variable $X_b - Y_b$ can take any value in the interval $|x| < b - 1$, where it is described by the density

$$h(x) = \int_{-\infty}^{+\infty} f(x+y) f(y)\,dy = \frac{1}{\log^2 b} \int_\alpha^\beta \frac{dy}{(x+y)y}.$$

Here the limits of integration are determined to satisfy $1 < y < b$ and $1 < x + y < b$. So, assuming for simplicity that $0 < x < b - 1$, the limits are

$$\alpha = 1, \quad \beta = b - x.$$

Writing $\frac{1}{(x+y)y} = \frac{1}{x}\left(\frac{1}{y} - \frac{1}{x+y}\right)$, we find that

$$h(x) = \frac{1}{x\log^2 b}\left(\log(y) - \log(x+y)\right)\Big|_{x=\alpha}^{\beta} = \frac{1}{x\log^2 b}\,\log\frac{(b-x)(x+1)}{b}.$$

It should also be clear that

$$h(0) = \frac{1}{\log^2 b}\int_1^b \frac{dy}{y^2} = \frac{1 - \frac{1}{b}}{\log^2 b}.$$

Using $\log \frac{(b-x)(x+1)}{b} < \log(x+1) < x$, we obtain that $h(x) < \frac{1}{\log^2 b} \le 1$, for $b \ge e^2$.

In this range, since $\frac{(b-x)(x+1)}{b} < b$, we also have that $h(x) \le \frac{1}{x \log b} \le \frac{1}{x}$. Hence, in view of the symmetry of the distribution of $X_b - Y_b$,

$$\begin{aligned} h(X_b - Y_b) &= 2 \int_0^{b-1} h(x) \log(1/h(x))\, dx \\ &\ge 2 \int_0^{b/2} h(x) \log x\, dx \\ &= \frac{2}{\log^2 b} \int_2^{b/2} \frac{\log x}{x} \log \frac{(b-x)(x+1)}{b}\, dx. \end{aligned}$$

But for $0 < x < b/2$,

$$\log \frac{(b-x)(x+1)}{b} > \log \frac{x+1}{2} > \log x - \log 2,$$

so

$$\begin{aligned} h(X_b - Y_b) &> \frac{2}{\log^2 b} \int_2^{b/2} \frac{\log^2 x - \log 2 \log x}{x}\, dx \\ &= \frac{2}{\log^2 b} \left(\frac{1}{3} (\log^3(b/2) - \log^3 2) - \frac{\log 2}{2} (\log^2(b/2) - \log^2 2) \right) \\ &> \frac{2}{\log^2 b} \left(\frac{1}{3} \log^3(b/2) - \frac{1}{2} \log^2(b/2) \right) \\ &\sim \frac{2}{3} \log b. \end{aligned}$$

Therefore, like on the previous step, $H(X_b - Y_b)$ is bounded from below by a function, which is equivalent to $b^{4/3}$. Thus, for large b, the inequality (14) may not hold either.

Theorem 1.3 is proved.

4 Remarks

For a random variable X having a density, consider the entropic distance from the distribution of X to normality

$$D(X) = h(Z) - h(X),$$

where Z is a normal random variable with parameters $\mathbb{E}Z = \mathbb{E}X$, $\mathrm{Var}(Z) = \mathrm{Var}(X)$. This functional is well-defined for the class of all probability distributions on the line with finite second moment, and in general $0 \le D(X) \le +\infty$.

The entropy power inequality implies that

$$D(X+Y) \le \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\, D(X) + \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\, D(X) \le \max(D(X), D(Y)), \tag{15}$$

where $\sigma_1^2 = \mathrm{Var}(X)$, $\sigma_2^2 = \mathrm{Var}(Y)$.

In turn, if X and Y are identically distributed, then Theorem 1.3 reads as follows: For any positive constant c, there exists a convex probability measure μ on $\mathbb{R}$ with X, Y independently distributed according to μ, with

$$D(X \pm Y) \le D(X) - c.$$

This may be viewed as a strengthened variant of (15). That is, in Theorem 1.3 we needed to show that both $D(X+Y)$ and $D(X-Y)$ may be much smaller than $D(X)$ in the additive sense. In particular, $D(X)$ has to be very large when c is large. For example, in our construction of the previous section

$$\mathbb{E}X_b = \frac{b-1}{\log b}, \quad \mathbb{E}X_b^2 = \frac{b^2-1}{2\log b},$$

which yields

$$D(X_b) \sim \frac{3}{2}\log b, \quad D(X_b + Y_b) \sim \frac{4}{3}\log b,$$

as $b \to +\infty$.

In [7, 8] a slightly different question, raised by M. Kac and H. P. McKean [20] (with the desire to quantify in terms of entropy the Cramer characterization of the normal law), has been answered. Namely, it was shown that $D(X+Y)$ may be as small as we wish, while $D(X)$ is separated from zero. In the examples of [8], $D(X)$ is of order 1, while for Theorem 1.3 it was necessary to use large values for $D(X)$, arbitrarily close to infinity. In addition, the distributions in [7, 8] are not convex.

Acknowledgements Sergey G. Bobkov was supported in part by the NSF grant DMS-1106530. Mokshay M. Madiman was supported in part by the NSF CAREER grant DMS-1056996.

References

1. S. Artstein-Avidan, V. Milman, Y. Ostrover, The M-ellipsoid, symplectic capacities and volume. Comment. Math. Helv. **83**(2), 359–369 (2008)
2. S. Bobkov, M. Madiman, Dimensional behaviour of entropy and information. C. R. Acad. Sci. Paris Sér. I Math. **349**, 201–204 (2011)
3. S. Bobkov, M. Madiman, Reverse Brunn-Minkowski and reverse entropy power inequalities for convex measures. J. Funct. Anal. **262**(7), 3309–3339 (2012)
4. S. Bobkov, M. Madiman, The entropy per coordinate of a random vector is highly constrained under convexity conditions. IEEE Trans. Inform. Theor. **57**(8), 4940–4954 (2011)
5. S.G. Bobkov, Large deviations and isoperimetry over convex probability measures. Electron. J. Probab. **12**, 1072–1100 (2007)
6. S.G. Bobkov, On Milman's ellipsoids and M-position of convex bodies, in *Concentration, Functional Inequalities and Isoperimetry*. Contemporary Mathematics, vol. 545 (American Mathematical Society, Providence, 2011), pp. 23–34
7. S.G. Bobkov, G.P. Chistyakov, F. Götze, Entropic instability of Cramer's characterization of the normal law, in *Selected Works of Willem van Zwet*. Sel. Works Probab. Stat. (Springer, New York, 2012), pp. 231–242 (2012)
8. S.G. Bobkov, G.P. Chistyakov, F. Götze, Stability problems in Cramér-type characterization in case of i.i.d. summands. Teor. Veroyatnost. i Primenen. **57**(4), 701–723 (2011)
9. C. Borell, Complements of Lyapunov's inequality. Math. Ann. **205**, 323–331 (1973)
10. C. Borell, Convex measures on locally convex spaces. Ark. Math. **12**, 239–252 (1974)
11. C. Borell, Convex set functions in d-space. Period. Math. Hungar. **6**(2), 111–136 (1975)
12. J. Bourgain, B. Klartag, V.D. Milman, Symmetrization and isotropic constants of convex bodies, in *Geometric Aspects of Functional Analysis (1986/87)*. Lecture Notes in Mathematics, vol. 1850 (Springer, Berlin, 2004), pp. 101–115
13. H.J. Brascamp, E.H. Lieb, On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. J. Funct. Anal. **22**(4), 366–389 (1976)
14. M. Costa, T.M. Cover, On the similarity of the entropy power inequality and the Brunn-Minkowski inequality. IEEE Trans. Inform. Theor. **IT-30**, 837–839 (1984)
15. A. Dembo, T. Cover, J. Thomas, Information-theoretic inequalities. IEEE Trans. Inform. Theor. **37**(6), 1501–1518 (1991)
16. B. Klartag, V.D. Milman, Geometry of log-concave functions and measures. Geom. Dedicata **112**, 169–182 (2005)
17. H. Koenig, N. Tomczak-Jaegermann, Geometric inequalities for a class of exponential measures. Proc. Am. Math. Soc. **133**(4), 1213–1221 (2005)
18. M. Madiman, I. Kontoyiannis, The Ruzsa divergence for random elements in locally compact abelian groups. Preprint (2013)
19. M. Madiman, On the entropy of sums, in *Proceedings of the IEEE Information Theory Workshop*, Porto, Portugal (IEEE, New York, 2008)
20. H.P. McKean Jr., Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. Arch. Rational Mech. Anal. **21**, 343–367 (1966)
21. V.D. Milman, An inverse form of the Brunn-Minkowski inequality, with applications to the local theory of normed spaces. C. R. Acad. Sci. Paris Sér. I Math. **302**(1), 25–28 (1986)
22. V.D. Milman, Entropy point of view on some geometric inequalities. C. R. Acad. Sci. Paris Sér. I Math. **306**(14), 611–615 (1988)
23. V.D. Milman, Isomorphic symmetrizations and geometric inequalities, in *Geometric Aspects of Functional Analysis (1986/87)*. Lecture Notes in Mathematics, vol. 1317 (Springer, Berlin, 1988), pp. 107–131
24. G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics, vol. 94 (Cambridge University Press, Cambridge, 1989)

25. A. Prékopa, Logarithmic concave measures with applications to stochastic programming. Acta Sci. Math. Szeged **32**, 301–316 (1971)
26. C.A. Rogers, G.C. Shephard, The difference body of a convex body. Arch. Math. (Basel) **8**, 220–233 (1957)
27. C.E. Shannon, A mathematical theory of communication. Bell System Tech. J. **27**, 379–423, 623–656 (1948)
28. A.J. Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon. Inform. Contr. **2**, 101–112 (1959)
29. S. Szarek, D. Voiculescu, Shannon's entropy power inequality via restricted Minkowski sums, in *Geometric Aspects of Functional Analysis*. Lecture Notes in Mathematics, vol. 1745 (Springer, Berlin, 2000), pp. 257–262

On Probability Measures with Unbounded Angular Ratio

G.P. Chistyakov

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract The angular ratio is an important characteristic of probability measures on $\mathbb{Z}$ in the theory of ergodic dynamical systems. Answering the J. Rosenblatt question, we describe probability measures whose spectrum is not inside some Stolz region at 1, i.e., which have unbounded angular ratio.

Keywords Angular ratio • Characteristic functions • Ergodic dynamical system

2010 *Mathematics Subject Classification.* Primary 37-XX, 60E10; secondary 60A10.

1 Introduction

Given an ergodic dynamical system $(X, \mathcal{B}, m, \tau)$ and a probability measure μ on the integers, define $\mu f(x) = \sum_{k=-\infty}^{\infty} \mu(k) f(\tau^k x)$ for all $f \in L^1(X)$. The almost everywhere (a.e.) convergence of the convolution powers $\mu^n f(x)$ depends on properties of μ. Bellow et al. [2] showed that if μ has $m_2(\mu) = \sum_{k=-\infty}^{\infty} k^2 \mu(k)$ finite and $m_1(\mu) = \sum_{k=-\infty}^{\infty} k\mu(k) = 0$, then for all $f \in L^p(X)$, $1 < p < \infty$, $\lim_{n\to\infty} f(x)$ exists for a.e. x. However, if $m_2(\mu)$ is finite and $m_1(\mu) \neq 0$, then there exists $E \in \mathcal{B}$ such that $\limsup_{n\to\infty} \mu^n 1_E(x) = 1$ a.e. and

G.P. Chistyakov (✉)
Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany
e-mail: chistyak@math.uni-bielefeld.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_5,

$\liminf_{n\to\infty} \mu^n 1_E(x) = 0$ a.e. In the case when $m_2(\mu)$ is infinite and $m_1(\mu) = 0$, Bellow et al. [2] gave as well examples of μ for which we have divergence and other examples which show that convergence is possible. In this problem boundedness and unboundedness of the angular ratio of the probability measure μ plays an important role.

Let μ be a probability measure on $\mathbb{Z}$ and $\hat{\mu}(\lambda) := \sum_{k=-\infty}^{\infty} \lambda^k \mu(k)$, $\lambda \in \mathbb{C}$, $|\lambda| = 1$, be its characteristic function. A probability measure μ on $\mathbb{Z}$ has a bounded angular ratio, if μ is strictly aperiodic (i.e., $|\hat{\mu}(\lambda)| < 1$ if $\lambda \in \mathbb{C}$, $|\lambda| = 1$, $\lambda \neq 1$) and there exist $\varepsilon > 0$, $K < \infty$ such that if $\lambda \in \mathbb{C}$, $|\lambda| = 1$, $\lambda \neq 1$, $|\lambda - 1| < \varepsilon$, then $\frac{|\hat{\mu}(\lambda)-1|}{1-|\hat{\mu}(\lambda)|} \leq K$. That is, if μ is strictly aperiodic, then μ has a bounded angular ratio if and only if $\sup_{|\lambda|=1,\,\lambda\neq 1} \frac{|\hat{\mu}(\lambda)-1|}{1-|\hat{\mu}(\lambda)|} < \infty$. Another characterization is that μ has a bounded angular ratio if and only if $\hat{\mu}(\{\lambda \in \mathbb{C} : |\lambda| = 1\})$ is contained in some Stolz region (see Bellow et al. [1]).

Bellow et al. [2] proved that if μ is strictly aperiodic probability measure on $\mathbb{Z}$ such that $m_1(\mu)$ exists and if $m_1(\mu) \neq 0$, then μ has a unbounded angular ratio. If $m_2(\mu) < \infty$, then μ has a bounded angular ratio if and only if $m_1(\mu) = 0$. However, if $m_2(\mu) = \infty$, μ can be strictly aperiodic, $m_1(\mu)$ can exist and $m_1(\mu) = 0$, but μ fails to have a bounded angular ratio.

Moreover Bellow et al. [2] largely used probability measures on $\mathbb{Z}$ such that $\lim_{\lambda\to 1} \frac{|\hat{\mu}(\lambda)-1|}{1-|\hat{\mu}(\lambda)|} = \infty$. In this paper we study the problem of describing of probability measures with such property and with the weaker property $\limsup_{\lambda\to 1} \frac{|\hat{\mu}(\lambda)-1|}{1-|\hat{\mu}(\lambda)|} = \infty$.

2 Results

Let X be a random variable on a probability space $(\Omega, \mathcal{B}, P)$ with a distribution function F and the characteristic function $\varphi(t) := \mathbb{E}e^{itX}$, $t \in \mathbb{R}$. Consider the function

$$R_\varphi(t,\alpha) = |\mathrm{Im}\,\varphi(t)|/(1-\mathrm{Re}\,\varphi(t))^\alpha, \quad t \in \mathbb{R},$$

where the parameter $\alpha \in [1/2, 1]$. The inequality $|\varphi(t)|^2 \leq 1$, $t \in \mathbb{R}$, implies the simple estimate

$$|\mathrm{Im}\,\varphi(t)| \leq \sqrt{2(1-\mathrm{Re}\,\varphi(t))}, \quad t \in \mathbb{R},$$

and we have the upper bound $R_\varphi(t, 1/2) \leq \sqrt{2}$, $t \in \mathbb{R}$. But the function $R_\varphi(t,\alpha)$ can be unbounded as $t \to 0$ for $\alpha \in (1/2, 1]$.

Bellow et al. [2] largely used distribution functions F such that

$$R_\varphi(t,1) \to \infty, \qquad t \to 0. \tag{1}$$

Note that if $\mathbb{E}X^2 < \infty$, the Taylor formula easily implies that (1) holds if and only if $\mathbb{E}X \neq 0$. Bellow et al. [2] proved that (1) holds if $\mathbb{E}|X| < \infty$ and

$\mathbb{E}X \neq 0$ and they constructed an example of X such that $\mathbb{E}X = 0$ and (1) holds. I. V. Ostrovskii proved (oral communication) that there exists a random variable X such that $\mathbb{E}|X|^\beta < \infty$ with some $\beta \in (1,2)$, $\mathbb{E}X = 0$, and

$$\limsup_{t\to 0} R_\varphi(t,1) = \infty.$$

He conjectured that random variables X such that $\mathbb{E}|X|^\beta < \infty$ with some $\beta \in (1,2)$ and $\mathbb{E}X = 0$ satisfy the relation

$$\liminf_{t\to 0} R_\varphi(t,1) < \infty. \tag{2}$$

In this note we give necessary and sufficient conditions in order that (1) holds. In particular, our result implies the Ostrovskii conjecture.

Let X be a random variable such that $\mathbb{E}|X|^\beta < \infty$ for some $\beta \in (1,2)$ or for all $\beta \in (1,2)$. Assume that $\mathbb{E}X = 0$. We construct random variables X such that the relation $\limsup_{t\to 0} R_\varphi(t,1) = \infty$ holds and try to describe the sets, where $R_\varphi(t,1)$ tends to ∞ when $t \to 0$.

In order to formulate our first result, we introduce the following notations:

$$W_F(x) = 1 - F(x+0) + F(-x), \quad V_F(x) = 1 - F(x+0) - F(-x), \quad x \geq 0;$$

$$U_{kF}(x) = \int_{[-x,x]} u^k \, dF(u), \quad x \geq 0, \ k = 1,2,\ldots.$$

Theorem 2.1. *Let $W_F(x) > 0$, $x > 0$, and $\alpha \in (1/2, 1]$. The relation*

$$\lim_{t\to 0} R_\varphi(t,\alpha) = \infty \tag{3}$$

holds if and only if

$$A_F(x,\alpha) := \frac{|U_{1F}(x)|}{x W_F(x)^\alpha} \to \infty, \quad x \to +\infty, \tag{4}$$

$$B_F(x,\alpha) := \frac{x^{2\alpha-1}|U_{1F}(x)|}{U_{2F}(x)^\alpha} \to \infty, \quad x \to +\infty. \tag{5}$$

Remark 2.2. The relation (4) is equivalent to the following one

$$\frac{\left|\int_0^x V_F(u)\,du\right|}{x W_F(x)^\alpha} \to \infty, \quad x \to +\infty. \tag{6}$$

Indeed, integrating by parts, we can write

$$U_{1F}(x) = \int_0^x V_F(u)\,du - x V_F(x).$$

Using this formula we easily obtain the equivalence of the relations (4) and (6).

Now we discuss consequences of Theorem 2.1.

Corollary 2.3. *Let $\alpha \in (1/2, 1)$. The relation (3) holds if and only if*

$$x W_F(x)^\alpha \to 0, \qquad x^{1-2\alpha} U_{2F}(x)^\alpha \to 0 \quad as \quad x \to +\infty \tag{7}$$

and $\mathbb{E}X \neq 0$.

Let us establish the form of Theorem 2.1 in the case $\alpha = 1$ under different assumptions with respect to the random variable X. We assume in the next corollaries that the relation (3) holds for $\alpha = 1$.

Corollary 2.4. *Let there exist $\beta \in (1, 2]$ such that $\mathbb{E}|X|^\beta < +\infty$. The relation (3) holds if and only if $\mathbb{E}X \neq 0$.*

Ostrovskii's conjecture (2) immediately follows from this corollary.

Corollary 2.5. *Let X be a random variable such that $\mathbb{E}|X| < \infty$ and let $\mathbb{E}X \neq 0$. Then (3) holds.*

This corollary is one of the above Bellow et al. results [2].

Denote

$$\hat{U}_{1F}(T) = \int_{|x|>T} |x| \, dF(x), \qquad T > 0.$$

Corollary 2.6. *Assume that $W_F(x) > 0$, $x > 0$. Let $\mathbb{E}X = 0$ and let (3) hold, then the function $\hat{U}_{1F}(T)$ is slowly varying as $T \to \infty$.*

Remark 2.7. We easily conclude from Corollary 2.6 that, for the random variables X such that $\mathbb{E}X = 0$ and (3) is valid, the following estimate holds

$$W_F(x) \le c_F h(x)/(x+1), \quad x > 0, \tag{8}$$

where c_F is a positive constant and $h(x)$ is a slowly varying functions as $x \to \infty$.

Corollary 2.8. *Assume that $W_F(x) > 0$, $x > 0$. Let $\mathbb{E}X = 0$ and let the relation*

$$\liminf_{T \to +\infty} \frac{|U_{1F}(T)|}{\hat{U}_{1F}(T)} > 0 \tag{9}$$

hold. In order that (3) holds it is necessary and sufficient that the function $\hat{U}_{1F}(T)$ should be slowly varying as $T \to \infty$.

Remark 2.9. The relation (9) holds for random variables X such that $X \ge a$ a.e. with some $a \in \mathbb{R}$.

Corollary 2.10. *Let $\mathbb{E}|X| = \infty$. If the relation (3) holds, then the function $Q_{1F}(T) = \int_{-T}^{T} |x| \, dF(x)$, $T > 0$, is slowly varying as $T \to \infty$.*

Remark 2.11. From Corollary 2.10 and from the relation (4) of Theorem 2.1 we conclude that, for the distribution function of a random variable X such that $\mathbb{E}|X| = \infty$ and (3) holds, the inequality (8) is valid.

Corollary 2.12. *Let $\mathbb{E}|X| = \infty$. Let the assumption*

$$\liminf_{T\to\infty} \frac{|U_{1F}(T)|}{Q_{1F}(T)} > 0 \tag{10}$$

hold. The relation (3) is valid if and only if $Q_{1F}(T)$ is a slowly varying function as $T \to \infty$.

Remark 2.13. Random variables X such that $X \geq a$ a.e., where $a \in \mathbb{R}$, satisfy the assumption (10).

Now we give the examples of random variables X such that $\mathbb{E}|X| < \infty$, $\mathbb{E}X = 0$ and random variables X such that $\mathbb{E}|X| = \infty$ with characteristic functions satisfying (3), where $\alpha = 1$.

Let $F_0(x) = F(x, 1, \beta, \gamma, c)$ be a stable distribution function with the characteristic function of the form

$$\varphi_0(t) = \exp\left\{i\gamma t - c|t|(1 - i\frac{2}{\pi}\beta\frac{t}{|t|}\log|t|)\right\}, \quad t \in \mathbb{R}, \tag{11}$$

where $\gamma \in \mathbb{R}$, $c > 0$ and $-1 \leq \beta \leq 1, \beta \neq 0$. It is easy to see that $R_{\varphi_0}(t, 1) \to \infty$ as $t \to 0$.

Proposition 2.14. *The characteristic function of a distribution function F from the domain of attraction of the stable distribution function F_0 satisfies the relation (3) with $\alpha = 1$.*

In order to construct the examples of desired random variables X it remains to recall the following well-known result (see [4], p. 76):

In order that a distribution function $F(x)$ belongs to the domain of attraction of the stable distribution function F_0, it is necessary and sufficient that, as $|x| \to \infty$,

$$F(x) = \frac{c_1 + o(1)}{-x}h(-x), \ x < 0, \quad 1 - F(x) = 1 - \frac{c_2 + o(1)}{x}h(x), \ x > 0,$$

where the function $h(x)$ is slowly varying as $x \to +\infty$ and c_1 and c_2 are constants with $c_1, c_2 \geq 0$, $c_1 \neq c_2$.

Now let us discuss another problem connected with Theorem 2.1 in the case $\alpha = 1$.

Let X be a random variable such that $\mathbb{E}|X|^\beta < \infty$ with some $\beta \in (1, 2)$. Assume that $\mathbb{E}X = 0$. Then, by Corollary 2.4, $\liminf_{t\to 0} R_\varphi(t, 1) < \infty$. Our nearest aim is to construct random variables X such that the relation $\limsup_{t\to 0} R_\varphi(t, 1) = \infty$ holds and to try to describe the sets, where $R_\varphi(t, 1)$ tends to ∞ when $t \to 0$.

Bellow et al. [2] raised the following question. Let $\varphi(t)$ be the characteristic function of a random variable X such that $\mathbb{E}|X|^\beta < \infty$ for all $\beta \in (0,2)$ and $\mathbb{E}X = 0$. Does there exist a sequence of intervals (α_k, β_k),

$$\beta_1 > \alpha_1 > \beta_2 > \alpha_2 > \cdots > \beta_k > \alpha_k \downarrow 0,$$

such that

$$\lim_{k\to\infty} \frac{|\arg\varphi(\beta_k) - \arg\varphi(\alpha_k)|}{1 - \min_{t\in(\alpha_k,\beta_k)} |\varphi(t)|} = \infty\,?$$

What may we say about the numbers α_k and β_k?

In order to answer this question we introduce the following notations. Let $L(x) > 0$, $x \geq 1$, be a slowly varying function at infinity and $L(x) \to \infty$ as $x \to \infty$. Let $\{x_k\}_{k=1}^\infty$ be a sequence of positive numbers such that $x_k \uparrow \infty$ and

$$D_k := \frac{L(x_{k+1})}{\sum_{m=1}^k L(x_m)} \to \infty, \quad k \to \infty. \tag{12}$$

From the representation for slowly varying functions (see [3], p. 282), we note that $x_{k+1}/x_k \geq D_k^{1/\varepsilon}$ for any fixed $\varepsilon > 0$ and $k \geq k_0(\varepsilon)$, and therefore from the assumption (12) it follows $x_{k+1}/x_k \to \infty$ as $k \to \infty$.

Let $\varphi(t)$ have the form

$$\varphi(t) = p_1 e^{-it} + p_2 \sum_{k=1}^\infty \frac{L(x_k)}{x_k^2} e^{ix_k t}, \quad t \in \mathbb{R}, \tag{13}$$

where

$$p_1 := p_2 \sum_{k=1}^\infty \frac{L(x_k)}{x_k^2} \quad \text{and} \quad p_2 := \Big(\sum_{k=1}^\infty \frac{L(x_k)}{x_k} + \sum_{k=1}^\infty \frac{L(x_k)}{x_k^2} \Big)^{-1}.$$

We see that $\varphi(t)$ is the characteristic function of a random variable X such that $\mathbb{E}|X|^\beta < \infty$ for all $0 < \beta < 2$, $\mathbb{E}X^2 = \infty$ and $\mathbb{E}X = 0$. Denote $\alpha_k := 10 x_{k+1}^{-1}$ and $\beta_k := R_k x_{k+1}^{-1}$, where $R_k > 10$ and $M_{k1} := \min\{R_k, D_k/R_k\} \to \infty$ as $k \to \infty$. Note that $\alpha_k < \beta_k < \alpha_{k-1}$ for all large $k \in \mathbb{N}$.

Theorem 2.15. *Let $\varphi(t)$ be the characteristic function of the form (13). Then we have, for considered above α_k and β_k and for all sufficiently large $k \in \mathbb{N}$,*

$$\frac{|\arg\varphi(\beta_k) - \arg\varphi(\alpha_k)|}{1 - \min_{t\in(\alpha_k,\beta_k)} |\varphi(t)|} \geq \frac{1}{11} M_{k1}. \tag{14}$$

In particular, if $L(x) = \log_2 x$, then choosing $x_k = 2^{2^{k \log_2 k}}$, we have $D_k = e^{1/\log 2} k(1 + o(1))$, $\alpha_k = 10 \cdot 2^{-2^{(k+1)\log_2(k+1)}}$ and $\beta_k = R_k \alpha_k$, where $R_k \to +\infty$ and $R_k = o(k)$.

In the case when there exists $\beta \in (1,2)$ such that $\mathbb{E}|X|^{\beta'} < \infty$ for all $\beta' < \beta$, $\mathbb{E}|X|^{\beta} = \infty$ and $\mathbb{E}X = 0$ we have the following result. Let $\{x_k\}_{k=1}^{\infty}$ be a sequence of positive numbers such that $x_1 = 10$, $x_k \uparrow \infty$ and $x_{k+1}/x_k \to \infty$. Let $\varphi(t)$ have the form

$$\varphi(t) = p_3 e^{-it} + p_4 \sum_{k=1}^{\infty} x_k^{-\beta} e^{ix_k t}, \quad t \in \mathbb{R}, \tag{15}$$

where

$$p_3 := p_4 \sum_{k=1}^{\infty} x_k^{1-\beta} \quad \text{and} \quad p_4 := \Big(\sum_{k=1}^{\infty} x_k^{1-\beta} + \sum_{k=1}^{\infty} x_k^{-\beta} \Big)^{-1}.$$

Denote $\alpha_k := 10 x_{k+1}^{-1}$ and $\beta_k := N_k^{-1} x_k^{-1}$, where $N_k (x_k/x_{k+1})^{\beta-1} \to +\infty$ and $N_k x_k / x_{k+1} \to 0$ as $k \to \infty$. Note that $\alpha_k < \beta_k < \alpha_{k-1}$ for all large $k \in \mathbb{N}$. Denote $M_{k2} := \min\{N_k^{-1}(x_{k+1}/x_k), N_k (x_k/x_{k+1})^{\beta-1}\}$. It is clear that $M_{k2} \to \infty$ as $k \to \infty$.

Theorem 2.16. *Let $\varphi(t)$ be the characteristic function of the form (15). Then we have, for considered above α_k and β_k and for all sufficiently large $k \in \mathbb{N}$,*

$$\frac{|\arg \varphi(\beta_k) - \arg \varphi(\alpha_k)|}{1 - \min_{t \in (\alpha_k, \beta_k)} |\varphi(t)|} \geq \frac{1}{13} M_{k2}. \tag{16}$$

Thus, we answered the Bellow, Jones and Rosenblatt question.

3 Proof of Theorem 2.1

In the first step we shall prove the sufficiency of the conditions (4) and (5). First we estimate from below $|\mathrm{Im}\,\varphi(t)|$. We assume that $t > 0$ and use the notation $T = 1/t$. Write the formula

$$\begin{aligned} \mathrm{Im}\,\varphi(t) &= \int_{[-\pi T, \pi T]} \sin(tx)\, dF(x) + \int_{|x| > \pi T} \sin(tx)\, dF(x) \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k+1}}{(2k+1)!} U_{2k+1,F}(\pi T) + \int_{|x| > \pi T} \sin(tx)\, dF(x). \end{aligned} \tag{17}$$

From this formula we obtain the desired lower bound

$$|\mathrm{Im}\,\varphi(t)| \geq t |U_{1F}(\pi T)| - \pi e^{\pi} t^2 U_{2F}(\pi T) - W_F(\pi T). \tag{18}$$

Using the inequality $\sin^2 u \leq u^2$, $u \in \mathbb{R}$, we deduce the upper bound

$$1 - \operatorname{Re}\varphi(t) \le \frac{1}{2}t^2 U_{2F}(\pi T) + 2W_F(\pi T).$$

By the inequality $(a+b)^\alpha \le a^\alpha + b^\alpha$, $a, b \ge 0$, $0 < \alpha \le 1$, this estimate implies

$$(1 - \operatorname{Re}\varphi(t))^\alpha \le \Big(\frac{1}{2}\Big)^\alpha t^{2\alpha} U_{2F}^\alpha(\pi T) + 2^\alpha W_F^\alpha(\pi T). \tag{19}$$

We conclude from the estimates (18) and (19) that, for small $t > 0$,

$$R_\varphi(t,\alpha) \ge \Big(\frac{2^\alpha}{\pi A_F(\pi T,\alpha)} + \frac{\pi^{2\alpha-1}}{2^\alpha B_F(\pi T,\alpha)}\Big)^{-1} - 2\pi e^\pi - \frac{1}{2^\alpha} \tag{20}$$

and, by (4) and (5) we get from (20) that $R_\varphi(t,\alpha) \to \infty$ as $t \to 0$. Thus we have proved the sufficiency of the assumptions of Theorem 2.1.

Let us prove the necessity of the assumptions (4) and (5). First we prove the necessity of the assumptions (4) and (5) in the case $\alpha = 1$. We will use the relations (3), (4) and (5) for $\alpha = 1$ until a special remark.

By (3), the function $\operatorname{Im}\varphi(t)$ does not vanish for sufficiently small $t > 0$. Let for definiteness $\operatorname{Im}\varphi(t) > 0$ (otherwise we consider $\overline{\varphi(t)}$). Then, by l'Hôpital's rule, we have

$$\frac{\frac{1}{t}\int_0^t \operatorname{Im}\varphi(u)\,du}{1 - \frac{1}{t}\int_0^t \operatorname{Re}\varphi(u)\,du} \to \infty, \quad t \to 0.$$

This means that the characteristic function $g(t) = \frac{1}{t}\int_0^t \varphi(u)\,du$, $t \in \mathbb{R}$, satisfies the relation (3) as well. The function $g(t)$ is the characteristic function of a unimodal distribution function $G(x)$ with the mode 0, i.e., $G(x)$ is convex in $x < 0$ and concave in $x > 0$. Therefore the function $G(x)$ is absolutely continuous and its density $p_G(x)$ is non-decreasing in $(-\infty, 0)$ and is non-increasing in $(0, +\infty)$. It is well-known (see [5]) that

$$F(x) = G(x) - xp_G(x), \qquad x \in \mathbb{R} \setminus \{0\}, \tag{21}$$

and $xp_G(x) \to 0$ as $x \to 0$.

First we shall prove that the relations (4) and (5) hold for the functions $A_G(1/t, 1)$ and $B_G(1/t, 1)$. Then, using these relations we deduce (4) and (5) for the functions $A_F(1/t, 1)$ and $B_F(1/t, 1)$.

Lemma 3.1. *The following lower bound holds*

$$\int_{|x|>\pi T} \sin^2(tx/2)\,dG(x) \ge \frac{1}{8}W_G(\pi T), \quad t > 0.$$

Proof. Let

$$E_1 = (\pi T, 5\pi T/3] \cup \Big(\cup_{k=1}^\infty [(2k+1/3)\pi T, (2k+2-1/3)\pi T]\Big), \quad E_2 = (\pi T, \infty) \setminus E_1.$$

Since the function $p_G(x),\ x > 0$, is non-increasing, by the choice of E_1 and E_2, we have

$$\int_{E_1} p_G(x)\,dx \geq \int_{E_2} p_G(x)\,dx.$$

Since $\sin^2(tx/2) \geq 1/4$ for $x \in E_1$, we easily obtain the estimate

$$\begin{aligned}\int_{x>\pi T} \sin^2(tx/2)\,dG(x) &\geq \int_{E_1} \sin^2(tx/2)\,dG(x)\\ &\geq \frac{1}{4}\int_{E_1} dG(x) \geq \frac{1}{8}\int_{x>\pi T} dG(x) = \frac{1}{8}(1 - G(\pi T)).\end{aligned}$$

We carry out the estimate of the integral over the set $\{x < -\pi T\}$ in the same way. The lemma is proved. □

Let us find a upper bound of $|\mathrm{Im}\, g(t)|$. Write the formula (17) for $\mathrm{Im}\, g(t)$. We obtain from this formula

$$|\mathrm{Im}\, g(t)| \leq t|U_{1G}(\pi T)| + \pi e^{\pi} t^2 U_{2G}(\pi T) + W_G(\pi T), \quad t > 0. \tag{22}$$

With the help of the inequality $\sin^2 u \geq (2u/\pi)^2,\ 0 \leq u \leq \pi/2$, and Lemma 3.1 we deduce the lower bound

$$1 - \mathrm{Re}\, g(t) \geq \frac{2}{\pi^2} t^2 U_{2G}(\pi T) + \frac{1}{4} W_G(\pi T), \quad t > 0. \tag{23}$$

The inequalities (22) and (23) imply

$$R_g(t,1) \leq \Big(\frac{1}{4\pi}\frac{1}{A_G(\pi T,1)} + \frac{2}{\pi}\frac{1}{B_G(\pi T,1)}\Big)^{-1} + \frac{1}{2}\pi^3 e^{\pi} + 4.$$

Since $R_g(t,1) \to \infty$ as $t \to 0$, we obtain the following relations

$$A_G(1/t,1) \to \infty, \quad t \to 0, \tag{24}$$

$$B_G(1/t,1) \to \infty, \quad t \to 0. \tag{25}$$

Let us show that (24) and (25) imply (4) and (5).

By Remark 2.2, the relation (24) is equivalent to the following one

$$\frac{t\big|\int_0^T V_G(x)\,dx\big|}{W_G(T)} \to \infty, \quad t \to 0. \tag{26}$$

In the sequel we need an analog of some well-known result (see [3], Ch. VIII, 9).

Lemma 3.2. *If the relation (26) holds for the distribution function $G(x)$, then the function $U_G^*(T) := \int_0^T V_G(x)\,dx$ is slowly varying as $T \to +\infty$.*

Remark 3.3. The definition of a slowly varying function includes the assumption of positivity of this function for large $T > 0$. In this paper we say that a function is slowly varying, if its modulus is slowly varying.

Proof. By (26), the function $U_G^*(T)$ does not change the sign for large T. Let $a > 1$ be a constant. The following formula holds

$$\log \frac{U_G^*(aT)}{U_G^*(T)} = T \int_1^a \frac{V_G(sT)}{U_G^*(sT)}\, ds.$$

Since $|V_G(w)| \le W_G(w)$, $w \in \mathbb{R}$, we obtain from the last formula the inequality

$$\left| \log \frac{U_G^*(aT)}{U_G^*(T)} \right| \le \int_1^a \frac{sT\, W_G(sT)}{|U_G^*(sT)|}\, \frac{ds}{s}. \tag{27}$$

By (26), the integrand on the right-hand side of the inequality (27) is bounded for large T and tends to 0 as $T \to +\infty$. By Lebesgue's theorem, we conclude that $U_G^*(aT)/U_G^*(T) \to 1$ as $T \to +\infty$, as was to be proved. □

By Lemma 3.2, the function $U_G^*(T)$ is slowly varying as $T \to +\infty$. The function $U_F^*(T)$ safes the sign for large T and is slowly varying. This follows from the following formula, which we obtain with the help of (21),

$$U_F^*(T) = \int_0^T V_G(x)\, dx + \int_{-T}^T x\, dG(x) = 2U_G^*(T) - T V_G(T), \tag{28}$$

and the relation (26). In the sequel we assume for definiteness that $U_F^*(T) > 0$ and therefore $U_G^*(T) > 0$ (otherwise we would use the function $-U_F^*(T)$). Write with the help of (21)

$$W_F(T) = W_G(T) + T(p_G(T) + p_G(-T)).$$

Since the function $\tilde{p}_G(x) := p_G(x) + p_G(-x)$, $x > 0$, is non-increasing, we have

$$W_F(T) \le W_G(T) + 2W_G(T/2) \le 3W_G(T/2). \tag{29}$$

In view of (28), (29) and (26), for large T, we obtain the relation

$$\frac{U_F^*(T)}{W_F(T)} \sim \frac{U_F^*(T/2)}{W_F(T)} \ge \frac{U_F^*(T/2)}{3W_G(T/2)} = \frac{2U_G^*(T/2) - V_G(T/2)T/2}{3W_G(T/2)} \ge \frac{U_G^*(T/2)}{3W_G(T/2)}$$

from which, by (26), (4) follows. Here we used the fact that the assumptions (4) and (6) are equivalent. In addition here and in the sequel the sign $\sim$ indicates that the ratio of the two sides tends to 1.

We now deduce the relation (5). Write with the help of (4) and (28), for large T,

$$U_{1F}(T) = U_F^*(T) - TV_F(T) \sim U_F^*(T) = 2U_G^*(T) - TV_G(T) \sim 2U_G^*(T);$$

$$U_{2F}(T) \le 2\int_0^T xW_F(x)\,dx = 2\int_0^T xW_G(x)\,dx + 2\int_{-T}^T x^2\,dG(x)$$
$$\le 3U_{2G}(T) + T^2W_G(T). \tag{30}$$

For large T, we obtain from these relations the following lower bound

$$\frac{T|U_{1F}(T)|}{U_{2F}(T)} \ge \frac{1}{3}\frac{T|U_G^*(T)|}{U_{2G}(T) + T^2W_G(T)}.$$

By the relations (24) and (25) the assumption (5) holds.

Thus we have proved the necessity of the assumptions of Theorem 2.1 in the case $\alpha = 1$.

Let us prove the necessity of the assumptions of Theorem 2.1 in the case $\alpha \in (1/2, 1)$. Now we consider the relations (3), (4)–(6) for $\alpha \in (1/2, 1)$.

Remark 3.4. The assumption (4) implies that the function $U_{1T}(T)$ is slowly varying when $T \to +\infty$. Indeed, as it was noticed in Remark 2.2, (4) is equivalent to (6). Then, by Lemma 3.2, we conclude that the function $U_F^*(T)$ is slowly varying. Since $U_{1F}(T) = U_F^*(T) - TV_F(T)$, by (4) for $\alpha = 1$ the function $U_{1F}(T)$ is slowly varying as well.

By l'Hôpital's rule, we deduce from (3)

$$\frac{\frac{1}{t}\int_0^t |\mathrm{Im}\,\varphi(u)|^{1/\alpha}\,du}{\frac{1}{t}\int_0^t (1 - \mathrm{Re}\,\varphi(u))\,du} \to \infty, \quad t \to 0. \tag{31}$$

In view of the inequality $(a+b+c)^{1/\alpha} \le 3^{1/\alpha}(a^{1/\alpha} + b^{1/\alpha} + c^{1/\alpha})$, $a, b, c \ge 0$, and applying (22) to $\mathrm{Im}\,\varphi(u)$, we get the upper bound

$$\int_0^t |\mathrm{Im}\,\varphi(u)|^{1/\alpha}\,du \le \int_0^t (u|U_{1F}(\pi/u)| + \pi e^{\pi}u^2U_{2F}(\pi/u) + W_F(\pi/u))^{1/\alpha}\,du \le I_1 + I_2 + I_3$$
$$:= 3^{1/\alpha}\int_0^t u^{1/\alpha}|U_{1F}(\pi/u)|^{1/\alpha}\,du + (3\pi e^{\pi})^{1/\alpha}\int_0^t u^{2/\alpha}U_{2F}(\pi/u)^{1/\alpha}\,du$$
$$+ 3^{1/\alpha}\int_0^t W_F(\pi/u)^{1/\alpha}\,du. \tag{32}$$

Since, as it is easy to see,

$$1 - \mathrm{Re}\,\varphi(u) \ge \frac{2}{\pi^2}u^2U_{2F}(\pi/u), \quad u > 0,$$

we have, for small $t > 0$,

$$TI_2 \le \left(\frac{3}{2}\pi^3 e^\pi\right)^{1/\alpha} T \int_0^t (1 - \operatorname{Re}\varphi(u))^{1/\alpha}\, du \le \left(\frac{3}{2}\pi^3 e^\pi\right)^{1/\alpha} T \int_0^t (1 - \operatorname{Re}\varphi(u))\, du$$
$$= \left(\frac{3}{2}\pi^3 e^\pi\right)^{1/\alpha} (1 - \operatorname{Re} g(t)). \tag{33}$$

The term I_3 admits the following upper bound $TI_3 \le 3^{1/\alpha} W_F(\pi T)$. On the other hand, by the estimates (23) and (29), we have $1 - \operatorname{Re} g(2t) \ge \frac{1}{4} W_G(\pi T/2) \ge \frac{1}{12} W_F(\pi T)$. Therefore we finally deduce, for small $t > 0$,

$$TI_3 \le 12 \cdot 3^{1/\alpha}(1 - \operatorname{Re} g(2t)) \le 48 \cdot 3^{1/\alpha}(1 - \operatorname{Re} g(t)). \tag{34}$$

From (31), by (23), (32)–(34), we obtain the relation

$$\frac{T \int_0^t u^{1/\alpha} |U_{1F}(\pi/u)|^{1/\alpha}}{t^2 U_{2G}(\pi T) + W_G(\pi T)} \to \infty, \quad t \to 0. \tag{35}$$

Note that the relation (3) with $\alpha \in (1/2, 1)$ implies (3) with $\alpha = 1$. Therefore, by Remark 3.4, the function $|U_{1F}(\pi T)|$ is slowly varying as $T \to +\infty$. Now we use the following well-known lemma [6]

Lemma 3.5. *Let a function $h(x)$ be slowly varying as $x \to +\infty$. Then*

$$\lim_{x \to +\infty} x^{1+\gamma} \int_x^\infty u^{-2-\gamma} \frac{h(u)}{h(x)}\, du = \frac{1}{1+\gamma} \quad \text{and} \quad \lim_{T \to +\infty} \frac{1}{T} \int_0^T \frac{h(x)}{h(T)}\, dx = 1,$$

where the parameter $\gamma \ge 0$.

Since the function $|U_{1F}(\pi T)|^{1/\alpha}$ is slowly varying as $T \to +\infty$, applying to it Lemma 3.5 with $\gamma = 1/\alpha$, we obtain the relation

$$T \int_0^t u^{1/\alpha} |U_{1F}(\pi/u)|^{1/\alpha}\, du \sim (1 + 1/\alpha)^{-1} t^{1/\alpha} |U_{1F}(\pi T)|^{1/\alpha}, \quad t \to 0.$$

Therefore the relation (35) is equivalent to the following two relations

$$\frac{t^{1/\alpha} |U_{1F}(\pi T)|^{1/\alpha}}{W_G(\pi T)} \to \infty, \quad t \to 0, \tag{36}$$

$$\frac{t^{-2+1/\alpha} |U_{1F}(\pi T)|^{1/\alpha}}{U_{2G}(\pi T)} \to \infty, \quad t \to 0. \tag{37}$$

Since $|U_{1F}(\pi T/2)|^{1/\alpha} \sim |U_{1F}(\pi T)|^{1/\alpha}$ as $T \to \infty$, the assumption (4) follows from (36) with the help of the inequality (29). The assumption (5) follows from

(37) with the help of the estimates (29), (30) and the relation (36). Theorem 2.1 is completely proved.

4 Discussion of Theorem 2.1

Proof of Corollary 2.3. It is obvious that the assertion of Corollary 2.3 holds in the case $W_F(x) = 0$ for $x > a > 0$.

Let $W_F(x) > 0$ for all $x > 0$. Note that if the assumptions (7) hold and $\mathbb{E}(X) \neq 0$, then the relations (4) and (5) hold as well and, by Theorem 2.1, the relation (3) is valid.

Let the relation (3) hold. By Theorem 2.1, the assumptions (4) and (5) hold. It follows from (4) and Remark 3.4 that $|U_{1F}(x)|$ is slowly varying function. Therefore we obtain from (4) the following upper bound, for sufficiently large $x \geq x_\varepsilon > 0$,

$$W_F(x) \leq x^{\varepsilon - 1/\alpha},$$

with some $0 < \varepsilon < (1-\alpha)/(2\alpha)$. This estimate implies that $\mathbb{E}|X|^{-2\varepsilon+1/\alpha} < \infty$ and therefore $\mathbb{E}|X| < \infty$. Let us show that $\mathbb{E}X \neq 0$. Assuming to the contrary $\mathbb{E}X = 0$, we would get

$$U_{1F}(T) = -\int_{|x|>T} x\, dF(x) \tag{38}$$

and we would have the estimate

$$|U_{1F}(T)| \leq t^{-1-2\varepsilon+1/\alpha}\mathbb{E}|X|^{-2\varepsilon+1/\alpha}, \quad t > 0.$$

It follows from this estimate that $|U_{1F}(x)|$ is not a slowly varying function (see [3], p. 277), a contradiction. Thus, $\mathbb{E}X \neq 0$. But then the assumptions (7) follow from the relations (4) and (5). Corollary 2.3 is completely proved. □

Proof of Corollary 2.4. First consider the case where $W_F(x) > 0$ for all $x > 0$. We verify the sufficiency of the assumption $\mathbb{E}X \neq 0$. Since $\mathbb{E}|X|^\beta < \infty$, we have $W_F(x) \leq x^{-\beta}\mathbb{E}|X|^\beta$, $x \geq 1$, and then it is easy to see under the assumption $\mathbb{E}X \neq 0$ that (4) holds. In addition, for considered random variables X the inequality

$$U_{2F}(T) \leq T^{2-\beta}\mathbb{E}|X|^\beta, \quad T > 0,$$

holds, therefore (5) holds as well.

Now let us prove the necessity of the assumption $\mathbb{E}X \neq 0$. If $\mathbb{E}X = 0$ then (38) holds and the upper bound

$$|U_{1F}(T)| \leq t^{\beta-1}\mathbb{E}|X|^\beta, \quad t > 0, \tag{39}$$

is true. By Remark 3.4, the function $U_{1F}(T)$ is slowly varying as $T \to +\infty$. This contradicts to the estimate (39). Hence the assumption $\mathbb{E}X = 0$ is false.

The proof of the corollary in the case $W_F(x) = 0$ for $|x| > a$ with some $a > 0$ easily follows from Taylor formula for the functions $\operatorname{Im}\varphi(t)$ and $1 - \operatorname{Re}\varphi(t)$. □

Proof of Corollary 2.5. Without loss of generality we assume that $W_F(x) > 0$, $x > 0$. If $\mathbb{E}|X| < \infty$, we have $W_F(x) = o(1/x)$, $x \to +\infty$. Using this relation and the assumption $\mathbb{E}X \neq 0$, it is easy to see that (4) holds. We have, for every fixed $N > 1$,

$$U_{2F}(T) = \int_{|x|<T/N} x^2\,dF(x) + \int_{T/N\le|x|\le T} x^2\,dF(x) \le \frac{T\mathbb{E}|X|}{N} + T\int_{|x|\ge T/N} |x|\,dF(x).$$

Therefore $tU_{2F}(T) \to 0$, as $t \to 0$, and (5) holds, if $\mathbb{E}X \neq 0$. Then, by Theorem 2.1, the relation (3) holds. □

Proof of Corollary 2.6. Let $\mathbb{E}X = 0$ and let (3) hold. Then, by Theorem 2.1, the relation (4) is true and we obtain the relation

$$t\hat{U}_{1F}(T)/W_F(T) \to \infty, \quad t \to 0.$$

Therefore the assertion of the corollary follows from the following well-known result (see [3], p. 281).

Lemma 4.1. *Let $W_F(x) > 0$, $x > 0$, and the following relation holds*

$$xW_F(x)/Z_F(x) \to 0, \quad x \to +\infty, \quad \textit{where} \quad Z_F(x) = \int_{u>x} W_F(u)\,du.$$

Then the function $Z_F(x)$ is slowly varying for $x \to +\infty$.

□

Proof of Corollary 2.8. The necessity of the assumption of the corollary follows from Corollary 2.6. Let us prove the sufficiency of this assumption. Note that the following formula

$$\frac{W_F(T)}{t\hat{U}_{1F}(T)} = 1 - T\int_T^\infty \frac{1}{x^2}\frac{\hat{U}_{1F}(x)}{\hat{U}_{1F}(T)}\,dx$$

holds. Since the function $\hat{U}_{1F}(x)$ is slowly varying, the right hand-side of this equality tends to zero as $T \to +\infty$. This follows from Lemma 3.5.

Taking into account the condition (9) we see that the relation (4) holds. In addition we note that

$$\frac{|U_{1F}(T)|}{\hat{U}_{1F}(T)} \cdot \frac{1}{B_F(t)} = \left|1 - t\int_0^T \frac{\hat{U}_{1F}(x)}{\hat{U}_{1F}(T)}\,dx\right|.$$

The right hand-side of this equality tends to 0 as $T \to +\infty$, by Lemma 3.5. Therefore, in view of (9), we obtain (4). Then, by Theorem 2.1, the property (3) holds. □

Proof of Corollary 2.10. Since $\mathbb{E}|X| = \infty$, we have $W_F(x) > 0$, $x > 0$. Let (3) hold. Then, by Theorem 2.1, the relation (4) holds and therefore, by (10), the relation

$$tQ_{1F}(T)/W_F(T) \to \infty, \quad t \to 0 \tag{40}$$

is true as well. The assertion of the corollary follows immediately from the following result (see [3], p. 281).

Lemma 4.2. *Let $X \geq 0$ almost surely. If the relation (40) holds, the function $Q_{1F}(T)$ is slowly varying for $T \to +\infty$.* □

□

Proof of Corollary 2.12. One can obtain a proof of this corollary in the same way as the proof of Corollary 2.8. Therefore we omit it. □

Proof of Proposition 2.14. The assertion of Corollary 2.14 immediately follows from the following Ibragimov and Linnik result [4], p. 85.

Theorem 4.3. *In order that the distribution with characteristic function $\varphi(t)$ belongs to the domain of attraction of the stable law whose characteristic function has the form (11), it is necessary and sufficient that, in the neighborhood of the origin,*

$$\varphi(t) = \exp\left\{i\gamma t - c|t|\tilde{h}(t)(1 - i\frac{2}{\pi}\beta\frac{t}{|t|}\log|t|)\right\},$$

where $\tilde{h}(t)$ is slowly varying as $t \to 0$.

□

5 Proof of Theorems 2.15 and 2.16

In order to prove Theorem 2.15 we need the following lemma.

Lemma 5.1. *Let $\varphi(t)$ be the characteristic function of the form (13) of a random variable X with a distribution function F. Then the inequalities*

$$R_\varphi(\beta_k, 1) \geq \left(\frac{2}{\pi A_F(\pi/\beta_k, 1)} + \frac{\pi}{2B_F(\pi/\beta_k, 1)}\right)^{-1} - 2\pi e^\pi - \frac{1}{2} \geq \frac{1}{5}M_{k1} - 3\pi e^\pi \tag{41}$$

hold for β_k from Theorem 2.15 and for all sufficiently large $k \in \mathbb{N}$.

Proof. The first of the inequalities (41) follows from (20) with $\alpha = 1$. By the definition of F and the assumption (12), we easily obtain the following estimates

$$\int_{\pi/\beta_k}^{\infty} x\,dF(x) \ge p_2 L(x_{k+1}) x_{k+1}^{-1}, \quad W_F(\pi/\beta_k) \le 2p_2 L(x_{k+1}) x_{k+1}^{-2} \quad \text{and}$$

$$\beta_k \int_{-1}^{\pi/\beta_k} x^2\,dF(x) \le \beta_k p_2 \sum_{m=1}^{k} L(x_m) + \beta_k.$$

These bounds imply, for $k \ge k_0$,

$$\Big(\frac{2}{\pi A_F(\pi/\beta_k,1)} + \frac{\pi}{2B_F(\pi/\beta_k,1)}\Big)^{-1} = \frac{\int_{\pi/\beta_k}^{\infty} x\,dF(x)}{\frac{1}{2}\beta_k \int_{-1}^{\pi/\beta_k} x^2\,dF(x) + \frac{2}{\beta_k} W_F(\pi/\beta_k)}$$

$$\ge p_2 \frac{L(x_{k+1}) x_{k+1}^{-1}}{\frac{1}{2}p_2\beta_k \sum_{m=1}^{k} L(x_m) + \beta_k + \frac{2}{\beta_k} 2p_2 L(x_{k+1}) x_{k+1}^{-2}} \ge \frac{1}{5} M_{k1},$$

as was to be proved. The lemma is proved. □

Proof of Theorem 2.15. Let $\varphi(t)$ be the characteristic function of the form (13) of a random variable X with a distribution function F. Note that the relation

$$\arg\varphi(t) = \arctan\frac{\operatorname{Im}\varphi(t)}{\operatorname{Re}\varphi(t)} = \frac{\operatorname{Im}\varphi(t)}{\operatorname{Re}\varphi(t)} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \Big(\frac{\operatorname{Im}\varphi(t)}{\operatorname{Re}\varphi(t)}\Big)^{2n}$$

holds for small t. Since

$$1 - \min_{t\in(\alpha_k,\beta_k)} |\varphi(t)| \le 1 - \min_{t\in(\alpha_k,\beta_k)} \operatorname{Re}\varphi(t)$$

and

$$\frac{\operatorname{Im}\varphi(t)}{\operatorname{Re}\varphi(t)} = \operatorname{Im}\varphi(t)\big(1 + (1-\operatorname{Re}\varphi(t)) + (1-\operatorname{Re}\varphi(t))^2 + \dots\big),$$

we have

$$\frac{|\arg\varphi(\beta_k) - \arg\varphi(\alpha_k)|}{1 - \min_{t\in(\alpha_k,\beta_k)} |\varphi(t)|} \ge \frac{|\operatorname{Im}\varphi(\beta_k) - \operatorname{Im}\varphi(\alpha_k)|}{1 - \min_{t\in(\alpha_k,\beta_k)} \operatorname{Re}\varphi(t)} - c, \tag{42}$$

where $c > 0$ is an absolute constant. In the sequel we denote all such constants by c. Now we note that

$$1 - \min_{t\in(\alpha_k,\beta_k)} \operatorname{Re}\varphi(t) = \max_{t\in(\alpha_k,\beta_k)} (1 - \operatorname{Re}\varphi(t)) \le \max_{t\in(\alpha_k,\beta_k)} \Big(\frac{t^2}{2}\int_{-1}^{\pi/t} x^2\,dF(x) + 2(1-F(\pi/t))\Big)$$

$$\le \frac{1}{2}\beta_k^2 \int_{-1}^{\pi/\alpha_k} x^2\,dF(x) + 2(1 - F(\pi/\beta_k)). \tag{43}$$

With the help of the formula

$$\operatorname{Im}\varphi(\beta_k) - \operatorname{Im}\varphi(\alpha_k) = 2\int_{-1}^{\infty} \sin\Big(\frac{\beta_k - \alpha_k}{2}x\Big)\cos\Big(\frac{\beta_k + \alpha_k}{2}x\Big)\,dF(x),$$

we obtain the lower bound

$$\begin{aligned}|\operatorname{Im}\varphi(\beta_k)-\operatorname{Im}\varphi(\alpha_k)| &= \Big|2\int_{-1}^{\infty}\sin\Big(\frac{\beta_k-\alpha_k}{2}x\Big)\,dF(x)\\ &\quad -4\int_{-1}^{\infty}\sin^2\Big(\frac{\beta_k+\alpha_k}{2}x\Big)\sin\Big(\frac{\beta_k-\alpha_k}{2}x\Big)\,dF(x)\Big|\\ &\geq (\beta_k-\alpha_k)\int_{-1}^{\pi/(\beta_k-\alpha_k)} x\,dF(x)-c(\beta_k-\alpha_k)^2\int_{-1}^{\pi/(\beta_k-\alpha_k)}x^2\,dF(x)\\ &\quad -(\beta_k+\alpha_k)^2\int_{-1}^{\pi/(\beta_k-\alpha_k)}x^2\,dF(x)-6(1-F(\pi/(\beta_k-\alpha_k))).\end{aligned}\tag{44}$$

We deduce from (43) and (44) that

$$\frac{|\operatorname{Im}\varphi(\beta_k)-\operatorname{Im}\varphi(\alpha_k)|}{1-\min_{t\in(\alpha_k,\beta_k)}\operatorname{Re}\varphi(t)}\geq(\beta_k-\alpha_k)\frac{\int_{\pi/(\beta_k-\alpha_k)}^{\infty}x\,dF(x)}{\frac{1}{2}\beta_k^2\int_{-1}^{\pi/\alpha_k}x^2\,dF(x)+2(1-F(\pi/\beta_k))}-c.\tag{45}$$

It is easy to see that, for the random variable X with a distribution function F satisfying the assumptions of Theorem 2.15, the following relations hold

$$\begin{aligned}(\beta_k-\alpha_k)\int_{\pi/(\beta_k-\alpha_k)}^{\infty}x\,dF(x)&\geq\frac{1}{2}\beta_k\int_{\pi/\beta_k}^{\infty}x\,dF(x),\\ \frac{1}{2}\beta_k^2\int_{-1}^{\pi/\alpha_k}x^2\,dF(x)&=\frac{1}{2}\beta_k^2\int_{-1}^{\pi/\beta_k}x^2\,dF(x).\end{aligned}\tag{46}$$

Using (45) and (46), we finally obtain

$$\begin{aligned}\frac{|\operatorname{Im}\varphi(\beta_k)-\operatorname{Im}\varphi(\alpha_k)|}{1-\min_{t\in(\alpha_k,\beta_k)}\operatorname{Re}\varphi(t)}&\geq\frac{1}{2}\frac{\beta_k\int_{\pi/\beta_k}^{\infty}x\,dF(x)}{\frac{1}{2}\beta_k^2\int_{-1}^{\pi/\beta_k}x^2\,dF(x)+2(1-F(\pi/\beta_k))}-c\\ &=\frac{1}{2}\Big(\frac{2}{\pi A_F(\pi/\beta_k,1)}+\frac{\pi}{2B_F(\pi/\beta_k,1)}\Big)^{-1}-c.\end{aligned}\tag{47}$$

The statement of the theorem follows from (41), (42) and (47). □

Now we need the following lemma.

Lemma 5.2. *Let $\varphi(t)$ be the characteristic function of the form (15) of a random variable X with a distribution function F. Then the inequalities*

$$R_\varphi(\beta_k,1)\geq\Big(\frac{2}{\pi A_F(\pi/\beta_k,1)}+\frac{\pi}{2B_F(\pi/\beta_k,1)}\Big)^{-1}-2\pi e^{\pi}-\frac{1}{2}\geq\frac{1}{6}M_{k2}-3\pi e^{\pi}\tag{48}$$

hold for β_k from Theorem 2.16 and for all sufficiently large $k\in\mathbb{N}$.

Proof. Let us prove (48). The first of the inequalities (48) follows from the bound (20) with $\alpha = 1$. By the definition of F, we easily obtain the following estimates

$$\int_{\pi/\beta_k}^{\infty} x\,dF(x) \geq p_4 x_{k+1}^{1-\beta}, \quad W_F(\pi/\beta_k) \leq 2p_4 x_{k+1}^{-\beta} \quad \text{and}$$

$$\beta_k \int_{-1}^{\pi/\beta_k} x^2\,dF(x) \leq 2p_4\beta_k x_k^{2-\beta} + p_3\beta_k.$$

These bounds imply

$$\Big(\frac{2}{\pi A_F(\pi/\beta_k,1)} + \frac{\pi}{2B_F(\pi/\beta_k,1)}\Big)^{-1} = \frac{\int_{\pi/\beta_k}^{\infty} x\,dF(x)}{\frac{1}{2}\beta_k \int_{-1}^{\pi/\beta_k} x^2\,dF(x) + \frac{2}{\beta_k} W_F(\pi/\beta_k)}$$

$$\geq p_4 \frac{x_{k+1}^{1-\beta}}{p_4\beta_k x_k^{2-\beta} + \frac{1}{2}p_3\beta_k + p_4\frac{4}{\beta_k}x_{k+1}^{-\beta}} \geq \frac{1}{6}M_{k2},$$

and we arrive at the assertion of the lemma. □

Proof of Theorem 2.16. We prove this theorem, using Lemma 5.2 and repeating the arguments of the proof of Theorem 2.15. □

Acknowledgements Research supported by SFB 701.

References

1. A. Bellow, R. Jones, J. Rosenblatt, *Almost Everywhere Convergence of Powers.* Almost everywhere convergence (Columbus, OH, 1988) (Academic, Boston, 1989), pp. 99–120
2. A. Bellow, R. Jones, J. Rosenblatt, Almost everywhere convergence of convolution powers. Ergod. Theor. Dyn. Syst. **14**(3), 415–432 (1994)
3. W. Feller, *An Introduction to Probability Theory and Its Applications.* vol. 2 (Wiley, New York, 1970)
4. I. Ibragimov, Yu. Linnik, *Independent and Stationary Sequences of Random Variables* (Wolters-Noordhoff publishing Groningen, The Netherlands, 1971)
5. E. Lukacs, *Characteristic Functions* (Grifin, London, 1970)
6. E. Seneta, *Regularly Varying Functions.* Lecture Notes in Mathematics, vol. 508 (Springer, Berlin, 1976)

CLT for Stationary Normal Markov Chains via Generalized Coboundaries

Mikhail Gordin

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stationary Markov chain with a stationary probability distribution μ on the state space of X and the transition operator $Q : L_2(\mu) \to L_2(\mu)$. Let $f \in L_2(\mu)$ be a function on the state space of X. The solvability in $L_2(\mu)$ of the *Poisson equation* $f = g - Qg$ implies that the stationary sequence $(f(X_n))_{n\in\mathbb{Z}}$ can be represented in the form

$$f(X_n) = \big(g(X_{n+1}) - (Qg)(X_n)\big) + \big(g(X_n) - g(X_{n+1})\big) = \eta_n + \zeta_n \ (n \in \mathbb{Z}).$$

Here $\eta = (\eta_n)_{n\in\mathbb{Z}}$ is a stationary sequence of square integrable *martingale differences*, and $\zeta = (\zeta_n)_{n\in\mathbb{Z}}$ is an L_2-*coboundary* that is a difference of two consecutive elements of a stationary sequence of square integrable random variables. This representation reduces the Central Limit Theorem (CLT) question for $(f(X_n))_{n\in\mathbb{Z}}$ to the well-studied case of martingale differences. However, in many situations the martingale approximation as a tool in limit theorems works well, though the above martingale-coboundary representation does not hold. In particular, if the transition operator Q is *normal* in $L_2(\mu)$, 1 is a simple eigenvalue of Q, and the assumptions

M. Gordin (✉)
St. Petersburg Division, V.A. Steklov Institute of Mathematics, 27 Fontanka emb., 191023 Saint Petersburg, Russia

Faculty of Mathematics and Mechanics, Saint Petersburg State University, Saint Petersburg, Russia

e-mail: gordin@pdmi.ras.ru

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_6,

(1) $\sigma_f^2 = \int_D \frac{1-|z|^2}{|1-z|^2}\rho_f\, dz < \infty$,
(2) $\lim_{n\to\infty} n^{-\frac{1}{2}} |\sum_{k=0}^{n-1} Q^k f|_2 = 0$

hold true for a real-valued function $f \in L_2(\mu)$, the Central Limit Theorem for $(f(X_n))_{n\in\mathbb{Z}}$ was established via the martingale approximation.

In the present paper we show that under condition (1) $(f(X_n))_{n\in\mathbb{Z}}$ admits a generalized form of the martingale-coboundary representation as the sum of a square integrable stationary martingale difference and a *generalized coboundary*. The latter is a stationary sequence of random variables which are increments of a stationary sequence of *m-functions* introduced in the paper. Furthermore, it turns out that assumption (2) means exactly that the generalized coboundary can be neglected in the limit. Connection with generalized solutions to the Poisson equation is also studied.

Keywords Generalized coboundary • Limit theorems • Markov chain • Martingale approximation • Normal transition operator • Poisson equation

2010 *Mathematics Subject Classification.* Primary 60F05, 60J05.

1 Introduction

Let $\eta = (\eta_n)_{n\in\mathbb{Z}}$ be a stationary (in the strict sense) sequence of integrable random variables. Assume also that η is a sequence of *martingale differences* that is for every n

$$\mathbb{E}(\eta_n|\eta_{n-1}, \eta_{n-2}, \dots) = 0.$$

Let, moreover, η be ergodic, real-valued and $\mathbb{E}\eta_n^2 < \infty$. Then, according to the classical results of Billingsley [1] and Ibragimov [11], the sequence η satisfies the Central Limit Theorem (CLT). It is known that under the same conditions also the Functional Central Limit Theorem (FCLT) and the Law of the Iterated Logarithm (including its functional form due to Strassen) are valid. Under appropriate assumptions some of these results extend to not necessarily stationary sequences or arrays of martingale differences.

A natural idea is to use a certain approximation by martingales (that is the sums of martingale differences) to establish limit theorems of the above-mentioned type for the sums of dependent random variables more general than martingale differences. More precisely, one needs to construct a martingale difference approximation of the random sequence in question and represent the error of this approximation in a form which allows us, under the appropriate normalization, to neglect by this error in the limit. In the stationary setup an approach to this problem was proposed in [7]

basing on the so-called martingale-coboundary representation. The latter means that a stationary sequence $\xi = (\xi_n)_{n\in\mathbb{Z}}$ admits the representation

$$\xi_n = \eta_n + \zeta_n, n \in \mathbb{Z}, \tag{1}$$

where $\eta = (\eta_n)_{n\in\mathbb{Z}}$ is a stationary sequence of *martingale differences*, and $\zeta = (\zeta_n)_{n\in\mathbb{Z}}$ is a so-called *coboundary* which can be written as

$$\zeta_n = \theta_n - \theta_{n-1}, n \in \mathbb{Z}, \tag{2}$$

with a certain stationary sequence $\theta = (\theta_n)_{n\in\mathbb{Z}}$. One says in this case that ζ *is a coboundary of* θ; we speak of a B*-coboundary* if each of θ_n in the above representation belongs to a certain Banach space B of random variables. It is assumed that the random sequences ξ, η, θ in this representation are defined on a common probability space so that they are jointly stationary. A convenient way to formulate this type of interrelation between random sequences (which does not lead to any loss of generality) is to assume that a probability preserving invertible map T acts on the basic probability space so that

$$\zeta_{n+1} = \zeta_n \circ T,\ \eta_{n+1} = \eta_n \circ T,\ \theta_{n+1} = \theta_n \circ T\ (n \in \mathbb{Z}).$$

As to the asymptotic distributions of the sums

$$\sum_{k=0}^{n-1} \xi_k, n \geq 1, \tag{3}$$

normalized by dividing by positive reals tending to ∞, it is clear that one can neglect by the contribution of the sequence ζ into these sums and extend to $\xi = \eta + \zeta$ certain limit theorems originally known to hold for the martingale difference η. To deduce the martingale-coboundary representation for a stationary sequence ξ, some conditions need to be imposed on ξ. These conditions are usually stated in terms of a *compatible filtration* $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ (that is a family of sub-σ-fields satisfying $\cdots \subseteq \mathcal{F}_{n-1} \subseteq \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \cdots$ and $T^{-1}\mathcal{F}_n = \mathcal{F}_{n+1}$) on the basic probability space. Specifying such a filtration is a standard prerequisite to develop the martingale approximation for stationary sequences. Given such a filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$, we need to distinguish between a general *non-adapted* sequence and an *adapted* sequence $\xi = (\xi_n)$ where ξ_n is $\mathcal{F}_n$-measurable for every $n \in \mathbb{Z}$. The latter case can be treated easier and is equivalent to the study of functions of a stationary Markov chain with a general measurable state space. Though only very special Markov chains emerge in this context, and, on the other hand, both adapted and non-adapted cases can be studied, basing on the martingale-coboundary representation and without any reference to markovianity [7], it is the whole class of general Markov chains where the application of the martingal-coboundary decomposition can be done in a very natural, simple and elegant way in terms of a condition related to the so-called

Poisson equation. More specifically, let $X = (X_n)_{n\in\mathbb{Z}}$, μ and $Q : L_2(\mu) \to L_2(\mu)$ be, respectively, a stationary Markov chain, its stationary probability distribution and its transition operator. Let $f \in L_2(\mu)$ be a function on the state space of X. Then the solvability in $g \in L_2$ of the *Poisson equation*

$$f = g - Qg \tag{4}$$

implies the applicability of the above mentioned martingale-coboundary representation to the stationary sequence $(f(X_n))_{n\in\mathbb{Z}}$ (see Abstract for an explanation; notice that the converse is also true). Moreover, in this case η turns out to be an L_2-*coboundary* (that is a coboundary of a square integrable sequence θ). In the context of limit theorems this was independently observed in [13] (for the particular case of Harris recurrent chains) and in [8] (for the general ergodic case).

Now we will explain the topic of the present paper. It was recognized during the last three decades that the martingale approximation as a tool in limit theorems is still effective in an area where the martingale-coboundary representation with an L_2-coboundary not always holds. This means that, under some assumptions, ζ in representation (1) needs not be an L_2-coboundary to make a contribution to (3) which is, having being divided by $\sqrt{n}$, negligible in the limit. The first CLT result of such kind was obtained for stationary Markov chains with *normal* transition operators [2,9] (recall that a bounded operator in a Hilbert space is said to be *normal* if it commutes with its adjoint). More specifically, let the chain $(X_n)_{n\in\mathbb{Z}}$ introduced above have a normal transition operator Q in $L_2(\mu)$ (we call such a chain *normal*, too). Assuming that 1 is a simple eigenvalue of Q and, for an $f \in L_2(\mu)$, the equation

$$f = (I - Q)^{1/2} g \tag{5}$$

(called the *fractional Poisson equation of order* $1/2$ [3]) has a solution $g \in L_2$, the CLT holds for $(f(X_n))$. Independently, under the same condition the CLT and the FCLT for stationary Markov chains with *selfadjoint* transition operators were established in [12]. Moreover, in the normal case the most general known condition for the CLT to hold was proposed in [10]. This compound condition consists of two assumptions which appear in Theorem 4.1 of the present paper as (1) and (2).

Later the CLT [14] and the FCLT [15] (see also [16] for an alternative proof) were established for stationary Markov chains with not necessarily normal transition operators under a ceratin hypothesis we call the *Maxwell-Woodroofe condition*. This condition (which we just mention without further discussion in the present paper) is stronger than the requirement that (5) is solvable in L_2, but is less restrictive than the assumption of the L_2-solvability of (4). These results were achieved by means of the martingale approximation based on relation (1). Obtaining bounds for the sequences ξ and ζ in (1) is somewhat tricky, especially in proofs of the FCLT. This impressive development, however, left open certain important questions, some part of which will be touched in the present paper under the assumption of normality. In our opinion, the key problem here is finding a suitable extension to a more general setup of the known relation between the Poisson equation and the martingale-coboundary

representation. This could clarify the structure of the sequence ζ and should be helpful, in particular, when one needs to show that this sequence is negligible. One may expect that the fractional Poisson equation (5) plays an important role in this problem. However, some known facts show that the relation between the solvability of (5) and the applicability of limit theorems is not so simple. For example, even for selfadjoint transition operators a natural fractional modification of the martingale-coboundary representation in general does not hold under the assumption that (5) is solvable in L_2 (see [4] for a counterexample). Further, as we mentioned above, for a function f of a normal Markov chain to satisfy the CLT, a weaker condition than the L_2-solvability of (5) is known (see [10] and the present paper). Moreover, without the assumption of normality the solvability of (5) in L_2 does no longer imply that the variances of sums (3) grow linearly in n [19].

These and other facts stimulate attempts to find a more precise substitute for the Poisson condition in the context of the CLT and other limit theorems. In this paper we analyze further the compound condition used in [10] to deduce the CLT for normal Markov chains. To make shorter the discussion of our approach, we only deal in this Introduction with those functions on the state space which are *completely nondeterministic*. It is this class of functions to which the study of the general case will be reduced at the cost of a certain additional assumption. We generalize the known relation between the Poisson equation (of degree one) and the martingale-coboundary representation. This is achieved by extending the class of admissible solutions of the Poisson equation along with extending the class of possible ingredients of the martingale-coboundary representation. Notice that commonly in the first case we deal with functions defined on the state space of a Markov chain, while in the second one we deal with functions on its path space which forms our basic probability space. Correspondingly, we are led to two kinds of extensions of the related L_2-spaces. We call their elements *t-functions* and *m-functions*, respectively (in general, they are not functions at all). We use the martingale decomposition with respect to a given filtration to construct the space of *m-functions* as an extension of the L_2-space on the basic probability space. To construct the space of t-functions we use a system of operators which can be very loosely described as compressions of the system of conditional expectations defined by the filtration mentioned above. In fact, the definition of this system of operators involves, along with the powers of Q and Q^*, the so-called *defect operators*. This way we arrive at expressions which are well-known in the theory of non-selfadjoint operators, in particular, in connection with dilations, characteristic functions and functional models. In the context of limit theorems, we finally obtain two conditions parallel to (1) and (2) in the abstract. The first of them requires, for an L_2-function on the state space, the solvability of the Poisson equation (of degree one) in *t-functions* (under the assumption of normality of Q this is exactly (1)). This condition guarantees (and is equivalent to) the generalized martingale-coboundary representation involving m-function. The conditions for the applicability of the CLT to a function f can be expressed in terms of the t-function solving the Poisson equation with f in the right hand side. Finally we obtain conditions for the CLT to

apply to a function f formulated entirely in terms of this function and the transition operator Q, without any reference to the path space of the Markov chain.

The main conclusion which can be done from the present paper is that in the normal case the (generalized) coboundary can be completely and explicitly restored in a very simple way from the martingale difference part of the martingale-coboundary representation. This martingale difference part (rather than the coboundary) seems to be the most natural functional parameter in the situation of the present paper.

A natural framework for some part of our considerations is given by the classical dilation theory of (not necessarily normal) contractions in Hilbert spaces [17, 18]. However, there are some probabilistic aspects which can not be treated in the framework of a purely Hilbert space theory (martingale difference nature of wandering subspaces, limit theorems). As to our approach to constructing extensions of Hilbert spaces in terms of filtrations, it can have some parallels in the analytic function theory in the unit disk. Clarifying these connections and considering the general non-normal case or other limit theorems require additional study and will not be discussed here.

The author thanks Dr. Holger Kösters and the anonymous referee for careful reading the first version of this paper and their suggestions which improved the paper.

2 Contractions, Transition Operators and Markov Chains

2.1 Some Notation

Let $(\mathcal{S}, \mathcal{M})$ and $Q : \mathcal{S} \times \mathcal{M} \to [0, 1]$ be a measurable space and a transition probability (= Markov kernel) on it. Assume that for Q there exists a stationary probability μ on $(\mathcal{S}, \mathcal{M})$ so that $\int_{\mathcal{S}} Q(s, A)\mu(ds) = \mu(A), A \in \mathcal{M}$. By the same symbol Q we will denote the transition (or Markov) operator defined on bounded measurable functions f by the relation $(Qf)(\cdot) = \int_{\mathcal{S}} f(s)Q(\cdot, ds)$. For every $p \in [1, \infty]$ the same formula defines in $L_p(\mu)$ an operator Q of norm 1 preserving positivity and acting identically on constants.

For every $n \in \mathbb{Z}$ denote by $\mathcal{S}_n$ a copy of $\mathcal{S}$, and set $\Omega = \prod_{n\in\mathbb{Z}} \mathcal{S}_n$. Assume that $X = (X_n)_{n\in\mathbb{Z}}$ is a stationary homogeneous Markov chain which has μ as the one-dimensional distribution and Q as the transition operator. The latter means that

$$\mathbb{E}(f(X_n)|X_{n-1}, X_{n-2}, \dots) = (Qf)(X_{n-1})$$

for every $n \in \mathbb{Z}$ and every bounded measurable f. We assume that the chain X is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such a way that, for every $n \in \mathbb{Z}$, $X_n(\omega) = s_n$, the n-th entry of $\omega = (\dots, s_{-1}, s_0, s_1, \dots)$, and also $\mathcal{F} = \sigma(X_n, n \in \mathbb{Z})$, the σ-field generated by all $X_n, n \in \mathbb{Z}$. We denote by T the $\mathbb{P}$-preserving bi-measurable invertible self-map of Ω uniquely determined by the relations $X_n(T(\cdot)) = X_{n+1}(\cdot), n \in \mathbb{Z}$. Starting with $(\mathcal{S}, \mathcal{M}), Q$ and μ, we can

always construct such a chain and related objects whenever $(\mathcal{S}, \mathcal{M})$ is a standard Borel space. It is known that the transformation T of $(\Omega, \mathcal{F}, \mathbb{P})$ is *ergodic* (that is there is no $A \in \mathcal{F}$ with $T^{-1}A = A$ and $\mathbb{P}(A)(1 - \mathbb{P}(A)) \neq 0$) if and only if 1 is a simple eigenvalue of the transition operator Q.

For sub-σ-fields $\mathcal{M}' \subseteq \mathcal{M}$ and $\mathcal{F}' \subseteq \mathcal{F}$ the standard notations $L_p(\mathcal{S}, \mathcal{M}', \mu)$ and $L_p(\Omega, \mathcal{F}', \mathbb{P})$ will be abbreviated to $L_p(\mathcal{M}', \mu)$ and $L_p(\mathcal{F}', \mathbb{P})$, or even to $L_p(\mu)$ or $L_p(\mathbb{P})$ if $\mathcal{M}' = \mathcal{M}$ or $\mathcal{F}' = \mathcal{F}$.

The Markov chain X generates on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ an increasing filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ and a decreasing filtration $(\mathcal{F}^n)_{n\in\mathbb{Z}}$ where $\mathcal{F}_n = \sigma(X_k, k \leq n)$ and $\mathcal{F}^n = \sigma(X_k, k \geq n)$, $n \in \mathbb{Z}$. These filtrations are compatible with T in the sense that $T^{-1}\mathcal{F}_n = \mathcal{F}_{n+1} \supseteq \mathcal{F}_n$ and $T^{-1}\mathcal{F}^n = \mathcal{F}^{n+1} \subseteq \mathcal{F}^n$ for $n \in \mathbb{Z}$. The increasing filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ can be completed to obtain $(\mathcal{F}_n)_{-\infty\leq n\leq\infty}$ by setting $\mathcal{F}_{-\infty} = \bigcap_{n\in\mathbb{Z}} \mathcal{F}_n$ and $\mathcal{F}_\infty = \bigvee_{n\in\mathbb{Z}} \mathcal{F}_n$. Analogously, $(\mathcal{F}^n)_{-\infty\leq n\leq\infty}$ is a completion of $(\mathcal{F}^n)_{n\in\mathbb{Z}}$ defined by setting $\mathcal{F}^{-\infty} = \bigvee_{n\in\mathbb{Z}} \mathcal{F}^n$ and $\mathcal{F}^\infty = \bigcap_{n\in\mathbb{Z}} \mathcal{F}^n$. The above filtrations give rise to the families $(\mathbb{E}_n)_{-\infty\leq n\leq\infty}$ and $(\mathbb{E}^n)_{-\infty\leq n\leq\infty}$ of conditional expectations. Let U and I be a unitary operator defined by $Uf = f \circ T$, $f \in L_2(\mu)$, and the identity operator, respectively. We denote by $|\cdot|_p$ and $||\cdot||_p$ the $L_p(\mu)$-norm and the $L_p(\mathbb{P})$-norm, respectively. The symbols $(\cdot, \cdot)$ and $||\cdot||$ denote the inner product in every Hilbert space and the norm in abstract Hilbert spaces.

Recall that a *contraction* is an operator in a Hilbert space whose norm is less than or equal to one. The transition operator Q in the situation described above defines a contraction in $L_2(\mu)$. Since the measurable map $X_0 : \Omega \to \mathcal{S}$ transforms the measure $\mathbb{P}$ to the measure μ, the mapping $L_2(\mu) \ni f \mapsto \tilde{f} \stackrel{\text{def}}{=} f \circ X_0 \in L_2(\mathbb{P})$ is an isometric embedding of $L_2(\mu)$ to $L_2(\mathbb{P})$. We have to emphasize that in many respects we just reproduce (or go in parallel to) well-known points from the dilation theory of contractions in Hilbert spaces [17, 18].

2.2 Normal Contractions

A bounded operator Q in a Hilbert space H satisfying the relation $QQ^* = Q^*Q$ is said to be *normal*. We are mostly interested in normal contractions. If Q is a normal contraction in H and $f \in H$, there exists such a unique measure ρ_f on the closed unit disk $D \subset \mathbb{C}$ that

$$(Q^m f, Q^n f) = \int_D z^m \bar{z}^n \rho_f(dz)$$

for every $m, n \geq 0$. In particular, if μ is a stationary probability measure for a transition probability Q, the transition operator $Q : L_2(\mu) \to L_2(\mu)$ has the norm 1, hence is a contraction. If, moreover, Q is normal, the above formula applies to Q and every $f \in L_2(\mu)$. The spectral theory of normal operators allows us to investigate the Poisson equation (4) for a normal Q in terms of ρ_f. In particular, (4) is solvable in $L_2(\mu)$ for an $f \in L_2(\mu)$ if and only if

$$\int_D \frac{1}{|1-z|^2}\,\rho_f(dz) < \infty. \tag{6}$$

Clearly, the latter condition implies that f is orthogonal to all fixed points of Q.

2.3 *Unitary Part of a Transition Operator and Its Deterministic σ-Field*

Let Q be a contraction in a Hilbert space H. It is known [17, 18] that the subspace $H_u \subseteq H$ defined by the relation

$$H_u = \{f \in H : \cdots = |Q^{*2} f| = |Q^* f| = |f| = |Qf| = |Q^2 f| = \ldots\} \tag{7}$$

reduces the operator Q so that $H = H_u \oplus H_{cnu}$, where H_u and H_{cnu} are completely Q-invariant (that is both Q- and Q^*-invariant), $Q|_{H_u}$ is a unitary operator, and H_u is the greatest subspace with such properties. The operators $Q_u = Q|_{H_u}$ and $Q_{cnu} = Q|_{H_{cnu}}$ are called the *unitary part* and the *completely non-unitary part* of Q, respectively. In this notation we have

$$Q = Q_u \oplus Q_{cnu}.$$

In the case of a *normal* contraction Q this decomposition can be immediately deduced by means of the projection-valued spectral measure P_Q of the operator Q. We say that a projection-valued measure *is concentrated* on a Borel set $A \subseteq \mathbb{C}$ if it vanishes on every Borel set disjoint with A; the *restriction* of such a measure P to a Borel set A is another such a (uniquely defined) measure which is concentrated on A and agrees with P on every Borel subset of A; we denote this measure by P^A. Using this terminology and notation, P_Q is concentrated on the closed unit disk D so that $P_Q = P_Q^D$. Let $K = \{|z| \in \mathbb{C} : z = 1\}$ and $D_0 = \{z \in \mathbb{C} : |z| < 1\}$ be the unit circle and the open unit disk. Then $H_u = P_Q(K)H$ and $H_{cnu} = P_Q(D_0)H$. Let P_Q^K and $P_Q^{D_0}$ be the restrictions of P_Q to K and to D_0, respectively. Then we have $P_Q = P_Q^K + P_Q^{D_0}$. By abuse of notation, we also have $P_Q^K = P_{Q_u}$ and $P_Q^{D_0} = P_{Q_{cnu}}$ (here P_{Q_u} and $P_{Q_{cnu}}$ are considered, due to the canonical inclusions of H_u and H_{cnu} in H, as measures with values in orthoprojections of H rather than of H_u or H_{cnu}).

For a normal contraction there exist simple criteria for the relations $f \in H_u$ and $f \in H_{cnu}$.

Proposition 2.1. *Let $Q : H \to H$ be a normal contraction, $H = H_u \oplus H_{cnu}$ the orthogonal decomposition defined above, $||\cdot||$ the norm in H and $f \in H$.*

Then $f \in H_u$ if and only if at least one of the relations $\lim_{n\to\infty} ||Q^n f|| = ||f||$, $\lim_{n\to\infty} ||Q^{*n} f|| = ||f||$ holds. In fact, in this case equalities in (7) take place.

Further, $f \in H_{cnu}$ if and only if at least one of the relations $\lim_{n\to\infty} ||Q^n f|| = 0$, $\lim_{n\to\infty} ||Q^{*n} f|| = 0$ holds. If so, both of these relations hold simultaneously.

Proof. Let $f = f_u + f_{cnu}$, where $f_u \in H_u$, $f_{cnu} \in H_{cnu}$, and let ρ_f be the spectral measure of f. Then we have

$$||Q^n f_u||^2 = ||Q^{*n} f_u||^2 = \int_K |z|^{2n} \rho_f(dz) = \rho_f(K) = ||f_u||^2,$$

$$||Q^n f_{cnu}||^2 = ||Q^{*n} f_{cnu}||^2 = \int_{D_0} |z|^{2n} \rho_f(dz) \underset{n\to\infty}{\downarrow} 0,$$

and

$$||Q^n f||^2 = ||Q^{*n} f||^2 = ||Q^n f_u||^2 + ||Q^n f_{cnu}||^2 \underset{n\to\infty}{\downarrow} ||f_u||^2.$$

These relations, along with the relation $||f||^2 = ||f_u||^2 + ||f_{cnu}||^2$, imply the assertions of the proposition. □

Let now $H = L_2(\mu)$ and Q be a transition operator with the stationary probability μ. Then, according to S. Foguel's theorem [5, 6], the subspace H_u is of the form $L_2(\mathcal{M}_{det}, \mu)$, where $\mathcal{M}_{det}$ is a sub-σ-field of $\mathcal{M}$ which we will call *deterministic*. (Caution: sometimes this term is used for the σ-fields related to the one-sided analogues of the condition (7)). Moreover, $Q|_{L_2(\mathcal{M}_{det},\mu)}$ defines a μ-preserving automorphism of $\mathcal{M}_{det}$, and $\mathcal{M}_{det}$ is the largest sub-σ-field of $\mathcal{M}$ with this property. In the Markov chain context we will use denotations H_{det} and H_{ndet} (from *deterministic* and *nondeterministic*) instead of H_u and H_{cnu}, respectively. The orthogonal projection P_{det} to $H_{det} = L_2(\mathcal{M}_{det}, \mu)$ coincides with the corresponding conditional expectation $\mathbb{E}^{\mathcal{M}_{det}} : L_2(\mu) \to L_2(\mathcal{M}_{det}, \mu)$; the range of the complementary projection $P_{ndet} = I - \mathbb{E}^{\mathcal{M}_{det}}$ is H_{ndet}. In the normal case the projection $P_{det} : L_2(\mu) \to L_2(\mathcal{M}_{det}, \mu)$ is exactly the spectral projection $P_Q(K)$ of the operator Q while the complementary projection P_{ndet} agrees with $P_Q(D_0)$.

Remark 2.2. For an $f \in L_2(\mu)$ the orthogonal decomposition

$$f = f_{det} + f_{ndet}$$

with $f_{det} = P_{det} f$ and $f_{ndet} = P_{ndet} f$ leads to the decomposition of the stationary random sequence $(f(X_n))_{n\in\mathbb{Z}}$ into the sum of the sequences $(f_{det}(X_n))_{n\in\mathbb{Z}}$ and $(f_{ndet}(X_n))_{n\in\mathbb{Z}}$, the second of them having zero conditional expectation given the first one. Without additional assumptions the sequence $(f_{det}(X_n))_{n\in\mathbb{Z}}$ may be an arbitrary stationary sequence of square-integrable variables whose influence to the behavior of $(f(X_n))_{n\in\mathbb{Z}}$ is out of our control. The sequence $(f_{ndet}(X_n))_{n\in\mathbb{Z}}$, unlike $(f_{det}(X_n))_{n\in\mathbb{Z}}$, admits some further analysis. Under the assumption of normality of Q some problems (such as the Central Limit Theorem) concerning $(f(X_n))_{n\in\mathbb{Z}}$ can be treated in terms of the spectral measures of the functions f, f_{det} and f_{ndet}. Notice that $\rho_f = \rho_{f_{det}} + \rho_{f_{ndet}}$, where $\rho_{f_{det}}$ and $\rho_{f_{ndet}}$ are concentrated on K and D_0, respectively.

2.4 Defect Operators and Defect Spaces of a Contraction

We present here the definition and some properties of the *defect operators* and the *defect spaces* of a contraction $Q : H \to H$ (see [17] and [18] for proofs and more details). The operators

$$D_Q = (I - Q^*Q)^{\frac{1}{2}}, D_{Q^*} = (I - QQ^*)^{\frac{1}{2}}$$

are called the *defect operators* of Q. These operators are self-adjoint non-negative (in the spectral sense) contractions, satisfying

$$QD_Q = D_{Q^*}Q,\ D_Q Q^* = Q^* D_{Q^*}.$$

The spaces

$$\mathcal{D}_Q = \overline{D_Q H},\ \mathcal{D}_{Q^*} = \overline{D_{Q^*} H}$$

are called *defect spaces* of Q. It follows from the above relations that

$$Q\mathcal{D}_Q \subseteq \mathcal{D}_{Q^*}, Q^*\mathcal{D}_{Q^*} \subseteq \mathcal{D}_Q.$$

In the case of a normal contraction Q the corresponding defect operators agree, and so are the defect subspaces. In this case the defect subspace is invariant with respect to both Q and Q^*, and the restriction $D_Q|_{H_{cnu}}$ of the defect operator to the completely non-unitary subspace is injective. Indeed, if $f \in H_{cnu}$ and $D_Q f = 0$ the spectral measure ρ_f is concentrated on D_0 by the first of these two relations, while by the second relation $(Qf, Qf) = (f, f)$; the latter means that ρ_f is concentrated on K, implying $\rho_f = 0$. Furthermore, it is easy to see from the consideration of spectral measures that $\mathcal{D}_Q = H_{cnu}$ if Q is a normal contraction.

Remark 2.3. When Q is a transition operator, its defect subspaces are in a natural unitary correspondence with the spaces of the forward and the backward martingale differences of the Markov chain X (see the next section of the paper; compare with [18], Sect. 3.2). □

3 Quasi-functions

3.1 Quasi-functions: m-Functions and t-Functions

Looking for a generalization of the martingale-coboundary representation and the Poisson equation, we need some more general objects than the $L_2(\mathbb{P})$-functions of the form $f(X_0)$ in the first case and $L_2(\mu)$-functions in the second one. The first problem is solved in terms of the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ determined by the Markov

chain, while the second one is treated in terms of the transition operator and some other auxiliary operators acting on $L_2(\mu)$-functions. In both cases we obtain a decomposition of an L_2-function into a series. Removing the requirement that the decomposition belongs to an L_2-function, we arrive at a class of objects which are given by their decompositions but, in general, are no longer functions. These objects called *quasi-functions* will be considered as elements of certain Banach spaces. These Banach spaces contain conventional L_2-spaces as dense subspaces such a way that every quasi-function can be represented in a canonical way as a limit of L_2-functions. Moreover, every operator we are interested in admits a canonical extension from L_2 to the appropriate Banach space. We will consider quasi-functions of two kinds. Quasi-functions of the first kind generalize conventional functions defined on the path space of the Markov chain under consideration and will be called *m-functions*; quasi-functions of the second kind, generalizing conventional functions defined on the state space of the Markov chain, will be called *t-functions*. It turns out that some conventional functions on the path space can have a martingale-coboundary representation in terms of *m-functions*; some of them are conventional L_2-functions, but some other are not. Also the Poisson equation for an L_2-function on the state space with no L_2-solution may be sometimes solved in *t-functions*.

As an introductory step, we start with considering the decompositions of functions from L_2-spaces.

3.2 Functions and m-Functions

For every $g \in L_2(\mu)$ we have the following martingale decomposition

$$\tilde{g} = \sum_{n=0}^{\infty} (\mathbb{E}_{-n} - \mathbb{E}_{-n-1})\tilde{g} + \mathbb{E}_{-\infty}\tilde{g}, \tag{8}$$

converging in the norm of $L_2(\mathbb{P})$. Rewriting the summands of (8) in terms of the operators Q, U and the embedding $g \mapsto \tilde{g}$, we have

$$\begin{gathered}(\mathbb{E}_{-n} - \mathbb{E}_{-n-1})\tilde{g} \\ = U^{-n}\mathbb{E}_0 U^n \tilde{g} - U^{-n-1}\mathbb{E}_0 U^{n+1}\tilde{g} = U^{-n}\mathbb{E}_0\mathbb{E}^n U^n \tilde{g} - U^{-n-1}\mathbb{E}_0\mathbb{E}^{n+1}U^{n+1}\tilde{g} \\ = U^{-n}\widetilde{Q^n g} - U^{-n-1}\widetilde{Q^{n+1}g}\end{gathered} \tag{9}$$

and, with the limits in the norm of $L_2(\mathbb{P})$,

$$\begin{aligned} &\mathbb{E}_{-\infty}\tilde{g} \\ &= \lim_{n\to\infty}\mathbb{E}_{-n}\tilde{g} = \lim_{n\to\infty}U^{-n}\mathbb{E}_0U^n\tilde{g} = \lim_{n\to\infty}U^{-n}\mathbb{E}_0\mathbb{E}^nU^n\tilde{g} \\ &= \lim_{n\to\infty}U^{-n}\widetilde{Q^ng} = \lim_{n\to\infty}U^{-n}\widetilde{Q^ng_{det}} \\ &= \tilde{g}_{det}. \end{aligned} \tag{10}$$

Deriving (9) and (10), we used the fact that $\tilde{g}$ is X_0-measurable, along with some standard properties of Markov chains. In (10) we also used the decomposition $g = g_{det} + g_{ndet}$, the relation $Q^ng_{ndet} \underset{n\to\infty}{\to} 0$ and the identity $U^{-1}\widetilde{Qg_{det}} = \widetilde{g_{det}}$ which can be explained as follows. Since the map $g \to \tilde{g}$ isometrically embeds $L_2(\mu)$ to $L_2(\mathbb{P})$, the subspace $H_{det} \subseteq L_2(\mu)$ is also embedded to $L_2(\mathbb{P})$. Furthermore, it can be easily verified that H_{det} is a completely invariant subspace of the unitary U, and that $U|_{H_{det}} = Q|_{H_{det}}$. Another way to express this is the relation $U^{-1}\widetilde{Qg_{det}} = \widetilde{g_{det}}$ used in (10). Moreover, this relation allows us to substitute $\tilde{g}$ by $\tilde{g}_{ndet}$ in the right-hand side of (9). With (10) and properly modified (9), the identity (8) can be rewritten as

$$\tilde{g} = \sum_{n=0}^{\infty}(U^{-n}\widetilde{Q^ng_{ndet}} - U^{-n-1}\widetilde{Q^{n+1}g_{ndet}}) + \tilde{g}_{det}. \tag{11}$$

Assuming now $g \in H_{ndet}$, we have the following martingale decomposition:

$$\tilde{g} = \sum_{n=0}^{\infty}U^{-n}(\widetilde{Q^ng} - U^{-1}\widetilde{Q^{n+1}g}). \tag{12}$$

Analyzing the right-hand side of (12), observe that all terms in this series are of the form $U^{-n}(\widetilde{r_n} - U^{-1}\widetilde{Qr_n})$, $r_n \in H_{ndet}$ $(n \in \mathbb{Z})$. Terms of such form are mutually orthogonal for different $n \in \mathbb{Z}$. Set

$$L_n = \overline{\{U^n(\tilde{r} - U^{-1}\widetilde{Qr}) : r \in H_{ndet}\}} \ (n \in \mathbb{Z})$$

and denote by M the closed subspace of $L_2(\mathbb{P})$ generated by all $L_n, n \in \mathbb{Z}$. In view of the mutual orthogonality of L_n we have

$$M = \oplus_{n\in\mathbb{Z}}L_n \tag{13}$$

(we use $\oplus$ both as a symbol of an exterior operation and also for the closed span of some orthogonal subspaces of a certain Hilbert space). The space M is a completely invariant subspace of the operator U. The operator $U|_M$ is unitarily equivalent to the two-sided shift operator, and every L_n is a wandering subspace for $U|_M$. From now on we will write U instead of $U|_M$. Denoting by $\bigvee$ the linear span of some set of liner subspaces, we also have

$$M = \overline{\bigvee_{n\in\mathbb{Z}} U^n H_{ndet}}. \tag{14}$$

Indeed, the left-hand side is contained in the right-hand one because of the obvious relation $L_n \subseteq \overline{U^n H_{ndet} \bigvee U^{n-1} H_{ndet}}$ $(n \in \mathbb{Z})$; the opposite inclusion is a consequence of (12) and the complete invariance of M with respect to U. We will also need the U^{-1}-invariant spaces

$$M_n = \oplus_{k\le n} L_k \left(= \overline{\bigvee_{k\le n} U^k H_{ndet}}\right), n \in \mathbb{Z}.$$

Since

$$||\tilde{h} - U^{-1}\widetilde{Qh}||^2 = ((I - Q^*Q)h, h),$$

setting for every $h \in H_{ndet}$

$$l(\tilde{h} - U^{-1}\widetilde{Qh}) = (I - Q^*Q)^{\frac{1}{2}}h$$

defines a unitary map $l : L_0 \to \mathcal{D}_Q$. As was observed in Sect. 2.4, for a normal transition operator Q we have $\mathcal{D}_Q = H_{ndet}$, and therefore l maps L_0 to H_{ndet}. Then the space M_0 is unitarily equivalent to the space of one-sided sequences of the elements of H_{ndet} via the correspondence

$$M_0 \ni \sum_{n\le 0} U^n p_n \leftrightarrow (\ldots, l(p_{-1}), l(p_0)) \in H_{ndet} \otimes l_2(\mathbb{Z}_-), \tag{15}$$

where $(p_n)_{n\le 0}$ is a sequence of elements of L_0 with $\sum_{n\le 0} ||p_n||_2^2 < \infty$ and $H_{ndet} \otimes l_2(\mathbb{Z}_-)$ denotes the Hilbert space tensor product of Hilbert spaces. The elements of $H_{ndet} \otimes l_2(\mathbb{Z}_-)$ are sequences $(\ldots, a_{-1}, a_0)$ with $a_n \in H_{ndet}$ $(n \le 0)$ and $\sum_{n\le 0} |a_n|_2^2 < \infty$. By this unitary equivalence the one-sided shift $\sigma : (\ldots, a_{-1}, a_0) \mapsto (\ldots, a_{-1}, a_0, 0)$ in the space $H_{ndet} \otimes l_2(\mathbb{Z}_-)$ corresponds to the isometric operator $U^{-1}|_{M_0}$, while the co-isometric inverse shift $\sigma^*: (\ldots, a_{-1}, a_0) \mapsto (\ldots, a_{-2}, a_{-1})$ corresponds to $(U^{-1}|_{M_0})^*$. Furthermore, since Q acts on the space H_{ndet}, we can define its coordinatewise action on $H_{ndet} \otimes l_2(\mathbb{Z}_-)$ by

$$Q(\ldots, a_{-1}, a_0) = (\ldots, Qa_{-1}, Qa_0).$$

We set $\hat{Q} = l^{-1}Ql : L_0 \to L_0$, and extend it (with the same notation and in agreement with (15)) to $\hat{Q} : M_0 \to M_0$ by setting $\hat{Q}(\sum_{n\le 0} U^n p_n) = \sum_{n\le 0} U^n \hat{Q} p_n$.

We are in position now to give a description of those elements of M_0 which are martingale decompositions (12) of certain $\tilde{g}$ with $g \in H_{ndet}$.

Proposition 3.1. *The following conditions on the series $\sum_{n\le 0} U^n p_n \in M_0$ are equivalent:*

(1) the series $\sum_{n\le 0} U^n p_n$ *represents a decomposition* (12) *of certain* $\tilde{g}$ *with* $g \in H_{ndet}$;
(2) there exists such $p \in L_0$ *that* $\sum_{n\ge 0} ||\hat{Q}^n p||_2^2 < \infty$ *and* $p_{-n} = \hat{Q}^n p$ *for every* $n \ge 0$;
(3) there exists such $r \in H_{ndet}$ *that* $\sum_{n\ge 0} |Q^n r|_2^2 < \infty$ *and for every* $n \ge 0$ $l(p_{-n}) = Q^n r$.

Proof. Conditions (2) and (3) are equivalent because l is a unitary operator and $\hat{Q} = l^{-1} Q l$. Let us show that (1) implies (3). According to (12), for the martingale decomposition of an $\tilde{g}$ with $g \in H_{ndet}$ we have for $n \ge 0$ $p_{-n} = \widetilde{Q^n g} - U^{-1}\widetilde{Q^{n+1} g}$, so that $l(p_{-n}) = Q^n (I - Q^* Q)^{\frac{1}{2}} g = Q^n r$, where $r = (I - Q^* Q)^{\frac{1}{2}} g$. This follows that $\sum_{n\ge 0} |Q^n r|_2^2 = \sum_{n\ge 0}((Q^*Q)^n (I - Q^* Q) g, g) = (g, g) < \infty$. Conversely, assuming (3), let $h \in H_{ndet}$ is such that $\sum_{n\ge 0} |Q^n r|_2^2 < \infty$. This is equivalent to

$$\int_D \frac{1}{1 - |z|^2}\, \rho_r(dz) < \infty,$$

which follows that there exists such $g \in H_{ndet}$ that $r = (I - Q^* Q)^{\frac{1}{2}} g$. Then in the martingale decomposition $\tilde{g} = \sum_{n\le 0} U^n p'_n$ we have for $n \le 0$ $p'_n = \widetilde{Q^{-n} g} - U^{-1}\widetilde{Q^{-n+1} g}$ or $l(p'_n) = Q^{-n}(I - Q^* Q)^{\frac{1}{2}} g = Q^{-n} r = l(p_n)$, and we conclude $p'_n = p_n$. □

Let $c_0(\mathbb{Z}_-)$ be the space of all complex sequences indexed by the elements of $\mathbb{Z}_-$ and tending to zero, $c_0(\mathbb{Z}_-)$ being supplied with the sup-norm. Then the injective tensor product $H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$ is the space of all sequences $\overline{a} = (\ldots, a_{-1}, a_0)$ with $a_n \in H_{ndet}$ for $n \le 0$, $|a_n|_2 \underset{n\to-\infty}{\to} 0$ and with the norm of $\overline{a} = (\ldots, a_{-1}, a_0)$ defined as $\sup_{n\le 0} |a_n|_2$. The space $H_{ndet} \otimes l_2(\mathbb{Z}_-)$, represented as a space of sequences of elements of H_{ndet}, can be in a natural way continuously and injectively mapped into $H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$. Notice that the shift operators σ_n and σ_n^* can be extended to $H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$. Observe that $\sigma^{*n}\overline{a} \underset{n\to\infty}{\to} 0$ for every $\overline{a} \in H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$. We can transfer this extension, via correspondence (15), to a space containing M_0. Elements of M_0 are sums $\sum_{n\le 0} U^n p_n$, where $p_n \in L_0, n \le 0$, and $\sum_{n\le 0} ||p_n||_2^2 < \infty$. Then the extended space denoted by M_0^{ext} and consisting of the formal sums $\sum_{n\le 0} U^n p_n$ where $p_n \in L_0, n \le 0$, $||p_n||_2 \underset{n\to-\infty}{\to} 0$; the norm of $\sum_{n\le 0} U^n p_n$ is defined as $\sup_{n\le 0} ||p_n||_2$. The operators $U^{-1}|_{M_0}$ and $(U^{-1}|_{M_0})^*$ admit obvious extensions to M_0^{ext} which we denote $U^{-1}|_{M_0^{ext}}$ and $(U^{-1}|_{M_0^{ext}})^*$. Analogously, every space M_n $(n \in \mathbb{Z})$ can be extended to the space M_n^{ext}. If we write the elements of M_n^{ext} as $\sum_{k\le n} U^k p_k$ with $p_k \in L_0 (k \le n)$, we obtain a growing sequence of subspaces of the space

$$M^{ext} = \{\sum_{n\in\mathbb{Z}} U^n p_n : p_n \in L_0 (n \in \mathbb{Z}), ||p_n||_2 \underset{|n|\to\infty}{\to} 0\}.$$

The space M^{ext} is an extension of M, and we will hold the notations U and U^{-1} for natural extensions of these operators from M to M^{ext}.

Definition 3.2. Elements of the Banach space M^{ext} are called *m-functions*.

Remark 3.3. There are also other operators which can be naturally extended from M to M^{ext}. For example, so are the projections $\mathbb{E}_n : \sum_{k\in\mathbb{Z}} U^k p_k \mapsto \sum_{k\le n} U^k p_k$, $n \in \mathbb{Z}$. □

3.3 Functions and t-Functions

The space of t-functions which we are going to define extends the space $H_{ndet} \subseteq L_2(\mu)$. Functions from H_{ndet} admit some decomposition; t-functions will be defined in terms of a similar decomposition. The next problem to solve will be how to embed the space of t-functions to the space of m-functions generalizing the embedding $g \mapsto g \circ X_0$ of the space H_{ndet} to $L_2(\mathbb{P})$. This also will be done in terms of the corresponding decompositions.

Taking in (11) the conditional expectation relative to X_0 (which is, in particular, the left inverse for the embedding $g \mapsto g \circ X_0$), we obtain

$$g = \sum_{n=0}^{\infty} (Q^{*n}Q^n - Q^{*(n+1)}Q^{n+1}) g_{ndet} + g_{det}. \tag{16}$$

Now we again assume that $g \in H_{ndet}$. Then we have

$$g = \sum_{n=0}^{\infty} (Q^{*n}Q^n - Q^{*(n+1)}Q^{n+1}) g \tag{17}$$

or

$$g = \sum_{n=0}^{\infty} Q^{*n}(I - Q^*Q)Q^n g, \tag{18}$$

where the series' converge in the norm $|\cdot|_2$.

Since Q is normal, we have

$$\begin{aligned}(g,g) &= \sum_{n=0}^{\infty} (Q^{*n}(I - Q^*Q)Q^n g, g) \\ &= \sum_{n=0}^{\infty} (Q^n(I - Q^*Q)^{\frac{1}{2}} g, Q^n(I - Q^*Q)^{\frac{1}{2}} g).\end{aligned} \tag{19}$$

Then, we have an isometric correspondence

$$H_{ndet} \ni g \leftrightarrow (\dots, Q(I-Q^*Q)^{\frac{1}{2}}g, (I-Q^*Q)^{\frac{1}{2}}g) \in H_{ndet} \otimes l_2(\mathbb{Z}_-) \qquad (20)$$

(the set $\mathbb{Z}_-$ rather than $\mathbb{Z}_+$ was chosen here for the future denotational convenience). According to Proposition 3.1, another description of the image of H_{ndet} in $H_{ndet} \otimes l_2(\mathbb{Z}_-)$ by the above correspondence is as follows:

$$\{(\dots, Qr, r) : r \in H_{ndet}, \sum_{n\geq 0} |Q^n r|_2^2 < \infty\}.$$

We will identify this image with H_{ndet}. The extension $H_{ndet}^{ext} \supseteq H_{ndet}$ is then defined by

$$H_{ndet}^{ext} = \{(\dots, Qr, r) : r \in H_{ndet}\}$$

(we consider H_{ndet}^{ext} as a subspace of $H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$).

Definition 3.4. Elements of the Banach spaces H_{ndet}^{ext} are called *t-functions*.

It is clear from this definition that every t-function is a sequence of functions from H_{ndet} with some additional properties; in case the corresponding series converges in $L_2(\mu)$ its sum gives an H_{ndet}-representative of the corresponding t-function; otherwise a t-function is a proper generalized function; anyway, a t-function is a limit (in the sense of $H_{ndet} \otimes_\epsilon c_0(\mathbb{Z}_-)$) of functions from H_{ndet}.

Let us turn now to the embedding of t-functions to m-functions. Reformulating the mapping $f \mapsto \tilde{f} = f \circ X_0$, $f \in H_{ndet}$, in terms of decompositions, we obtain

$$\begin{aligned} &H_{ndet} \ni g \leftrightarrow (\dots, (I-Q^*Q)^{\frac{1}{2}}g, Q(I-Q^*Q)^{\frac{1}{2}}g, (I-Q^*Q)^{\frac{1}{2}}g) \\ &\mapsto \sum_{n\leq 0} U^n l^{-1}\big(Q^{-n}(I-Q^*Q)^{\frac{1}{2}}g\big) = \tilde{g} \in M_0. \end{aligned} \qquad (21)$$

This embedding can be described differently as

$$(\dots, Qr, r) \mapsto \widetilde{(\dots, Qr, r)} = \sum_{n\leq 0} U^n l^{-1}\big(Q^{-n}r\big), \qquad (22)$$

which makes sense both for $H_{ndet} \to \widetilde{H_{ndet}} \subseteq M_0$ and for $H_{ndet}^{ext} \to \widetilde{H_{ndet}^{ext}} \subseteq M_0^{ext}$.

Proposition 3.5. *Let $Q : L_2(\mu) \to L_2(\mu)$ be a normal transition operator for a stationary Markov chain X and $f \in H_{ndet}$ have the spectral measure ρ_f. Then the following conditions on the function f are equivalent:*

(1) $f = g - Qg$ with some $g \in H_{ndet}^{ext}$;
(2) $\tilde{f} = h + \tilde{g} - U\tilde{g}$ with some $g \in H_{ndet}^{ext}$ and $h \in M_1$ such that $\mathbb{E}_0 h = 0$;
(3) $r \stackrel{def}{=} (I-Q)^{-1}(I-Q^*Q)^{\frac{1}{2}} f \in H_{ndet}$;
(4) $\sigma_f^2 \stackrel{def}{=} |r|_f^2 = \int_D \frac{1-|z|^2}{|1-z|^2}\rho_f\, dz < \infty$.

Moreover, deducing (2) from (1) or (3) we can always set $h = U\tilde{g} - \widetilde{Qg}$; also, we have $||h||_2^2 = |r|_2^2 = \sigma_f^2$.

Proof. Let (1) holds true. Then $\tilde{g} = \sum_{n\le 0} U^n l^{-1}\big(Q^{-n}r\big)$ for some $r \in H_{ndet}$, and

$$\tilde{f} = \tilde{g} - \widetilde{Qg} = U\tilde{g} - \widetilde{Qg} + \tilde{g} - U\tilde{g} = h + \tilde{g} - U\tilde{g}, \tag{23}$$

where

$$h = U\tilde{g} - \widetilde{Qg} = \sum_{n\le 0} U^{n+1} l^{-1}\big(Q^{-n}r\big) - \sum_{n\le 0} U^n l^{-1}\big(Q^{-n+1}r\big) = U^1 l^{-1}\big(r\big) \in L_1, \tag{24}$$

and (2) follows. To establish (2) $\to$ (1), apply $\mathbb{E}_0$ to the relation (2), obtaining $\tilde{f} = \tilde{g} - \mathbb{E}_0 U\tilde{g}$; then check $\mathbb{E}_0 U\tilde{g} = \widetilde{Qg}$ by means of the represention $\tilde{g} = \sum_{n\le 0} U^n l^{-1}\big(Q^{-n}r\big)$ with some $r \in H_{ndet}$.

Let us show now that (1) and (3) are equivalent. The relation (1) holds if and only if for some $r \in H_{ndet}$ and every $n \ge 0$ $Q^n Q(I - Q^*Q)^{\frac{1}{2}} f = Q^n(I - Q)r$. But this is equivalent to $Q(I - Q^*Q)^{\frac{1}{2}} f = (I - Q)r$ or to $r = (I - Q)^{-1}(I - Q^*Q)^{\frac{1}{2}} f$ which is equivalent to (3). Since $f \in H_{ndet}$ and H_{ndet} is invariant with respect to Q, such $r \in H_{ndet}$ exists if and only if (4) holds. The last assertions follow from (23) and (24). □

Remark 3.6 (Unicity and reality). It is easy to see that the equation $f = g - Qg$ may have at most one solution $g \in H_{ndet}^{ext}$.

Functions we consider are in general complex-valued; so functions and quasi-functions constitute Banach spaces over $\mathbb{C}$. The involutive conjugation in these spaces is well-defined, its fixed points are said to be real. The operators Q, Q^*, their spectral projections and conditional expectations $\mathbb{E}_n (n \in \mathbb{Z})$ preserve the reality of functions and quasi-functions. In view of this, for example, in the orthogonal decomposition $f = f_{det} + f_{ndet}$ the summands f_{det} and f_{ndet} are real-valued provided that so is f. These facts and the unicity imply that the solution of the Poisson equation with a real right-hand side must be real; also for a real function the ingredients of the martingale-coboundary representation must be real. Notice that for the spectral measure ρ_f of a real-valued function f with respect to the operator Q the real axis is the symmetry axis. □

4 The CLT

In addition to assumptions of Sect. 2 (including the normality of the transition operator) we assume that 1 is a simple eigenvalue of the operator Q. It is known [10] that this implies (and is equivalent to) the ergodicity of the shift transformation T.

We give now an alternative proof of a version of the Central Limit Theorem for a stationary normal Markov chain (Thm. 7.1 in [10]). Let $N(m, \sigma^2)$ be the normal

law with the mean value m and the variance σ^2, degenerate if $\sigma^2 = 0$. As above, D denotes the closed unit disk in $\mathbb{C}$.

Theorem 4.1. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary homogeneous Markov chain which has a probability measure μ as the one-dimensional distribution and a normal operator $Q : L_2(\mu) \to L_2(\mu)$ as the transition operator. Assume that the eigenvalue 1 of Q is simple. Let a real-valued function $f \in L_2(\mu)$ with the spectral measure ρ_f satisfy the conditions*

(1) $\sigma_f^2 = \int_D \frac{1-|z|^2}{|1-z|^2}\rho_f\, dz < \infty$,

(2) $\lim_{n\to\infty} n^{-\frac{1}{2}} \left|\sum_{k=0}^{n-1} Q^k f\right|_2 = 0$.

Then the random variables $\left(n^{-\frac{1}{2}} \sum_{k=0}^{n-1} f(X_k)\right)_{n\geq 1}$ converge in distribution to the normal law $N(0, \sigma_f^2)$. Moreover,

$$\lim_{n\to\infty} n^{-1} \left|\left|\sum_{k=0}^{n-1} f(X_k)\right|\right|_2^2 = \sigma_f^2. \tag{25}$$

Proof. Let us first reduce the proof to the case $f \in H_{ndet}$. Since the decomposition $L_2(\mu) = H_{det} \oplus H_{ndet}$ reduces the operator Q, the assumption (2) implies for $f = f_{det} + f_{ndet}$ ($f_{det} \in H_{det}$, $f_{ndet} \in H_{ndet}$) that

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \left|\sum_{k=0}^{n-1} Q^k f_{det}\right|_2 = 0 \tag{26}$$

and

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \left|\sum_{k=0}^{n-1} Q^k f_{ndet}\right|_2 = 0. \tag{27}$$

Since $Q|_{H_{det}}$ is a unitary operator which agrees, after embedding $\widetilde{H_{det}}$ to $L_2(\mathbb{P})$, with U, for every $n \geq 1$ we have

$$\left|\sum_{k=0}^{n-1} Q^k f_{det}\right|_2 = \left|\left|\sum_{k=0}^{n-1} f_{det}(X_k)\right|\right|_2,$$

which follows

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \left|\left|\sum_{k=0}^{n-1} f_{det}(X_k)\right|\right|_2 = 0. \tag{28}$$

It is therefore clear that the random variables $(n^{-\frac{1}{2}} \sum_{k=0}^{n-1} f_{det}(X_k), n \geq 1)$ converges to 0 both in probability and in the norm $||\cdot||_2$. By this reason we will assume $f \in H_{ndet}$ in the rest of the proof.

In view of the assumption (1) and Proposition 3.5, f admits the representation $f = g - Qg$ with some $g \in H^{ext}_{ndet}$, and we have

$$\sum_{k=0}^{n-1} f(X_k) = \sum_{k=0}^{n-1} U^k(U\tilde{g} - \widetilde{Qg}) + \tilde{g} - U^n\tilde{g}.$$

Here $(U^k(U\tilde{g} - \widetilde{Qg}))_{k\geq 0}$ is a stationary ergodic sequence of martingal differences whose variance is, by Proposition 3.5, σ_f^2. Then, in view of the Billingsley-Ibragimov theorem, we only need to show that

$$n^{-1}||\tilde{g} - U^n\tilde{g}||_2^2 \underset{n\to\infty}{\to} 0. \tag{29}$$

We have with an $r \in H_{ndet}$ from (3) in Proposition 3.5

$$\begin{aligned}
n^{-1}||\tilde{g} - U^n\tilde{g}||_2^2 &= n^{-1}\Big|\Big|\sum_{k\leq 0} U^k l^{-1}(Q^{-k}r) - \sum_{k\leq 0} U^{n+k} l^{-1}(Q^{-k}r)\Big|\Big|_2^2 \\
&= n^{-1}\Big|\Big|\sum_{0\leq k\leq n-1} U^{n-k} l^{-1}(Q^k r)\Big|\Big|_2^2 + n^{-1}\Big|\Big|\sum_{k\leq 0} U^k l^{-1}(Q^{-k}r) - \sum_{k\leq 0} U^k l^{-1}(Q^{n-k}r)\Big|\Big|_2^2 \\
&= n^{-1}\sum_{k=0}^{n-1} |Q^k r|_2^2 + n^{-1}\Big|\Big|\sum_{k\leq 0} U^k l^{-1}(Q^{-k}r) - \sum_{k\leq 0} U^k l^{-1}(Q^{n-k}r)\Big|\Big|_2^2 \\
&= n^{-1}\sum_{k=0}^{n-1} |Q^k r|_2^2 + n^{-1}|g - Q^n g|_2^2 = n^{-1}\sum_{k=0}^{n-1} |Q^k r|_2^2 + n^{-1}\Big|\sum_{k=0}^{n-1} Q^k f\Big|_2^2.
\end{aligned} \tag{30}$$

The summands of the last sum tend to zero: the first one because so does $|Q^n r|_2$ and the second one by assumption (2) of the theorem. This completes the proof. □

Acknowledgements The author was supported by grants NS-1216.2012.1 and RFFI 10-01-00242-a.

References

1. P. Billingsley, The Lindeberg-Lévy theorem for martingales. Proc. Am. Math. Soc. **12**, 788–792 (1961)
2. Y. Derriennic, M. Lin, Sur le théorème limite central de Kipnis et Varadhan pour les chaînes réversibles ou normales. C. R. Acad. Sci. Paris Sér. I Math. **323** (9), 1053–1057 (1996)
3. Y. Derriennic, M. Lin, Fractional Poisson equations and ergodic theorems for fractional coboundaries. Israel J. Math. **123**, 93–130 (2001)
4. Y. Derriennic, M. Lin, The central limit theorem for Markov chains with normal transition operators, started at a point. Probab. Theor. Relat. Fields **119**(4), 508–528 (2001)

5. S.R. Foguel, On order preserving contractions. Israel J. Math. **1**, 54–59 (1963)
6. S.R. Foguel, *The Ergodic Theory of Markov Processes.* Van Nostrand Mathematical Studies, No. 21 (Van Nostrand Reinhold Co., New York, 1969)
7. M.I. Gordin, On the central limit theorem for stationary processes (Russian). Dokl. Akad. Nauk SSSR **188**(4), 739–741; Transl.: Soviet Math. Dokl. **10**, 1174–1176 (1969)
8. M.I. Gordin, B.A. Lifshits, Central limit theorem for stationary Markov processes (Russian). Dokl. Akad. Nauk SSSR **239**, 766–767. Transl.: Soviet Math. Dokl. **19**(2), 392–394 (1978)
9. M. Gordin, B. Lifshits, *A Remark On a Markov Process with Normal Transition Operator* (Russian). Third International Conference on Probability Theory and Mathematical Statistics, June 1981, Vilnius. Abstracts of Communications, vol. 1: A-K (1981), pp. 147–148
10. M.I. Gordin, B.A. Lifshits, The central limit theorem for Markov processes with normal transition operator and applications to random walks on compact Abelian groups. Sect. IV.7–IV.9 in A.N. Borodin, I.A. Ibragimov: Limit theorems for functionals of random walks (Russian). Trudy Mat. Inst. Steklov, vol. **195** (1994); transl. Proc. Steklov Inst. Math. **195** (American Mathematical Society, Providence, 1995)
11. I.A. Ibragimov, A central limit theorem for a class of dependent random variables (Russian). Teor. Verojatnost. i Primenen. **8**, 89–94 (1963)
12. C. Kipnis, S.R.S. Varadhan, Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Comm. Math. Phys. **104**(1), 1–19 (1986)
13. N. Maigret, Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. Ann. Inst. H. Poincaré Sect. B (N.S.), **14**(4), 425–440 (1978)
14. M. Maxwell, M. Woodroofe, Central limit theorems for additive functionals of Markov chains. Ann. Probab. **28**(2), 713–724 (2000)
15. M. Peligrad, S. Utev, A new maximal inequality and invariance principle for stationary sequences. Ann. Probab. **33**(2), 798–815 (2005)
16. M. Peligrad, S. Utev, W.B. Wu, A maximal $\mathbb{L}_p$-inequality for stationary sequences and its applications. Proc. Am. Math. Soc. **135**(2), 541–550 (2007) (electronic)
17. B. Sz.-Nagy, C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space.* Translated from the French and revised (North-Holland, Amsterdam, 1970)
18. B. Sz.-Nagy, *Unitary Dilations of Hilbert Space Operators and Related Topics.* Expository Lectures from the CBMS Regional Conference held at the University of New Hampshire, Durham, NH, June 7–11, 1971; Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics, No. 19 (American Mathematical Society, Providence, 1974)
19. O. Zhao, M. Woodroofe, On martingale approximations. Ann. Appl. Probab. **18**(5), 1831–1847 (2008)

Operator-Valued and Multivariate Free Berry-Esseen Theorems

Tobias Mai and Roland Speicher

Dedicated to Professor Friedrich Götze on the occasion of his 60th birthday

Abstract We address the question of a Berry-Esseen type theorem for the speed of convergence in a multivariate free central limit theorem. For this, we estimate the difference between the operator-valued Cauchy transforms of the normalized partial sums in an operator-valued free central limit theorem and the Cauchy transform of the limiting operator-valued semicircular element. Since we have to deal with in general non-self-adjoint operators, we introduce the notion of matrix-valued resolvent sets and study the behavior of Cauchy transforms on them.

Keywords Free Berry-Esseen • operator valued • multivariate • linearization trick • matrix valued spectrum

2010 *Mathematics Subject Classification.* 46L54, 60F05, 47A10.

1 Introduction

In classical probability theory the famous Berry-Esseen theorem gives a quantitative statement about the order of convergence in the central limit theorem. It states in its simplest version: If $(X_i)_{i\in\mathbb{N}}$ is a sequence of independent and identically distributed random variables with mean 0 and variance 1, then the distance between

T. Mai · R. Speicher (✉)
Saarland University, Fachbereich Mathematik, Postfach 151150, 66041 Saarbrücken, Germany
e-mail: mai@math.uni-sb.de; speicher@math.uni-sb.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_7,

$S_n := \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ and a normal variable γ of mean 0 and variance 1 can be estimated in terms of the Kolmogorov distance Δ by

$$\Delta(S_n, \gamma) \leq C \frac{1}{\sqrt{n}} \rho,$$

where C is a constant and ρ is the absolute third moment of the variables X_i. The question for a free analogue of the Berry-Esseen estimate in the case of one random variable was answered by Chistyakov and Götze in [2] (and independently, under the more restrictive assumption of compact support of the X_i, by Kargin [10]): If $(X_i)_{i \in \mathbb{N}}$ is a sequence of free and identically distributed variables with mean 0 and variance 1, then the distance between $S_n := \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ and a semicircular variable s of mean 0 and variance 1 can be estimated as

$$\Delta(S_n, s) \leq c \frac{|m_3| + \sqrt{m_4}}{\sqrt{n}},$$

where $c > 0$ is an absolute constant and m_3 and m_4 are the third and fourth moment, respectively, of the X_i.

In this paper we want to present an approach to a multivariate version of a free Berry-Esseen theorem. The general idea is the following: Since there is up to now no suitable replacement of the Kolmorgorov metric in the multivariate case, we will, in order to describe the speed of convergence of a d-tuple $(S_n^{(1)}, \ldots, S_n^{(d)})$ of partial sums to the limiting semicircular family $(s_1, \ldots, s_d)$, consider the speed of convergence of $p(S_n^{(1)}, \ldots, S_n^{(d)})$ to $p(s_1, \ldots, s_d)$ for any self-adjoint polynomial p in d non-commuting variables. By using the linearization trick of Haagerup and Thorbjørnsen [5,6], we can reformulate this in an operator-valued setting, where we will state an operator-valued free Berry-Esseen theorem. Because estimates for the difference between scalar-valued Cauchy transforms translate by results of Bai [1] to estimates with respect to the Kolmogorov distance, it is convenient to describe the speed of convergence in terms of Cauchy transforms. On the level of deriving equations for the (operator-valued) Cauchy transforms we can follow ideas which are used for dealing with speed of convergence questions for random matrices; here we are inspired in particular by the work of Götze and Tikhomirov [4], but see also [1].

Since the transition from the multivariate to the operator-valued setting leads to operators which are, even if we start from self-adjoint polynomials p, in general not self-adjoint, we have to deal with (operator-valued) Cauchy transforms defined on domains different from the usual ones. Since most of the analytic tools fail in this generality, we have to develop them along the way.

As a first step in this direction, the present paper (which is based on the unpublished preprint [13]) leads finally to the proof of the following theorem:

Theorem 1.1. *Let $(\mathcal{C}, \tau)$ be a non-commutative C^*-probability space with τ faithful and put $\mathcal{A} := \mathrm{M}_m(\mathbb{C}) \otimes \mathcal{C}$ and $E := \mathrm{id} \otimes \tau$. Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of non-zero elements in the operator-valued probability space $(\mathcal{A}, E)$. We assume:*

- *All X_i's have the same $*$-distribution with respect to E and their first moments vanish, i.e. $E[X_i] = 0$.*
- *The X_i are $*$-free with amalgamation over $\mathrm{M}_m(\mathbb{C})$ (which means that the $*$-algebras $\mathcal{X}_i$, generated by $\mathrm{M}_m(\mathbb{C})$ and X_i, are free with respect to E).*
- *We have $\sup_{i\in\mathbb{N}} \|X_i\| < \infty$.*

Then the sequence $(S_n)_{n\in\mathbb{N}}$ defined by

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \qquad n \in \mathbb{N}$$

converges to an operator-valued semicircular element s. Moreover, we can find $\kappa > 0$, $c > 1$, $C > 0$ and $N \in \mathbb{N}$ such that

$$\|G_s(b) - G_{S_n}(b)\| \leq C \frac{1}{\sqrt{n}} \|b\| \qquad \text{for all } b \in \Omega \text{ and } n \geq N,$$

where

$$\Omega := \left\{ b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa,\ \|b\| \cdot \|b^{-1}\| < c \right\}$$

and where G_s and G_{S_n} denote the operator-valued Cauchy transforms of s and of S_n, respectively.

Applying this operator-valued statement to our multivariate problem gives the following main result on a multivariate free Berry Esseen theorem.

Theorem 1.2. *Let $(x_i^{(k)})_{k=1}^d$, $i \in \mathbb{N}$, be free and identically distributed sets of d self-adjoint non-zero random variables in some non-commutative C^*-probability space $(\mathcal{C}, \tau)$, with τ faithful, such that the conditions*

$$\tau(x_i^{(k)}) = 0 \qquad \text{for } k = 1, \ldots, d \text{ and all } i \in \mathbb{N}$$

and

$$\sup_{i\in\mathbb{N}} \max_{k=1,\ldots,d} \|x_i^{(k)}\| < \infty$$

are fulfilled. We denote by $\Sigma = (\sigma_{k,l})_{k,l=1}^d$, where $\sigma_{k,l} := \tau(x_i^{(k)} x_i^{(l)})$, their joint covariance matrix. Moreover, we put

$$S_n^{(k)} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i^{(k)} \qquad \text{for } k = 1, \ldots, d \text{ and all } n \in \mathbb{N}.$$

Then $(S_n^{(1)}, \ldots, S_n^{(d)})$ converges in distribution to a semicircular family $(s_1, \ldots, s_d)$ of covariance Σ. We can quantify the speed of convergence in the following way. Let p be a (not necessarily self-adjoint) polynomial in d non-commutating variables and put

$$P_n := p(S_n^{(1)}, \ldots, S_n^{(d)}) \qquad \text{and} \qquad P := p(s_1, \ldots, s_d).$$

Then, there are constants $C > 0$, $R > 0$ and $N \in \mathbb{N}$ (depending on the polynomial) such that

$$|G_P(z) - G_{P_n}(z)| \leq C \frac{1}{\sqrt{n}} \qquad \text{for all } |z| > R \text{ and } n \geq N,$$

where G_P and G_{P_n} denote the scalar-valued Cauchy transform of P and of P_n, respectively.

In the case of a self-adjoint polynomial p, we can consider the distribution measures μ_n and μ of the operators P_n and P from above, which are probability measures on $\mathbb{R}$. Moreover, let $\mathcal{F}_{\mu_n}$ and $\mathcal{F}_{\mu}$ be their cumulative distribution functions. In order to deduce estimates for the Kolmogorov distance

$$\Delta(\mu_n, \mu) = \sup_{x \in \mathbb{R}} |\mathcal{F}_{\mu_n}(x) - \mathcal{F}_{\mu}(x)|$$

one has to transfer the estimate for the difference of the scalar-valued Cauchy transforms of P_n and P from near infinity to a neighborhood of the real axis. A partial solution to this problem was given in the appendix of [14], which we will recall in Sect. 4. But this leads to the still unsolved question, whether $p(s_1, \ldots, s_d)$ has a continuous density. We conjecture that the latter is true for any self-adjoint polynomial in free semicirculars, but at present we are not aware of a proof of that statement.

The paper is organized as follows. In Sect. 2 we recall some basic facts about holomorphic functions on domains in Banach spaces. The tools to deal with matrix-valued Cauchy transform will be presented in Sect. 3. Section 4 is devoted to the proof of Theorems 1.1 and 1.2.

2 Holomorphic Functions on Domains in Banach Spaces

For reader's convenience, we briefly recall the definition of holomorphic functions on domains in Banach spaces and we state the theorem of Earle-Hamilton, which will play a major role in the subsequent sections.

Definition 2.1. Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be two complex Banach spaces and let $D \subseteq X$ be an open subset of X. A function $f : D \to Y$ is called

- **Strongly holomorphic**, if for each $x \in D$ there exists a bounded linear mapping $Df(x) : X \to Y$ such that

$$\lim_{y \to 0} \frac{\|f(x+y) - f(x) - Df(x)y\|_Y}{\|y\|_X} = 0.$$

- **Weakly holomorphic**, if it is locally bounded and the mapping

$$\lambda \mapsto \phi(f(x + \lambda y))$$

is holomorphic at $\lambda = 0$ for each $x \in D$, $y \in Y$ and all continuous linear functionals $\phi : Y \to \mathbb{C}$.

An important theorem due to Dunford says, that a function on a domain (i.e. an open and connected subset) in a Banach space is strongly holomorphic if and only if it is weakly holomorphic. Hence, we do not have to distinguish between both definitions.

Definition 2.2. Let D be a nonempty domain in a complex Banach space $(X, \|\cdot\|)$ and let $f : D \to D$ be a holomorphic function. We say, that $f(D)$ **lies strictly inside** D, if there is some $\epsilon > 0$ such that

$$B_\epsilon(f(x)) \subseteq D \qquad \text{for all } x \in D$$

holds, whereby we denote by $B_r(y)$ the open ball with radius r around y.

The remarkable fact, that strict holomorphic mappings are strict contractions in the so-called Carathéodory-Riffen-Finsler metric, leads to the following theorem of Earle-Hamilton (cf. [3]), which can be seen as a holomorphic version of Banach's contraction mapping theorem. For a proof of this theorem and variations of the statement we refer to [7].

Theorem 2.3 (Earle-Hamilton, 1970). *Let $\emptyset \neq D \subseteq X$ be a domain in a Banach space $(X, \|\cdot\|)$ and let $f : D \to D$ be a bounded holomorphic function. If $f(D)$ lies strictly inside D, then f has a unique fixed point in D.*

3 Matrix-Valued Spectra and Cauchy Transforms

The statement of the following lemma is well-known and quite simple. But since it turns out to be extremely helpful, it is convenient to recall it here.

Lemma 3.1. *Let $(A, \|\cdot\|)$ be a complex Banach-algebra with unit 1. If $x \in A$ is invertible and $y \in A$ satisfies $\|x - y\| < \sigma \frac{1}{\|x^{-1}\|}$ for some $0 < \sigma < 1$, then y is invertible as well and we have*

$$\|y^{-1}\| \leq \frac{1}{1-\sigma}\|x^{-1}\|.$$

Proof. We can easily check that

$$\sum_{n=0}^{\infty} \left(x^{-1}(x-y)\right)^n x^{-1}$$

is absolutely convergent in A and gives the inverse element of y. Moreover we get

$$\|y^{-1}\| \leq \sum_{n=0}^{\infty} \left(\|x^{-1}\|\|x-y\|\right)^n \|x^{-1}\| < \frac{1}{1-\sigma}\|x^{-1}\|,$$

which proves the stated estimate. □

Let $(\mathcal{C}, \tau)$ be a non-commutative C^*-probability space, i.e., $\mathcal{C}$ is a unital C^*-algebra and τ is a unital state (positive linear functional) on $\mathcal{C}$; we will always assume that τ is faithful. For fixed $m \in \mathbb{N}$ we define the operator-valued C^*-probability space $\mathcal{A} := \mathrm{M}_m(\mathbb{C}) \otimes \mathcal{C}$ with conditional expectation

$$E := \mathrm{id}_m \otimes \tau : \ \mathcal{A} \to \mathrm{M}_m(\mathbb{C}), \ b \otimes c \mapsto \tau(c)b,$$

where we denote by $\mathrm{M}_m(\mathbb{C})$ the C^*-algebra of all $m \times m$ matrices over the complex numbers $\mathbb{C}$. Under the canonical identification of $\mathrm{M}_m(\mathbb{C}) \otimes \mathcal{C}$ with $\mathrm{M}_m(\mathcal{C})$ (matrices with entries in $\mathcal{C}$), the expectation E corresponds to applying the state τ entrywise in a matrix. We will also identify $b \in \mathrm{M}_m(\mathbb{C})$ with $b \otimes 1 \in \mathcal{A}$.

Definition 3.2. For $a \in \mathcal{A} = \mathrm{M}_m(\mathcal{C})$ we define the **matrix-valued resolvent set**

$$\rho_m(a) := \{b \in \mathrm{M}_m(\mathbb{C}) \mid b - a \text{ is invertible in } \mathcal{A}\}$$

and the **matrix-valued spectrum**

$$\sigma_m(a) := \mathrm{M}_m(\mathbb{C}) \backslash \rho_m(a).$$

Since the set $\mathrm{GL}(\mathcal{A})$ of all invertible elements in $\mathcal{A}$ is an open subset of $\mathcal{A}$ (cf. Lemma 3.1), the continuity of the mapping

$$f_a : \ \mathrm{M}_m(\mathbb{C}) \to \mathcal{A}, \ b \mapsto b - a$$

implies, that the matrix-valued resolvent set $\rho_m(a) = f_a^{-1}(\mathrm{GL}(\mathcal{A}))$ of an element $a \in \mathcal{A}$ is an open subset of $\mathrm{M}_m(\mathbb{C})$. Hence, the matrix-valued spectrum $\sigma_m(a)$ is always closed.

Although the behavior of this matrix-valued generalizations of the classical resolvent set and spectrum seems to be quite similar to the classical case (which is of course included in our definition for $m = 1$), the matrix valued spectrum is in general not bounded and hence not a compact subset of $\mathrm{M}_m(\mathbb{C})$. For example, we have for all $\lambda \in \mathbb{C}$, that

$$\sigma_m(\lambda 1) = \{b \in \mathrm{M}_m(\mathbb{C}) \mid \lambda \in \sigma_{\mathrm{M}_m(\mathbb{C})}(b)\},$$

i.e. $\sigma_m(\lambda 1)$ consists of all matrices $b \in \mathrm{M}_m(\mathbb{C})$ for which λ belongs to the spectrum $\sigma_{\mathrm{M}_m(\mathbb{C})}(b)$. Particularly, $\sigma_m(\lambda 1)$ is unbounded for $m \geq 2$.

In the following, we denote by $\mathrm{GL}_m(\mathbb{C}) := \mathrm{GL}(\mathrm{M}_m(\mathbb{C}))$ the set of all invertible matrices in $\mathrm{M}_m(\mathbb{C})$.

Lemma 3.3. *Let $a \in \mathcal{A}$ be given. Then for all $b \in \mathrm{GL}_m(\mathbb{C})$ the following inclusion holds:*

$$\left\{\lambda b \mid \lambda \in \rho_{\mathcal{A}}(b^{-1}a)\right\} \subseteq \rho_m(a)$$

Proof. Let $\lambda \in \rho_{\mathcal{A}}(b^{-1}a)$ be given. By definition of the usual resolvent set this means that $\lambda 1 - b^{-1}a$ is invertible in $\mathcal{A}$. It follows, that

$$\lambda b - a = b\left(\lambda 1 - b^{-1}a\right)$$

is invertible as well, and we get, as desired, $\lambda b \in \rho_m(a)$. □

Lemma 3.4. *For all $0 \neq a \in \mathcal{A}$ we have*

$$\left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \frac{1}{\|a\|}\right\} \subseteq \rho_m(a)$$

and

$$\sigma_m(a) \cap \mathrm{GL}_m(\mathbb{C}) \subseteq \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| \geq \frac{1}{\|a\|}\right\}.$$

Proof. Obviously, the second inclusion is a direct consequence of the first. Hence, it suffices to show the first statement.

Let $b \in \mathrm{GL}_m(\mathbb{C})$ with $\|b^{-1}\| < \frac{1}{\|a\|}$ be given. It follows, that $h := 1 - b^{-1}a$ is invertible, because

$$\|1 - h\| = \|b^{-1}a\| \leq \|b^{-1}\| \cdot \|a\| < 1.$$

Therefore, we can deduce, that also

$$b - a = b\left(1 - b^{-1}a\right) \tag{1}$$

is invertible, i.e. $b \in \rho_m(a)$. This proves the assertion. □

The main reason to consider matrix-valued resolvent sets is, that they are the natural domains for matrix-valued Cauchy transforms, which we will define now.

Definition 3.5. For $a \in \mathcal{A}$ we call

$$G_a : \rho_m(a) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto E\left[(b-a)^{-1}\right]$$

the **matrix-valued Cauchy transform** of a.

Note that G_a is a continuous function (and hence locally bounded) and induces for all $b_0 \in \rho_m(a)$, $b \in \mathrm{M}_m(\mathbb{C})$ and bounded linear functionals $\phi : \mathcal{A} \to \mathbb{C}$ a function

$$\lambda \mapsto \phi\big(G_a(b_0 + \lambda b)\big),$$

which is holomorphic in a neighborhood of $\lambda = 0$. Hence, G_a is weakly holomorphic and therefore (as we have seen in the previous section) strongly holomorphic as well.

Because the structure of $\rho_m(a)$ and therefore the behavior of G_a might in general be quite complicated, we restrict our attention to a suitable restriction of G_a. In this way, we will get some additional properties of G_a.

The first restriction enables us to control the norm of the matrix-valued Cauchy transform on a sufficiently nice subset of the matrix-valued resolvent set.

Lemma 3.6. *Let $0 \neq a \in \mathcal{A}$ be given. For $0 < \theta < 1$ the matrix valued Cauchy transform G_a induces a mapping*

$$G_a : \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \theta \cdot \frac{1}{\|a\|}\right\} \to \left\{b \in \mathrm{M}_m(\mathbb{C}) \mid \|b\| < \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|}\right\}.$$

Proof. Lemma 3.4 (c) tells us, that the open set

$$U := \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \theta \cdot \frac{1}{\|a\|}\right\}$$

is contained in $\rho_m(a)$, i.e. G_a is well-defined on U. Moreover, we get from (1)

$$(b-a)^{-1} = \big(1 - b^{-1}a\big)^{-1} b^{-1} = \sum_{n=0}^{\infty} \big(b^{-1}a\big)^n b^{-1}$$

and hence

$$\|G_a(b)\| \leq \|(b-a)^{-1}\| \leq \|b^{-1}\| \sum_{n=0}^{\infty} \big(\|b^{-1}\|\|a\|\big)^n < \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|} \tag{2}$$

for all $b \in U$. This proves the claim. □

To ensure, that the range of G_a is contained in $\mathrm{GL}_m(\mathbb{C})$, we have to shrink the domain again.

Lemma 3.7. *Let $0 \neq a \in \mathcal{A}$ be given. For $0 < \theta < 1$ and $c > 1$ we define*

$$\Omega := \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \theta \cdot \frac{1}{\|a\|},\ \|b\| \cdot \|b^{-1}\| < c\right\}$$

and

$$\Omega' := \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b\| < \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|}\right\}.$$

If the condition

$$\frac{\theta}{1-\theta} < \frac{\sigma}{c}$$

is satisfied for some $0 < \sigma < 1$*, then the matrix-valued Cauchy transform* G_a *induces a mapping* $G_a : \Omega \to \Omega'$ *and we have the estimates*

$$\|G_a(b)\| \leq \|(b-a)^{-1}\| < \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|} \qquad \text{for all } b \in \Omega \tag{3}$$

and

$$\|G_a(b)^{-1}\| < \frac{1}{1-\sigma} \cdot \|b\| \qquad \text{for all } b \in \Omega. \tag{4}$$

Proof. For all $b \in \Omega$ we have

$$G_a(b) - b^{-1} = E\big[(b-a)^{-1} - b^{-1}\big] = E\Big[\sum_{n=1}^{\infty} \big(b^{-1}a\big)^n b^{-1}\Big],$$

which enables us to deduce

$$\|G_a(b) - b^{-1}\| \leq \|b^{-1}\| \sum_{n=1}^{\infty} \big(\|b^{-1}\|\|a\|\big)^n \leq \frac{\theta}{1-\theta} \cdot \|b^{-1}\| < \frac{\theta}{1-\theta} \cdot \frac{c}{\|b\|} < \sigma \cdot \frac{1}{\|b\|}.$$

Using Lemma 3.1, this implies $G_a(b) \in \mathrm{GL}_m(\mathbb{C})$ and (4). Since we already know from (2) in Lemma 3.6, that (3) holds, it follows $G_a(b) \in \Omega'$ and the proof is complete. □

Remark 3.8. Since domains of our holomorphic functions should be connected it is necessary to note, that for $\kappa > 0$ and $c > 1$

$$\Omega = \big\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa,\ \|b\| \cdot \|b^{-1}\| < c\big\}$$

and for $r > 0$

$$\Omega' = \big\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b\| < r\big\}$$

are pathwise connected subsets of $\mathrm{M}_m(\mathbb{C})$. Indeed, if $b_1, b_2 \in \mathrm{GL}_m(\mathbb{C})$ are given, we consider their polar decomposition $b_1 = U_1 P_1$ and $b_2 = U_2 P_2$ with unitary matrices $U_1, U_2 \in \mathrm{GL}_m(\mathbb{C})$ and positive-definite Hermitian matrices $P_1, P_2 \in \mathrm{GL}_m(\mathbb{C})$ and define (using functional calculus for normal elements in the C^*-algebra $\mathrm{M}_m(\mathbb{C})$)

$$\gamma :\ [0,1] \to \mathrm{GL}_m(\mathbb{C}),\ t \mapsto U_1^{1-t} P_1^{1-t} U_2^t P_2^t.$$

Then γ fulfills $\gamma(0) = b_1$ and $\gamma(1) = b_2$, and $\gamma([0,1])$ is contained in Ω and Ω' if b_1, b_2 are elements of Ω and Ω', respectively.

Since the matrix-valued Cauchy transform is a solution of a special equation (cf. [8, 12]), we will be interested in the following situation:

Corollary 3.9. *Let $\eta : \mathrm{GL}_m(\mathbb{C}) \to \mathrm{M}_m(\mathbb{C})$ be a holomorphic function satisfying*

$$\|\eta(w)\| \leq M\|w\| \qquad \text{for all } w \in \mathrm{GL}_m(\mathbb{C})$$

for some $M > 0$. Moreover, we assume that

$$bG_a(b) = 1 + \eta(G_a(b))G_a(b) \qquad \text{for all } b \in \Omega$$

holds. Let $0 < \theta, \sigma < 1$ and $c > 1$ be given with

$$\frac{\theta}{1-\theta} < \sigma \min\left\{\frac{1}{c}, \frac{\|a\|^2}{M}\right\}$$

and let Ω and Ω' be as in Lemma 3.7.

Then, for fixed $b \in \Omega$, the equation

$$bw = 1 + \eta(w)w, \qquad w \in \Omega' \tag{5}$$

has a unique solution, which is given by $w = G_a(b)$.

Proof. Let $b \in \Omega$ be given. For all $w \in \Omega'$ we get

$$\|\eta(w)\| \leq M\|w\| \leq \frac{\theta}{1-\theta} \cdot \frac{M}{\|a\|}$$

and therefore

$$\|b^{-1}\eta(w)\| \leq \|b^{-1}\|\|\eta(w)\| \leq \frac{\theta}{1-\theta} \cdot \frac{M}{\|a\|^2} \cdot \theta < \theta\sigma < 1.$$

This means, that $1 - b^{-1}\eta(w)$ and hence $b - \eta(w)$ is invertible with

$$\begin{aligned}\|(b-\eta(w))^{-1}\| &\leq \|b^{-1}\|\|(1-b^{-1}\eta(w))^{-1}\| \\ &\leq \|b^{-1}\| \sum_{n=0}^{\infty} \|b^{-1}\eta(w)\|^n \\ &< \frac{\theta}{1-\theta\sigma} \cdot \frac{1}{\|a\|},\end{aligned}$$

and shows, that we have a well-defined and holomorphic mapping

$$\mathcal{F} : \Omega' \to \mathrm{M}_m(\mathbb{C}),\ w \mapsto (b - \eta(w))^{-1}$$

with

$$\|\mathcal{F}(w)\| = \|(b - \eta(w))^{-1}\| < \frac{\theta}{1-\theta\sigma} \cdot \frac{1}{\|a\|} < \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|}$$

and therefore $\mathcal{F}(w) \in \Omega'$.

Now, we want to show that $\mathcal{F}(\Omega')$ lies strictly inside Ω'. We put

$$\epsilon := \min\left\{\frac{1}{2} \cdot \frac{1}{\|b\| + \sigma\|a\|},\ \left(1 - \frac{1-\theta}{1-\theta\sigma}\right) \cdot \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|}\right\} > 0$$

and consider $w \in \Omega'$ and $u \in \mathrm{M}_m(\mathbb{C})$ with $\|u - \mathcal{F}(w)\| < \epsilon$. At first, we get

$$\|b - \eta(w)\| \leq \|b\| + \|\eta(w)\| \leq \|b\| + \frac{M}{\|a\|} \cdot \frac{\theta}{1-\theta} \leq \|b\| + \sigma\|a\|$$

and thus

$$\|u - (b - \eta(w))^{-1}\| = \|u - \mathcal{F}(w)\| < \epsilon \leq \frac{1}{2} \cdot \frac{1}{\|b\| + \sigma\|a\|} \leq \frac{1}{2} \cdot \frac{1}{\|b - \eta(w)\|},$$

which shows $u \in \mathrm{GL}_m(\mathbb{C})$, and secondly

$$\begin{aligned}\|u\| &= \|u - (b - \eta(w))^{-1}\| + \|\mathcal{F}(w)\| \\ &< \epsilon + \frac{1-\theta}{1-\theta\sigma} \cdot \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|} \\ &< \frac{\theta}{1-\theta} \cdot \frac{1}{\|a\|}\end{aligned}$$

which shows $u \in \Omega'$.

Let now $w \in \Omega'$ be a solution of (5). This implies that

$$w^{-1}\mathcal{F}(w) = w^{-1}(b - \eta(w))^{-1} = \big(bw - \eta(w)w\big)^{-1} = 1,$$

and hence $\mathcal{F}(w) = w$. Since $\mathcal{F} : \Omega' \to \Omega'$ is holomorphic on the domain Ω' and $\mathcal{F}(\Omega')$ lies strictly inside Ω', it follows by the Theorem of Earle-Hamilton, Theorem 2.3, that $\mathcal{F}$ has exactly one fixed point. Because $G_a(b)$ (which is an element of Ω' by Lemma 3.7) solves (5) by assumption and hence is already a fixed point of $\mathcal{F}$, it follows $w = G_a(b)$ and we are done. □

Remark 3.10. Let $(\mathcal{A}', E')$ be an arbitrary operator-valued C^*-probability space with conditional expectation $E' : \mathcal{A}' \to \mathrm{M}_m(\mathbb{C})$. This provides us with a unital (and continuous) $*$-embedding $\iota : \mathrm{M}_m(\mathbb{C}) \to \mathcal{A}'$. In this section, we only considered the special embedding

$$\iota : \mathrm{M}_m(\mathbb{C}) \to \mathcal{A},\ b \mapsto b \otimes 1,$$

which is given by the special structure $\mathcal{A} = \mathrm{M}_m(\mathbb{C}) \otimes \mathcal{C}$. But we can define matrix-valued resolvent sets, spectra and Cauchy transforms also in this more general framework. To be more precise, we put for all $a \in \mathcal{A}'$

$$\rho_m(a) := \{b \in \mathrm{M}_m(\mathbb{C}) \mid \iota(b) - a \text{ is invertible in } \mathcal{A}'\}$$

and $\sigma_m(a) := \mathrm{M}_m(\mathbb{C}) \backslash \rho_m(a)$ and

$$G_a : \rho_m(a) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto E'\big[(\iota(b) - a)^{-1}\big].$$

We note, that all the results of this section stay valid in this general situation.

4 Multivariate Free Central Limit Theorem

4.1 Setting and First Observations

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence in the operator-valued probability space $(\mathcal{A}, E)$ with $\mathcal{A} = \mathrm{M}_m(\mathcal{C}) = \mathrm{M}_m(\mathbb{C}) \otimes \mathcal{C}$ and $E = \mathrm{id} \otimes \tau$, as defined in the previous section. We assume:

- All X_i's have the same $*$-distribution with respect to E and their first moments vanish, i.e. $E[X_i] = 0$.
- The X_i are $*$-free with amalgamation over $\mathrm{M}_m(\mathbb{C})$ (which means that the $*$-algebras $\mathcal{X}_i$, generated by $\mathrm{M}_m(\mathbb{C})$ and X_i, are free with respect to E).
- We have $\sup_{i\in\mathbb{N}} \|X_i\| < \infty$.

If we define the linear (and hence holomorphic) mapping

$$\eta : \mathrm{M}_m(\mathbb{C}) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto E[X_i b X_i],$$

we easily get from the continuity of E, that

$$\|\eta(b)\| \leq \Big(\sup_{i\in\mathbb{N}} \|X_i\|\Big)^2 \|b\| \qquad \text{for all } b \in \mathrm{M}_m(\mathbb{C})$$

holds. Hence we can find $M > 0$ such that $\|\eta(b)\| < M\|b\|$ holds for all $b \in \mathrm{M}_m(\mathbb{C})$. Moreover, we have for all $k \in \mathbb{N}$ and all $b_1, \ldots, b_k \in \mathrm{M}_m(\mathbb{C})$

$$\sup_{i\in\mathbb{N}} \|E[X_i b_1 X_i \ldots b_k X_i]\| \leq \Big(\sup_{i\in\mathbb{N}} \|X_i\|\Big)^{k+1} \|b_1\| \cdots \|b_k\|.$$

Since $(X_i)_{i\in\mathbb{N}}$ is a sequence of centered free non-commutative random variables, Theorem 8.4 in [15] tells us that the sequence $(S_n)_{n\in\mathbb{N}}$ defined by

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \qquad n \in \mathbb{N}$$

converges to an operator-valued semicircular element s. Moreover, we know from Theorem 4.2.4 in [12] that the operator-valued Cauchy transform G_s satisfies

$$bG_s(b) = 1 + \eta(G_s(b))G_s(b) \qquad \text{for all } b \in U_r,$$

where we put $U_r := \{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < r\} \subseteq \rho_m(s)$ for all suitably small $r > 0$.

By Proposition 7.1 in [9], the boundedness of the sequence $(X_i)_{i\in\mathbb{N}}$ guarantees boundedness of $(S_n)_{n\in\mathbb{N}}$ as well. In order to get estimates for the difference between the Cauchy transforms G_s and G_{S_n} we will also need the fact, that $(S_n)_{n\in\mathbb{N}}$ is bounded away from 0. The precise statement is part of the following lemma, which also includes a similar statement for

$$S_n^{[i]} := S_n - \frac{1}{\sqrt{n}} X_i = \frac{1}{\sqrt{n}} \sum_{\substack{j=1 \\ j\neq i}}^{n} X_j \qquad \text{for all } n \in \mathbb{N} \text{ and } 1 \leq i \leq n.$$

Lemma 4.1. *In the situation described above, we have for all $n \in \mathbb{N}$ and all $1 \leq i \leq n$*

$$\|S_n\| \geq \|\alpha\|^{\frac{1}{2}} \qquad \textit{and} \qquad \|S_n^{[i]}\| \geq \sqrt{1 - \frac{1}{n}}\|\alpha\|^{\frac{1}{2}},$$

where $\alpha := E[X_i^* X_i] \in \mathrm{M}_m(\mathbb{C})$.

Proof. By the $*$-freeness of $X_1, X_2, \ldots$, we have

$$E[X_i^* X_j] = E[X_i^*] \cdot E[X_j] = 0, \qquad \text{for } i \neq j$$

and thus

$$\|S_n\|^2 = \|S_n^* S_n\| \geq \|E[S_n^* S_n]\| = \frac{1}{n} \left\| \sum_{i,j=1}^{n} E[X_i^* X_j] \right\| = \|\alpha\|.$$

Similarly

$$\begin{aligned} \|S_n^{[i]}\|^2 &= \|(S_n^{[i]})^* S_n^{[i]}\| \\ &\geq \|E[(S_n^{[i]})^* S_n^{[i]}]\| \\ &= \left\| E[S_n^* S_n] - \frac{1}{n} E[X_i^* X_i] \right\| \\ &= \frac{n-1}{n}\|\alpha\|, \end{aligned}$$

which proves the statement. □

We define for $n \in \mathbb{N}$

$$R_n : \rho_m(S_n) \to \mathcal{A},\ b \mapsto \big(b - S_n\big)^{-1}$$

and for $n \in \mathbb{N}$ and $1 \leq i \leq n$

$$R_n^{[i]} : \rho_m(S_n^{[i]}) \to \mathcal{A},\ b \mapsto \big(b - S_n^{[i]}\big)^{-1}.$$

Lemma 4.2. *For all $n \in \mathbb{N}$ and $1 \leq i \leq n$ we have*

$$R_n(b) = R_n^{[i]}(b) + \frac{1}{\sqrt{n}} R_n^{[i]}(b) X_i R_n^{[i]}(b) + \frac{1}{n} R_n(b) X_i R_n^{[i]}(b) X_i R_n^{[i]}(b) \tag{6}$$

and

$$R_n(b) = R_n^{[i]}(b) + \frac{1}{\sqrt{n}} R_n^{[i]}(b) X_i R_n(b) \tag{7}$$

for all $b \in \rho_m(S_n) \cap \rho_m(S_n^{[i]})$.

Proof. We have

$$\begin{aligned}
\big(b - S_n\big) R_n(b) \big(b - S_n^{[i]}\big) &= b - S_n^{[i]} \\
&= \big(b - S_n\big) + \frac{1}{\sqrt{n}} \big(b - S_n^{[i]}\big) R_n^{[i]}(b) X_i \\
&= \big(b - S_n\big) + \frac{1}{\sqrt{n}} \big(b - S_n\big) R_n^{[i]}(b) X_i + \frac{1}{n} X_i R_n^{[i]}(b) X_i,
\end{aligned}$$

which leads, by multiplication with $R_n(b) = (b - S_n)^{-1}$ from the left and with $R_n^{[i]}(b) = (b - S_n^{[i]})^{-1}$ from the right, to (6).

Moreover, we have

$$\big(b - S_n^{[i]}\big) R_n(b) \big(b - S_n\big) = b - S_n^{[i]} = \big(b - S_n\big) + \frac{1}{\sqrt{n}} X_i,$$

which leads, by multiplication with $R_n(b) = (b - S_n)^{-1}$ from the right and with $R_n^{[i]}(b) = (b - S_n^{[i]})^{-1}$ from the left, to equation (7). □

Obviously, we have

$$G_n := G_{S_n} = E \circ R_n \qquad \text{and} \qquad G_n^{[i]} := G_{S_n^{[i]}} = E \circ R_n^{[i]}.$$

4.2 Proof of the Main Theorem

During this subsection, let $0 < \theta, \sigma < 1$ and $c > 1$ be given, such that

$$\frac{\theta}{1-\theta} < \sigma \min\left\{\frac{1}{c}, \frac{\|\alpha\|}{M}\right\} \tag{8}$$

holds. For all $n \in \mathbb{N}$ we define

$$\kappa_n := \theta \min\left\{\frac{1}{\|s\|}, \frac{1}{\|S_n\|}, \frac{1}{\|S_n^{[1]}\|}, \ldots, \frac{1}{\|S_n^{[n]}\|}\right\}$$

and

$$\Omega_n := \left\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa_n,\ \|b\| \cdot \|b^{-1}\| < c\right\}.$$

Lemma 3.4 shows, that Ω_n is a subset of $\rho_m(S_n)$.

Theorem 4.3. *For all $2 \leq n \in \mathbb{N}$ the function G_n satisfies the following equation*

$$\Lambda_n(b)G_n(b) = 1 + \eta(G_n(b))G_n(b), \qquad b \in \Omega_n,$$

where

$$\Lambda_n : \Omega_n \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto b - \Theta_n(b)G_n(b)^{-1},$$

with a holomorphic function

$$\Theta_n : \Omega_n \to \mathrm{M}_m(\mathbb{C})$$

satisfying

$$\sup_{b\in\Omega_n} \|\Theta_n(b)\| \leq \frac{C}{\sqrt{n}}$$

with a constant $C > 0$, independent of n.

Proof. (i) Let $n \in \mathbb{N}$ and $b \in \rho_m(S_n)$ be given. Then we have

$$S_n R_n(b) = bR_n(b) - (b - S_n)R_n(b) = bR_n(b) - 1$$

and hence

$$E[S_n R_n(b)] = E\big[bR_n(b) - 1\big] = bG_n(b) - 1.$$

(ii) Let $n \in \mathbb{N}$ be given. For all

$$b \in \rho_{m,n} := \rho_m(S_n) \cap \bigcap_{i=1}^{n} \rho_m(S_n^{[i]})$$

we deduce from the formula in (6), that

$$
\begin{aligned}
E[S_n R_n(b)] &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[X_i R_n(b)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(E\big[X_i R_n^{[i]}(b)\big] + \frac{1}{\sqrt{n}} E\big[X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] \\
&\quad + \frac{1}{n} E\big[X_i R_n(b) X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(E\big[X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] + \frac{1}{\sqrt{n}} E\big[X_i R_n(b) X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(E\big[X_i G_n^{[i]}(b) X_i\big] G_n^{[i]}(b) + \frac{1}{\sqrt{n}} E\big[X_i R_n(b) X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big(\eta(G_n^{[i]}(b)) G_n^{[i]}(b) + r_{n,1}^{[i]}(b) \Big),
\end{aligned}
$$

where

$$
r_{n,1}^{[i]} : \rho_m(S_n) \cap \rho_m(S_n^{[i]}) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto \frac{1}{\sqrt{n}} E\big[X_i R_n(b) X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big].
$$

There we used the fact, that, since the $(X_j)_{j \in \mathbb{N}}$ are free with respect to E, also X_i is free from $R_n^{[i]}$, and thus we have

$$
E\big[X_i R_n^{[i]}(b)\big] = E[X_i] E\big[R_n^{[i]}(b)\big] = 0
$$

and

$$
E\big[X_i R_n^{[i]}(b) X_i R_n^{[i]}(b)\big] = E\big[X_i E\big[R_n^{[i]}(b)\big] X_i\big] E\big[R_n^{[i]}(b)\big].
$$

(iii) Taking (7) into account, we get for all $n \in \mathbb{N}$ and $1 \leq i \leq n$

$$
G_n(b) = E\big[R_n(b)\big] = E\big[R_n^{[i]}(b)\big] + \frac{1}{\sqrt{n}} E\big[R_n^{[i]}(b) X_i R_n(b)\big] = G_n^{[i]}(b) - r_{n,2}^{[i]}(b)
$$

and therefore

$$
G_n^{[i]}(b) = G_n(b) + r_{n,2}^{[i]}(b)
$$

for all $b \in \rho_m(S_n) \cap \rho_m(S_n^{[i]})$, where we put

$$
r_{n,2}^{[i]} :\ \rho_m(S_n) \cap \rho_m(S_n^{[i]}) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto -\frac{1}{\sqrt{n}} E\big[R_n^{[i]}(b) X_i R_n(b)\big].
$$

(iv) The formula in (iii) enables us to replace $G_n^{[i]}$ in (ii) by G_n. Indeed, we get

$$\begin{aligned} E[S_n R_n(b)] &= \frac{1}{n}\sum_{i=1}^{n}\Big(\eta(G_n^{[i]}(b))G_n^{[i]}(b) + r_{n,1}^{[i]}(b)\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\eta\big(G_n(b) + r_{n,2}^{[i]}(b)\big)\big(G_n(b) + r_{n,2}^{[i]}(b)\big) + r_{n,1}^{[i]}(b)\Big) \\ &= \eta(G_n(b))G_n(b) + \frac{1}{n}\sum_{i=1}^{n} r_{n,3}^{[i]}(b) \end{aligned}$$

for all $b \in \rho_{m,n}$, where the function

$$r_{n,3}^{[i]} : \ \rho_m(S_n) \cap \rho_m(S_n^{[i]}) \to \mathrm{M}_m(\mathbb{C})$$

is defined by

$$r_{n,3}^{[i]}(b) := \eta(G_n(b))r_{n,2}^{[i]}(b) + \eta(r_{n,2}^{[i]}(b))G_n(b) + \eta(r_{n,2}^{[i]}(b))r_{n,2}^{[i]}(b) + r_{n,1}^{[i]}(b).$$

(v) Combining the results from (i) and (iv), it follows

$$bG_n(b) - 1 = E[S_n R_n(b)] = \eta(G_n(b))G_n(b) + \Theta_n(b),$$

where we define

$$\Theta_n : \ \rho_{m,n} \to \mathrm{M}_m(\mathbb{C}), \ b \mapsto \frac{1}{n}\sum_{i=1}^{n} r_{n,3}^{[i]}(b).$$

Due to (8), Lemmas 3.4 and 3.7 show that $\Omega_n \subseteq \rho_{m,n}$ and $G_n(b) \in \mathrm{GL}_m(\mathbb{C})$ for $b \in \Omega_n$. This gives

$$\big(b - \Theta_n(b)G_n(b)^{-1}\big)G_n(b) = 1 + \eta(G_n(b))G_n(b)$$

and hence, as desired, for all $b \in \Omega_n$

$$\Lambda_n(b)G_n(b) = 1 + \eta(G_n(b))G_n(b).$$

(v) The definition of Ω_n gives, by Lemma 3 and by Lemma 4.1, the following estimates

$$\|G_n(b)\| \leq \|R_n(b)\| \leq \frac{\theta}{1-\theta} \cdot \frac{1}{\|S_n\|} \leq \frac{\theta}{1-\theta} \cdot \frac{1}{\|\alpha\|^{\frac{1}{2}}}, \qquad b \in \Omega_n$$

and

$$\|G_n^{[i]}(b)\| \leq \|R_n^{[i]}(b)\| \leq \frac{\theta}{1-\theta} \cdot \frac{1}{\|S_n^{[i]}\|} \leq \frac{\theta}{1-\theta} \cdot \frac{1}{\sqrt{1-\frac{1}{n}}\|\alpha\|^{\frac{1}{2}}}, \qquad b \in \Omega_n.$$

Therefore, we have for all $b \in \Omega_n$ by (ii)

$$\|r_{n,1}^{[i]}(b)\| \leq \frac{1}{\sqrt{n}}\|X_i\|^3 \|R_n(b)\| \|R_n^{[i]}(b)\|^2 \leq \frac{1}{\sqrt{n}} \frac{n}{n-1} \Big(\frac{\theta}{1-\theta} \frac{1}{\|\alpha\|^{\frac{1}{2}}}\Big)^3 \|X_i\|^3$$

and by (iii)

$$\|r_{n,2}^{[i]}(b)\| \leq \frac{1}{\sqrt{n}}\|X_i\| \|R_n(b)\| \|R_n^{[i]}(b)\| \leq \frac{1}{\sqrt{n-1}} \Big(\frac{\theta}{1-\theta} \frac{1}{\|\alpha\|^{\frac{1}{2}}}\Big)^2 \|X_i\|$$

and finally by (iv)

$$\begin{aligned}
\|r_{n,3}^{[i]}(b)\| &\leq 2M\|G_n(b)\| \|r_{n,2}^{[i]}(b)\| + M\|r_{n,2}^{[i]}(b)\|^2 + \|r_{n,1}^{[i]}(b)\| \\
&\leq \frac{1}{\sqrt{n-1}} \Big(\frac{\theta}{1-\theta} \frac{1}{\|\alpha\|^{\frac{1}{2}}}\Big)^3 \|X_i\| \cdot \\
&\quad \Big(2M + \frac{1}{\sqrt{n-1}} M \Big(\frac{\theta}{1-\theta} \frac{1}{\|\alpha\|^{\frac{1}{2}}}\Big) \|X_i\| + \sqrt{\frac{n}{n-1}} \|X_i\|^2\Big) \\
&\leq \frac{C}{\sqrt{n}}
\end{aligned}$$

for all $b \in \Omega_n$, where $C > 0$ is a constant, which is independent of n. Hence, it follows from (v) that

$$\sup_{b \in \Omega_n} \|\Theta_n(b)\| \leq \frac{C}{\sqrt{n}}.$$

□

The definition of Ω_n ensures, that

$$G := G_s : \ \rho_m(s) \to \mathrm{M}_m(\mathbb{C})$$

satisfies

$$bG(b) = 1 + \eta(G(b))G(b) \qquad \text{for all } b \in \Omega,$$

where

$$\Omega := \Big\{ b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \theta \cdot \frac{1}{\|s\|}, \ \|b\| \cdot \|b^{-1}\| < c \Big\} \supseteq \Omega_n.$$

We choose

$$0 < \gamma < \frac{c-1}{c+1} \qquad \text{and} \qquad 0 < \theta^* < (1-\gamma)\theta \tag{9}$$

(note, that $0 < \gamma < 1$) and we put $c^* := c - (1+c)\gamma$, which fulfills clearly $1 < c^* < c$. Since we have $\theta^* < \theta$ and $c^* < c$, we see

$$\frac{\theta^*}{1-\theta^*} c^* < \frac{\theta}{1-\theta} c < \sigma$$

and hence

$$\frac{\theta^*}{1-\theta^*} < \frac{\sigma}{c^*}. \tag{10}$$

Finally, we define

$$\kappa_n^* := \theta^* \min\left\{ \frac{1}{\|s\|}, \frac{1}{\|S_n\|}, \frac{1}{\|S_n^{[1]}\|}, \ldots, \frac{1}{\|S_n^{[n]}\|} \right\}$$

and

$$\Omega_n^* := \left\{ b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa_n^*,\ \|b\| \cdot \|b^{-1}\| < c^* \right\} \subseteq \Omega_n.$$

Corollary 4.4. *There exists $N \in \mathbb{N}$ such that*

$$\Lambda_n(\Omega_n^*) \subseteq \Omega_n \qquad \textit{for all } n \geq N.$$

Proof. Since we have by Theorem 4.3

$$\sup_{b \in \Omega_n} \|\Theta_n(b)\| \leq \frac{C}{\sqrt{n}}$$

for all $2 \leq n \in \mathbb{N}$, we can choose an $N \in \mathbb{N}$ such that

$$\sup_{b \in \Omega_n} \|\Theta_n(b)\| \leq \frac{\gamma}{c^*}(1-\sigma)$$

holds for all $n \geq N$. Now, we get for all $b \in \Omega_n^*$:

(i) $\Lambda_n(b)$ is invertible: Since (4) gives

$$\|G_n(b)^{-1}\| \leq \frac{1}{1-\sigma}\|b\| \qquad \text{for all } b \in \Omega_n,$$

we immediately get

$$\|\Lambda_n(b) - b\| \leq \|\Theta_n(b)\|\|G_n(b)^{-1}\| < \gamma \frac{\|b\|}{c^*} < \gamma \frac{1}{\|b^{-1}\|} < \frac{1}{\|b^{-1}\|}$$

(ii) We have $\|\Lambda_n(b)^{-1}\| < \kappa_n$: Using Lemma 3.1, we get from (i) that

$$\|\Lambda_n(b)^{-1}\| \leq \frac{1}{1-\gamma}\|b^{-1}\| < \frac{\kappa_n^*}{1-\gamma} < \kappa_n.$$

iii) We have $\|\Lambda_n(b)\|\|\Lambda_n(b)^{-1}\| < c$: Using

$$\|\Lambda_n(b) - b\| < \gamma\frac{\|b\|}{c^*}$$

from (i) and

$$\|\Lambda_n(b)^{-1}\| < \frac{1}{1-\gamma}\|b^{-1}\|$$

from (ii), we get

$$\begin{aligned}\|\Lambda_n(b)\|\|\Lambda_n(b)^{-1}\| &\leq \big(\|b\| + \|\Lambda_n(b) - b\|\big)\|\Lambda_n(b)^{-1}\| \\ &< \Big(1 + \frac{\gamma}{c^*}\Big)\frac{1}{1-\gamma} \cdot \|b\|\|b^{-1}\| \\ &< \frac{c^* + \gamma}{1-\gamma} < c.\end{aligned}$$

Finally, this shows $\Lambda_n(b) \in \Omega_n$. □

Corollary 4.5. *For all $n \geq N$ we have*

$$G_n(b) = G(\Lambda_n(b)) \qquad \text{for all } b \in \Omega_n^*.$$

Proof. For all $n \in \mathbb{N}$ we define

$$\Omega_n' := \Big\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b\| < \frac{\kappa_n}{1-\theta}\Big\}.$$

Let $n \geq N$ and $b \in \Omega_n^*$ be given. We know, that

$$\Lambda_n(b)G(\Lambda_n(b)) = 1 + \eta(G(\Lambda_n(b)))G(\Lambda_n(b))$$

holds, i.e. $w = G(\Lambda_n(b)) \in \Omega_n'$ is a solution of the equation

$$\Lambda_n(b)w = 1 + \eta(w)w, \qquad w \in \Omega_n'.$$

Combining (8) with Lemma 4.1, we get

$$\frac{\theta}{1-\theta} < \sigma \min\Big\{\frac{1}{c}, \frac{\|\alpha\|}{M}\Big\} \leq \sigma \min\Big\{\frac{1}{c}, \frac{\|S_n\|^2}{M};\ n \in \mathbb{N}\Big\}.$$

Hence, the equation above has, by Theorem 3.9, the unique solution $w = G_n(b) \in \Omega_n'$. This implies, as desired, $G_n(b) = G(\Lambda_n(b))$. □

Corollary 4.6. *For all $n \geq N$ we have*

$$\|G(b) - G_n(b)\| \leq C' \frac{1}{\sqrt{n}} \|b\| \qquad \text{for all } b \in \Omega_n^*,$$

where $C' > 0$ is a constant independent of n.

Proof. For all $b \in \Omega_n^* \subseteq \Omega_n \subseteq \Omega$ we have

$$\begin{aligned} G(b) - G_n(b) &= G(b) - G(\Lambda_n(b)) \\ &= E\big[(b-s)^{-1} - (\Lambda_n(b)-s)^{-1}\big] \\ &= E\big[(b-s)^{-1}(\Lambda_n(b)-b)(\Lambda_n(b)-s)^{-1}\big] \end{aligned}$$

and therefore by (4), which gives

$$\|G_n(b)^{-1}\| \leq \frac{1}{1-\sigma}\|b\| \qquad \text{for all } b \in \Omega_n^*,$$

and (since $\Lambda_n(b) \in \Omega_n \subseteq \Omega$) by (3)

$$\begin{aligned} \|G(b) - G_n(b)\| &\leq \|(b-s)^{-1}\| \cdot \|\Lambda_n(b) - b\| \cdot \|(\Lambda_n(b)-s)^{-1}\| \\ &\leq \Big(\frac{\theta}{1-\theta} \cdot \frac{1}{\|s\|}\Big)^2 \cdot \|\Theta_n(b)\| \cdot \|G_n(b)^{-1}\| \\ &\leq C' \frac{1}{\sqrt{n}} \|b\|, \end{aligned}$$

where

$$C' := \frac{C}{1-\sigma}\Big(\frac{\theta}{1-\theta} \cdot \frac{1}{\|s\|}\Big)^2 > 0.$$

This proves the corollary. □

We recall, that the sequence $(X_i)_{i\in\mathbb{N}}$ is bounded, which implies boundedness of the sequence $(S_n)_{n\in\mathbb{N}}$ as well. This has the important consequence, that

$$\kappa_n^* = \theta^* \min\Big\{\frac{1}{\|s\|}, \frac{1}{\|S_n\|}, \frac{1}{\|S_n^{[1]}\|}, \ldots, \frac{1}{\|S_n^{[n]}\|}\Big\} \geq \kappa^*$$

for some $\kappa^* > 0$. If we define

$$\Omega^* := \Big\{b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa^*,\ \|b\| \cdot \|b^{-1}\| < c^*\Big\},$$

we easily see $\Omega^* \subseteq \Omega_n^*$ for all $n \in \mathbb{N}$. Hence, by renaming Ω^* to Ω etc., we have shown our main Theorem 1.1.

We conclude this section with the following remark about the geometric structure of subsets of $\mathrm{M}_m(\mathbb{C})$ like Ω.

Lemma 4.7. *For $\kappa > 0$ and $c > 1$ we consider*

$$\Omega := \left\{ b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa,\ \|b\| \cdot \|b^{-1}\| < c \right\}.$$

For $\lambda, \mu \in \mathbb{C}\backslash\{0\}$ we define

$$\Lambda(\lambda, \mu) := \begin{pmatrix} \lambda & 0 & \dots & 0 \\ 0 & \mu & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu \end{pmatrix} \in \mathrm{GL}_m(\mathbb{C}).$$

If $\frac{1}{\kappa} < |\mu|$ holds, we have $\Lambda(\lambda, \mu) \in \Omega$ for all

$$\max\left\{ \frac{1}{\kappa}, \frac{|\mu|}{c} \right\} < |\lambda| < c|\mu|. \tag{11}$$

Particularly, we have for all $|\lambda| > \frac{1}{\kappa}$, that $\lambda 1 \in \Omega$.

Proof. Let $\mu \in \mathbb{C}\backslash\{0\}$ with $\frac{1}{\kappa} < |\mu|$ be given. For all $\lambda \in \mathbb{C}\backslash\{0\}$, which satisfy (11), we get

$$\|\Lambda(\lambda, \mu)^{-1}\| = \|\Lambda(\lambda^{-1}, \mu^{-1})\| = \max\left\{ |\lambda|^{-1}, |\mu|^{-1} \right\} < \kappa.$$

and

$$\begin{aligned} \|\Lambda(\lambda, \mu)\| \cdot \|\Lambda(\lambda, \mu)^{-1}\| &= \max\left\{ |\lambda|, |\mu| \right\} \cdot \max\left\{ |\lambda|^{-1}, |\mu|^{-1} \right\} \\ &= \begin{cases} |\mu||\lambda|^{-1}, & \text{if } |\lambda| < |\mu| \\ |\lambda||\mu|^{-1}, & \text{if } |\lambda| \geq |\mu| \end{cases} \\ &< c, \end{aligned}$$

which implies $\Lambda(\lambda, \mu) \in \Omega$. In particular, for $\lambda \in \mathbb{C}\backslash\{0\}$ with $|\lambda| > \frac{1}{\kappa}$ we see that $\mu = \lambda$ fulfills (11) and it follows $\lambda 1 = \Lambda(\lambda, \lambda) \in \Omega$. □

4.3 Application to Multivariate Situation

4.3.1 Multivariate Free Central Limit Theorem

Let $(x_i^{(k)})_{k=1}^d$, $i \in \mathbb{N}$, be free and identically distributed sets of d self-adjoint non-zero random variables in some non-commutative C^*-probability space $(\mathcal{C}, \tau)$, with τ faithful, such that

$$\tau(x_i^{(k)}) = 0 \qquad \text{for } k = 1, \ldots, d \text{ and all } i \in \mathbb{N}$$

and

$$\sup_{i \in \mathbb{N}} \max_{k=1,\ldots,d} \|x_i^{(k)}\| < \infty. \tag{12}$$

We denote by $\Sigma = (\sigma_{k,l})_{k,l=1}^d$, where $\sigma_{k,l} := \tau(x_i^{(k)} x_i^{(l)})$, their joint covariance matrix. Moreover, we put

$$S_n^{(k)} := \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i^{(k)} \qquad \text{for } k = 1, \ldots, d \text{ and all } n \in \mathbb{N}.$$

We know (cf. [11]), that $(S_n^{(1)}, \ldots, S_n^{(d)})$ converges in distribution as $n \to \infty$ to a semicircular family $(s_1, \ldots, s_d)$ of covariance Σ. For notational convenience we will assume that $s_1, \ldots, s_d$ live also in $(\mathcal{C}, \tau)$; this can always be achieved by enlarging $(\mathcal{C}, \tau)$.

Using Proposition 2.1 and Proposition 2.3 in [6], for each polynomial p of degree g in d non-commuting variables vanishing in 0, we can find $m \in \mathbb{N}$ and $a_1, \ldots, a_d \in \mathrm{M}_m(\mathbb{C})$ such that

$$\lambda 1 - p(S_n^{(1)}, \ldots, S_n^{(d)}) \quad \text{and} \quad \lambda 1 - p(s_1, \ldots, s_d)$$

are invertible in $\mathcal{C}$ if and only if

$$\Lambda(\lambda, 1) - S_n \quad \text{and} \quad \Lambda(\lambda, 1) - s,$$

respectively, are invertible in $\mathcal{A} = \mathrm{M}_m(\mathcal{C})$. The matrices $\Lambda(\lambda, 1) \in \mathrm{M}_m(\mathbb{C})$ were defined in Lemma 4.7, and S_n and s are defined as follows:

$$S_n := \sum_{k=1}^d a_k \otimes S_n^{(k)} \in \mathcal{A} \qquad \text{for all } n \in \mathbb{N}$$

and

$$s := \sum_{k=1}^d a_k \otimes s_k \in \mathcal{A}.$$

If we also put

$$X_i := \sum_{k=1}^{d} a_k \otimes x_i^{(k)} \in \mathcal{A} \qquad \text{for all } i \in \mathbb{N},$$

then we have

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

We note, that the sequence $(X_i)_{i\in\mathbb{N}}$ is $*$-free with respect to the conditional expectation $E : \mathcal{A} = \mathrm{M}_m(\mathcal{C}) \to \mathrm{M}_m(\mathbb{C})$ and that all the X_i's have the same $*$-distribution with respect to E and that they satisfy $E[X_i] = 0$. In addition, (12) implies $\sup_{i\in\mathbb{N}} \|X_i\| < \infty$. Hence, the conditions of Theorem 1.1 are fulfilled. But before we apply it, we note that $(S_n)_{n\in\mathbb{N}}$ converges in distribution (with respect to E) to s, which is an $\mathrm{M}_m(\mathbb{C})$-valued semicircular element with covariance mapping

$$\eta : \mathrm{M}_m(\mathbb{C}) \to \mathrm{M}_m(\mathbb{C}),\ b \mapsto E[sbs],$$

which is given by

$$\eta(b) = E[sbs] = \sum_{k,l=1}^{d} \mathrm{id} \otimes \tau[(a_k \otimes s_k)(b \otimes 1)(a_l \otimes s_l)] = \sum_{k,l=1}^{d} a_k b a_l \sigma_{k,l}.$$

Now, we get from Theorem 1.1 constants $\kappa^* > 0$, $c^* > 0$ and $C' > 0$ and $N \in \mathbb{N}$ such that we have for the difference of the operator-valued Cauchy transforms

$$G_s(b) := E[(b-s)^{-1}] \quad \text{and} \quad G_{S_n}(b) := E[(b-S_n)^{-1}]$$

the estimate

$$\|G_s(b) - G_{S_n}(b)\| \leq C' \frac{1}{\sqrt{n}} \|b\| \qquad \text{for all } b \in \Omega^* \text{ and } n \geq N,$$

where we put

$$\Omega^* := \left\{ b \in \mathrm{GL}_m(\mathbb{C}) \mid \|b^{-1}\| < \kappa^*,\ \|b\| \cdot \|b^{-1}\| < c^* \right\}.$$

Moreover, Proposition 2.3 in [6] tells us

$$\left(\lambda 1 - p(S_n^{(1)}, \ldots, S_n^{(d)})\right)^{-1} = (\pi \otimes \mathrm{id}_{\mathcal{C}})\left((\Lambda(\lambda, 1) - S_n)^{-1}\right)$$

and

$$\left(\lambda 1 - p(s_1, \ldots, s_d)\right)^{-1} = (\pi \otimes \mathrm{id}_{\mathcal{C}'})\left((\Lambda(\lambda, 1) - s)^{-1}\right),$$

where $\pi : \mathrm{M}_m(\mathbb{C}) \to \mathbb{C}$ is the mapping given by $\pi((a_{i,j})_{i,j=1,\ldots,m}) := a_{1,1}$. Since $\tau \circ (\pi \otimes \mathrm{id}_{\mathcal{C}}) = \pi \circ E$, this implies a direct connection between the operator-valued Cauchy transforms of S_n and s and the scalar-valued Cauchy transforms of

$P_n := p(S_n^{(1)}, \ldots, S_n^{(d)})$ and $P := p(s_1, \ldots, s_d)$, respectively. To be more precise, we get

$$G_{P_n}(\lambda) := \tau[(\lambda - P_n)^{-1}] = \pi\big(G_{S_n}(\Lambda(\lambda, 1))\big)$$

and

$$G_P(\lambda) := \tau[(\lambda - P)^{-1}] = \pi\big(G_s(\Lambda(\lambda, 1))\big)$$

for all $\lambda \in \rho_{\mathcal{C}}(P_n)$ and $\lambda \in \rho_{\mathcal{C}}(P)$, respectively.

If we choose $\mu \in \mathbb{C}$ such that $|\mu| > \frac{1}{\kappa^*}$ holds, it follows from Lemma 4.7, that $\Lambda(\lambda, \mu) \in \Omega^*$ is fulfilled for all $\lambda \in A(\mu)$, where $A(\mu) \subseteq \mathbb{C}$ denotes the open set of all $\lambda \in \mathbb{C}$ satisfying (11), i.e.

$$A(\mu) := \left\{\lambda \in \mathbb{C} \mid \max\left\{\frac{1}{\kappa^*}, \frac{|\mu|}{c^*}\right\} < |\lambda| < c^*|\mu|\right\}.$$

If we apply Propositions 2.1 and 2.2 in [6] to the polynomial $\frac{1}{\mu^g} p$ (which corresponds to the operators $\frac{1}{\mu} S_n$ and $\frac{1}{\mu} S$), we easily deduce that

$$\lambda 1 - \frac{1}{\mu^{g-1}} p(S_n^{(1)}, \ldots, S_n^{(d)}) \quad \text{and} \quad \lambda 1 - \frac{1}{\mu^{g-1}} p(s_1, \ldots, s_d)$$

are invertible in $\mathcal{C}$ if and only if

$$\Lambda(\lambda, \mu) - S_n \quad \text{and} \quad \Lambda(\lambda, \mu) - S_n,$$

respectively, are invertible in $\mathcal{A}$. Moreover, we have

$$\mu^{g-1} G_{P_n}(\lambda \mu^{g-1}) = \pi\big(G_{S_n}(\Lambda(\lambda, \mu))\big)$$

and

$$\mu^{g-1} G_P(\lambda \mu^{g-1}) = \pi\big(G_s(\Lambda(\lambda, \mu))\big)$$

for all $\lambda \in \rho_{\mathcal{C}}(\frac{1}{\mu^{g-1}} P_n)$ and $\lambda \in \rho_{\mathcal{C}}(\frac{1}{\mu^{g-1}} P)$, respectively.

Particularly, for all $\lambda \in A(\mu)$ we get $\Lambda(\lambda, \mu) \in \Omega^*$ and hence $\lambda \in \rho_{\mathcal{C}}(\frac{1}{\mu^{g-1}} P_n) \cap \rho_{\mathcal{C}}(\frac{1}{\mu^{g-1}} P)$ for all $n \geq N$. Therefore, Theorem 1.1 implies

$$\begin{aligned}
|\mu|^{g-1} |G_P(\lambda \mu^{g-1}) - G_{P_n}(\lambda \mu^{g-1})| &= \left|\pi\big(G_s(\Lambda(\lambda, \mu)) - G_{S_n}(\Lambda(\lambda, \mu))\big)\right| \\
&\leq \left\|G_s(\Lambda(\lambda, \mu)) - G_{S_n}(\Lambda(\lambda, \mu))\right\| \\
&\leq C' \frac{1}{\sqrt{n}} \|\Lambda(\lambda, \mu)\| \\
&\leq C' \frac{1}{\sqrt{n}} \max\{|\lambda|, |\mu|\} \\
&\leq C' c^* |\lambda| \frac{1}{\sqrt{n}}
\end{aligned}$$

and hence

$$|G_P(\lambda\mu^{g-1}) - G_{P_n}(\lambda\mu^{g-1})| \leq C'c^* \frac{1}{\sqrt{n}} |\lambda\mu^{g-1}|.$$

This means, that

$$|G_P(z) - G_{P_n}(z)| \leq C'c^* \frac{1}{\sqrt{n}} |z|$$

holds for all $z \in \mathbb{C}$ with $\frac{z}{\mu^{g-1}} \in A(\mu)$ and all $n \geq N$. By definition of $A(\mu)$, we particularly get

$$|G_P(z) - G_{P_n}(z)| \leq C \frac{1}{\sqrt{n}} \qquad \text{for all } \frac{1}{c^*}|\mu|^g < |z| < c^*|\mu|^g \text{ and } n \geq N,$$

where we put $C := C'(c^*)^2|\mu|^g > 0$. Since $z \mapsto G_P(z) - G_{P_n}(z)$ is holomorphic on $\{z \in \mathbb{C} \mid |z| > R\}$ for $R := \frac{1}{c^*}|\mu|^g > 0$ and extends holomorphically to ∞, the maximum modulus principle gives

$$|G_P(z) - G_{P_n}(z)| \leq C \frac{1}{\sqrt{n}} \qquad \text{for all } |z| > R \text{ and } n \geq N.$$

This shows Theorem 1.2 in the case of a polynomial p vanishing in 0. For a general polynomial p, we consider the polynomial $\tilde{p} = p - p_0$ with $p_0 := p(0, \ldots, 0)$, which leads to the operators $\tilde{P} = P - p_0 1$ and $\tilde{P}_n = P_n - p_0 1$. Since we can apply the result above to $\tilde{p}$ and since the Cauchy transforms G_P and G_{P_n} are just translations of $G_{\tilde{P}}$ and $G_{\tilde{P}_n}$, respectively, the general statement follows easily.

4.3.2 Estimates in Terms of the Kolmogorov Distance

In the classical case, estimates between scalar-valued Cauchy transforms can be established (for self-adjoint operators) in all of the upper complex plane and lead then to estimates in terms of the Kolmogorov distance. In the case treated above, we have a statement about the behavior of the difference between two Cauchy transforms only near infinity. Even in the case, where our operators are self-adjoint, we still have to transport estimates from infinity to the real line, and hence we can not apply the results of Bai [1] directly. A partial solution to this problem was given in the appendix of [14] with the following theorem, formulated in terms of probability measures instead of operators. There we use the notation G_μ for the Cauchy transform of the measure μ, and put

$$D_R^+ := \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0,\ |z| > R\}.$$

Theorem 4.8. *Let μ be a probability measure with compact support contained in an interval $[-A, A]$ such that the cumulative distribution function $\mathcal{F}_\mu$ satisfies*

$$|\mathcal{F}_\mu(x+t) - \mathcal{F}_\mu(x)| \leq \rho|t| \qquad \text{for all } x, t \in \mathbb{R}$$

for some constant $\rho > 0$. Then for all $R > 0$ and $\beta \in (0,1)$ we can find $\Theta > 0$ and $m_0 > 0$ such that for any probability measure ν with compact support contained in $[-A, A]$, which satisfies

$$\sup_{z \in D_R^+} |G_\mu(z) - G_\nu(z)| \leq e^{-m}$$

for some $m > m_0$, the Kolmogorov distance $\Delta(\mu,\nu) := \sup_{x\in\mathbb{R}} |\mathcal{F}_\mu(x) - \mathcal{F}_\nu(x)|$ fulfills

$$\Delta(\mu,\nu) \leq \Theta \frac{1}{m^\beta}.$$

Obviously, this leads to the following questions: First, the stated estimate for the speed of convergence in terms of the Kolmogorov distance is far from the expected one. We hope to improve this result in a future work. Furthermore, in order to apply this theorem, we have to ensure that $p(s_1, \ldots, s_d)$ has a continuous density. As mentioned in the introduction, it is a still unsolved problem, whether this is always true for any self-adjoint polynomials p.

Acknowledgements This project was initiated by discussions with Friedrich Götze during the visit of the second author at the University of Bielefeld in November 2006. He thanks the Department of Mathematics and in particular the SFB 701 for its generous hospitality and Friedrich Götze for the invitation and many interesting discussions. A preliminary version of this paper appeared as preprint [13] of SFB 701. Research of T. Mai supported by funds of Roland Speicher from the Alfried Krupp von Bohlen und Halbach-Stiftung; research of R. Speicher partially supported by a Discovery grant from NSERC (Canada) and by a Killam Fellowship from the Canada Council for the Arts.

The second author also thanks Uffe Haagerup for pointing out how ideas from [6] can be used to improve the results from an earlier version of this paper.

References

1. Z.D. Bai, Convergence rate of expected spectral distributions of large-dimensional random matrices: Part I. Wigner matrices. Ann. Probab. **21**, 625–648 (1993)
2. G.P. Chistyakov, F. Götze, Limit theorems in free probability theory I. Ann. Probab. **1**(1), 54–90 (2008)
3. C.J. Earle, R.S. Hamilton, *A Fixed Point Theorem for Holomorphic Mappings*. Global Analysis (Proc. Sympos. Pure Math., Vol. XVI, Berkeley, CA, 1968) (American Mathematical Society, Providence, 1970), pp. 61–65

4. F. Götze, A. Tikhomirov, *Limit Theorems for Spectra of Random Matrices with Martingale Structure*. Stein's Method and Applications, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., vol. 5 (Singapore University Press, Singapore, 2005), pp. 181–193
5. U. Haagerup, S. Thorbjørnsen, A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. Ann. Math. **162**, 711–775 (2005)
6. U. Haagerup, H. Schultz, S. Thorbjørnsen, A random matrix approach to the lack of projections in $C^*_{\mathrm{red}}(\mathbb{F}_2)$. Adv. Math. **204**, 1–83 (2006)
7. L.A. Harris, Fixed points of holomorphic mappings for domains in Banach spaces. Abstr. Appl. Anal. **2003**(5), 261–274 (2003)
8. J.W. Helton, R.R. Far, R. Speicher, Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints. Int. Math. Res. Not. (22), Article ID rnm086, 15 (2007)
9. M. Junge, Embedding of the operator space OH and the logarithmic 'little Grothendieck inequality'. Invent. math. **161**, 225–286 (2005)
10. V. Kargin, Berry-Esseen for free random variables. J. Theor. Probab. **20**, 381–395 (2007)
11. R. Speicher, A new example of independence and white noise. Probab. Theor. Relat. Fields **84**, 141–159 (1990)
12. R. Speicher, Combinatorial theory of the free product with amalgamation and operator-valued free probability theory. Mem. Am. Math. Soc. **132**(627), x+88 (1998)
13. R. Speicher, On the rate of convergence and Berry-Esseen type theorems for a multivariate free central limit theorem. SFB 701. Preprint (2007)
14. R. Speicher, C. Vargas, Free deterministic equivalents, rectangular random matrix models and operator-valued free probability theory. Random Matrices: Theor. Appl. **1**(1), 1150008, 26 (2012)
15. D. Voiculescu, Operations on certain non-commutative operator-valued random variables. Astérisque **232**, 243–275 (1995). Recent advances in operator algebras (Orléans, 1992)

A Characterization of Small and Large Time Limit Laws for Self-normalized Lévy Processes

Ross Maller and David M. Mason

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday.

Abstract We establish asymptotic distribution results for self-normalized Lévy processes at small and large times that are analogs of those of Chistyakov and Götze [Ann. Probab. 32:28–77, 2004] for self-normalized sums.

Keywords Domain of Attraction of Normal Distribution • Large Times • Lévy Process • Quadratic Variation • Self-Normalized • Small Times • Stable Laws

AMS 2000 Subject Classifications: 60F05, 60F17, 60G51.

1 Introduction and Statements of Main Results

Let $\xi, \xi_1, \xi_2, \ldots,$ be i.i.d. nondegenerate random variables with common distribution function F. For each $n \geq 1$ let $\mathbb{S}_n = \xi_1 + \cdots + \xi_n$ and $\mathbb{V}_n = \xi_1^2 + \cdots + \xi_n^2$. Consider the self-normalized sum

$$\mathbb{T}_n := \mathbb{S}_n / \sqrt{\mathbb{V}_n}. \tag{1}$$

R. Maller
Centre for Mathematical Analysis, Mathematical Sciences Institute, Australian National University, Canberra ACT, Australia
e-mail: Ross.Maller@anu.edu.au

D.M. Mason (✉)
Department of Applied Economics and Statistics, 213 Townsend Hall, Newark, DE 19716, USA
e-mail: davidm@udel.edu

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_8,

(Here and elsewhere $0/0 := 0$.) Giné et al. [6] proved that $\mathbb{T}_n$ converges in distribution to a standard normal rv Z if and only if F is in the domain of attraction of a normal law, written $F \in D(N)$, and $E\xi = 0$. (We write "rv" to mean "random variable" throughout.) This verified part of a conjecture of Logan et al. [9]. (Later Mason [13] provided an alternate proof.) Chistyakov and Götze [3] established the rest of the Logan et al. conjecture by completely characterizing when $\mathbb{T}_n$ converges in distribution to a non-degenerate rv Y such that $P\{|Y| = 1\} \neq 1$. A bit later, as a by-product of the study of a seemingly unrelated problem, Mason and Zinn [14] found a simple proof of the full Logan et al. conjecture assuming symmetry.

Theorem 1.1 of Chistyakov and Götze [3] implies the following: one has $\mathbb{T}_n \xrightarrow{\mathrm{D}} Y$, where $P(|Y| = 1) = 0$, if and only if there exists a sequence of norming constants $b_n > 0$ such that either

$$b_n^{-1}\mathbb{S}_n \xrightarrow{\mathrm{D}} Z, \tag{2}$$

where Z is a standard normal rv, or for some $0 < \alpha < 2$,

$$b_n^{-1}\mathbb{S}_n \xrightarrow{\mathrm{D}} U^\alpha, \tag{3}$$

where U^α is a strictly stable rv as defined in the Appendix. In case (2), $Y \stackrel{\mathrm{D}}{=} Z$ and in case (3), $Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}$. The V^α rv arises in the distributional limit in (4), which is implied by (3):

$$\left(b_n^{-1}\mathbb{S}_n, b_n^{-2}\mathbb{V}_n\right) \xrightarrow{\mathrm{D}} (U^\alpha, V^\alpha)\,. \tag{4}$$

For details see the Appendix.

Our aim is to prove analogs of the Chistyakov and Götze [3] result for a Lévy process X_t, $t \geq 0$, at small times ($t \searrow 0$) and large times ($t \to \infty$). To state our results we must first fix notation. We abbreviate "infinitely divisible" to "inf. div." throughout. Let $(\Omega, \mathcal{F}, P)$ be a probability space carrying a real-valued Lévy process $(X_t)_{t\geq 0}$ having nondegenerate inf. div. characteristic function

$$E e^{\mathrm{i}\theta X_t} = e^{t\Psi(\theta)}, \quad \theta \in \mathbb{R}, \tag{5}$$

where

$$\Psi(\theta) = \mathrm{i}\gamma\theta - \frac{1}{2}\sigma^2\theta^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}\theta x} - 1 - \mathrm{i}\theta x \mathbf{1}_{\{|x|\leq 1\}}\right) \Pi(\mathrm{d}x), \tag{6}$$

$\gamma \in \mathbb{R}$, $\sigma^2 \geq 0$, and Π is a measure on $\mathbb{R} \setminus \{0\}$ with $\int_{\mathbb{R}\setminus\{0\}}(x^2 \wedge 1)\Pi(\mathrm{d}x)$ finite.

We say that X_t has *canonical triplet* (γ, σ^2, Π) and X_1 is inf. div. with *triplet* (γ, σ^2, Π). The tails $\overline{\Pi}(x)$ and $\overline{\Pi}^{\pm}(x)$ of Π are defined by

$$\overline{\Pi}^{-}(x) = \Pi\{(-\infty, -x)\},\ \overline{\Pi}^{+}(x) = \Pi\{(x, \infty)\},\ \text{and}\ \overline{\Pi}(x) = \overline{\Pi}^{+}(x) + \overline{\Pi}^{-}(x), \quad x > 0. \tag{7}$$

Assume throughout that $\sigma^2 + \overline{\Pi}(0+) > 0$, otherwise X degenerates to a constant drift.

Let $(\Delta X_t)_{t\geq 0}$, with $\Delta X_t = X_t - X_{t-}$, $X_{0-} = 0$, denote the jump process of X, and consider the Lévy process

$$V_t = \sigma^2 t + \sum_{0<s\leq t} (\Delta X_s)^2,\ t > 0. \tag{8}$$

V is a subordinator with drift σ^2 and Lévy measure satisfying $\overline{\Pi}_V(x) = \overline{\Pi}(\sqrt{x})$, $x > 0$. By Theorem 2.1 of Maller and Mason [10] the joint characteristic function of (X_t, V_t) is given by

$$Ee^{\mathrm{i}(\theta_1 X_t + \theta_2 V_t)}$$
$$= \exp\left\{ \mathrm{i}t\left(\theta_1\gamma + \theta_2\sigma^2\right) - t\theta_1^2\sigma^2/2 + t\int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\theta_1 x + \theta_2 x^2)} - 1 - \mathrm{i}\theta_1 x \mathbf{1}_{\{|x|\leq 1\}}\right)\Pi(\mathrm{d}x)\right\}.$$

We shall say that (X_t, V_t) has triplet (γ, σ^2, Π).

Here is our small time ($t \searrow 0$) analog of the Chistyakov and Götze [3] result.

Theorem 1.1. *Let X_t, $t \geq 0$, be a Lévy process satisfying $\overline{\Pi}(0+) = \infty$. Assume that*

$$X_t/\sqrt{V_t} \xrightarrow{\mathrm{D}} Y, \text{ as } t \searrow 0, \tag{9}$$

where Y is a finite rv with $P(|Y| = 1) = 0$. Then either Y is standard normal or

$$Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}, \tag{10}$$

where (U^α, V^α) is a strictly stable pair of index α for some $0 < \alpha < 2$, as in (4).

Here is our large time analog ($t \to \infty$) of the Chistyakov and Götze [3] result.

Theorem 1.2. *Let X_t, $t \geq 0$, be a Lévy process satisfying $\sigma^2 + \overline{\Pi}(0+) > 0$. We have*

$$X_t/\sqrt{V_t} \xrightarrow{\mathrm{D}} Y, \text{ as } t \to \infty, \tag{11}$$

where $P(|Y| = 1) = 0$, if and only if either X_1 has expectation 0 and is in the domain of attraction of a normal law as $t \to \infty$, in which case $Y \stackrel{\mathrm{D}}{=} Z$, or X_1 is in the domain of attraction of a strictly stable law in the sense that for a sequence of positive norming constants b_n,

$$b_n^{-1}\left\{X_{(1)} + \cdots + X_{(n)}\right\}$$

converges in distribution to a nondegenerate strictly stable law of index $0 < \alpha < 2$, where $X_{(1)}, X_{(2)}, \ldots$, are i.i.d. as X_1, in which case Y is as in (10).

Remark 1. Chistyakov and Götze [3] also show that $\mathbb{T}_n$ converges in distribution to a non-degenerate rv Y such that $P\{|Y| = 1\} = 1$ if and only if $P\{|\xi| > x\}$ is slowly varying at infinity. We do not have such a complete picture for X_t as $t \searrow 0$. Assuming $\overline{\Pi}(0+) = \infty$, the proof of Lemma 5.3 of Maller and Mason [12] shows that if $\overline{\Pi}$ is slowly varying at zero then $|X_t|/\sqrt{V_t} \xrightarrow{\mathrm{D}} 1$. However, we do not know whether the converse is true, except in the case when X_t is symmetric. In this case, Maller and Mason [10] prove that whenever $\overline{\Pi}(0+) = \infty$, $X_t/\sqrt{V_t} \xrightarrow{\mathrm{D}} Y$, as $t \searrow 0$, where Y is equal to 1 or -1 with probability $1/2$ each if and only if $\overline{\Pi}$ is slowly varying at zero. Further results are given in Theorem 3.4 of Maller and Mason [12]. Analogous statements can be said about the case $t \to \infty$.

1.1 Some Needed Technical Results

To prove Theorems 1.1 and 1.2 we shall need to establish a number of technical results about $X_t/\sqrt{V_t}$, which are also of independent interest. To do this we must introduce some more notation and definitions. We will use some truncated mean and variance functions, defined for $x > 0$ by

$$v(x) = \gamma - \int_{x<|y|\le 1} y\Pi(\mathrm{d}y),\ V(x) = \sigma^2 + \int_{|y|\le x} y^2\Pi(\mathrm{d}y),\ \text{and } U(x) = \sigma^2 + 2\int_0^x y\overline{\Pi}(y)\mathrm{d}y. \tag{12}$$

These functions are finite for all $x > 0$ by virtue of the properties of the Lévy measure Π, which further imply that $\lim_{x\searrow 0} x^2\overline{\Pi}(x) = 0$, and $\lim_{x\searrow 0} xv(x) = 0$.

By the *relative compactness* of a real-valued stochastic process $(S_t)_{t\ge 0}$, as $t \to \infty$, we will mean that it satisfies

$$\lim_{x\to\infty} \limsup_{t\to\infty} P(|S_t| > x) = 0,$$

or, equivalently, every sequence $t_k \to \infty$ contains a subsequence $t_{k'} \to \infty$ with $S_{t_{k'}}$ converging in distribution to an a.s. finite rv. If in addition each such subsequential limit is not degenerate at a constant, we say that S_t is *stochastically compact*, as $t \to \infty$.

By the *Feller class at 0* we will mean the class of Lévy processes which are stochastically compact at 0 after norming and centering; that is, those for which there are nonstochastic functions $a(t)$, $b(t) > 0$ (where, *throughout, $b(t)$ will be assumed positive, but not, a priori, monotone),* such that every sequence $t_k \searrow 0$ contains a subsequence $t_{k'} \searrow 0$ with

$$\left(X_{t_{k'}} - a(t_{k'})\right)/b(t_{k'}) \xrightarrow{\mathrm{D}} Y',\ \text{as } k' \to \infty, \tag{13}$$

where Y' is a finite nondegenerate rv, a.s. (The prime on Y' denotes that in general it will depend on the choice of subsequence $t_{k'}$.) We describe this kind of convergence as "$X_t \in FC$ at 0".

It was shown in Maller and Mason [10] that when the relation (13) holds (with Y' not degenerate at a constant) then it must be the case that Y' is an inf. div. rv, and $b(t_{k'}) \to 0$ as $t_{k'} \searrow 0$.

Closely related is the *centered Feller class at 0*. This is the class of Lévy processes which are stochastically compact at 0, after norming, but with no centering function needed; that is, those for which there is a nonstochastic function $b(t) > 0$ such that every sequence $t_k \searrow 0$ contains a subsequence $t_{k'} \searrow 0$ with

$$X_{t_{k'}}/b(t_{k'}) \xrightarrow{\mathrm{D}} Y', \text{ as } k' \to \infty, \tag{14}$$

where Y' is a finite, nondegenerate, necessarily inf. div., rv, a.s. We describe this as "$X_t \in FC_0$ at 0".

The classes FC and FC_0 "at infinity" are defined in exactly the same ways, but with the subsequences tending to infinity rather than to 0. Maller and Mason [11,12] have carried out a thorough study of FC and FC_0 at 0 and infinity and have obtained a number of useful analytic equivalences in terms of the Lévy measure Π of X_t.

The following propositions connect the self-normalized and compactness ideas, and will be essential ingredients in the proofs of Theorems 1.1 and 1.2.

Proposition 1.1. *Suppose $X_t/\sqrt{V_t}$ is relatively compact as $t \searrow 0$, $\overline{\Pi}(0+) = \infty$ and no subsequential limit has positive mass at ± 1. Then $X \in FC_0$ as $t \searrow 0$, or, equivalently, by Theorem 2.3 of Maller and Mason [12]*

$$\limsup_{x \searrow 0} \frac{x^2\overline{\Pi}(x) + x|\nu(x)|}{V(x)} < \infty. \tag{15}$$

Proposition 1.2. *Suppose $X_t/\sqrt{V_t}$ is relatively compact as $t \to \infty$ and no subsequential limit has positive mass at ± 1. Then $X \in FC_0$ as $t \to \infty$, or, equivalently, by Theorem 1 (ii) of Maller and Mason [11]*

$$\limsup_{x \to \infty} \frac{x^2\overline{\Pi}(x) + x|\nu(x)|}{V(x)} < \infty. \tag{16}$$

These two propositions will be proved in a separate section.

2 Proofs of Theorems

The proofs will require the following two limit theorems, which we state as Lemmas 2.1 and 2.2. Recall that $(X_t)_{t\geq 0}$ is Lévy with canonical triplet (γ, σ^2, Π). Suppose that, for a sequence of integers $n_k \to \infty$ and positive constants $B(n_k)$,

$$X_{n_k}/B(n_k) \xrightarrow{\mathrm{D}} U, \tag{17}$$

where U is inf. div. with triplet (b, a, Λ), $b \in \mathbb{R}$ and $a \geq 0$. Notice that necessarily $B(n_k) \to \infty$.

Each rv $X_{n_k}/B(n_k)$ is inf. div. with triplet $\left(b_{n_k}, a_{n_k}, \Lambda_{n_k}\right)$, where

$$b_{n_k} = n_k \gamma / B(n_k), a_{n_k} = n_k \sigma^2 / B^2(n_k) \text{ and } \Lambda_{n_k}(\mathrm{d}x) = n_k \Pi(\mathrm{d}x / B(n_k)).$$

Moreover, $\left(X_{n_k}/B(n_k), V_{n_k}/B^2(n_k)\right)$ has joint characteristic function

$$Ee^{\mathrm{i}(\theta_1 X_{n_k}/B(n_k) + \theta_2 V_{n_k}/B^2(n_k))}$$
$$= \exp\left\{\mathrm{i}\left(\theta_1 b_{n_k} + \theta_2 a_{n_k}\right) - \theta_1^2 a_{n_k}/2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\theta_1 x + \theta_2 x^2)} - 1 - \mathrm{i}\theta_1 x \mathbf{1}_{\{|x|\leq 1\}}\right) \Lambda_{n_k}(\mathrm{d}x)\right\}.$$

Lemma 2.1. *Whenever (17) holds,*

$$\left(X_{n_k}/B(n_k), V_{n_k}/B^2(n_k)\right) \xrightarrow{\mathrm{D}} (U, W), \tag{18}$$

where (U, W) has joint characteristic function

$$Ee^{\mathrm{i}(\theta_1 U + \theta_2 W)}$$
$$= \exp\left\{\mathrm{i}(\theta_1 b + \theta_2 a) - \theta_1^2 a/2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\theta_1 x + \theta_2 x^2)} - 1 - \mathrm{i}\theta_1 x \mathbf{1}_{\{|x|\leq 1\}}\right) \Lambda(\mathrm{d}x)\right\}. \tag{19}$$

Proof. For each $h > 0$ let

$$a^h = a + \int_{0<|x|\leq h} x^2 \Lambda(\mathrm{d}x) \text{ and } b^h = b - \int_{h<|x|\leq 1} x \Lambda(\mathrm{d}x), \tag{20}$$

and let $a_{n_k}^h$ and $b_{n_k}^h$ be defined as in (20) with Λ replaced by Λ_{n_k}, a by a_{n_k}, and b by b_{n_k}. According to Theorem 15.14 of Kallenberg [8], (17) happens if and only if

$$\Lambda_{n_k} \text{ converges vaguely to } \Lambda \text{ on } \mathbb{R} \setminus \{0\} \tag{21}$$

and for any $h > 0$ such that $\Lambda\{|x| = h\} = 0$,

$$a_{n_k}^h \to a^h \text{ and } b_{n_k}^h \to b^h. \tag{22}$$

By the vague convergence of Λ_{n_k} to Λ, and since $a_{n_k}^h \to a^h$, we also have, for any $r > 2$,

$$\int_{0<|x|\leq h} x^r \Lambda_{n_k}(\mathrm{d}x) \to \int_{0<|x|\leq h} x^r \Lambda(\mathrm{d}x). \tag{23}$$

To verify (23), take $r > 2$ and $0 < \delta < h$, with δ and h continuity points of Λ, and write

$$\int_{0<|x|\le\delta} x^r \Lambda_{n_k}(dx) \le \delta^{r-2} \int_{0<|x|\le\delta} x^2 \Lambda_{n_k}(dx)$$
$$\le \delta^{r-2} \left(a_{n_k} + \int_{0<|x|\le h} x^2 \Lambda_{n_k}(dx) \right).$$

Thus

$$\lim_{\delta \searrow 0} \limsup_{k\to\infty} \int_{0<|x|\le\delta} x^r \Lambda_{n_k}(dx) \le \lim_{\delta\searrow 0} \delta^{r-2} a^h = 0.$$

We also have by vague convergence of Λ_{n_k} to Λ,

$$\lim_{k\to\infty} \int_{\delta<|x|\le h} x^r \Lambda_{n_k}(\mathrm{d}x) = \int_{\delta<|x|\le h} x^r \Lambda(\mathrm{d}x).$$

Write $L = \Lambda \circ T^{-1}$, where $T(x) = \left(x, x^2\right)$. Now on account of (21) we can readily infer that $\Lambda_{n_k} \circ T^{-1}$ converges vaguely to L on $\overline{\mathbb{R}}^2 \setminus \{(0,0)\}$. Thus by using the bivariate version of Theorem 15.14 of Kallenberg [8], we get after a little algebra that (18) holds with (U, W) having characteristic function

$$\exp\left\{ -\theta_1^2 a/2 + \mathrm{i}\left(b^h\theta_1 + a\theta_2\right) + \int_{\mathbb{R}\setminus\{0\}} \left(\exp\left(\mathrm{i}\left(\theta_1 x + \theta_2 x^2\right)\right) - 1 - \mathrm{i}\theta_1 x \mathbf{1}\{|x| \le h\}\right) \Lambda(\mathrm{d}x) \right\}, \tag{24}$$

for any $h > 0$ such that $\Lambda\{|x| = h\} = 0$. Note that in applying the bivariate version of Theorem 15.14 of Kallenberg [8], we get, using $a^h_{n_k} \to a^h$ and (23), that

$$\begin{pmatrix} a^h_{n_k} & \int_{0<|x|\le h} x^3 \Lambda_{n_k}(\mathrm{d}x) \\ \int_{0<|x|\le h} x^3 \Lambda_{n_k}(\mathrm{d}x) & \int_{0<|x|\le h} x^4 \Lambda_{n_k}(\mathrm{d}x) \end{pmatrix} \to$$

$$\begin{pmatrix} a^h & \int_{0<|x|\le h} x^3 \Lambda(\mathrm{d}x) \\ \int_{0<|x|\le h} x^3 \Lambda(\mathrm{d}x) & \int_{0<|x|\le h} x^4 \Lambda(\mathrm{d}x) \end{pmatrix}$$

and

$$\begin{pmatrix} b^h_{n_k} \\ a_{n_k} + \int_{0<|x|\le 1} x^2 \Lambda_{n_k}(\mathrm{d}x) \end{pmatrix} \to \begin{pmatrix} b^h \\ a + \int_{0<|x|\le 1} x^2 \Lambda(\mathrm{d}x) \end{pmatrix}.$$

We see then that the resulting limiting infinitely divisible vector has in its defining characteristic function the Lévy measure $L = \Lambda \circ T^{-1}$ on $\overline{\mathbb{R}}^2 \setminus \{(0,0)\}$ and the matrix and constant vector, respectively,

$$\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} b \\ a + \int_{0<|x|\le 1} x^2 \Lambda(\mathrm{d}x) \end{pmatrix}.$$

For very similar details see the proof of Lemma 4 of Giné and Mason [5]. Since

$$b^h = b - \int_{h<|x|\le 1} x\Lambda\,(\mathrm{d}x)\,,$$

we get that the characteristic function (24) is equal to (19). □

As in Sect. 1, let $\xi_1, \xi_2, \ldots,$ be i.i.d. nondegenerate random variables with cumulative distribution function F, and for each integer $n \ge 1$ denote the sums $\mathbb{S}_n = \sum_{i=1}^n \xi_i$ and $\mathbb{V}_n = \sum_{i=1}^n \xi_i^2$. Suppose that there exist a subsequence $\{n_k\} \subset \{n\}$ and norming constants $B\,(n_k)$ such that

$$\mathbb{S}_{n_k}/B\,(n_k) \xrightarrow{\mathrm{D}} U, \tag{25}$$

where U is an inf. div. rv with triplet (b, a, Λ).

Lemma 2.2. *Whenever (25) holds,*

$$\left(\mathbb{S}_{n_k}/B\,(n_k)\,, \mathbb{V}_{n_k}/B^2\,(n_k)\right) \xrightarrow{\mathrm{D}} (U, W)\,, \tag{26}$$

where (U, W) *has joint characteristic function (19).*

Proof. By Corollary 15.16 of Kallenberg [8], (25) occurs if and only if

$$n_k \mathcal{L}\,(\xi/B\,(n_k)) \text{ converges vaguely to } \Lambda \text{ on } \mathbb{R}\setminus\{0\} \tag{27}$$

and for any $h > 0$ such that $\Lambda\{|x| = h\} = 0$

$$n_k E\left[(\xi/B\,(n_k))^2\,\mathbf{1}\,\{|\xi/B\,(n_k)| \le h\}\right] \to a^h \tag{28}$$

and

$$n_k E\left[(\xi/B\,(n_k))\,\mathbf{1}\,\{|\xi/B\,(n_k)| \le h\}\right] \to b^h,$$

where a^h and b^h are defined as in (22). Also, as in the proof of Lemma 2.1, for every $r > 2$,

$$n_k E\left[(\xi/B\,(n_k))^r\,\mathbf{1}\,\{|\xi/B\,(n_k)| \le h\}\right] \to \int_{0<|x|\le h} x^r\Lambda\,(\mathrm{d}x)\,,$$

and similarly as in the proof of Lemma 2.1, $n_k \mathcal{L}\left(\xi/B\,(n_k)\,, \xi^2/B^2\,(n_k)\right)$ converges vaguely to the measure L on $\overline{\mathbb{R}}^2 \setminus \{(0,0)\}$. □

Remark 2. In the sequel we shall only need the special case of Lemma 2.2 when $\xi_1, \xi_2, \ldots,$ are i.i.d $\xi = X_1$, where X_1 is inf. div. with canonical triplet (γ, σ^2, Π).

2.1 Proof of Theorem 1.2

It is more efficient to prove Theorem 1.2 first. By Proposition 1.2, whenever

$$X_t/\sqrt{V_t} \xrightarrow{\mathrm{D}} Y, \text{ as } t \to \infty,$$

where Y does not put positive mass ± 1, then X_t is centered stochastically compact at infinity with a norming function b_t. This implies by Lemmas 2.1 and 2.2 that if $\xi_1, \ldots, \xi_m$ are i.i.d. X_1, there exists a positive norming b_m such that

$$\left(X_m/b_m, V_m/b_m^2\right) \text{ and } \left(\mathbb{S}_m/b_m, \mathbb{V}_m/b_m^2\right)$$

have the same nondegenerate subsequential distributional limits as $m \to \infty$. Thus both

$$X_m/\sqrt{V_m} \text{ and } \mathbb{S}_m/\sqrt{\mathbb{V}_m}$$

converge in distribution to the same nondegenerate rv Y that does not put positive mass on ± 1. The proof of Theorem 1.2 now follows from the Chistyakov and Götze [3] result. □

2.2 Proof of Theorem 1.1

Assume that (9) holds, where $P\{|Y| = 1\} = 0$. We know by Proposition 1.1 that this forces X_t to be centered stochastically compact at 0. Thus there exists a norming function a_t such that every subsequence t_k converging to zero contains a further subsequence s_n with

$$\left(X_{ts_n}/a_{s_n}, V_{ts_n}/a_{s_n}^2\right)_{t\geq 0} \xrightarrow{\mathrm{D}} (U_t, W_t)_{t\geq 0}, \text{ as } n \to \infty, \tag{29}$$

where the Lévy process (U, W), which may depend on the subsequence s_n, has joint characteristic function

$$Ee^{\mathrm{i}(\theta_1 U_t + \theta_2 W_t)}$$
$$= \exp\left\{\mathrm{i}t\left(\theta_1 b + \theta_2 a\right) - t\theta_1^2 a/2 + t\int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\theta_1 x + \theta_2 x^2)} - 1 - \mathrm{i}\theta_1 x \mathbf{1}_{\{|x|\leq 1\}}\right)\Lambda(\mathrm{d}x)\right\},$$

with $b \in \mathbb{R}$, $a \geq 0$ and Λ being a Lévy measure on $\mathbb{R}\setminus\{0\}$. See [12] Theorem 2.3 for the functional convergence in (29).

We first claim that if U contains a normal component, i.e. $a > 0$, this forces Y to be standard normal. This will be a consequence of the following lemma.

Lemma 2.3. *If U contains a normal component ($a \neq 0$), then one can find a subsequence $t' \searrow 0$ such that*

$$\left(X_{t'}/a_{t'}, V_{t'}/a_{t'}^2\right) \xrightarrow{\mathrm{D}} (aZ,\ a), \tag{30}$$

where Z is standard normal.

Proof. Using (29), we see that, for each fixed $m > 1$,

$$\left(X_{s_n/m}/\left(a_{s_n}/\sqrt{m}\right), V_{s_n/m}/\left(a_{s_n}^2/m\right)\right) \xrightarrow{\mathrm{D}} (\sqrt{m}U_{1/m}, mW_{1/m}), \text{ as } n \to \infty,$$

where $(\sqrt{m}U_{1/m}, mW_{1/m})$ has characteristic function

$$= \exp\left\{\mathrm{i}\left(\frac{\theta_1 b}{\sqrt{m}} + \theta_2 a\right) - \theta_1^2 a/2 + \frac{1}{m}\int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1 - \mathrm{i}\sqrt{m}\theta_1 x \mathbf{1}_{\{|x|\le 1\}}\right)\Lambda(\mathrm{d}x)\right\},$$

which as we will show converges as $m \to \infty$ to $\exp\left\{-\theta_1^2 a/2 + \mathrm{i}\theta_2 a\right\}$. Thus with some abuse of notation we can extract a sequence s_{n_k}/m_k converging to 0 so that as $k \to \infty$,

$$\left(X_{s_{n_k}/m_k}/\left(a_{s_{n_k}}/\sqrt{m_k}\right), V_{s_{n_k}/m_k}/\left(a_{s_{n_k}}^2/m_k\right)\right) \xrightarrow{\mathrm{D}} (aZ,\ a), \tag{31}$$

having characteristic function $\exp\left\{-\theta_1^2 a/2 + \mathrm{i}\theta_2 a\right\}$. Actually to show (31) it remains to prove that

$$\lim_{m\to\infty} \frac{1}{m}\int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1 - \mathrm{i}\sqrt{m}\theta_1 x \mathbf{1}_{\{|x|\le 1\}}\right)\Lambda(\mathrm{d}x) = 0. \tag{32}$$

To see why (32) is true notice that

$$\limsup_{m\to\infty} \frac{1}{m}\left|\int_{1/\sqrt{m}\le|x|<\infty} \left(e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1\right)\Lambda(\mathrm{d}x)\right| \le \limsup_{m\to\infty} \frac{2}{m}\Lambda\left(1/\sqrt{m} \le |x| < \infty\right) = 0.$$

Also for all $0 < \delta < 1$

$$\begin{aligned}
\limsup_{m\to\infty} \frac{1}{m}\left|\int_{1/\sqrt{m}\le|x|\le 1} \mathrm{i}\sqrt{m}x\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x)\right| &\le \limsup_{m\to\infty} \frac{1}{\sqrt{m}}\int_{1/\sqrt{m}\le|x|\le\delta} |x|\,\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x) \\
&\quad + \limsup_{m\to\infty} \frac{1}{\sqrt{m}}\left|\int_{\delta\le|x|\le 1} x\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x)\right| \\
&\le \limsup_{m\to\infty} \int_{1/\sqrt{m}\le|x|\le\delta} x^2\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x) \\
&= \int_{0<|x|\le\delta} x^2\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x).
\end{aligned}$$

Since $\delta > 0$ can be made arbitrarily small we get

$$\lim_{m\to\infty} \frac{1}{m} \left| \int_{1/\sqrt{m}\le|x|\le 1} \mathrm{i}\sqrt{m}x\mathbf{1}_{\{|x|\le 1\}}\Lambda(\mathrm{d}x) \right| = 0. \tag{33}$$

Thus to complete the proof of (32) it suffices to show that

$$\lim_{m\to\infty} \frac{1}{m} \int_{0<|x|\le 1/\sqrt{m}} \left(e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1 - \mathrm{i}\sqrt{m}\theta_1 x\mathbf{1}_{\{|x|\le 1\}} \right) \Lambda(\mathrm{d}x) = 0. \tag{34}$$

Now the LHS of (34) does not exceed

$$\frac{1}{m} \int_{0<|x|\le 1/\sqrt{m}} \left| e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1 - \mathrm{i}\left(\sqrt{m}\theta_1 x\mathbf{1}_{\{|x|\le 1\}} + m\theta_2 x^2\right) \right| \Lambda(\mathrm{d}x)$$
$$+ \int_{0<|x|\le 1/\sqrt{m}} \theta_2 x^2 \Lambda(\mathrm{d}x),$$

and for some $C > 0$

$$\frac{1}{m} \int_{0<|x|\le 1/\sqrt{m}} \left| e^{\mathrm{i}(\sqrt{m}\theta_1 x + m\theta_2 x^2)} - 1 - \mathrm{i}\left(\sqrt{m}\theta_1 x\mathbf{1}_{\{|x|\le 1\}} + m\theta_2 x^2\right) \right| \Lambda(\mathrm{d}x)$$
$$\le \frac{C}{m} \int_{0<|x|\le 1/\sqrt{m}} \left| \sqrt{m}\theta_1 x + m\theta_2 x^2 \right|^2 \Lambda(\mathrm{d}x),$$

which for some $D > 0$ depending on θ_1 and θ_2 is

$$\le D \int_{0<|x|\le 1/\sqrt{m}} x^2 \Lambda(\mathrm{d}x).$$

Since the limit of this as $m \to \infty$ is 0, we have shown (34), which together with (33) gives (32). □

Hence if there exists a nonzero normal component in the characteristic function of U in (29) for convergence of $(X_t/a_t, V_t/a_t^2)$ along a subsequence s_n, then Y must be standard normal. Since in this case for some subsequence t' we have $X_{t'}/\sqrt{V_{t'}} \xrightarrow{\mathrm{D}} Z$, as $t' \searrow 0$ we get $Y = Z$ in (9).

From now on we shall assume that Y is not standard normal, which means that $a = 0$ in the characteristic function of the (U, W) appearing in (29) for convergence of $(X_t/a_t, V_t/a_t^2)$ along a subsequence s_n. Also note that (U, W) may be different for different subsequences. However, in all cases, by assumption (9),

$$U_t/\sqrt{W_t} \stackrel{\mathrm{D}}{=} Y, \text{ for each } t > 0.$$

Note that Theorem 2.1 (iii) of Maller and Mason [12] implies that $P\{W > 0\} = 1$. Moreover, it can be shown that for any integer $m \geq 1$,

$$\frac{U_m}{\sqrt{W_m}} \stackrel{\mathrm{D}}{=} \frac{U_{(1)} + \cdots + U_{(m)}}{\sqrt{W_{(1)} + \cdots + W_{(m)}}} \stackrel{\mathrm{D}}{=} Y, \tag{35}$$

where $\big(U_{(1)}, W_{(1)}\big), \ldots, \big(U_{(m)}, W_{(m)}\big)$ are i.i.d. (U_1, W_1). To see this, observe that for any fixed integer $m \geq 1$,

$$\big(X_{ms_n}/a_{s_n}, V_{ms_n}/a_{s_n}^2\big) \xrightarrow{\mathrm{D}} \big(U_{(1)} + \cdots + U_{(m)}, W_{(1)} + \cdots + W_{(m)}\big) \stackrel{\mathrm{D}}{=} (U_m, V_m)\,.$$

We claim that

$$U_t/\sqrt{W_t} \xrightarrow{\mathrm{D}} Y, \text{ as } t \to \infty. \tag{36}$$

This follows from (35) combined with the facts that

$$W_{m+1} - W_m \stackrel{\mathrm{D}}{=} W_{(1)} = O_p(1),\ W_{(1)} + \cdots + W_{(m+1)} \xrightarrow{\mathrm{P}} \infty$$

and

$$\sup_{m<t\leq m+1} |U_t - U_m| \stackrel{\mathrm{D}}{=} \sup_{0<t\leq 1} |U_t| = O_p(1),$$

which together imply that

$$\sup_{m<t\leq m+1} \left|U_t/\sqrt{W_t} - U_m/\sqrt{W_m}\right| \xrightarrow{\mathrm{P}} 0.$$

Therefore we can apply Theorem 1.2 combined with the fact that Y is not standard normal to conclude that U is in the domain of attraction of a strictly stable law of index $0 < \alpha < 2$, and thus $Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}$. □

Remark 3. We do not have such a complete picture of the distributional limits of $X_t/\sqrt{V_t}$ as $t \searrow 0$ as Chistyakov and Götze [3] obtained for self-normalized sums in their Theorem 1.1. All we can say is that if

$$X_t/\sqrt{V_t} \xrightarrow{\mathrm{D}} Y, \text{ as } t \searrow 0,$$

where Y does not place positive mass on any constant then either $Y \stackrel{\mathrm{D}}{=} Z$ or $Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}$. Only in the case when $Y \stackrel{\mathrm{D}}{=} Z$ do we know that this happens if and only if for some norming function b_t,

$$X_t/b_t \xrightarrow{\mathrm{D}} Z, \text{ as } t \searrow 0.$$

This was proved in Theorem 2.4 of [12]. The story for the case $Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}$ is not complete. Presently all that we can infer is that if for some $0 < \alpha < 2$,

$$X_t/b_t \stackrel{\mathrm{D}}{\longrightarrow} U^\alpha, \text{ as } t \searrow 0,$$

then

$$X_t/\sqrt{V_t} \stackrel{\mathrm{D}}{\longrightarrow} U^\alpha/\sqrt{V^\alpha}, \text{ as } t \searrow 0.$$

However, right now, we cannot go the other way, except under the assumption of symmetry. See [10].

3 Proofs of Propositions 1.1 and 1.2

3.1 *Proof of Proposition 1.1*

Suppose $X_t/\sqrt{V_t}$ is relatively compact as $t \downarrow 0$ and no subsequential limit has positive mass at ± 1. Then by Theorem 3.1 of [12] we have

$$\limsup_{x \downarrow 0} \frac{x|\nu(x)|}{U(x)} = \limsup_{x \downarrow 0} \frac{x|\nu(x)|}{x^2\overline{\Pi}(x) + V(x)} < \infty, \tag{37}$$

while by Proposition 5.5 of [12] we have

$$\limsup_{x \downarrow 0} \frac{x^2\overline{\Pi}(x)}{V(x)} < \infty, \tag{38}$$

since if (38) fails then there is a subsequential limit rv of $X_t/\sqrt{V_t}$ (as $t \downarrow 0$) with positive mass at ± 1. Now (37) and (38) imply

$$\limsup_{x \downarrow 0} \frac{x|\nu(x)|}{V(x)} < \infty$$

which together with (38) gives (15). □

3.2 *Proof of Proposition 1.2*

The proof of Proposition 1.2 will be a consequence of the following two propositions and theorem, which are the large time analogs of their small time versions given in Propositions 5.1 and 5.5 and Theorem 3.1 of [12].

Recall the definition of $U(x)$ given in (12). Note that, after integrating by parts,

$$U(x) = V(x) + x^2\overline{\Pi}(x),\ x > 0. \tag{39}$$

The function $U(x)$ is continuous, in fact, differentiable, at each $x > 0$, with

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{U(x)}{x^2}\right) = \frac{-2\left(U(x) - x^2\overline{\Pi}(x)\right)}{x^3}.$$

Further,

$$U(x) - x^2\overline{\Pi}(x) = \sigma^2 + 2\int_0^x y\left(\overline{\Pi}(y) - \overline{\Pi}(x)\right)\mathrm{d}y \geq 2\int_0^x y\left(\overline{\Pi}(y) - \overline{\Pi}(x)\right)\mathrm{d}y.$$

The right-hand side here could be 0 only if $\overline{\Pi}(y)$ is constant on $(0, x]$, and since $\overline{\Pi}(\infty) = 0$, as long as $\overline{\Pi}(x) > 0$ for some $x > 0$, we see that $x^{-2}U(x)$ is strictly decreasing for large enough x, and $x^{-2}U(x) \to \infty\mathbf{1}_{\{\sigma^2>0\}} + \overline{\Pi}(0+)\mathbf{1}_{\{\sigma^2=0\}} > 0$ as $x \searrow 0$, while $x^{-2}U(x) \searrow 0$ as $x \nearrow \infty$.

In view of the monotonicity of $x^{-2}U(x)$ just established, for each $\lambda > 0$, once t is large enough, depending on λ, for $x^{-2}U(x) < \infty\mathbf{1}_{\{\sigma^2>0\}} + \overline{\Pi}(0+)\mathbf{1}_{\{\sigma^2=0\}}$, the function

$$b_\lambda(t) := \inf\{x > 0 : x^{-2}U(x) \leq (\lambda t)^{-1}\}$$

is finite, positive, is such that $b_\lambda(t) \to \infty$ as $t \to \infty$, and is such that

$$\frac{tU(b_\lambda(t))}{b_\lambda^2(t)} = \frac{1}{\lambda}. \tag{40}$$

Further, $x^{-2}U(x)$ has no intervals of constancy once x is large enough, because of its strict monotonicity, so $b_\lambda(t)$ is continuous and strictly increasing for each $\lambda > 0$, for large enough t.

In the sequel we shall often use the following decomposition, or a variant of it (see [15], Theorem 19.2, p. 120, or Eq. (6.1) of [4]):

$$X_t = t\nu(b) + \sigma Z_t + X_t^{(S,b)} + X_t^{(B,b)},\ t \geq 0,\ b > 0, \tag{41}$$

where Z_t is a standard Brownian motion, $X_t^{(S,b)}$ is the compensated sum of "small" jumps, i.e.

$$X_t^{(S,b)} = \text{a.s.}\lim_{\varepsilon \downarrow 0}\left(\sum_{0<s\leq t}\Delta X_s\mathbf{1}_{\{\varepsilon<|\Delta X_s|\leq b\}} - t\int_{\varepsilon<|x|\leq b} x\Pi(\mathrm{d}x)\right),\ t \geq 0,$$

and $X_t^{(B,b)}$ is the "big" jumps, i.e.,

$$X_t^{(B,b)} = \sum_{0<s\le t} \Delta X_s \mathbf{1}_{\{|\Delta X_s|>b\}},\ t \ge 0.$$

Further, the processes $(Z_t)_{t\ge 0}$, $(X_t^{(S,b)})_{t\ge 0}$ and $(X_t^{(B,b)})_{t\ge 0}$ are all independent.

Theorem 3.1. *We have that*

$$\frac{X_t}{\sqrt{V_t}} \text{ is relatively compact as } t\to\infty \text{ if and only if } \limsup_{x\to\infty} \frac{x|\nu(x)|}{U(x)} < \infty. \quad (42)$$

We will deduce Theorem 3.1 from the following analogue of Theorem 2 of [7] and Corollary 3.1 is immediate from it.

Proposition 3.1. *There is a nonstochastic function $a(t)$ such that*

$$(X_t - a(t))/\sqrt{V_t} \text{ is relatively compact as } t\to\infty \quad (43)$$

if and only if

$$\limsup_{t\to\infty} \frac{|t\nu(b_\lambda(t)) - a(t)|}{b_\lambda(t)} < \infty, \quad (44)$$

for all small, and hence, all, $\lambda > 0$.

Corollary 3.1 (Corollary to Proposition 3.1).

(i) $(X_t - t\nu(b_\lambda(t))/\sqrt{V_t}$ *is always relatively compact as* $t\to\infty$*, for any* $\lambda > 0$.
(ii) If X_t *is symmetric, then* $X_t/\sqrt{V_t}$ *is always relatively compact as* $t\to\infty$.

Proof of Proposition 3.1.

(i) First suppose $EX_1^2 < \infty$. From (40) we see that $b_\lambda(t) \asymp \sqrt{t}$ as $t\to\infty$. The convergences

$$\frac{X_t - tEX_1}{\sqrt{t\,\mathrm{Var}X_1}} \xrightarrow{\mathrm{D}} N(0,1)$$

and

$$V_t/t \xrightarrow{\mathrm{P}} EX_1^2,$$

as $t\to\infty$ can be found in [2, 15] or see [4]. So (43) holds with $a(t) = tEX_1$, and with this choice,

$$\frac{t\,|\nu(b_\lambda(t)) - a(t)|}{b_\lambda(t)} = \frac{\left|t\int_{|y|>b_\lambda(t)} y\Pi(\mathrm{d}y)\right|}{b_\lambda(t)} \le \frac{t\int_{|y|>0} y^2\Pi(\mathrm{d}y)}{b_\lambda^2(t)} = O(1).$$

So Proposition 3.1 is true when $EX_1^2 < \infty$.

(ii) Next suppose $EX_1^2 = \infty$. In particular, this means $\overline{\Pi}(x) > 0$ for all $x > 0$. Fix $\lambda > 0$ and then take $t > 0$ big enough, depending on λ, for $b_\lambda(t) > 1$.

From (41) with $b = 1$ we have, for $t > 0$,

$$\begin{aligned} X_t &= t\nu(1) + \sigma Z_t + X_t^{(S,1)} + X_t^{(B,1)} \\ &= t\gamma + X_t^{(B,1)} + O_P(\sqrt{t}), \text{ as } t \to \infty, \end{aligned}$$

because $X_t^{(S,1)}$ is a mean 0, finite variance Lévy process. Also

$$\begin{aligned} V_t &= \sigma^2 t + \sum_{0<s\le t} (\Delta X_s)^2 \mathbf{1}_{\{|\Delta X_s| \le 1\}} + \sum_{0<s\le t} (\Delta X_s)^2 \mathbf{1}_{\{|\Delta X_s| > 1\}} \\ &=: \sigma^2 t + V_t^{(S,1)} + V_t^{(B,1)} \\ &= V_t^{(B,1)} + O_P(t), \text{ as } t \to \infty. \end{aligned}$$

This is true because $V_t^{(S,1)}$ is a Lévy process with finite mean, so $V_t^{(S,1)}/t = O_P(1)$ as $t \to \infty$ by the weak law of large numbers. But $V_t^{(B,1)}/t \xrightarrow{\text{P}} \infty$ as $t \to \infty$ since $EX_1^2 = \infty$, so $V_t/V_t^{(B,1)} \xrightarrow{\text{P}} 1$ as $t \to \infty$, and thus from

$$\begin{aligned} \frac{X_t^{(B,1)} + t\gamma - a(t)}{\sqrt{V_t^{(B,1)}}} &= \frac{X_t - a(t)}{\sqrt{V_t}} \sqrt{\frac{V_t}{V_t^{(B,1)}}} + O_P\left(\sqrt{\frac{t}{V_t^{(B,1)}}}\right) \\ &= \frac{X_t - a(t)}{\sqrt{V_t}} (1 + o_P(1)) + o_P(1), \text{ as } t \to \infty, \end{aligned} \tag{45}$$

we see that (43) holds if and only if

$$\left(X_t^{(B,1)} - \widetilde{a}(t)\right) / \sqrt{V_t^{(B,1)}} \text{ is relatively compact as } t \to \infty,$$

for some $\widetilde{a}(t) = a(t) - t\gamma$. So we can ignore small jumps in X and assume X is compound Poisson with no drift, no Brownian component, and all jumps exceeding 1 in magnitude.

Thus in (12) we take $\gamma = \sigma^2 = 0$, and can write

$$X_t = \sum_{i=1}^{N_t} J_i, \tag{46}$$

and

$$V_t = \sum_{0<s\le t} (\Delta X_s)^2 \mathbf{1}_{\{|\Delta X_s| > 1\}} = \sum_{i=1}^{N_t} J_i^2, \tag{47}$$

for $(J_i)_{i=1,2,\ldots}$ i.i.d. and distributed as $\Pi(\mathrm{d}x)\mathbf{1}_{\{|x|>1\}}/\overline{\Pi}(1)$, and $(N_t)_{t\geq 0}$ independently distributed as a Poisson process with rate $\overline{\Pi}(1)$.

We decompose X_t as

$$X_t = T_t(\lambda) + R_t(\lambda), \tag{48}$$

where

$$T_t(\lambda) := \sum_{i=1}^{N_t} J_i \mathbf{1}_{\{|J_i|\leq b_\lambda(t)\}} = \sum_{i=1}^{N_t} J_i \mathbf{1}_{\{1<|J_i|\leq b_\lambda(t)\}},$$

and

$$R_t(\lambda) := \sum_{i=1}^{N_t} J_i \mathbf{1}_{\{|J_i|>b_\lambda(t)\}}. \tag{49}$$

Then we can calculate

$$E(T_t(\lambda)) = t\overline{\Pi}(1)\int_{1<|x|\leq b_\lambda(t)} x\Pi(\mathrm{d}x)/\overline{\Pi}(1) = t\nu(b_\lambda(t)),$$

and

$$\mathrm{Var}(T_t(\lambda)) = t\int_{1<|x|\leq b_\lambda(t)} x^2\Pi(\mathrm{d}x) \leq tU(b_\lambda(t)).$$

We can thus write, for any $L > 0$,

$$\begin{aligned} &P\left(|t\nu(b_\lambda(t)) - a(t)| > 3L^2 b_\lambda(t)\right) \\ &\leq P\left(|T_t(\lambda) - ET_t(\lambda)| > L^2 b_\lambda(t)\right) + P\left(|T_t(\lambda) - a(t)| > 2L^2 b_\lambda(t)\right), \end{aligned} \tag{50}$$

and we proceed by estimating the quantities on the right-hand side of (50).

By Chebyshev's inequality, for any $L > 0$, $K > 0$,

$$\begin{aligned} P\left(|T_t(\lambda) - ET_t(\lambda)| > LKb_\lambda(t)\right) &\leq \frac{\mathrm{Var}(T_t(\lambda))}{L^2K^2b_\lambda^2(t)} \\ &\leq \frac{tU(b_\lambda(t))}{L^2K^2b_\lambda^2(t)} = \frac{1}{L^2K^2\lambda}. \end{aligned} \tag{51}$$

With $K = L$ this gives a bound for the first term on the right-hand side of (50). The second term on the right-hand side of (50) does not exceed

$$P\left(|T_t(\lambda) - a(t)| > 2L^2 b_\lambda(t), Lb_\lambda(t) \geq \sqrt{V_t}\right) + P\left(Lb_\lambda(t) < \sqrt{V_t}\right). \tag{52}$$

Let

$$U_t(\lambda) := \sum_{i=1}^{N_t} J_i^2 \wedge b_\lambda^2(t). \tag{53}$$

It's not hard to check that

$$E\,(U_t(\lambda)) = \lambda^{-1}b_\lambda^2(t) - tV(1). \tag{54}$$

On the event $\{\max_{1\le i\le N_t}|J_i| \le b_\lambda(t)\}$ we have $V_t = U_t(\lambda)$, so

$$\begin{aligned}
P\left(\sqrt{V_t} > Lb_\lambda(t)\right) &\le P\left(V_t > L^2b_\lambda^2(t),\ \max_{1\le i\le N_t}|J_i| \le b_\lambda(t)\right)\\
&\qquad + 1 - P\left(\max_{1\le i\le N_t}|J_i| \le b_\lambda(t)\right)\\
&\le P\left(U_t(\lambda) > L^2b_\lambda^2(t)\right) + 1 - \sum_{n\ge 0}P^n(|J_1| \le b_\lambda(t))P(N_t = n)\\
&\le \frac{E(U_t(\lambda))}{L^2b_\lambda^2(t)} + 1 - \sum_{n\ge 0}e^{-t\overline{\Pi}(1)}(t\overline{\Pi}(1)P(|J_1| \le b_\lambda(t)))^n/n!\\
&= \frac{\left(\lambda^{-1}b_\lambda^2(t) - tV(1)\right)}{L^2b_\lambda^2(t)} + 1 - e^{-t\overline{\Pi}(1)P(|J_1|>b_\lambda(t))}\\
&\le \lambda^{-1}L^{-2} + 1 - e^{-t\overline{\Pi}(b_\lambda(t))} \le \lambda^{-1}L^{-2} + 1 - e^{-\lambda^{-1}}.
\end{aligned} \tag{55}$$

(In the last inequality, recall that $x^2\overline{\Pi}(x) \le U(x)$, so $t\,\overline{\Pi}(b_\lambda(t)) \le 1/\lambda$.) (55) gives a bound for the second term in (52).

Next, to estimate $R_t(\lambda)$ in (49), put

$$S_t(\lambda) := \sum_{i=1}^{N_t}\mathbf{1}_{\{|J_i|>b_\lambda(t)\}}.$$

Then by Cauchy-Schwarz,

$$|R_t(\lambda)|^2 \le \left(\sum_{i=1}^{N_t}J_i^2\right)\left(\sum_{i=1}^{N_t}\mathbf{1}_{\{|J_i|>b_\lambda(t)\}}\right) = V_tS_t(\lambda).$$

Thus, for $L > 0$,

$$\begin{aligned}
P\left(|R_t(\lambda)| > L\sqrt{V_t}\right) &\le P(S_t(\lambda) > L^2) \le L^{-2}E(S_t(\lambda))\\
&= L^{-2}(t\overline{\Pi}(1))\overline{\Pi}(b_\lambda(t))/\overline{\Pi}(1) \le L^{-2}\lambda^{-1}.
\end{aligned} \tag{56}$$

So (cf. (48)) we see that the first term in (52) does not exceed

$$
\begin{aligned}
&P\left(|T_t(\lambda) - a(t)| > 2L\sqrt{V_t}\right) \\
&\le P\left(|X_t - a(t)| > L\sqrt{V_t}\right) + P(|R_t(\lambda)| > L\sqrt{V_t}) \\
&\le P\left(|X_t - a(t)| > L\sqrt{V_t}\right) + L^{-2}\lambda^{-1}, \qquad (57)
\end{aligned}
$$

by (56). Going back to (50), we put together (52) with $K = L$, and the bounds in (55) and (57), to deduce that

$$
\begin{aligned}
&P\left(|t\nu(b_\lambda(t)) - a(t)| > 3L^2 b_\lambda(t)\right) = \mathbf{1}_{\{|t\nu(b_\lambda(t))-a(t)|>3L^2b_\lambda(t)\}} \\
&\le \lambda^{-1}L^{-4} + \lambda^{-1}L^{-2} + 1 - e^{-\lambda^{-1}} + P\left(|X_t - a(t)| > L\sqrt{V_t}\right) + L^{-2}\lambda^{-1}.
\end{aligned}
$$

Choose L so large that $\lambda^{-1}L^{-4} + 2L^{-2}\lambda^{-1} < e^{-\lambda^{-1}}/2$. Then

$$
\mathbf{1}_{\{|t\nu(b_\lambda(t))-a(t)|>3L^2b_\lambda(t)\}} \le P(|X_t - a(t)| > L\sqrt{V_t}) + 1 - e^{-\lambda^{-1}}/2.
$$

Now assume (43), i.e., that $|X_t - a(t)|/\sqrt{V_t}$ is relatively compact as $t \to \infty$. Letting $t \to \infty$ then $L \to \infty$ gives

$$
\lim_{L\to\infty} \limsup_{t\to\infty} \mathbf{1}_{\{|t\nu(b_\lambda(t))-a(t)|>3L^2b_\lambda(t)\}} \le 1 - e^{-\lambda^{-1}}/2 < 1.
$$

So, for $L \ge$ some $L_0(\lambda) > 0$, and $t \ge$ some $t_0(L, \lambda) > 0$, we have $\mathbf{1}_{\{|t\nu(b_\lambda(t))-a(t)|>3L^2b_\lambda(t)\}} < 1$, hence $|t\nu(b_\lambda(t)) - a(t)| \le 3L^2 b_\lambda(t)$. Thus for $t \ge t_0(L_0, \lambda)$,

$$
\frac{|t\nu(b_\lambda(t)) - a(t)|}{b_\lambda(t)} \le 3L_0^2,
$$

which implies (44).

For the converse, suppose (44) holds for all sufficiently small $\lambda > 0$. Fix such a $\lambda \in (0, 1)$. Therefore by (44) we can find a $t_\lambda > 0$ such that

$$
c(\lambda) := \sup_{t\ge t_\lambda} \frac{|t\nu(b_\lambda(t)) - a(t)|}{b_\lambda(t)} < \infty. \qquad (58)
$$

Then for all $t \ge t_\lambda$ and any $L > 0$

$$
\begin{aligned}
&P\left(|X_t - a(t)| > 3L\sqrt{V_t}\right) \le P\left(|X_t - t\nu(b_\lambda(t))| > 3L\sqrt{V_t} - c(\lambda)b_\lambda(t)\right) \\
&\le P\left(|X_t - t\nu(b_\lambda(t))| > 2L\sqrt{V_t}\right) + P\left(L\sqrt{V_t} \le c(\lambda)b_\lambda(t))\right). \qquad (59)
\end{aligned}
$$

To deal with the first term on the right-hand side, take $K \in (0, \lambda^{-1/2})$, and suppose $t > 0$ is so large that $tV(1) < (\lambda^{-1} - K^2)b_\lambda^2(t)$. This is possible since

$b_\lambda^2(t)/t = \lambda U(b_\lambda(t)) \to \infty$ as $t \to \infty$. Recall that $X_t = T_t(\lambda) + R_t(\lambda)$ by (48), where $ET_t(\lambda) = t\nu(b_\lambda(t))$, and argue as follows:

$$\begin{aligned} P\left(|X_t - t\nu(b_\lambda(t))| > 2L\sqrt{V_t}\right) &\le P\left(|T_t(\lambda) - t\nu(b_\lambda(t))| > L\sqrt{V_t}\right) \\ &\quad + P\left(|R_t(\lambda)| > L\sqrt{V_t}\right) \\ &\le P\left(|T_t(\lambda) - t\nu(b_\lambda(t))| > LKb_\lambda(t)\right) + P\left(\sqrt{V_t} \le Kb_\lambda(t)\right) \\ &\quad + P\left(|R_t(\lambda)| > L\sqrt{V_t}\right) \\ &\le \frac{1}{L^2K^2\lambda} + P\left(\sqrt{V_t} \le Kb_\lambda(t)\right) + \frac{1}{\lambda L^2} \quad \text{(by (51) and (56))}. \end{aligned} \tag{60}$$

Recall (53) and note that $V_t \ge U_t(\lambda)$, so for $t > 0$

$$P\left(\sqrt{V_t} > Kb_\lambda(t)\right) \ge P\left(U_t(\lambda) > K^2b_\lambda^2(t)\right).$$

Using a second moment version of Wald's lemma we see that

$$\begin{aligned} \mathrm{Var}\left(U_t\left(\lambda\right)\right) &\le t\overline{\Pi}(1)E\left(J_1^4 \wedge b_\lambda^4\left(t\right)\right) \\ &= tb_\lambda^4\left(t\right)\overline{\Pi}\left(b_\lambda\left(t\right)\right) + t\int_{1\le|y|\le b_\lambda(t)} y^4\Pi\left(dy\right) \\ &\le tb_\lambda^2\left(t\right)\left(b_\lambda^2\left(t\right)\overline{\Pi}\left(b_\lambda\left(t\right)\right) + \int_{1\le|y|\le b_\lambda(t)} y^2\Pi\left(dy\right)\right) \\ &\le tb_\lambda^2\left(t\right)U\left(b_\lambda\left(t\right)\right) = \lambda^{-1}b_\lambda^4\left(t\right). \end{aligned}$$

Recall that we keep $K^2 < \lambda^{-1}$ and $tV(1) < (\lambda^{-1} - K^2)b_\lambda^2(t)$. There is a one-sided Chebyshev inequality of the form $P(Y - EY < x) \ge x^2/(x^2 + \mathrm{Var}Y)$, for any rv Y and $x > 0$ (e.g., [1], p. 70). Apply this with $Y = -U_t(\lambda)$, recalling that $E(U_t(\lambda)) = \lambda^{-1}b_\lambda^2(t) - tV(1)$, by (54), to get

$$\begin{aligned} P(\sqrt{V_t} > Kb_\lambda(t)) &\ge P\left(U_t(\lambda) > K^2b_\lambda^2(t)\right) \\ &= P\left(-U_t(\lambda) + E(U_t(\lambda)) < (\lambda^{-1} - K^2)b_\lambda^2(t) - tV(1)\right) \\ &\ge \frac{\left((\lambda^{-1} - K^2)b_\lambda^2(t) - tV(1)\right)^2}{\left((\lambda^{-1} - K^2)b_\lambda^2(t) - tV(1)\right)^2 + \mathrm{Var}(U_t(\lambda))} \\ &\ge \frac{(1 - K^2\lambda)^2b_\lambda^4(t) - 2t\lambda(1 - K^2\lambda)b_\lambda^2(t)V(1)}{b_\lambda^4(t)(1 + \lambda) + t^2\lambda^2V^2(1)}. \end{aligned} \tag{61}$$

For the second term on the right-hand side of (59), use (61) with K replaced by $K_\lambda := c(\lambda)L^{-1}$, with $L > c(\lambda)\sqrt{\lambda}$, so $K_\lambda < \lambda^{-1/2}$. Thus, finally,

$$\begin{aligned} &P(|X_t - a(t)| > 3L\sqrt{V_t}) \\ &\le \frac{1}{\lambda L^2 K^2} + \left[1 - \frac{(1-K^2\lambda)^2 b_\lambda^4(t) - 2t\lambda(1-K^2\lambda)b_\lambda^2(t)V(1)}{b_\lambda^4(t)(1+\lambda) + t^2\lambda^2 V^2(1)}\right] \\ &\quad + \left[1 - \frac{(1-K_\lambda^2\lambda)^2 b_\lambda^4(t) - 2t\lambda(1-K_\lambda^2\lambda)b_\lambda^2(t)V(1)}{b_\lambda^4(t)(1+\lambda) + t^2\lambda^2 V^2(1)}\right] + \frac{1}{\lambda L^2}. \end{aligned}$$

Let $t \to \infty$, recalling that $t = o(b_\lambda^2(t))$, then $L \to \infty$, noting that $K_\lambda = c(\lambda)L^{-1} \to 0$, then let $K \downarrow 0$, to see that

$$\lim_{L\to\infty} \limsup_{t\to\infty} P(|X_t - a(t)| > 3L\sqrt{V_t}) \le \frac{2\lambda}{1+\lambda}.$$

Then let $\lambda \downarrow 0$ to get (43). □

Proof of Theorem 3.1. Suppose $X_t/\sqrt{V_t}$ is relatively compact but there is a sequence $x_k \to \infty$ such that

$$\frac{x_k|\nu(x_k)|}{U(x_k)} \to \infty, \text{ as } k \to \infty.$$

Let $t_k = x_k^2/U(x_k)$, so $x_k = b_1(t_k)$, in the notation of (40). Then

$$\frac{t_k|\nu(b_1(t_k))|}{b_1(t_k)} = \frac{t_k|\nu(x_k)|}{x_k} = \frac{x_k|\nu(x_k)|}{U(x_k)} \to \infty,$$

which contradicts (44) with $a(t) = 0$ and $\lambda = 1$.

Conversely, suppose $\limsup_{x\to\infty} x|\nu(x)|/U(x) < c < \infty$. Then with $a(t) \equiv 0$ we have

$$\limsup_{t\to\infty} \frac{|a(t) - t\nu(b_\lambda(t))|}{b_\lambda(t)} = \limsup_{t\to\infty} \frac{b_\lambda(t)|\nu(b_\lambda(t))|}{\lambda U(b_\lambda(t))} \le \frac{c}{\lambda},$$

so (44) holds with $a(t) = 0$, and $X_t/\sqrt{V_t}$ is relatively compact as $t \to \infty$, by Proposition 3.1. □

Proposition 3.2. *Suppose $T_t := X_t/\sqrt{V_t}$ is relatively compact as $t \to \infty$, and also that*

$$\limsup_{x\to\infty} x^2\overline{\Pi}(x)/V(x) = \infty.$$

Then there is a sequence $t_k \to \infty$ such that

$$\lim_{\delta\downarrow 0} \limsup_{t_k\to\infty} P(||T_{t_k}| - 1| \le \delta) > 0. \tag{62}$$

Proof of Proposition 3.2: Assume that T_t is relatively compact as $t \to \infty$ and let

$$R(x) := x^2 \overline{\Pi}(x)/V(x). \tag{63}$$

Suppose $\limsup_{x\to\infty} R(x) = \infty$. This implies that $EX_1^2 = \infty$, thus by (45) again we need only deal with the big jump process. So we take X and V as in (46) and (47), and set $\gamma = \sigma^2 = 0$ in (12).

We first show the existence of a sequence $\zeta_k \to \infty$ such that

$$\lim_{k\to\infty} \inf_{0<\lambda_1\le\lambda\le\lambda_2} R(\lambda\zeta_k) = \infty, \text{ for each } 0 < \lambda_1 < 1 < \lambda_2 < \infty. \tag{64}$$

To this end, fix $0 < \lambda_1 < 1 < \lambda_2 < \infty$, choose $c_n \uparrow \infty$ such that $R(c_n) \uparrow \infty$, let $n_1 = \min\{m \ge 1 : c_m > 2, \text{ and } R(c_m) > 2^3\}$, and then for $k = 1, 2, \ldots$, set

$$n_{k+1} = \min\{m > n_k : c_m/2^{k+1} > 2^{-k} c_{n_k} + k + 1, \text{ and } R(c_m) > 2^{3(k+1)}\}.$$

Then put $\zeta_k = 2^{-k} c_{n_k}$, so that $\zeta_k \to \infty$ as $k \uparrow \infty$. Note that $R(x)/x^2$ is nonincreasing on $(0, \infty)$. Choose $\lambda \in [1, \lambda_2]$ and k such that $2^k \ge \lambda_2$. Then $\lambda \le 2^k$ and

$$\frac{R(\lambda\zeta_k)}{\lambda^2} = \frac{\zeta_k^2 R(\lambda\zeta_k)}{(\lambda\zeta_k)^2} \ge \frac{\zeta_k^2 R(2^k\zeta_k)}{(2^k\zeta_k)^2} = 2^{-2k} R(2^k\zeta_k) = 2^{-2k} R(c_{n_k}) \ge 2^k,$$

so $\inf_{1\le\lambda\le\lambda_2} R(\lambda\zeta_k) \ge 2^k \to \infty$. Thus $R(\zeta_k) \ge 2^k \to \infty$, and for $\lambda \in [\lambda_1, 1]$, $\lambda_1^2 R(\zeta_k) \le \lambda^2 R(\zeta_k) \le R(\lambda\zeta_k)$, so $\inf_{\lambda_1\le\lambda\le 1} R(\lambda\zeta_k) \to \infty$. Hence (64) holds.

Recall that $U(x) \ge x^2\overline{\Pi}(x)$ for all $x > 0$, so for $\lambda > 0$,

$$0 \le 1 - \frac{(\lambda\zeta_k)^2\overline{\Pi}(\lambda\zeta_k)}{U(\lambda\zeta_k)} = \frac{V(\lambda\zeta_k)}{U(\lambda\zeta_k)} \le \frac{V(\lambda\zeta_k)}{(\lambda\zeta_k)^2\overline{\Pi}(\lambda\zeta_k)} = \frac{1}{R(\lambda\zeta_k)} \to 0, \tag{65}$$

uniformly in $\lambda \in [\lambda_1, \lambda_2]$, where $0 < \lambda_1 < 1 < \lambda_2 < \infty$. Now

$$\int_{\lambda_1}^{\lambda_2} \frac{(s\zeta_k)^2\overline{\Pi}(s\zeta_k)}{U(s\zeta_k)} \frac{\mathrm{d}s}{s} = \frac{1}{2}\int_{\lambda_1\zeta_k}^{\lambda_2\zeta_k} \frac{\mathrm{d}U(s)}{U(s)} = \frac{1}{2}\log\left(\frac{U(\lambda_2\zeta_k)}{U(\lambda_1\zeta_k)}\right) \tag{66}$$

(recall that $U(x)$ is continuous at each $x > 0$). The left-hand side of the last expression tends to $\int_{\lambda_1}^{\lambda_2} \mathrm{d}s/s = \log(\lambda_2/\lambda_1)$, so we have $U(\lambda_2\zeta_k)/U(\lambda_1\zeta_k) \to (\lambda_2/\lambda_1)^2$. Then by (65), $\overline{\Pi}(\lambda_2\zeta_k)/\overline{\Pi}(\lambda_1\zeta_k) \to 1$, and so we deduce

$$\lim_{k\to\infty} \sup_{0<\lambda_1\le\lambda\le\lambda_2} \left| \frac{\overline{\Pi}(\lambda\zeta_k)}{\overline{\Pi}(\zeta_k)} - 1 \right| = 0, \text{ for each } 0 < \lambda_1 < \lambda_2. \tag{67}$$

Recall the representation of X_t in (46) and the definition of the J_i. Let $J_{N_t}^{(1)}$ be any J_i which is largest in modulus among $J_1, \ldots, J_{N_t}$, and let $J_{N_t}^{(2)}$ denote any term among $J_1, \ldots, J_{N_t}$ of second largest modulus.

Define

$$^{(1)}\widetilde{X}_t := X_t - J_{N_t}^{(1)} = \sum_{i=1}^{N_t} J_i - J_{N_t}^{(1)}$$

and

$$^{(1)}V_t := V_t - |J_{N_t}^{(1)}|^2.$$

Put $t_k := 1/\overline{\Pi}(\zeta_k)$, so that $t_k \to \infty$. For $\delta > 0$ and $0 < \lambda_1 < \lambda_2$, define the events

$$A_k := \left\{ |J_{N_{t_k}}^{(2)}| \le \lambda_1 \zeta_k < \lambda_2 \zeta_k < |J_{N_{t_k}}^{(1)}| \right\},$$

$$B_k(\delta) := \left\{ |^{(1)}\widetilde{X}_{t_k}| > \delta |J_{N_{t_k}}^{(1)}| \right\},$$

and

$$C_k(\delta) := \left\{ ^{(1)}V_{t_k} > \delta^2 |J_{N_{t_k}}^{(1)}|^2 \right\}.$$

In the following, we will keep $\lambda_1 \zeta_k > 1$. A straightforward calculation gives

$$P(A_k) = t_k \overline{\Pi}(\lambda_2 \zeta_k) e^{-t_k \overline{\Pi}(\lambda_1 \zeta_k)} =: \rho_k(\lambda_1, \lambda_2), \text{ say.} \tag{68}$$

Also, on A_k, we have

$$B_k(\delta) \subseteq \left\{ \left| \sum_{i=1}^{N_{t_k}} J_i \mathbf{1}_{\{|J_i| \le \lambda_1 \zeta_k\}} \right| > \delta \lambda_2 \zeta_k \right\} = \left\{ \left| \sum_{i=1}^{N_{t_k}} J_i^k \right| > \delta \lambda_2 \zeta_k \right\}, \tag{69}$$

where $J_i^k := J_i \mathbf{1}_{\{|J_i| \le \lambda_1 \zeta_k\}}$. Note that

$$E(J_1^k) = \int_{1 < |x| \le \lambda_1 \zeta_k} x \Pi(\mathrm{d}x) / \overline{\Pi}(1),$$

and $E(N_{t_k}) = t_k \overline{\Pi}(1)$, so we can write

$$\sum_{i=1}^{N_{t_k}} J_i^k = \sum_{i=1}^{N_{t_k}} \left(J_i^k - E(J_1^k) \right) + \left(N_{t_k} - E(N_{t_k}) \right) E(J_1^k) + t_k \int_{1 < |x| \le \lambda_1 \zeta_k} x \Pi(\mathrm{d}x). \tag{70}$$

Now since T_t is assumed relatively compact, we have by (42)

$$x|\nu(x)| \le M(x^2 \overline{\Pi}(x) + V(x)),$$

for $x \geq$ some x_0, for some $M \in (0,\infty)$. Further, by (65), we have $V(\lambda\zeta_k) \leq (\lambda\zeta_k)^2\overline{\Pi}(\lambda\zeta_k)$, for k large, uniformly in $\lambda \in [\lambda_1, \lambda_2]$. Thus for k large, firstly,

$$\begin{aligned}\left|t_k \int_{1<|x|\leq\lambda_1\zeta_k} x\Pi(\mathrm{d}x)\right| &= t_k|\nu(\lambda_1\zeta_k)| \text{ (recall } \gamma = 0 \text{ in (12))} \\ &\leq 2Mt_k(\lambda_1\zeta_k)\overline{\Pi}(\lambda_1\zeta_k) \\ &\leq 4M\lambda_1\zeta_k \quad \text{(by (67), and } t_k = 1/\overline{\Pi}(\zeta_k)) \\ &\leq (\delta/2)\lambda_2\zeta_k,\end{aligned}$$

for $\lambda_2 > \lambda_1$ large enough. Secondly,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{N_{t_k}} \left(J_i^k - E(J_1^k)\right)\right) &= E(N_{t_k})\mathrm{Var}(J_1^k) \\ &\leq t_k\overline{\Pi}(1)\int_{1<|x|\leq\lambda_1\zeta_k} x^2\Pi(\mathrm{d}x)/\overline{\Pi}(1) \\ &\leq t_k V(\lambda_1\zeta_k).\end{aligned}$$

Third, using Cauchy-Schwarz,

$$\begin{aligned}\mathrm{Var}\left[\left(N_{t_k} - E(N_{t_k})\right)E(J_1^k)\right] &= t_k\overline{\Pi}(1)\left(\int_{1<|x|\leq\lambda_1\zeta_k} x\Pi(\mathrm{d}x)/\overline{\Pi}(1)\right)^2 \\ &\leq t_k\int_{1<|x|\leq\lambda_1\zeta_k} x^2\Pi(\mathrm{d}x) \\ &\leq t_k V(\lambda_1\zeta_k).\end{aligned}$$

Putting the three estimates into (70) and using Chebyshev's inequality, we find that for $\delta_1 > 0$ and $\lambda_1 > \lambda_2$ large enough

$$\begin{aligned}&P(B_k(\delta_1) \cap A_k) \\ &\leq P\left(\left|\sum_{i=1}^{N_{t_k}} \left(J_i^k - E(J_1^k)\right) + \left(N_{t_k} - E(N_{t_k})\right)E(J_1^k)\right| > (\delta_1/2)\lambda_2\zeta_k\right) \\ &\leq \frac{8t_k V(\lambda_1\zeta_k)}{(\delta_1\lambda_2\zeta_k)^2}.\end{aligned}$$

By a similar argument as in (69) and Markov's inequality we get for $\delta_2 > 0$

$$P(C_k(\delta_2) \cap A_k) \le P\left(\sum_{i=1}^{N_{t_k}} |J_i^k|^2 > (\delta_2\lambda_2\zeta_k)^2\right)$$
$$\le \frac{E(N_{t_k})E(J_1^k)^2}{(\delta_2\lambda_2\zeta_k)^2} \le \frac{t_k V(\lambda_1\zeta_k)}{(\delta_2\lambda_2\zeta_k)^2}.$$

Putting these together gives

$$P\left(\{B_k(\delta_2) \cap A_k\} \cup \{C_k(\delta_1) \cap A_k\}\right) \le \left(\frac{1}{\delta_1^2} + \frac{1}{\delta_2^2}\right)\frac{8t_k V(\lambda_1\zeta_k)}{(\lambda_2\zeta_k)^2} \tag{71}$$
$$=: \eta_k(\delta_1, \delta_2), \text{ say.}$$

Now, since $V_t = {}^{(1)}V_t + |J_{N_t}^{(1)}|^2 \ge |J_{N_t}^{(1)}|^2$, we can write, for $\delta > 0$,

$$P\left(\left|\frac{|X_{t_k}|}{\sqrt{V_{t_k}}} - 1\right| > \delta\right)$$
$$= P\left(\left||X_{t_k}| - \sqrt{V_{t_k}}\right| > \delta\sqrt{V_{t_k}}\right)$$
$$\le P\left(\left\{\left||X_{t_k}| - \sqrt{V_{t_k}}\right| > \delta|J_{N_{t_k}}^{(1)}|,\ {}^{(1)}V_{t_k} \le (\delta/2)^2|J_{N_{t_k}}^{(1)}|^2\right\}\right.$$
$$\left. \cup \left\{{}^{(1)}V_{t_k} > (\delta/2)^2|J_{N_{t_k}}^{(1)}|^2\right\}\right).$$

The latter does not exceed

$$P\left(\left\{\left|X_{t_k} - J_{N_{t_k}}^{(1)}\right| > \delta|J_{N_{t_k}}^{(1)}|/2\right\} \cup \left\{{}^{(1)}V_{t_k} > (\delta/2)^2|J_{N_t}^{(1)}|^2\right\}\right); \tag{72}$$

because, $\sqrt{{}^{(1)}V_{t_k}} \le (\delta/2)|J_{N_{t_k}}^{(1)}|$, thus $|\sqrt{V_{t_k}} - |J_{N_{t_k}}^{(1)}|| \le (\delta/2)|J_{N_{t_k}}^{(1)}|$, together with

$$\left||X_{t_k}| - \sqrt{V_{t_k}}\right| > \delta|J_{N_{t_k}}^{(1)}| > \left|\sqrt{V_{t_k}} - |J_{N_{t_k}}^{(1)}|\right|,$$

imply

$$\left|X_{t_k} - J_{N_{t_k}}^{(1)}\right|$$
$$\ge \left||X_{t_k}| - |J_{N_{t_k}}^{(1)}|\right| \ge \left|\left||X_{t_k}| - \sqrt{V_{t_k}}\right| - \left|\sqrt{V_{t_k}} - |J_{N_{t_k}}^{(1)}|\right|\right|$$
$$= \left||X_{t_k}| - \sqrt{V_{t_k}}\right| - \left|\sqrt{V_{t_k}} - |J_{N_{t_k}}^{(1)}|\right|$$
$$\ge (\delta - \delta/2)|J_{N_{t_k}}^{(1)}| = \delta|J_{N_t}^{(1)}|/2.$$

Observe that (72) does not exceed $P(B_k(\delta/2) \cup C_k(\delta/2))$. Argue that, by (71) and (68),

$$\begin{aligned} P(B_k(\delta_1) \cup C_k(\delta_2)) &\le P\left(\{B_k(\delta_1) \cap A_k\} \cup \{C_k(\delta_2)) \cap A_k\}\right) \\ &\quad + 1 - P(A_k) \\ &\le \eta_k(\delta_1, \delta_2) + 1 - \rho_k(\lambda_1, \lambda_2). \end{aligned}$$

Thus by (72)

$$P\left(\left|\frac{|X_{t_k}|}{\sqrt{V_{t_k}}} - 1\right| > \delta\right) \le \eta_k(\delta/2, \delta/2) + 1 - \rho_k(\lambda_1, \lambda_2). \tag{73}$$

Now by (65) and (67)

$$t_k V(\lambda_1 \zeta_k) = o\left(t_k \zeta_k^2 \overline{\Pi}(\lambda_1 \zeta_k)\right) = o(\zeta_k^2),$$

so $\eta_k(\delta_1, \delta_2) \to 0$ as $k \to \infty$, while, by (67), $\rho_k(\lambda_1, \lambda_2) \to e^{-1}$ as $k \to \infty$. Letting $k \to \infty$ in (73) gives

$$\limsup_{t_k \to \infty} P\left(\left|\frac{|X_{t_k}|}{\sqrt{V_{t_k}}} - 1\right| > \delta\right) \le 1 - e^{-1} < 1,$$

so (62) holds. □

We are now ready to complete the proof of Proposition 1.2. Suppose $X_t/\sqrt{V_t}$ is relatively compact as $t \to \infty$ and no subsequential limit has positive mass at ± 1. Then by Theorem 3.1 we have

$$\limsup_{x \to \infty} \frac{x|\nu(x)|}{U(x)} = \limsup_{x \to \infty} \frac{x|\nu(x)|}{x^2 \overline{\Pi}(x) + V(x)} < \infty, \tag{74}$$

while by Proposition 3.2 we have

$$\limsup_{x \to \infty} \frac{x^2 \overline{\Pi}(x)}{V(x)} < \infty, \tag{75}$$

since if (75) fails then there is a subsequential limit rv of $X_t/\sqrt{V_t}$ with positive mass at ± 1. Now (74) and (75) imply

$$\limsup_{x \to \infty} \frac{x|\nu(x)|}{V(x)} < \infty, \tag{76}$$

which together with (75) gives (16). □

3.2.1 Comments on Proofs of Propositions 3.1 and 3.2

The proofs of Propositions 3.1 and 3.2 parallel very closely, after notational changes, those of Propositions 5.1 and 5.5 of [12], which are their small time versions. Therefore for the sake of brevity, but at the sacrifice of readability, we could have replaced the foregoing proofs by the following road maps:

For the proof of Proposition 3.1 proceed exactly as it is given above until right before equation (50), and then continue on as in the proof of Proposition 5.1 of [12] starting at its equation (5.8) with the role of ε suppressed, i.e. $T_t(\varepsilon,\lambda)$, $R_t(\varepsilon,\lambda)$, $S_t(\varepsilon,\lambda)$, $N_t(\varepsilon,\lambda)$, $U_t(\varepsilon,\lambda)$, $V_t(\varepsilon)$, $V(\varepsilon)$ and $X_t(\varepsilon)$, are replaced by $T_t(\lambda)$, $R_t(\lambda)$, $S_t(\lambda)$, $N_t(\lambda)$, $U_t(\lambda)$, V_t, $V(1)$ and X_t, respectively. Also replace $t\nu(\varepsilon)$ by 0 and $\alpha_t(\varepsilon,\lambda)$ by $T_t(\lambda)$, and use the definition of $c(\lambda)$ given in (58).

From (65) the proof of Proposition 3.2 goes exactly as that of Proposition 5.5 of [12] beginning from its equation (5.38) with the role of ε suppressed in the notation analogously as it was done in the proof of Proposition 1.1 and with $t_k \searrow 0$ changed to $t_k \to \infty$, $t_k \nu(t_k)$ to 0 and $\varepsilon < \lambda_1 \xi_k$ to $1 < \lambda_1 \xi_k$.

4 Appendix: Strictly Stable Bivariate Laws

Theorem 1.1 of [3] says that $\mathbb{T}_n \xrightarrow{\mathrm{D}} Y$, where $P\{|Y| = 1\} = 0$, if and only there exists a sequence of norming constants b_n such that either $b_n^{-1}\mathbb{S}_n \xrightarrow{\mathrm{D}} Z$ or ξ is in the domain of attraction of a stable law of index $0 < \alpha < 2$. Moreover, in the normal case $E\xi = 0$, in the case $1 < \alpha < 2$, $E\xi = 0$ and in the case $\alpha = 1$, ξ is in the domain of attraction of Cauchy's law and Feller's condition holds, that is,

$$\lim_{n\to\infty} E\sin(\xi/b_n) \text{ exists and is finite.}$$

This means in the stable law of index $0 < \alpha < 2$ case that necessarily for some function L slowly varying at infinity,

$$1 - F(x) := P\{|\xi| > x\} = x^{-\alpha} L(x), \text{ for } x > 0,$$

and some $0 \le p \le 1$, as $x \to \infty$,

$$P\{\xi > x\}/P\{|\xi| > x\} \to p \text{ and } P\{\xi < -x\}/P\{|\xi| > x\} \to 1 - p. \tag{77}$$

By applying Theorem 15.14 of [8] much as we did in the proofs of Lemmas 2.1 and 2.2, we can show that with a norming sequence b_n of the form

$$b_n \sim F^{-1}(1/n) = n^{1/\alpha} L^*(1/n),$$

with L^* slowly varying at zero, one has $\left(b_n^{-1}\mathbb{S}_n, b_n^{-2}\mathbb{V}_n\right) \xrightarrow{\mathrm{D}} (U^\alpha, V^\alpha)$, where U^α has characteristic function with Lévy measure Λ on $\mathbb{R} \setminus \{0\}$ of the form:

$$\overline{\Lambda}(y) = \overline{\Lambda}^{+}(y) + \overline{\Lambda}^{-}(y) := c_1 y^{-\alpha} + c_2 y^{-\alpha},\ y > 0,$$

for some $c_1 \geq 0$ and $c_2 \geq 0$, with at least one non zero and $c_1/(c_1 + c_2) = p$ as in (77). Moreover, (U^α, V^α) has joint characteristic function

$$\varphi_\alpha(s,t) = E \exp\left(\mathrm{i}sU^\alpha + \mathrm{i}tV^\alpha\right),$$

of the following form with

$$K_{r,l}(y) = \begin{cases} r, & y > 0 \\ 0, & y = 0, \\ l, & y < 0 \end{cases}$$

where $r = c_1\alpha$ and $l = c_2\alpha$:

Case 1: $0 < \alpha < 1$

$$\varphi_\alpha(s,t) = \exp\left(\int_{-\infty}^{\infty} \left(\exp\left(\mathrm{i}sy + \mathrm{i}ty^2\right) - 1\right) \frac{K_{r,l}(y)}{|y|^{1+\alpha}} \mathrm{d}y\right).$$

Case 2: $\alpha = 1$

$$\varphi_1(s,t) = \exp\left(\int_{-\infty}^{\infty} \left(\exp\left(\mathrm{i}ty^2\right)\cos(sy) - 1\right) \frac{1}{y^2} \mathrm{d}y\right).$$

Case 3: $1 < \alpha < 2$

$$\varphi_\alpha(s,t) = \exp\left(\int_{-\infty}^{\infty} \left(\exp\left(\mathrm{i}sy + \mathrm{i}ty^2\right) - 1 - \mathrm{i}sy\right) \frac{K_{r,l}(y)}{|y|^{1+\alpha}} \mathrm{d}y\right).$$

An easy calculation verifies that for all $n \geq 1$ and $0 < \alpha < 2$

$$\varphi_\alpha^n\left(s/n^{1/\alpha}, t/n^{2/\alpha}\right) = \varphi_\alpha(s,t).$$

Thus if $\left(U_1^\alpha, V_1^\alpha\right), \ldots, \left(U_n^\alpha, V_n^\alpha\right)$ are i.i.d. (U^α, V^α) then for all $n \geq 1$

$$\left(n^{-1/\alpha}\sum_{i=1}^{n} U_i^\alpha, n^{-2/\alpha}\sum_{i=1}^{n} V_i^\alpha\right) \stackrel{\mathrm{D}}{=} (U^\alpha, V^\alpha).$$

This says that (U^α, V^α) is a *strictly bivariate stable random vector* and in the stable $0 < \alpha < 2$ case $\mathbb{S}_n/\sqrt{\mathbb{V}_n} \xrightarrow{\mathrm{D}} Y$, where $Y \stackrel{\mathrm{D}}{=} U^\alpha/\sqrt{V^\alpha}$.

Acknowledgements Research of Ross Maller was partially supported by ARC Grant DP1092502. Research of David M. Mason was partially supported by NSF Grant DMS–0503908.

References

1. F. Beichelt, *Stochastic Processes in Science, Engineering and Finance* (CRC Press LLC, Boca Raton, 2006)
2. J. Bertoin, *Lévy Processes* (Cambridge University Press, Cambridge, 1996)
3. G.P. Chistyakov, F. Götze, Limit distributions of studentized sums. Ann. Probab. **32**, 28–77 (2004)
4. R.A. Doney, R.A. Maller, Stability and attraction to Normality for Lévy processes at zero and infinity. J. Theor. Probab. **15**, 751–792 (2002)
5. E. Giné, D.M. Mason, On the LIL for self-normalized sums of IID random variables. J. Theor. Probab.**11**, 351–370 (1998)
6. E. Giné, F. Götze, D.M. Mason, When is the student t-statistic asymptotically standard normal? Ann. Probab.**25**, 1514–1531 (1997)
7. P.S. Griffin, Tightness of the Student T-statistic. Electron. Comm. Probab. **7**, 181–190 (2002)
8. O. Kallenberg, *Foundations of Modern Probability*, 2nd edn. (Springer, Berlin, 2001)
9. B.F. Logan, C.L. Mallows, S.O. Rice, L. Shepp, Limit distributions of self-normalized sums. Ann. Probab. **1**, 788–809 (1973)
10. R.A. Maller, D.M. Mason, Convergence in distribution of Lévy processes at small times with self-normalisation. Acta Sci. Math. (Szeged) **74**, 315–347 (2008)
11. R.A. Maller, D.M. Mason, Stochastic compactness of Lévy processes, in *Proceedings of High Dimensional Probability V*, Luminy, France, 2008, ed. by C. Houdré, V. Kolthchinskii, M. Peligrad, D. Mason. I.M.S. Collections, High Dimensional Probability V: The Luminy Volume, vol. 5 (Institute of Mathematical Statistics, Beachwood, 2009), pp. 239–257
12. R.A. Maller, D.M. Mason, Small-time compactness and convergence behavior of deterministically and self-normalised Lévy processes. Trans. Am. Math. Soc. **362**, 2205–2248 (2010)
13. D.M. Mason, The asymptotic distribution of self-normalized triangular arrays. J. Theor. Probab. **18**, 853–870 (2005)
14. D.M. Mason, J. Zinn, When does a randomly weighted self-normalized sum converge in distribution? Electron. Comm. Probab. **10**, 70–81 (2005)
15. K. Sato, *Lévy Processes and Infinitely Divisible Distributions* (Cambridge University Press, Cambridge, 1999)

Part III
Statistics and Combinatorics

A Nonparametric Theory of Statistics on Manifolds

Rabi Bhattacharya

Dedicated to Friedrich Götze on the Occasion of his Sixtieth Birthday

Abstract An expository account of the recent theory of nonparametric inference on manifolds is presented here, with outlines of proofs and examples. Much of the theory centers around Fréchet means; but functional estimation and classification methods using nonparametric Bayes theory are also indicated. Applications in paleomagnetism, morphometrics and medical diagnostics illustrate the theory.

Keywords Fréchet mean • intrinsic inference • extrinsic inference • shape spaces

2010 *Mathematics Subject Classification.* Primary 62G20; Secondary 62G05, 62G10, 62H35, 62P10

1 Introduction

Statistical inference on manifolds such as circles and spheres has a long history, dating back at least to early twentieth century. But a great deal of activity was inspired by the seminal 1953 paper of R.A. Fisher on the shifts of the earth's magnetic poles over geological time scales. Statistical inference on landmarks based shape manifolds, which are of special interest in this article, came later and owes

R. Bhattacharya (✉)
Department of Mathematics, The University of Arizona, Tucson, AZ 85721, USA
e-mail: rabi@math.arizona.edu

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_9,

much of its development to the pioneering work of Kendall [32–34], providing the appropriate geometric foundation for these spaces, and to Bookstein [13–15], who in a somewhat different vein created methodologies for applications of statistics of shapes to biology and medical imaging. We must also mention the work of Karcher [31] on the uniqueness of Fréchet means of probability measures on Riemannian manifolds, and the work of Ziezold [46] on the almost sure convergence properties of Fréchet mean sets on metric spaces. Parametric inference for these spaces grew quite rapidly during the past two decades or so. In addition to the work already mentioned, important contributions were made by many authors, such as Kent [36, 37], Goodall [26], Dryden and Mardia [18], Prentice and Mardia [43], and others. A comprehensive account of this theory with extensive references to original work until 1998 may be found in the book by Dryden and Mardia [19].

In the present article we provide an expository account of the recent nonparametric theory on general manifolds, with special emphasis on shape manifolds. This theory is largely based on the notion of the *Fréchet mean* of a probability measure Q, namely, the minimizer, if unique, of the expected squared distance from a point on the manifold. If the distance on the manifold M is the geodesic distance with respect to a Riemannian metric, the Fréchet mean is said to be *intrinsic*. If the distance is the Euclidean distance inherited from an embedding of M in a Euclidean space, then the Fréchet mean is called *extrinsic*. Hendriks and Landsman [27, 28], provided asymptotics of the extrinsic mean on regular submanifolds of Euclidean spaces, with the embedding given by the inclusion map. Independently of this, a theory of extrinsic inference for Fréchet means on general manifolds originated in the 1998 dissertation of Patrangenaru, and further developed in [10, 11]. The latter articles also provided a general theory of intrinsic inference. While the emphasis in applications in the latter articles are to the sphere S^d and Kendall's planar shape spaces, embeddings of projective shape spaces and $3D$ shape spaces and inference for Fréchet means on them are developed in [2,3,8,9,39]. Further progress in both intrinsic and extrinsic inference may be found in [4, 5] and in the monograph Bhattacharya and Bhattacharya [6]. Our goal here is to present the core of this emerging field in a reasonably accessible manner.

Because references to Bhattacharya and Patrangenaru and Bhattacharya and Bhattacharya occur frequently, we would henceforth refer to them as BP and BB, respectively.

Here is an outline of the contents of the paper. In Sect. 2, basic properties of Fréchet means on metric spaces are established, including consistency (Theorem 2.1), and a general result on the asymptotic distribution of sample Fréchet means on manifolds (Theorem 2.5). The latter turns out to be crucial for intrinsic inference developed in later sections. Consistency and asymptotic distribution of extrinsic sample means are established in Sect. 3 (Theorems 3.1, 3.3), while Sect. 4 provides the corresponding results for intrinsic sample means (Theorem 4.1). The groundwork for statistical inference on general manifolds is laid in Sect. 5, including the construction of confidence regions and two-sample and match pair tests, based on the asymptotic Normal and chisquare distributions derived in earlier sections. Section 6 describes the geometries of landmarks based shape spaces. Here an

observation, called a k-ad, consists of k landmarks, chosen with expert help, on an object of interest such as a brain scan, or some other digital image. The goal may be medical diagnosis, classifying biological species and subspecies, or computer vision/robotics. Because of differences in equipments used and/or their positioning relative to the object while recording images, etc., one considers for analysis the k-ad modulo an appropriate Lie group of transformations. In particular, the *similarity shape* of a k-ad is its orbit (or maximal invariant) under Euclidean rigid motions of translation and rotation, as well as scaling. The space of such shapes of k-ads in $\mathbb{R}^m$ is *Kendall's shape space* $\Sigma_m^k (k > m)$. For $m = 2$, it is more convenient for analytical and computational purposes to represent the k points of the k-ad in $\mathbb{R}^2$ as points in the complex plane. The planar shape space Σ_2^k can then be identified with the *complex projective space* $\mathbb{C}P^{k-2}$, which is a manifold of considerable interest in differential geometry. Its natural Riemannian structure is described in Sect. 6.1.1. Sect. 6.1.2 considers the intrinsic geometry of Σ_m^k in dimensions $m > 2$. Unfortunately, here the Lie group action is not free, resulting in orbits of different dimensions in different regions of Σ_m^k. If one removes the regions of singularity, the manifold is no longer complete in the Riemannian metric, and its curvature grows unboundedly as one approaches the singular sets, making inference difficult. For some recent progress in overcoming this in extending principal components analysis to Riemannian manifolds, see [29]. Sub-section 6.1.2 is devoted to the extrinsic geometry of Σ_2^k under the so-called *Veronese-Whitney embedding*, which is equivariant under the unitary group $SU(k-1)$.

As a matter of *notation*, a k-ad x in $\mathbb{R}^m$ is represented as an $m \times k$ *matrix*, with the k points appearing as k column vectors in $\mathbb{R}^m$. The *transpose* of a matrix A is expressed as A^t.

Section 6.2 defines the so-called *reflection similarity shape* $r\sigma(x)$ of a k-ad x in $\mathbb{R}^m$, identified with the orbit of the centered and scaled k-ad z under the group $O(m)$ of all orthogonal transformations. When restricted to the non-singular part of Σ_m^k using only k-ads each of which is of full rank m, the *reflection-similarity shape* space $R\Sigma_m^k$ is a manifold, although not complete. But its extrinsic analysis is facilitated by the embedding $r\sigma(x) \to z^t z$ into the space $S(k, \mathbb{R})$ of all symmetric $k \times k$ matrices, or into $S(k-1, \mathbb{R})$ if one reduces the k-ad to a $(k-1)$-ad by Helmertization to remove translation. Here z is the *preshape* of x obtained by scaling (to norm 1) the translated, or Helmertized k-ad. This new shape space and its embedding were originally introduced by Bandulasiri and Patrangenaru [8], and also arrived at independently by Dryden et al. [20]. The geometry and extrinsic inference for it was further developed in [2, 3, 6, 9]. This is a significant step in the analysis of 3D shapes. In the remaining two Sects. 6.3 and 6.4 we introduce affine and projective shapes. These are of much importance in problems of scene recognition and machine vision.

A proper extrinsic analysis requires a good *equivariant embedding*, whereby a reasonably large isometric group action on a Riemannian manifold M is replicated on its image (in an N-dimensional Euclidean space E^N), by the action of a subgroup of the general linear group $GL(N, \mathbb{R})$, via a group homomorphism. Often this latter is also a group of isometries on the image of M under the embedding when endowed with the metric tensor induced from E^N. This helps preserve much of the geometry of M. In view of this, in most examples of data analysis the results of extrinsic and intrinsic inference turn out to be nearly identical although they are based on different methodologies. The embeddings of the shape spaces considered in this article are equivariant under appropriately large group actions.

To illustrate the general theory, in Sect. 7 we develop in some detail intrinsic and extrinsic inference procedures on two specific manifolds—the sphere S^d and the planar shape space Σ_2^k. Section 8 provides a brief introduction to density estimation and classification using the nonparametric Bayes theory. Finally, Sect. 9 provides three examples of data analysis using the nonparametric theory presented in this article, and contrasts these, where possible, with results of parametric inference carried out in the literature. As is well recognized, nonparametric methods provide inference whose validity is model independent, while parametric models may be miss-specified and lead to conclusions not quite right. However, this advantage is often accompanied by larger confidence regions and smaller powers of tests. It, therefore, comes as a pleasant surprise that in most examples where data are available and for which parametric inference has been carried out, the model-independent procedures for shape spaces described in this article yield sharper inference-narrower confidence regions and much smaller p-values -than their parametric counterparts.

Finally, mention should also be made of the work of Ellingson et al. [21] for the estimation of the extrinsic mean of distributions of planar contours representing continuous planar shapes, via an infinite dimensional version of the Veronese-Whitney embedding of Σ_2^k.

We conclude this section with a sketch of the estimation of the extrinsic mean on $M = S^d$. Here the embedding J is the inclusion map of S^d into $\mathbb{R}^{d+1}$. The extrinsic mean μ_E of Q on S^d is given by $\mu_E = \mu^J/|\mu^J|$, where μ^J is the mean of Q viewed as a measure on $\mathbb{R}^{d+1}$. We assume $\mu^J \neq 0$, which is the necessary and sufficient condition for the uniqueness of the extrinsic mean on S^d. The extrinsic sample mean of i.i.d. observations $X_1, \cdots, X_n$ is, similarly, $\hat{\mu}_E = \bar{X}/|\bar{X}|$, where $\bar{X} = (X_1 + \cdots + X_n)/n$. It is easy to check that when $\hat{\mu}_E$ and μ_E are viewed as vectors in $\mathbb{R}^{d+1}$, $\hat{\mu}_E$ is asymptotically Normal $N(\mu_E, \Sigma/n)$, where the $(d+1) \times (d+1)$ matrix Σ is singular, since $\hat{\mu}_E$ lies nearly on $T_{\mu_E}(S^d)$-the tangent space of S^d at μ_E. The tangential component of $\hat{\mu}_E$, expressed in d coordinates with respect to a chosen orthonormal basis of $T_{\mu_E}(S^d)$, has the asymptotic distribution $N(0, \Sigma_1/n)$. Here Σ_1 is a $d \times d$ matrix which is nonsingular if the covariance matrix of Q (on R^{d+1}) is nonsingular. One may use this result for estimation and testing on S^d. For details of this and for intrinsic inference on S^d see Example 7.1.

2 Asymptotic Distribution Theory for Fréchet Means

Let (M, ρ) be a metric space and Q a probability measure on the Borel sigma-field of M. Consider a *Fréchet function* of Q defined by

$$F(x) = \int \rho^{\alpha}(x, y)Q(dy), \quad x \in M, \tag{1}$$

for some $\alpha \geq 1$. We will be mostly concerned with the case $\alpha = 2$. Assume that F is finite at least for one x. A minimizer of F, if unique, serves as a measure of location of Q. In general, the set C_Q of minimizers of F is called the *Fréchet mean set* of Q. In the case the minimizer is unique, one says that the *Fréchet mean exists* and refers to it as the *Fréchet mean* of Q. If $X_1, \cdots, X_n$ are i.i.d observations with common distribution Q, the Fréchet mean set and the Fréchet mean of the empirical $Q_n = 1/n \sum_{1 \leq j \leq n} \delta_{X_j}$ are named the *sample Fréchet mean set* and the *sample Fréchet mean*, respectively. For a reason which will be clear from the result below, in the case the Fréchet mean of Q exists, a (every) measurable selection from C_{Q_n} is taken to be a sample Fréchet mean.

The following is a general result on Fréchet mean sets C_Q and C_{Q_n} of Q and Q_n and *consistency* of the sample Fréchet mean.

Theorem 2.1 (Ziezold [46], BP [10], BB [6]). *Let M be a metric space such that every closed and bounded subset of M is compact. Suppose $\alpha \geq 1$ in (1) and $F(x)$ is finite for some x. Then (a) the Fréchet mean set C_Q is nonempty and compact, and (b) given any $\epsilon > 0$, there exists a positive integer valued random variable $N = N(\omega, \epsilon)$ and a P-null set $\Omega(\epsilon)$ such that*

$$C_{Q_n} \subseteq C_Q^{\epsilon} = \{x \in M : \rho(x, C_Q) < \epsilon\} \ \forall n \geq N, \forall \omega \in (\Omega(\epsilon))^c. \tag{2}$$

(c) In particular, if the Fréchet mean of Q exists then the sample Fréchet mean, taken as a measurable selection from C_{Q_n}, converges almost surely to it.

Remark 2.2. Unfortunately, it does not seem possible in general to estimate the Fréchet mean set C_Q consistently by C_{Q_n}, that is, the Hausdorff distance between the two does not necessarily go to zero with probability one, as n goes to infinity. Consider, for example, the simple case of $M = S^d$, with ρ as the chord distance and $\alpha = 2$. Take an absolutely continuous Q for which C_Q is not a singleton, as would be the case for the uniform distribution in particular. It is easy to see that the sample Fréchet mean set C_{Q_n} is, with probability one, a singleton.

Unless stated otherwise, we will assume in this article that the manifold M is *connected and satisfies the property that its closed bounded subsets are compact*. Obviously this is true if M is compact. The assumption also holds for all Riemannian manifolds which are complete under the geodesic distance, by the Hopf-Rinow theorem (see [17], pp. 146–147).

Remark 2.3. It has been shown by Karcher [31] for the case $\alpha = 2$ in (1) that, if the Fréchet function of Q is finite, then on a Riemannian manifold M with non-positive sectional curvature the Fréchet mean always exists as a unique minimizer.

We give a proof of Theorem 2.1 for a compact metric M, which is the case in many of the applications of interest here. Part (a) is then trivially true. For part (b), for each $\epsilon > 0$, write

$$\eta = \inf\{F(x) : x \in M\} \equiv F(q) \ \forall\ q \in C_Q,$$
$$\eta + \delta(\epsilon) = \inf\{F(x) : x \in M \backslash C_Q^\epsilon\}. \tag{3}$$

If $C_Q^\epsilon = M$, then (2) trivially holds. Consider the case $C_Q^\epsilon \neq M$, so that $\delta(\epsilon) > 0$.

Let $F_n(x)$ be the Fréchet function of Q_n, namely,

$$F_n(x) = \frac{1}{n} \sum_{1 \leq j \leq n} \rho^\alpha(x, X_j).$$

Now use the elementary inequality,

$$|\rho^\alpha(x, y) - \rho^\alpha(x', y)| \leq \alpha\rho(x, x')[\rho^{\alpha-1}(x, y) + \rho^{\alpha-1}(x', y)] \leq c\alpha\rho(x, x'),$$

with $c = 2\max\{\rho^{\alpha-1}(x, y), x, y \in M\}$, to obtain

$$|F(x) - F(x')| \leq c\alpha\rho(x, x'),\ |F_n(x) - F_n(x')| \leq c\alpha\rho(x, x'),\ \forall x, x'. \tag{4}$$

For each $x \in M \backslash C_Q^\epsilon$ find $r = r(x, \epsilon) > 0$ such that $c\alpha\rho(x, x') < \delta(\epsilon)/4\ \forall x'$ within a distance r from x. Let $m = m(\epsilon)$ of these balls with centers $x_1, \cdots, x_m$ (in $M \backslash C_Q^\epsilon$) cover $M \backslash C_Q^\epsilon$. By the SLLN, there exist integers $N_i = N_i(\omega)$ such that, outside a P-null set $\Omega_i(\epsilon)$, $|F_n(x_i) - F(x_i)| < \delta(\epsilon)/4\ \forall\ n \geq N_i (i = 1, \cdots, m)$. Let $N' = \max\{N_i : i = 1, \cdots, m\}$. If $n > N'$, then for every i and all x in the ball with center x_i and radius $r(x_i, \epsilon)$,

$$F_n(x) > F_n(x_i) - \delta(\epsilon)/4 > F(x_i) - \delta(\epsilon)/4 - \delta(\epsilon)/4$$
$$\geq \eta + \delta(\epsilon) - \delta(\epsilon)/2 = \eta + \delta(\epsilon)/2.$$

Next choose a point $q \in C_Q$ and find $N'' = N''(\omega)$, again by the SLLN, such that, if $n \geq N''$ then $|F_n(q) - F(q)| < \delta(\epsilon)/4$ and, consequently, $F_n(q) < \eta + \delta(\epsilon)/4$, outside of a P-Null set $\Omega''(\epsilon)$. Hence (2) follows with $N = \max\{N', N''\}$ and $\Omega(\epsilon) = \{\cup\Omega_i(\epsilon) : i = 1, \cdots, m\} \cup \Omega''(\epsilon)$. Part (c) is an immediate consequence of part (b).

Remark 2.4. For a compact metric space M, the conclusions of Theorem 2.1 hold for a *generalized Fréchet function* F by letting the integrand in (1) be an arbitrary continuous function $f(x, y)$ on $M \times M$ instead of $\rho^\alpha(x, y)$. Only a slight modification of the above proof is required for this.

For noncompact M, the proof of Theorem 2.1 is a little more elaborate and may be found in [6] or, for the case $\alpha = 2$, in [10].

We now proceed to derive the asymptotic distribution of sample Fréchet means on a d-dimensional differentiable manifold M. Let Q be a probability measure on M such that $Q(U) = 1$ for some open subset U of M which is C^2 diffeomorphic to an open set V of $\mathbb{R}^d$.

Consider a generalized Fréchet function F on U:

$$F(p) = \int_U f(p, p')Q(dp'), \ p \in U, \tag{5}$$

where $f : U \times U \to \mathbb{R}$, and the integral is finite for all p in U. Assume that F is twice differentiable in a neighborhood of the minimizer μ of F, assumed unique, and let μ_n be a consistent Fréchet sample mean. Let $\phi : U \to V$ be a C^2 diffeomorphism. Write $h(x, y) = f(\phi^{-1}x, \phi^{-1}y)$ for $x, y \in V$. Then $\nu = \phi(\mu)$ and $\nu_n = \phi(\mu_n)$ are the Fréchet minimizers of $Q \circ \phi^{-1}$ and $Q_n \circ \phi^{-1}$, respectively, of the Fréchet functions

$$H(x) = \int_V h(x, y)Q \circ \phi^{-1}(dy), \tag{6}$$

$$H_n(x) = \int_V h(x, y)Q_n \circ \phi^{-1}(dy) = \frac{1}{n}\sum_{j=1}^{n} h(x, Y_j), \ x \in V,$$

where $Y_j = \phi(X_j)$. Write $\psi^r(x, y) = D_r h(x, y) = (\partial/\partial x_r)h(x, y)$ $(r = 1, \cdots, d)$ and and let D stand for the *gradient*. For example, $D\psi^r(x, y)$ is the vector $(D_1\psi^r(x, y), \cdots, D_d\psi^r(x, y))$. By assumption, H is twice differentiable in a neighborhood of $\phi(\mu)$ and a Taylor expansion yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{1 \le j \le n} \psi^r(\nu_n, Y_j) \\ &= \frac{1}{\sqrt{n}} \sum_{1 \le j \le n} \psi^r(\nu, Y_j) + \left[\frac{1}{n} \sum_{1 \le j \le n} D\psi^r(\nu, Y_j) + \epsilon_{n,r} \right] \cdot \sqrt{n}(\nu_n - \nu), \end{aligned} \tag{7}$$

where $\cdot$ denotes inner product in $\mathbb{R}^d$ and, for some $\theta_{n,r}$ lying on the line segment joining ν_n and ν,

$$\epsilon_{n,r} = \frac{1}{n} \sum_{1 \le j \le n} D\psi^r(\theta_{n,r}, Y_j) - \frac{1}{n} \sum_{1 \le j \le n} D\psi^r(\nu, Y_j).$$

The following result, which is a slight extension of Theorem 2.1 in [11], now follows from (7).

Theorem 2.5. *Let Q be a probability measure on a d-dimensional manifold M. Assume that*

(i) there exists an open subset U of M such that $Q(U) = 1$,
(ii) for a given function f on $U \times U$, the generalized Fréchet function F of Q in (5) is finite and has a unique minimizer μ in U, and there is a neighborhood of μ on which $p \to f(p, p')$ is twice continuously differentiable for every p',
(iii) there exists a C^2-diffeomorphism $\phi : U \to V$ where V is an open subset of $\mathbb{R}^d$ such that for the function $\psi^r(x, y) = D_r h(x, y) = (\partial/\partial x_r) f(\phi^{-1}x, \phi^{-1}y)$ on $V \times V$ one has $E(\psi^r(\nu, Y_1))^2 < \infty \ \forall\ r = 1, \cdots, d$, with $\nu = \phi(\mu)$ and Y_1 having the distribution $Q \circ \phi^{-1}$,
(iv) $\sup\{E|D\psi^r(\nu, Y_1) - D\psi^r(y, Y_1)| : |y - \nu| \le \epsilon\} \to 0$, as $\epsilon \downarrow 0$, and, finally,
(v) the $d \times d$ matrix $\wedge = ((ED_s\psi^r(\nu, Y_j))) \equiv ((ED_s D_r h(\nu, Y_j)))$ is nonsingular.

Then $\nu_n = \phi(\mu_n)$ has the asymptotic distribution given by

$$\sqrt{n}(\nu_n - \nu) \to N(0, \wedge^{-1}\Sigma\wedge) \ \text{in distribution as } n \to \infty, \tag{8}$$

where Σ is the covariance matrix of $(\psi^r(\nu, Y_j), r = 1, \cdots, d)$.

Remark 2.6. Suppose M is a Riemannian manifold and Q a probability on M. If $q \in M$ and $C(q)$ is the cut locus of q (see Sect. 4 for definition), and if $Q(M \backslash C(q)) = 1$, then one may take U in Theorem 2.5 to be $M \backslash C(q)$. The inverse exponential map on $M \backslash C(q)$ may be taken to be the required diffeomorphism ϕ on $U = M \backslash C(q)$ onto its image V in the tangent space $T_q M$.

Note that $Q(M \backslash C(q)) = 1$ if Q is absolutely continuous with respect to a volume measure on M (see [24], p. 141).

Remark 2.7. On a Riemannian manifold M the Fréchet mean of Q for the case $f(p, p') = \rho^2(p, p')$ with geodesic distance ρ is called the *intrinsic mean* of Q. For manifolds M of nonnegative curvature, a recent criterion due to Afsari [1] under which Q is known to have an intrinsic mean is that the support of Q lie in a geodesic ball of radius $r^*/2$ where $r^* = \min\{inj(M), \pi/\sqrt{\bar{C}}\}$, $inj(M)$ being the injectivity radius of M (see Sect. 4), and $\bar{C}$ the least upper bound of sectional curvatures of M (see [31, 38]). Hence one may take U in Theorem 2.5 to be this geodesic ball in this case. For manifolds of non-positive curvature, the intrinsic mean always exists provided the Fréchet function is finite [31].

Remark 2.8. On a general differentiable manifold M, it is often useful and convenient to consider the *extrinsic mean* of Q which is the minimizer, if unique, with respect to the Euclidean distance ρ induced by an appropriate equivariant embedding of M in a Euclidean space E^N. For the case $\alpha = 2$ in (1), a broad verifiable necessary and sufficient condition for the existence of a unique minimizer is often available (see the next section). If the assumptions of Theorem 2.5 hold then

one may still apply it to the extrinsic sample mean, as would be the case, e.g., of the sphere $S^d = \{x \in \mathbb{R}^{d+1} : |x|^2 = 1\}$ with the embedding given by the inclusion map in $\mathbb{R}^{d+1}$, if one takes ϕ to be the inverse exponential map on $S^d \backslash \{-p_0\}$ for a suitable point p_0. But a more broadly applicable CLT for the sample Fréchet mean is provided in the next section (see Theorem 3.3).

3 Asymptotic Distribution of the Extrinsic Sample Mean on a Manifold

Let M be a d-dimensional differentiable manifold and Q a probability measure on it. Consider an embedding $J : M \to E^N$, where E^N is an N-dimensional real vector space, which we may identify with $\mathbb{R}^N$. The extrinsic distance ρ_E on M with respect to the embedding is given by the induced Euclidean distance on $J(M) : \rho_E(p, q) = |J(p) - J(q)|$, where $|.|$ denotes the norm on E^N and $<, >$ denotes the inner product. Letting $Q^J = Q \circ J^{-1}$ denote the induced distribution on E^N, and μ^J its mean, the Fréchet function on the image $J(M)$ of M is given by

$$
\begin{aligned}
F^J(x) &= \int |x - y|^2 Q^J(dy) = \int |x - \mu^J - (y - \mu^J)|^2 Q^J(dy) \qquad (9)\\
&= |x - \mu^J|^2 + \int |y - \mu^J|^2 Q^J(dy) + 2 < x - \mu^J, \int (y - \mu^J) Q^J(dy) >\\
&= |x - \mu^J|^2 + \int |y - \mu^J|^2 Q^J(dy), \qquad (x \in J(M))
\end{aligned}
$$

the integration being over E^N. The last sum is minimized (over $J(M)$) by taking x as the *orthogonal projection* $P(\mu^J)$ of μ^J on $J(M)$, i.e., the point in $J(M)$, if unique, which is at the minimum Euclidean distance from μ^J. Hence we have the following useful result.

Theorem 3.1 (Patrangenaru [40], Hendriks and Landsman [28], BP [10]). *Assume that the projection $P(\mu^J)$ is unique. Then the extrinsic mean of Q is $\mu_E = J^{-1} P(\mu^J)$.*

Remark 3.2. It is known that the set of points x of non-uniqueness of the projection $x \to P(x)$ on E^N (onto $J(M)$) has Lebesgue measure zero [10]. As an example, consider the case $M = S^d$, and the embedding in $\mathbb{R}^{d+1}$ given by the inclusion map. Then the only point of non-uniqueness of the projection map P is the origin 0 in $\mathbb{R}^{d+1}$, in which case the extrinsic mean set is all of S^d. The projection P in this case is defined by $P(x) = x/|x|$ for $x \neq 0$. Thus the extrinsic mean of Q on the sphere is $\mu_E = \mu^J/|\mu^J|$, which exists if and only if the Euclidean mean μ^J of the induced distribution Q^J on E^N is nonzero.

We now derive the asymptotic distribution of the extrinsic sample mean $\hat{\mu}_E$. Let $Y_i = J(X_i)$, where X_i $(i = 1, \cdots, n)$ are i.i.d. observations from the distribution Q on M. The mean of the probability $Q_n^J = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ induced on E^N by the empirical $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ on M is $\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n) = \int y Q_n^J(dy)$. Then $J(\widehat{\mu}_E) = P(\overline{Y})$. By calculus, and the CLT,

$$n^{1/2}\left[P(\bar{Y}) - P(\mu^J)\right] = n^{1/2}[(\text{Jacob } P)_{\mu^J}(\overline{Y} - \mu^J)] + o_p(1) \to N(0, C), \quad (10)$$

in distribution as $n \to \infty$. Here $(\text{Jacob } P)_x$ is the $N \times N$ Jacobian matrix of P (at x) considered as a transformation on $\mathbb{R}^N \approx E^N$ into $\mathbb{R}^N$, and $C = (\text{Jacob } P)_{\mu^J}\Sigma(\text{Jacob}P)_{\mu^J}^t$, Σ being the $N \times N$ covariance matrix of Y_1. Since P maps a neighborhood V of μ^J into the image manifold $J(M)$ of dimension d (smaller than N), the rank of $(\text{Jacob } P)_{\mu^J}$ is d, and the asymptotic distribution in (10) is singular. For purposes of inference it is therefore important to consider the differential $d_{\mu^J}P$ of P at μ^J as a map on the N-dimensional tangent space $T_{\mu^J}(\mathbb{R}^N) \approx \mathbb{R}^N$ into the d-dimensional tangent space $T_{P(\mu^J)}(J(M))$ of the manifold $J(M)$ at $P(\mu^J)$, rather than as a map on $T_{\mu^J}(\mathbb{R}^N)$ into $T_{P(\mu^J)}(\mathbb{R}^N)$ as considered in (10). Consider a standard basis, or frame, $\{e_i : i = 1, \cdots, N\}$ of $E^N \approx \mathbb{R}^N$ (in which Y_j's are expressed) and an orthonormal basis (frame) $\{F_1(y), \cdots, F_d(y)\}$ of the tangent space $T_y(J(M))$ for y in a neighborhood of $P(\mu^J)$ in $J(M)$.

Theorem 3.3 (BP [11], BB [6]). *Assume that the extrinsic mean is unique and the projection operator P is continuously differentiable in a neighborhood of μ^J. Then one has*

$$n^{1/2}(d_{\mu^J}P)(\overline{Y} - \mu^J) \to N(0, \Gamma) \text{ in distribution,} \quad (11)$$

with $\Gamma = B\Sigma B^t$, and

$$n^{1/2}(d_{\bar{Y}}P)(\overline{Y} - \mu^J) \to N(0, \Gamma) \text{ in distribution.} \quad (12)$$

Here $B = B(\mu^J) = ((b_{ij}(\mu^J)))$ is the $d \times N$ matrix of $d_{\mu^J}P$ with respect to an orthonormal basis $\{e_i : i = 1, \cdots, N\}$ of $T_{\mu^J}E^N \approx \mathbb{R}^N$ and a smooth orthonormal basis $\{F_1(P(\mu^J)), \cdots, F_d(P(\mu^J))\}$ of $T_{P(\mu^J)}(J(M))$, i.e., for y in a neighborhood of $P(\mu^J)$ in $J(M)$, $(d_xP)e_i = \sum_j b_{ji}(P(x))F_j(P(x))$.

Note that (12) follows from (11) using a Slutsky type argument.

Remark 3.4. If, for $z \in J(M)$, one views $T_z(J(M))$ as a subspace of $T_z(E^N)$ spanned by $\{e_i : i = 1, \cdots, N\}$, then $B(P(y)) = F(P(y))(\text{Jacob } P)_y$, where the $d \times N$ matrix $F(P(y))$ has row vectors $F_1(P(y)), \cdots, F_d(P(y))$ which form an orthonormal basis of $T_{P(y)}(J(M))$.

Remark 3.5. The matrix Γ in (11) is nonsingular if the support of the distribution of $(d_{\mu^J}P)(Y_i - \mu^J)$ does not lie in a subspace of $T_{P(\mu^J)}(J(M))$ (of dimension smaller than d). In particular, this is the case if Q has an absolutely continuous component with respect to the volume measure on M.

Example 3.1. Consider the sphere $S^d = \{x \in \mathbb{R}^{d+1} : |x|^2 = 1\}$ and the inclusion map as the embedding J. Then $P(x) = x/|x|$ $(x \neq 0)$. It is not difficult to check that the Jacobian matrix of the projection, considered as a map on $\mathbb{R}^{d+1}$ into $\mathbb{R}^{d+1}$, is given by

$$(\text{Jacob } P)_x = |x|^{-1}[I_{d+1} - |x|^{-2}(xx^t)], \ (x \neq 0). \tag{13}$$

Let $A(x)$ be a $d \times (d+1)$ matrix whose rows form an orthonormal basis of $T_x(S^d) = \{v \in \mathbb{R}^{d+1} : x^t v = 0\}$. Then the differential of $P(x)$, as a map on $\mathbb{R}^{d+1}$ into S^d, is expressed in coordinates of this basis as $(d_x P)u = A(x)(\text{Jacob } P)_x u (u \in \mathbb{R}^{d+1})$. The left sides of (11) and (12) are then obtained by letting $x = \mu^J$ and $\overline{X}$, respectively, and $u = \overline{Y} - \mu^J$. For $d = 2$, and $x = (x^1, x^2, x^3)^t \neq (0, 0, \pm 1)^t$, and $x^3 \neq 0$, one may choose the two rows of $A(x)$ as $(-x^2, x^1, 0)/\sqrt{(x^2)^2 + (x^1)^2}$ and $((x^1, x^2, -((x^2)^2 + (x^1)^2)/x^3)/c$, where c normalizes the second vector to unity. For $x = (0, 0, \pm 1)$, one may simply take the basis vectors of $T_x(S^2)$ as (1,0,0), and (0,1,0). If $x^3 = 0$ and $x^1 \neq 0$, $x^2 \neq 0$, then the second vector in the basis may be taken as (0,0,1). Permuting the indices, all cases are now covered.

4 Asymptotic Distribution of the Intrinsic Sample Mean and the Role of Curvature

In this section we apply Theorem 2.5 to the *intrinsic mean* μ_I on a Riemannian manifold M with metric tensor g. That is, μ_I is the Fréchet mean with respect to the geodesic distance $\rho = \rho_g$ (with $\alpha = 2$ in (1)).

On the tangent space $T_p(M)$ at p of a complete Riemannian manifold M, one defines the *exponential map* $Exp_p : T_p(M) \to M$, by letting $Exp_p(v)$ be the point $q = \gamma(|v|)$ reached at time $t = |v|$ by the unit speed geodesic $\gamma(t) = \gamma(t; p, v)$ with $\gamma(0) = p$ and initial speed $\dot{\gamma}(0) = v/|v|$ if $v \neq 0$, and $Exp_p(0) = p$. For each unit vector v in $T_p(M)$, let $t_0 = t_0(p, v)$ be the supremum of all t such that the unit speed geodesic $\gamma(.; p, v)$ is length minimizing on $[0, t]$. Then $\gamma(t_0; p, v)$ is called a *cut point* of p and the set of all cut points of p (as v varies over all unit vectors in $T_p(M)$) is called the *cut locus* of p and denoted by $C(p)$. For $q \in M \backslash C(p)$, the inverse $Exp_p^{-1}(q)$ of the exponential map is defined as $v = v(q) \in T_p(M)$ such that $Exp_p(v) = q$. It is known that Exp_p^{-1} is a diffeomorphism on $M \backslash C(p)$ onto its image in $T_p(M)$, which is homeomorphic to an open ball in $T_p(M)$ with center 0 (p. 271) [17]. The quantity $inj(M) = \inf\{\rho_g(p, C(p)); p \in M\}$ is called the *injectivity radius* of M. The inverse exponential map $Exp_p^{-1}(q)$ gives rise to the so-called *normal coordinates* of q (with pole p), $q \in M \backslash C(p)$, when expressed in terms of an orthonormal basis of $T_p(M)$.

Let Q be a probability with support contained in a geodesic ball $B_r(p)$ of radius r centered at p. If a unique minimizer of the Fréchet function $F(q) = \int \rho_g^2(q, q')Q(dq')$, $q \in B_r(p)$, exists (in $B_r(p)$), it is called a *local intrinsic mean*

of Q in $B_r(p)$. We will denote by $\bar{C}$ the least upper bound of sectional curvatures of M, if this l.u.b. is positive, and zero if the l.u.b is negative or zero. Part (a) of the following theorem, which is an extension of Theorem 2.2 in [11], and Theorem 4.2 in [4], now follows from Theorem 2.5. Part (b) is derived in [5]. For its notation we use the order $A \geq B$ for symmetric $d \times d$ matrices A, B to mean that $A - B$ is nononegative definite. The function f appearing in (16) is defined as

$$f(x) = \begin{cases} 1 & \text{if } \bar{C} = 0 \\ \sqrt{\bar{C}}x \cos(\sqrt{\bar{C}}x)/\sin(\sqrt{\bar{C}}x) & \text{if } \bar{C} > 0 \\ \sqrt{\bar{C}}x \cosh(\sqrt{\bar{C}}x)/\sinh(\sqrt{\bar{C}}x) & \text{if } \bar{C} < 0 \end{cases} \tag{14}$$

with $\bar{C}$, as defined earlier, the l.u.b. of the sectional curvatures of M if positive, or zero otherwise. Theorem 2.5 is a CLT for the local intrinsic sample mean μ_n around the local intrinsic mean μ_I of a probability Q, based on i.i.d. observations $X_1, \cdots, X_n$ with common distribution Q. Actually we look at vector valued $Y_i = \phi(X_i)$, where ϕ is the inverse exponential map Exp_p^{-1} on an appropriate open subset of $T_p(M)$, and derive a CLT for $\nu_n = \phi(\mu_n)$ around $\nu = \phi(\mu_I)$. Estimation of μ_I is then achieved via ϕ^{-1}.

Theorem 4.1. *Let Q have support in a geodesic ball $B_r(p)$ with $\overline{B_r(p)} \subset M \backslash C(p)$.*

Assume the following conditions (A1)–(A5)*:*

(A1) The local intrinsic mean μ_I exists in $B_r(p)$.

(A2) Let ϕ denote the inverse exponential map Exp_p^{-1}, $h(z, y) = \rho_g^2(\phi^{-1}z, \phi^{-1}y)$, with $z, y \in V = Exp_p^{-1}(B_r(p))$ expressed in normal coordinates with respect to an orthonormal basis of $T_p(M)$; then $z \to h(z, y)$ is twice continuously differentiable for all y.

(A3) With $\psi^{(r)}(z, y) \equiv D^r h(z, y) = (\partial/\partial z^r) d_g^2(\phi^{-1}z, \phi^{-1}y)$, one has $E(\psi^{(r)}(\nu, Y_1))^2 < \infty \ \forall r = 1, \cdots, d$, where $\nu = Exp_p^{-1}(\mu_I)$ and Y_1 has the distribution $Q \circ \phi^{-1}$.

(A4) One has $\sup\{E|D(\psi^{(r)}(y, Y_1) - D(\psi^{(r)}(\nu, Y_1)| : |y - \nu| \leq \epsilon\} \to 0$ as $\epsilon \downarrow 0$.

(A5) $\Lambda = ((ED_s\psi^{(r)}(\nu, Y_1))) \equiv ((\{ED_s D_r h(z, Y_1)\}_{z=\nu}))$ is nonsingular.

Then,

(a) Denoting by μ_n the local intrinsic sample mean, $\phi(\mu_n)$ has the asymptotic distribution given by

$$\sqrt{n}\,[\phi(\mu_n) - \phi(\mu_I)] \to N\left(0, \Lambda^{-1}\tilde{\Sigma}\Lambda^{-1}\right) \tag{15}$$

in distribution as $n \to \infty$, where $\tilde{\Sigma} = Cov(\{\psi^{(r)}(\nu, Y_1) : r = 1, \cdots, d\})$.

(b) If one takes $p = \mu_I$, then $\nu = 0$, and

(i) $\psi^{(r)}(0, y) = -2y^r$

(ii) $E(Y_1) = \int y Q \circ \phi^{-1}(dy) = 0$,

(iii) $\tilde{\Sigma} = 4Cov(Y_1) = 4E(Y_1 Y_1^t)$,
(iv) The matrix $\Lambda = ((\Lambda_{rs}))_{1\le r,s\le d}$ *satisfies the order relation*

$$\Lambda \ge ((2E(([1 - f(|Y_1|)]/|Y_1|^2)Y_1^r Y_1^s + f(|Y_1|)\delta_{rs})))_{1\le r,s\le d}, \qquad (16)$$

with equality in (16) *in the case of constant sectional curvature.*

Remark 4.2. If Q has a density component with respect to the volume measure on M, then $\tilde{\Sigma}$ is nonsingular.

Remark 4.3. It has been proved by W.S. Kendall [35] that if the support of Q is contained in $B_{r^*/2}(p)$ where $r^* = \min\{inj(M), \pi/\sqrt{C}\}$, then a local Fréchet mean μ_I of Q exists in $B_{r^*/2}(p)$. The result of Afsari [1] shows that this μ_I is the global minimizer on M. If, in addition, the support of Q is contained in $B_{r^*/2}(\mu_I)$, then all the assumptions of Theorem 4.1 are satisfied [5, 6]. For manifolds with nonpositive curvature, the central limit theorem (15) holds for all Q, provided the Fréchet function of Q is finite and $E|Y_1|^2 < \infty$ (see [11], Remark 2.2)

Example 4.1. The sphere $S^d = \{x \in \mathbb{R}^{d+1} : |x|^2 = 1\}$ is a compact Riemannian manifold under the metric induced by the inclusion map. Its geodesics are the big circles, the geodesic starting at p with an initial velocity v being given by $\gamma(t; p, v) = (\cos t|v|)p + (\sin t|v|)v/|v|$, with $v \in T_p(S^d) = \{v \in \mathbb{R}^{d+1} : p^t v = 0\}$. The cut locus of p is $C(p) = \{-p\}$. The exponential map and its inverse are given by

$$Exp_p(0) = p, Exp_p(v) = \cos(|v|)p + \sin(|v|)v/|v|,\ v \ne 0,\ (v \in T_p(S^d)); \qquad (17)$$

$$Exp_p^{-1}(p) = 0, Exp_p^{-1}(q) = \arccos(p^t q)/(1 - (p^t q)^2)^{1/2}[q - (p^t q)p],\ (q \ne p, -p).$$

The *geodesic distance* between p and q is $\rho_g(p, q) = \arccos(p^t q) \in [0, \pi]$, so that the injectivity radius is $inj(S^d) = \pi$. Because of isotropy, the sectional curvature is the same for every section of $T_p(S^d)$, for all p, and the unit sphere has therefore the constant curvature 1. Thus the quantity r^* appearing in Remark 4.3 has the value π, so that the conclusions of Theorem 4.1 hold if the support of Q is contained in $B_{\pi/2}(p)$ as well as in $B_{\pi/2}(\mu_I)$. Here the function f in (14) is $f(u) = u(\cos u)/(\sin u)$. The normal coordinates $y^1, \cdots, y^d$ at μ_I of $x = Exp_{\mu_I}(y)$, where y is expressed as $y = y^1 v_1 + \cdots + y^d v_d$ with respect to an orthonormal basis $\{v_r : r = 1, \cdots, d\}$ of $T_{\mu_I}(S^d)$, are now given by (see (17)):

$$y_r = \arccos(\mu_I^t x)/(1 - (\mu_I^t x)^2)^{1/2} x^t v_r,\ (r = 1, \cdots, d), x \in S^d. \qquad (18)$$

Now $\Lambda_{r,s}$ is computed from its definition in (A5), with Y_{1r} given by the right hand side of (18) obtained by substituting X_1 (with distribution Q) for x, and, similarly, Y_{1s} is obtained by changing r to s.

5 Nonparametric Inference on General Manifolds

Theorems 2.5 and 3.3 allow us to construct nonparametric confidence regions for intrinsic and extrinsic means of probability measures Q on a manifold M, and to carry out nonparametric two-sample tests for the equality of such means of two distributions Q_1 and Q_2 on M. The latter tests are really meant to distinguish Q_1 from Q_2. On high dimensional spaces, such as the shape spaces of main interest here, the means are generally good indices for this purpose, as the data examples in Sect. 9 show.

For the construction of an extrinsic confidence region for the extrinsic mean μ_E of Q one may use the corresponding region for μ^J using (11) or (12) and then transform by J^{-1}. In general (12) is simpler to use. The following asymptotic chisquare distribution is an easy consequence:

$$n\left[(d_{\bar{Y}}P)(\bar{Y}-\mu^J)\right]^t(\hat{B}\hat{\Sigma}\hat{B}^t)^{-1}\left[(d_{\bar{Y}}P)(\bar{Y}-\mu^J)\right]\to\chi^2_d \text{ in distribution,} \quad (19)$$

where χ^2_d is the chisquare distribution with d degrees of freedom. Here $\hat{B}=B(\bar{Y})$ estimates $B=B(\mu^J)$, and $\hat{\Sigma}$ is the sample covariance matrix of $Y_1,\cdots,Y_n$. The statistic does not depend on the choice of the orthonormal basis of $T_{\bar{Y}}(J(M))$ for computing $\hat{B}$. The relation (19) may be used to construct a confidence region for the extrinsic mean μ_E. Bootstrapping, which leads to a smaller order of coverage error in the case of an absolutely continuous Q, may not always be feasible if N is large and the sample size n is not sufficiently large to ensure that, with high probability, the bootstrap sample is not degenerate.

Turning to the (local) intrinsic mean μ_I of Q, (15) leads to the asymptotic chisquare distribution

$$n[\phi(\mu_n)-\phi(\mu_I)]^t\hat{\Lambda}\hat{\tilde{\Sigma}}^{-1}\hat{\Lambda}[\phi(\mu_n)-\phi(\mu_I)]\to\chi^2(d) \quad (20)$$

in distribution as $n\to\infty$, where $\hat{}$ denotes an estimate with Q replaced by the empirical Q_n; that is, the distribution $Q\circ\phi^{-1}$ of Y_1 is replaced by $Q_n\circ\phi^{-1}=n^{-1}\sum_{1\le i\le n}\delta_{Y_i}$. This leads to a confidence region for μ_I. One arbitrariness here is the choice of the point p in computing ϕ. It seems reasonable to take p close to μ_n. Another idea is to use $p=\mu_I$, in which case $\phi(\mu_I)=0$. To use (20) in this case to obtain a confidence region would be computationally more intensive. It would involve finding those values p such that, with $\phi=Exp_p^{-1}$, the left side in (20) (with $\phi(\mu_I)=0$) is smaller than $\chi^2_d(1-\alpha)$, the $(1-\alpha)$-th quantile of χ^2_d. This requires computing the quantities in (20), including $\phi(\mu_n)$, for each p. But, unlike the case of the extrinsic mean where the ambient vector space E^N has generally a large dimension and the bootstrap estimate of the covariance matrix Σ tends to be singular, $\tilde{\Sigma}$ in (20) is a $d\times d$ matrix. If Q is absolutely continuous, which is a reasonable assumption in most shape data, the bootstrap construction of the confidence region will tend to have a smaller coverage error than the one using χ^2_d.

We next consider the the two-sample problem of distinguishing two distributions Q_1 and Q_2 on M, based on two independent samples of sizes n_1 and n_2, respectively: $\{Y_{j1} = J(X_{j1}) : j = 1, \cdots, n_1\}, \{Y_{j2} = J(X_{j2}) : j = 1, \cdots, n_2\}$. Hence the proper null hypothesis is $H_0 : Q_1 = Q_2$. For high dimensional M it is often sufficient to test if the two Fréchet means are equal. For the extrinsic procedure, again consider an embedding J into E^N. Write μ_i for μ_i^J for the population means and $\bar{Y}_i$ for the corresponding sample means on E^N $(i = 1, 2)$. Let $n = n_1 + n_2$, and assume $n_1/n \to p_1, n_2/n \to p_2 = 1 - p_1, 0 < p_i < 1 (i = 1, 2)$, as $n \to \infty$. If $\mu_1 \neq \mu_2$ then $Q_1 \neq Q_2$. One may then test $H_0 : \mu_1 = \mu_2 (= \mu$,say). Since N is generally quite large compared to d, the direct test for $H_0 : \mu_1 = \mu_2$ based on $\bar{Y}_1 - \bar{Y}_2$ is generally not a good test. Instead, we compare the two extrinsic means μ_{E1} and μ_{E2} of Q_1 and Q_2 and test for their equality. This is equivalent to testing if $P(\mu_1) = P(\mu_2)$. Then, by (12), assuming H_0,

$$n^{1/2} d_{\bar{Y}} P(\bar{Y}_1 - \bar{Y}_2) \to N(0, B(p_1 \Sigma_1 + p_2 \Sigma_2) B^t) \tag{21}$$

in distribution, as $n \to \infty$.

Here $\bar{Y} = p_1 \bar{Y}_1 + p_2 \bar{Y}_2$ is the *pooled estimate* of the common mean $\mu_1 = \mu_2 = \mu$, say, $B = B(\mu)$ (see (11)), and Σ_1, Σ_2 are the covariance matrices of Y_{j1} and Y_{j2}. This leads to the asymptotic chisquare statistic below:

$$n[d_{\bar{Y}} P(\bar{Y}_1 - \bar{Y}_2)]^t [\hat{B}(p_1 \hat{\Sigma}_1 + p_2 \hat{\Sigma}_2) \hat{B}^t]^{-1} [d_{\bar{Y}} P(\bar{Y}_1 - \bar{Y}_2)] \to \chi^2_d \tag{22}$$

in distribution, as $n \to \infty$.

Here $\hat{B} = B(\bar{Y})$, $\hat{\Sigma}_i$ is the sample covariance matrix of Y_{ji}. One rejects the null hypothesis H_0 at a level of significance $1 - \alpha$ if and only if the observed value of the left side of (22) exceeds $\chi^2_d(1 - \alpha)$.

For the two-sample intrinsic test, let μ_{I1}, μ_{I2} denote the intrinsic means of Q_1 and Q_2 and consider $H_0 : \mu_{I1} = \mu_{I2}$. Denoting by μ_{n1}, μ_{n2} the intrinsic sample means, (15) implies that, under H_0,

$$n^{1/2}[\phi(\mu_{n1}) - \phi(\mu_{n2})] \to N(0, p_1 \Lambda_1^{-1} \tilde{\Sigma}_1 \Lambda_1^{-1} + p_2 \Lambda_2^{-1} \tilde{\Sigma}_2 \Lambda_2^{-1}) \tag{23}$$

in distribution,

where $\phi = Exp_p^{-1}$ for some convenient p in M, and Λ_i, $\tilde{\Sigma}_i$ are as in Theorem 4.1 with the empirical Q_{ni} in place of $Q_i (i = 1, 2)$. For p choose μ_n on the geodesic from μ_{n1} to μ_{n2} with $P_g(\mu_n, \mu_{n1}) = p_2 P_g(\mu_{n1}, \mu_{n2})$, and with this choice we write $\hat{\phi}$ for ϕ. The test then rejects $H_0 : Q_1 = Q_2$, if

$$n[\hat{\phi}(\mu_{n1}) - \hat{\phi}(\mu_{n2})]^t [p_1 \hat{\Lambda}_1^{-1} \hat{\tilde{\Sigma}}_1 \hat{\Lambda}_1^{-1} + p_2 \hat{\Lambda}_2^{-1} \hat{\tilde{\Sigma}}_2 \hat{\Lambda}_2^{-1}]^{-1} [\hat{\phi}(\mu_{n1}) - \hat{\phi}(\mu_{n2})] > \chi^2_d(1 - \alpha). \tag{24}$$

Finally, consider a *match pair problem* with i.i.d. observations (X_{j1}, X_{j2}) having the distribution Q on the product manifold $M \times M$. If J is an embedding of M into E^N, then $\tilde{J}(x, y) = (J(x), J(y))$ is an embedding of $M \times M$ into $E^N \times E^N$. Let μ_{E1}, μ_{E2} be the extrinsic means of the (marginal) distributions Q_1 and Q_2 of X_{j1} and X_{j2}, respectively. Once again, we are interested in testing $H_0 : Q_1 = Q_2$ by checking if $\mu_{E1} = \mu_{E2}$. Note that the extrinsic mean of Q is $\tilde{\mu}_E = (\mu_{E1}, \mu_{E2})$. If $\bar{Y}_1$, $\bar{Y}_2$ are the sample means of $Y_{j1} = J(X_{j1})$, $Y_{j2} = J(X_{j2})$, $j = 1, \cdots, n$, on E^N with $E(Y_{j1}) = \mu_1$ and $E(Y_{j2}) = \mu_2$, and $\bar{\bar{Y}} = (\bar{Y}_1, \bar{Y}_2)$, then the extrinsic sample mean in the image space $\tilde{J}(M \times M)$ is $(P(\bar{Y}_1), P(\bar{Y}_2))$. Also, write $\bar{Y} = (\bar{Y}_1 + \bar{Y}_2)/2$. Under H_0, $\mu_1 = \mu_2 = \mu$, say, and one has

$$n^{1/2} d_{\bar{Y}}(P(\bar{Y}_1) - P(\bar{Y}_2)) \to N(0, \Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}). \tag{25}$$

On the right, Σ_{11} and Σ_{22} are the $d \times d$ covariance matrices of $(d_\mu P)(Y_{j1} - \mu_1)$ and $(d_\mu P)(Y_{j2} - \mu_2)$, while Σ_{12} is the $d \times d$ cross covariance matrix of $(d_\mu P)(Y_{j1} - \mu_1)$ and $(d_\mu P)(Y_{j2} - \mu_2)$, and $\Sigma_{21} = \Sigma_{12}'$. As above, one derives a chisquare test for H_0, using (25) and sample estimates of the covariance matrices.

6 Geometry of Shape Spaces and Equivariant Embeddings

The manifolds of main interest to us are shape spaces of landmarks based k-ads. A k-ad is a set of k labeled landmarks, $k > m$, not all the same, measured on an object or scene of interest. In general, the k-ad $(x_1, \cdots, x_k)$ is a k-tuple of points in $\mathbb{R}^m$, represented as an $m \times k$ matrix, although only $m = 2$ and 3 are of practical interest for the most part. The shape of a k-ad is the k-ad modulo a Lie group of transformations or, equivalently, it is the maximal invariant, or orbit, of the k-ad under this group. The appropriate Lie group depends on the particular statistical goal and the way the measurement of a k-ad may vary, for example, because of differences in equipment, the position and angle from which the observations are taken or recorded, etc.

6.1 Kendall's Similarity Shape Space $\boldsymbol{\Sigma_m^k}$

The similarity shape of a k-ad $x = (x_1, \cdots, x_k)$ in $\mathbb{R}^m$, not all points the same, is its orbit under the group generated by translations, scaling and rotations. Writing $\bar{x} = (x_1 + \cdots + x_k)/k$, $< \bar{x} >= (\bar{x}, \cdots, \bar{x})$, the effect of translation is removed by looking at $(x_1 - \bar{x}, \cdots, x_k - \bar{x}) = x - < \bar{x} >$, which lies in the $mk - m$ dimensional hyperplane L of $\mathbb{R}^{mk}$ made up of $m \times k$ matrices with the m row sums all equal to zero. To get rid of scale, one looks at $u = (x - < \bar{x} >)/|x - < \bar{x} > |$, where $|.|$ is the usual norm in $\mathbb{R}^{mk}$. This translated and scaled k-ad is called the

preshape of the k-ad. It lies on the unit sphere in L, and is isomorphic to $S^{m(k-1)-1}$. An alternative representation of the preshape is obtained as $p = xH/|xH|$, where H is the $k \times (k-1)$ Helmert matrix comprising $k-1$ column vectors forming an orthonormal basis of $1^{\perp}$, namely, the subspace of $\mathbb{R}^k$ orthogonal to $(1, \cdots, 1)^t$. A standard H has the j-th column given by $(a(j), \cdots, a(j), -ja(j), 0, \cdots, 0)^t$, where $a(j) = [j(j+1)]^{-1/2}$ $(j = 1, \cdots, k-1)$. Then p is an $m \times (k-1)$ matrix of norm one. The shape $\sigma(x) = \sigma(p)$ of x is then identified with the orbit of p under all rotations:

$$\sigma(x) = \sigma(p) = \{Ap : A \in SO(m)\}, \tag{26}$$
$$[SO(m) = \{\mathrm{A} : AA^t = I_m, det(A) = 1\}].$$

$SO(m)$ is called the *special orthogonal group* acting on $\mathbb{R}^m$. The set of all shapes $\sigma(x)$ is Kendall's *similarity shape* space Σ_m^k, $R > m$.

6.1.1 Intrinsic Geometry of Σ_2^k

For the case $m = 2$, it is convenient to regard a k-ad $x = ((x_1, y_1), \cdots, (x_k, y_k))$ as a k-tuple $z = (z_1, \cdots, z_k)$ of numbers $z_1 = x_1 + iy_1, \cdots, z_k = x_k + iy_k$ in the complex plane $\mathbb{C}$, and let $p = (z - < \bar{z} >)/|z - < \bar{z} > |$. Then the shape of x, or z, is identified with the orbit O_p,

$$\sigma(z) = \sigma(p) = \{e^{i\theta} p : \theta \in (-\pi, \pi]\} = O_p. \tag{27}$$

One may equivalently, consider the shape as the orbit $\{\lambda((z - < \bar{z} >) : \lambda \in \mathbb{C}\}$. That is, the shape of x, or z, is identified with a complex line passing through the origin in the subspace of $\mathbb{C}^k$ of complex dimension $k-1$ defined by $\tilde{L} = \{q = (q_1, \cdots, q_k) \in \mathbb{C}^k \backslash \{0\} : q_1 + \cdots + q_k = 0\} \approx \mathbb{C}^{k-1} \backslash \{0\}$. The shape space is then identified with the *complex projective space* $\mathbb{C}P^{k-2}$, of (real) dimension $2k-4$.

Note that $\{e^{i\theta} : \theta \in (-\pi, \pi]\}$ is a 1-dimensional compact group $G(\approx S^1)$ of isometries of the preshape sphere $\mathbb{C}S^{k-1} = \{q = (q_1, \cdots, q_k) : |q| = 1, q_1 + \cdots + q_k = 0\}$. By Helmertization, we will use the *representation* of $\mathbb{C}S^{k-1}$ as $\{p = (p_1, \cdots, p_{k-1}) \in \mathbb{C}^{k-1} : |p| = 1\}$, which is isomorphic to S^{2k-3}, and $\Sigma_2^k = \mathbb{C}S^{k-1}/G$. Recall that the metric tensor on $S^{2k-3} \approx \mathbb{C}S^{k-1}$ is that inherited from the inclusion map into $\mathbb{R}^{2(k-1)} = \{(x_1, y_1, x_2, y_2, \cdots, x_{k-1}, y_{k-1}) : (x_j, y_j) \in \mathbb{R}^2 \ \forall j\} \approx \mathbb{C}^{k-1} = \{(z_1, z_2, \cdots, z_{k-1}) : z_j = x_j + iy_j \in \mathbb{C} \ \forall j\}$, namely, $\langle v, w \rangle = Re(vw^*)$, when v,w are expressed as complex $1 \times (k-1)$ matrices (row vectors) in $\mathbb{C}S^{k-1}$. The projection map is then $\pi : p \to \sigma(p)$. The vertical subspace V_p is obtained by differentiating the curve $\theta \to e^{i\theta} p$, say at $\theta = 0$, yielding ip. That is, $V_p = \{cip : c \in \mathbb{R}\}$. Thus the horizontal subspace is $H_p = \{\tilde{v} : Re(p\tilde{v}^*) = 0, Re((ip)\tilde{v}^*) = 0\} = \{\tilde{v} : p\tilde{v}^* = 0\}$. The geodesics $\gamma(t; \sigma(p), v)$ for $v = (d_p\pi)\tilde{v}$ (for $\tilde{v}$ in H_p), and the exponential map $Exp_{\sigma(p)}$ on Σ_2^k are specified by this isometry between $T_{\sigma(p)}(\Sigma_2^k)$ and H_p for all shapes $\sigma(p)$

(see Example 4.1). Thus, identifying vectors v in H_p with vectors v in $T_{\sigma(p)}(\Sigma_2^k)$, one obtains

$$T_{\sigma(p)}(\Sigma_2^k) = \{v = (d_p\pi)\tilde{v} : \forall v \text{ such that } p\tilde{v}^* = 0\} \tag{28}$$
$$Exp_{\sigma(p)}0 = \sigma(p),\ Exp_{\sigma(p)}v = \sigma(\cos(|\tilde{v}|)p + \sin(|\tilde{v}|)\tilde{v}/|\tilde{v}|)\ (v \neq 0,\ p\tilde{v}^* = 0);$$
$$\gamma(t;\sigma(p),v) = \sigma((\cos t)p + (\sin t)\tilde{v}/|\tilde{v}|),\ (t \in \mathbb{R}, p\tilde{v}^* = 0),\ v \neq 0.$$

Denoting by ρ_{gs} and ρ_g the geodesic distances on $\mathbb{C}S^{k-1}$ and Σ_2^k, respectively, and recalling that (see Example 4.1) $\rho_{gs}(p,q) = \arccos(Repq^*)$, one has

$$\begin{aligned}\rho_g(\sigma(p),\sigma(q)) &= \inf\{\rho_{gs}(p',q') : p' \in O_p, q' \in O_q\} \\ &= \inf\{\arccos(Ree^{i\theta}pq^*) : \theta \in [0,2\pi)\} \\ &= \arccos(|pq^*|) \in [0,\pi/2].\end{aligned} \tag{29}$$

It follows that the geodesics are periodic with period π, and the cut locus of $\sigma(p)$ is $\{\sigma(q) : \text{all q such that} \arccos(|pq^*|) = \pi/2\}$, and that the injectivity radius of Σ_2^k is $\pi/2$. The inverse exponential map is given by $Exp_{\sigma(p)}^{-1}(\sigma(q)) = v$, where $v = (d_p\pi)\tilde{v}$ ($\tilde{v} \in H_p$), and $\tilde{v}$ satisfies (Use (17) with the representation of S^{2k-3} as $\mathbb{C}S^{k-1}$)

$$\begin{aligned}\tilde{v} &= Exp_p^{-1}(qe^{i\theta}) \\ &= [\arccos(Re(pq^*e^{-i\theta})](1 - [Re(pq^*e^{-i\theta})]^2)^{-1/2}qe^{i\theta} - (pq^*e^{-i\theta})p,\end{aligned} \tag{30}$$

where θ is so chosen as to minimize $\rho_{gs}(p,qe^{i\theta}) = \arccos(Re(pq^*e^{-i\theta}))$. That is, $(pq^*e^{-i\theta}) = |pq^*|$, or $e^{i\theta} = pq^*/|pq^*|$ (for $pq^* \neq 0$, i.e., for $\sigma(q)$ not in $C(\sigma(p))$.

Hence, writing $\rho = (\arccos)\rho_g(\sigma(p),\sigma(q))$, $\rho \neq 0$, one has

$$\begin{aligned}\tilde{v} &= [\arccos(|pq^*|)](1 - |pq^*|^2)^{-1/2}\{(pq^*/|pq^*|)q - |pq^*|p\} \\ &= [\rho/\sin\rho]\{qe^{i\theta} - (\cos\rho)p\}\ (e^{i\theta} = pq^*/\cos\rho).\end{aligned} \tag{31}$$

This horizontal vector $\tilde{v}(\in H_p)$ represents $Exp_{\sigma(p)}^{-1}(\sigma(q)) = v$.

The sectional curvature of Σ_2^k at a section generated by two orthonormal vector fields $\tilde{W}_1$ and $\tilde{W}_2$ is $1 + 3\cos^2\phi$ where $\cos\phi = \langle U_1, iU_2\rangle$, U_1 and U_2 being the horizontal lifts of $\tilde{W}_1$ and $\tilde{W}_2$ (see [17]).

6.1.2 Extrinsic Geometry of Σ_2^k Induced by an Equivariant Embedding

One problem with carrying out an intrinsic analysis of the Fréchet mean is that no broad sufficient condition is known for its existence (i.e., of the uniqueness of the minimizer of the corresponding Fréchet function). Also, often such an analysis, assuming uniqueness, is computationally much more intensive than an extrinsic analysis. However, for an extrinsic analysis to be very effective one should choose a good embedding which retains as many geometrical features of the shape manifold as possible without making it cumbersome. Let Γ be a Lie group acting on a differentiable manifold M, and denote by $GL(N, \mathbb{R})$ the linear group of nonsingular transformations on a Euclidean space E^N of dimension N onto itself. An embedding J on M into E^N is said to be Γ-*equivariant* if there exists a group homomorphism $\Phi : \gamma \to \phi_\gamma$ of Γ into $GL(N, \mathbb{R})$ such that $J(\gamma p)) = \phi_\gamma(Jp)$ $\forall p \in M, \gamma \in \Gamma$. Often, when there is a natural Riemannian structure on M, Γ is a group of isometries of M. Consider the so-called *Veronese-Whitney embedding* J of Σ_2^k into the (real) vector space $S(k-1, \mathbb{C})$ of all $(k-1) \times (k-1)$ Hermitian matrices $B = B^*$, defined by

$$J\sigma(p) = p^* p \ \ [\sigma(p) = \{e^{i\theta} p, \theta \in [0, 2\pi), p \in \mathbb{C}S^{k-1}]. \tag{32}$$

The Euclidean inner product on $S(k-1, \mathbb{C})$, considered as a real vector space, is given by $\langle B, C \rangle = Re(\text{Trace}(BC^*))$. Let $SU(k-1)$ denote the special unitary group of all $(k-1) \times (k-1)$ unitary matrices A (i.e., $A^*A = I, det(A) = 1$) acting on $S(k-1, \mathbb{C})$ by $B \to A^*BA$. Then the embedding (32) is Γ-equivariant, with $\Gamma = \{\gamma_A : A \in SU(k-1)\}$ and the group action on Σ_2^k given by: $\gamma_A \sigma(p) = \sigma(pA)$. For $J\sigma(pA) = A^* p^* pA = \phi(\gamma_A)(J\sigma(p))$, say, where the group homomorphism on Γ onto $SU(k-1)$ is given by $\gamma_A \to \phi(\gamma_A) : \phi(\gamma_A)B = A^*BA$. Note that $SU(k-1)$ is a group of isometries of $S(k-1, \mathbb{C})$. If Σ_2^k is given the metric tensor inherited from $S(k-1, \mathbb{C})$ by the embedding (32), then the embedding is isometric as well as equivariant.

A *size-and-shape similarity shape* $s\sigma(z)$ is defined for Helmertized k-ads $z = (z_1, \cdots, z_{k-1})$ as its orbit under $SO(m)$. An equivariant embedding for it is $s\sigma(z) \to z^*z/|z|$, on the *size-and-shape-similarity shape* space $S\Sigma_2^k$ into $S(k-1, \mathbb{C})$.

6.2 Reflection Similarity Shape Space $R\Sigma_m^k$, $m > 2$

For $m > 2$, let $\tilde{N}S^{m(k-1)-1}$ be the subset of the centered preshape sphere $S^{m(k-1)-1}$ whose points p span $\mathbb{R}^m$, i.e., which, as $m \times (k-1)$ matrices, are of full rank. We define the *reflection similarity shape* of the k-ad as

$$r\sigma(p) = \{Ap : A \in O(m)\} \ (p \in \tilde{N}S^{m(k-1)-1}), \tag{33}$$

where $O(m)$ is the set of all $m \times m$ orthogonal matrices $A : AA^t = I_m$, $det(A) = \pm 1$. The set $\{r\sigma(p) : p \in \tilde{N}S^{m(k-1)-1}\}$ is the *reflection similarity shape space* $R\Sigma_m^k = \tilde{N}S^{m(k-1)-1}/O(m)$. Since $\tilde{N}S^{m(k-1)-1}$ is an open subset of the sphere $S^{m(k-1)-1}$, it is a Riemannian manifold. Also $O(m)$ is a compact Lie group acting on it. Hence there is a unique Riemannian structure on $R\Sigma_m^k$ such that the projection map $p \to O(p)$ is a Riemannian submersion.

We next consider a useful embedding of $R\Sigma_m^k$ into the vector space $S(k-1, \mathbb{R})$ of all $(k-1) \times (k-1)$ real symmetric matrices (see [3, 8, 9, 20]). Define

$$J(r\sigma(p)) = p^t p \ (p \in \tilde{N}S^{m(k-1)-1}), \tag{34}$$

with p an $m \times (k-1)$ matrix with norm one. Note that the right side is a function of $r\sigma(p)$. Here the elements p of the preshape sphere are Helmertized. To see that this is an embedding, we first show that J is one-to-one on $R\Sigma_m^k$ into $S(k-1, \mathbb{R})$. For this note that if $J(r\sigma(p))$ and $J(r\sigma(q))$ are the same, then the Euclidean distance matrices $((|p_i - p_j|))_{1 \le i \le j \le k-1}$ and $((|q_i - q_j|))_{1 \le i \le j \le k-1}$ are equal. Since p and q are centered, by geometry this implies that $q_i = Ap_i (i = 1, \cdots, k-1)$ for some $A \in O(m)$, i.e., $r\sigma(p) = r\sigma(q)$. We omit the proof that the differential dJ is also one-to-one. It follows that the embedding is equivariant with respect to a group action isomorphic to $O(k-1)$.

For $m > 2$, a *size-and-reflection shape* $sr\sigma(z)$ of a Helmertized k-ad z in $\mathbb{R}^m$ of full rank m is given by its orbit under the group $O(m)$. The space of all such shapes is the size-and-reflection shape space $SR\Sigma_m^k$. An $O(k-1)$-equivariant embedding of $SR\Sigma_m^k$ into $S(k-1, \mathbb{R})$ is : $J(sr\sigma(z)) = z^t z/|z|$.

6.3 *Affine Shape Space* $A\Sigma_m^k$

Let $k > m+1$. Consider the set of all k-ads in $\mathbb{R}^m$, with full rank m as $m \times k$ matrices. The affine shape of a k-ad x may be identified with its orbit under all affine transformations:

$$\sigma(x) = \{Ax + c : A \in GL(m, \mathbb{R}), c \in \mathbb{R}^m\}. \tag{35}$$

If the k-ad is centered as $u = x - < \bar{x} >$, then the affine shape of x, or of u, is given by

$$\sigma(x) = \sigma(u) = \{Au : A \in GL(m, \mathbb{R})\}, \ (u \text{ centered } k-ad \text{ of rank } m). \tag{36}$$

The space of all such affine shapes is the *affine shape space* $A\Sigma_m^k$. Note that two Helmertized k-ads u and v (as $m \times (k-1)$ matrices of full rank) have the same shape if and only if the rows of u and v span the same m-dimensional subspace of $\mathbb{R}^{k-1}$. Hence we can identify $A\Sigma_m^k$ with the Grasmannian $G_m(k-1)$, namely, the set of all m-dimensional subspaces of $\mathbb{R}^{k-1}$ (Sparr [45]. For the Grassmann manifold, refer to

Boothby [16], pp. 63, 168, 362, 363). For extrinsic analysis on $A\Sigma_m^k \approx G_m(k-1)$, consider the embedding of $A\Sigma_m^k$ into $S(k-1,\mathbb{R})$ given by

$$J(\sigma(u)) = FF^t, \tag{37}$$

where $F = (f_1 \cdots f_m)$ is a $(k-1) \times m$ matrix and $\{f_1, \cdots, f_m\}$ is an orthonormal basis of the m-dimensional subspace L, say, of $\mathbb{R}^{k-1}$ spanned by the rows of u. Note that the $(k-1) \times (k-1)$matrix FF^t is idempotent and is the matrix of orthogonal projection of $\mathbb{R}^{k-1}$ onto L. It is independent of the orthonormal basis chosen. The embedding is $O(k-1)$-equivariant under the group action $\sigma(u) \to \sigma(uO)$ $(O \in O(k-1))$ on $A\Sigma_m^k$, with $O(k-1)$ acting on $S(k-1,\mathbb{R})$ by $A \to OAO^t$.

6.4 Projective Shape Space $P\Sigma_m^k$

First, recall that the real projective space $\mathbb{R}P^m$ is the space of all lines through the origin in $\mathbb{R}^{m+1}$. Its elements are $[p] = \{\lambda p : \lambda \in \mathbb{R}\backslash\{0\}\}$ for all $p \in \mathbb{R}^{m+1}\backslash\{o\}$. It is also conveniently represented as the quotient S^m/G where G is the two-point group $\{e, -e\}$, e being the identity map and $-ep = -p$ $(p \in S^m)$. That is, a line through p is identified with $\{p/|p|, -p/|p|\}$ ($p \in \mathbb{R}^{m+1}\backslash\{o\}$). As a consequence, there is a unique Riemannian metric tensor on $\mathbb{R}P^m = S^m/G$ such that $p \to \{p, -p\}$ is a Riemannian submersion, with $\langle u, v\rangle_{\mathbb{R}P^m} = u^t v$ for all vectors u, v in $T_{[p]}\mathbb{R}P^m$. The geodesic distance is given by $\rho_g([p],[q]) = \arccos(|p^t q|) \in [0, \pi/2]$, and the cut locus of $[p]$ is $C([p]) = \{[q] : \cos(|p^t q|) = \pi/2\}$, so that the injectivity radius of $\mathbb{R}P^m$ is $\pi/2$. Its sectional curvature is constant $+1$ (as it is of S^m). The exponential map of $T_{[p]}\mathbb{R}P^m$ (and its inverse on $\mathbb{R}P^m\backslash(C([p]))$ can be easily expressed in terms of those for the sphere S^m. We will use [ ] for both representations.

The so-called *Veronese-Whitney embedding* of $\mathbb{R}P^m$ into $S(m+1,\mathbb{R})$ is given by

$$J([p]) = pp^t, \ (p = (p_1, \cdots, p_{m+1})^t \in S^m). \tag{38}$$

It is clearly $O(m+1)$-equivariant, with the group action on $\mathbb{R}P^m$ as : $A[p] = [Ap]$ $(A \in O(m+1))$.

Turning to landmarks based projective shapes, assume $k > m+2$. *A frame of* $\mathbb{R}P^m$ is a set of $m+2$ ordered points $([p_1], \cdots, [p_{m+2}])$ such that every subset of $m+1$ of these points spans $\mathbb{R}P^m$, i.e., every subset of $m+1$ points of $\{p_1, \cdots, p_{m+2}\}$ spans $\mathbb{R}^{m+1}$. The *standard frame* of $\mathbb{R}P^m$ is $([e_1], [e_2], \cdots, [e_{m+1}], [e_1 + e_2 + \cdots + e_{m+1}])$, where e_i $(\in \mathbb{R}^{m+1})$ has 1 in the ith position and zeros elsewhere. A k-ad $y = (y_1, \cdots, y_k) = ([p_1], \cdots, [p_k]) \in (\mathbb{R}P^m)^k$ is in *general position* if there exist $i_1 < i_2 < \cdots < i_{m+2}$ such that $(y_{i_1}, \cdots, y_{i_{m+2}})$ is a frame of $\mathbb{R}P^m$. A *projective transformation* α on $\mathbb{R}P^m$ is defined by

$$\alpha[p] = [Ap], \ (p \in \mathbb{R}^{m+1}\backslash\{0\}) \tag{39}$$

where $A \in GL(m+1,\mathbb{R})$. The usual operation of matrix multiplication on $GL(m+1,\mathbb{R})$ then leads to a corresponding group of projective transformations on $\mathbb{R}P^m$. This is the *projective group* $PGL(m)$. Note that, for a given A in $GL(m+1,\mathbb{R})$, cA determines the same element of $PGL(m)$ for all $c \neq 0$. The *projective shape* of a k-ad $y = (y_1, \cdots, y_k) = ([p_1], \cdots, [p_k]) \in (\mathbb{R}P^m)^k$ in general position is its orbit under $PGL(m)$:

$$\sigma(y) = \{\alpha y \equiv (\alpha[p_1], \cdots, \alpha[p_k]) : \alpha \in PGL(m)\}, \quad (y = ([p_1], \cdots, [p_k] \text{ in general position}). \tag{40}$$

The *projective shape space* $PG\Sigma_m^k$ is the set of all projective shapes of k-ads in general position. Following Mardia and Patrangenaru [39] and Patrangenaru et al. [41], we will consider a particular dense open subset of $PG\Sigma_m^k$. Fix a set of $m+2$ indices $I = \{i_j : j = 1, \cdots, m+2\}$, $1 \leq i_1 < i_2 < \cdots < i_{m+2} \leq k$. Define $PG_I\Sigma_m^k$ as the set of shapes $\sigma(y)$ in $PG\Sigma_m^k$, $y = (y_1, \cdots, y_k) = ([p_1], \cdots, [p_k])$, such that every subset of $m+1$ points of $\{[p_{i_j}], j = 1, \cdots, m+2\}$ spans $\mathbb{R}P^m$.

The shape space $PG_I\Sigma_m^k$ (with $I = \{1, 2, \cdots, m+2\}$) may be identified with $(\mathbb{R}P^m)^{k-m-2}$ (see [39]). It has been shown in [12] that the full projective shape space $PG\Sigma_m^k$ in a differentiable manifold.

7 Inference on Shape Spaces

In this section we indicate how Theorems 2.5, 3.3, 4.1 and the inference procedures for general manifolds described in Sect. 5 may be applied to shape spaces, using the sphere S^d and the planar shape space Σ_2^k as illustrations.

For intrinsic analysis, consider the function $h(z, y) = \rho_g^2(Exp_p z, Exp_p y)$ for z, y in T_pM, with an appropriate choice of p. One first needs to express explicitly the quantities $D_r h(z, y)$, $D_r D_s h(z, y)$ in normal coordinates at p, i.e., at $z = 0 \equiv Exp_p^{-1} p$. (See Theorem 4.1.) For this let $\gamma(s)$ be a geodesic starting at p, and $m \in M$. Define the *parametric surface* $c(s,t) = Exp_m(tExp_m^{-1}\gamma(s))$, $s \in [0, \epsilon)$, $\epsilon > 0$ small. Note that $c(s, 0) = m$ for all s, $c(s, 1) = \gamma(s)$, and that, for all fixed $s \in [0, \epsilon)$, $t \to c(s,t)$ is a geodesic starting at m and reaching $\gamma(s)$ at $t = 1$. Writing $T(s,t) = (\partial/\partial t)c(s,t)$, $S(s,t) = (\partial/\partial s)c(s,t)$, one then has $S(s, 0) = 0$, $S(s, 1) = \dot{\gamma}(s)$. Also, $\langle T(s,t), T(s,t)\rangle$ does not depend on t and, therefore,

$$\rho_g^2(\gamma(s), m) = \int_0^1 \langle T(s,t), T(s,t)\rangle dt. \tag{41}$$

Differentiating this respect to s and recalling the symmetry $(D/\partial s)T(s,t) = (D/\partial t)S(s,t)$ on a parametric surface (see [17, p. 68, Lemma 3.4]), and $(D/\partial t)T(s,t) = 0$, one has

$$(d/ds)\rho_g^2(\gamma(s),m) = 2\int_0^1 \langle (D/\partial s)T(s,t), T(s,t)\rangle dt \tag{42}$$
$$= 2\int_0^1 \langle (D/\partial t)S(s,t), T(s,t)\rangle dt = 2\int_0^1 (d/dt)\langle S(s,t), T(s,t)\rangle dt$$
$$= 2\langle S(s,1), T(s,1)\rangle = -2\langle \dot{\gamma}(s), Exp_{\gamma(s)}^{-1} m\rangle.$$

Setting $s = 0$ in (42) and letting $\dot{\gamma}(0) = v_r$, with $\{v_r : r = 1, \cdots, d\}$ an orthonormal basis of T_pM, one shows that the normal coordinates y_r of m (i.e., the coordinates of $y = Exp_p^{-1}m$ with respect to $\{v_r : r = 1, \cdots, d\}$) satisfy

$$-2y^r \equiv -2\langle Exp_p^{-1}m, v_r\rangle = [(d/ds)\rho_g^2(\gamma(s),m)]_{s=0}. \tag{43}$$

From this one gets

$$D_r h(0,y) = -2y^r (r = 1, \cdots, d). \tag{44}$$

If $Q(C(p)) = 0$, then writing $\tilde{Q}$ for the distribution induced from Q by the map Exp_p^{-1} on T_pM, the Fréchet function may be expressed as

$$F(q) = \int \rho_g^2(q,m)Q(dm) = \int h(z,y)\tilde{Q}(dy) = \tilde{F}(z),\ (z = Exp_p^{-1}q). \tag{45}$$

Since a (local) minimum of this is attained at $q = \mu_I$, $\tilde{F}$ must satisfy a first order condition $D_r\tilde{F}(z) = 0$ at $z = \nu$. In particular, letting $p = \mu_I$ and, consequently, $\nu = 0$, one has $\int D_r h(0,y)\tilde{Q}(dy) = 0$, so that (44) yields

$$\int y^r \tilde{Q}(dy) = 0\ (r = 1, \cdots, d),\ (\tilde{Q} = Q \circ \phi^{-1}, \phi = Exp_{\mu_I}^{-1}). \tag{46}$$

Note that (44) and (46) are the relations stated in Theorem 4.1(b)(i),(ii).

By Theorem 4.1, the asymptotic distribution of the sample intrinsic mean μ_n is that of $\phi^{-1}(\nu_n)$, where $\phi = Exp_p^{-1}$, and (see (7))

$$\sqrt{n}(\nu_n - \nu) \simeq \Lambda^{-1}[(1/\sqrt{n})\sum_{1\le j\le n} Dh(\nu, Y_j)],\ (\Lambda_{rs} = ED_rD_sh(\nu, Y_1), 1 \le r,s \le d), \tag{47}$$

with $Y_j = \phi(X_j)$, where X_j are i.i.d. with distribution Q. By (44), the right side of (47) simplifies to $\Lambda^{-1}[-2(1/\sqrt{n})\sum_{1\le j\le n} Y_j]$, if $p = \mu_I$ (and $\nu = 0$).

Example 7.1 (Confidence region for the intrinsic/extrinsic mean of Q on the sphere S^d). Let μ_I be the intrinsic mean of Q on S^d. Given n i.i.d. observations $X_1, \cdots, X_n$ on S^d with common distribution Q, let μ_n be the intrinsic sample mean. Write $\phi = Exp_{\mu_I}^{-1}$, and $\phi_p = Exp_p^{-1}$, so that $\phi_{\mu_I} = \phi$. By Theorem 4.1,

$$\sqrt{n}[\phi(\mu_n) - \phi(\mu_I)) = \sqrt{n}\phi(\mu_n) \to N(0, \Lambda^{-1}\tilde{\Sigma}\Lambda^{-1}) \text{ in distribution as } n \to \infty, \tag{48}$$

where the $d \times d$ matrices Λ and $\tilde{\Sigma}$ are given by

$$\tilde{\Sigma} = 4Cov(\phi(X_1)), \tag{49}$$

$$\Lambda_{rs} = 2E[(1 - (X_1^t\mu_I)^2)^{-1}\{1 - (1 - (X_1^t\mu_I)^2)^{-1/2} \cdot (X_1^t\mu_I)\arccos(X_1^t\mu_I)\}(X_1^t v_r)(X_1^t v_s)$$
$$+ (1 - (X_1^t\mu_I)^2)^{-1/2} \cdot (X_1^t\mu_I)(\arccos(X_1^t\mu_I)))\delta_{rs}], 1 \le r, s \le d.$$

Here $\{v_r : 1 \le r \le d\}$ is an orthonormal basis of $T_{\mu_I}S^d$.
A confidence region for μ_I, of asymptotic level $1 - \alpha$, is then given by

$$\{p \in S^d : n\phi_p(\mu_n)^t\hat{\Lambda}_p\hat{\tilde{\Sigma}}_p^{-1}\hat{\Lambda}_p\phi_p(\mu_n) \le \chi_d^2(1-\alpha)\}, \tag{50}$$

where Λ_p, $\tilde{\Sigma}_p$ are obtained by replacing μ_I by p in the expressions for Λ and $\tilde{\Sigma}$ in (49). The 'hat' ($\hat{}$) indicates that the expectations are computed under the empirical Q_n, rather than Q. As mentioned in Sect. 5, it would be computationally simpler to choose a particular $p = p_0$, say, and let $\phi = Exp_{p_0}^{-1}$. Then (20) yields a simpler confidence region:

$$\{p \in S^d : n[\phi(\mu_n) - \phi(p)]^t\hat{\Lambda}_{p_0}\tilde{\Sigma}_{p_0}^{-1}\hat{\Lambda}_{p_0}[\phi(\mu_n)) - \phi(\mu_p)] \le \chi_d^2(1-\alpha)\}. \tag{51}$$

We now turn to the distribution of the extrinsic mean $\bar{X}/|\bar{X}|$. The $(d+1)\times(d+1)$ Jacobian matrix $(Jacob)_x P$ of the projection map $P : x \to x/|x|$, viewed as a map on $\mathbb{R}^{d+1}\backslash\{0\}$ into $\mathbb{R}^{d+1}$, is given by (13). Let $B(x)$ be the $d \times (d+1)$ matrix of the differential $d_x P$ (on $T_x\mathbb{R}^{d+1}$ into $T_{P(x)}S^d = \{u \in \mathbb{R}^{d+1} : P(x)^t u = 0\}$) whose d rows form an orthonormal basis of $T_{P(x)}S^d$. Then the differential of the projection map is

$$(d_x P)u = [B(x)(Jacob)_x P]u. \tag{52}$$

If $\mu = EX_1 \neq 0$, then, by (19), a confidence region for the extrinsic mean $\mu/|\mu|$ is given by

$$\{x/|x| \in S^d : n[(d_{\bar{X}}P)(\bar{X} - x)]^t(\hat{B}\hat{\Sigma}\hat{B}^t)^{-1}[(d_{\bar{X}}P)(\bar{X} - x)] \le \chi_d^2(1-\alpha)\}. \tag{53}$$

Here $\hat{B} = B(\bar{X})$, $\Sigma = Cov(X_1)$, and $\hat{\Sigma}$ is obtained by replacing Q by Q_n in computing expectations.

Example 7.2 (Inference on the planar shape space Σ_2^k). To apply Theorem 4.1, we use (47) where $\phi = Exp_{\sigma(p)}^{-1}$ and p is a suitable point in $\mathbb{C}S^{k-1}$. To derive a computable expression for Λ, write the geodesic γ in the parametric surface $c(s,t)$ as $\gamma = \pi \circ \tilde{\gamma}$, where $\tilde{\gamma}$ is a geodesic in $\mathbb{C}S^{k-1}$ starting at $\tilde{\mu} \in \pi^{-1}\{\mu_I\}$. Then, with $\tilde{T}(s,1) = (d_{\gamma(s)}\pi^{-1})T(s,1)$,

$$(d/ds)\rho_g^2(\gamma(s),m) = 2 < T(s,1), \dot{\gamma}(s) >= 2 < \tilde{T}(s,1), \dot{\tilde{\gamma}}(s) >, \tag{54}$$
$$(d^2/ds^2)\rho_g^2(\gamma(s),m) = 2 < D_s\tilde{T}(s,1), \dot{\tilde{\gamma}}(s) > .$$

The final inner products are in $T\mathbb{C}S^{k-1}$, namely, $\langle \tilde{v}, \tilde{w} \rangle = Re(\tilde{v}\tilde{w}^*)$. Note that $\tilde{T}(s,1) = -Exp^{-1}_{\tilde{\gamma}(s)} q,\ q \in \pi^{-1}m$, may be expressed by (30) and (31) as

$$\tilde{T}(s,1) = -(\rho(s)/\sin\rho(s))[e^{i\theta(s)}q - (\cos\rho(s))\tilde{\gamma}(s)], \tag{55}$$

where $\rho(s) = \rho_g(\gamma(s),m)$ and $e^{i\theta(s)} = (1/\cos\rho(s))\tilde{\gamma}(s)q^*$. The covariant derivative $D_s\tilde{T}(s,1)$ is the projection of $(d/ds)\tilde{T}(s,1)$ onto $H_{\tilde{\gamma}(s)}$. Since $\langle \tilde{\mu}, \dot{\tilde{\gamma}}(0)\rangle = 0$, (54) then yields

$$[(d^2/ds^2)\rho_g^2(\gamma(s),m)]_{s=0} = 2\langle [(d/ds)\tilde{T}(s,1)]_{s=0}, \dot{\tilde{\gamma}}(0)\rangle. \tag{56}$$

Differentiating (55) one obtains

$$[(d/ds)\tilde{T}(s,1)]_{s=0} = [(d/ds)(\rho(s)\cos\rho(s))/\sin\rho(s))]_{s=0}\tilde{\mu} \tag{57}$$
$$+ [(\rho(s)cos\rho(s))/sin\rho(s))]_{s=0}\dot{\tilde{\gamma}}(0) - [(d/ds)(\rho(s)/(\cos\rho(s))(\sin\rho(s))]_{s=0}(\tilde{\mu}q^*)q$$
$$- [\rho(s)/(\cos\rho(s))(\sin\rho(s))]_{s=0}(\dot{\tilde{\gamma}}(0)q*)q.$$

From (54), $2\rho(s)\dot{\rho}(s) = 2\langle \tilde{T}(s,1), \tilde{\gamma}'(s)\rangle$, which along with (55) leads to

$$[(d/ds)\rho(s)]_{s=0} = -(1/\sin r)\langle(\tilde{\mu}q^*/\cos r)q, \dot{\tilde{\gamma}}(0)\rangle,\ (r = \rho_g(m,\mu_I)). \tag{58}$$

One then gets (see BB [5,6])

$$\langle [(d/ds)\tilde{T}(s,1)]_{s=0}, \dot{\tilde{\gamma}}(0)\rangle = \{(r\cos r)/(\sin r)\}|\dot{\tilde{\gamma}}(0)|^2 \tag{59}$$
$$- \{(1/\sin^2 r) - (r\cos r)/\sin^3 r\}(Re(x))^2 + r/((\sin r)(\cos r))(Im(x))^2,$$
$$(x = e^{i\theta}q\dot{\tilde{\gamma}}(0)^*, e^{i\theta} = \tilde{\mu}q^*/\cos r).$$

One can check that the right side of (59) depends only on $\pi(\tilde{\mu})$ and not any particular choice of $\tilde{\mu}$ in $\pi^{-1}\{\mu_I\}$.

Now let $\{v_1, \cdots, v_{k-2}, iv_1, \cdots, iv_{k-2}\}$ be an orthonormal basis of $T_{\sigma(p)}\Sigma_2^k$ where we identify Σ_2^k with $\mathbb{C}P^{k-2}$, and choose the unit vectors $v_r = (v_r^1, \cdots, v_r^{k-1})$, $r = 1, \cdots, k-2$, to have zero imaginary parts and satisfy the conditions $p^*v_r = 0$, $v_r^t v_s = 0$ for $r \neq s$.

Suppose now that $\sigma(p) = \mu_I$, i.e., $\gamma(0) = \mu_I$. If $\dot{\gamma}(0) = v$, then $\gamma(s) = Exp_{\mu_I}(sv)$, so that $\rho_g^2(\gamma(s),m) = h(sv,y)$ with $y = Exp^{-1}_{\mu_I}m$. Then, expressing v in terms of the orthonormal basis,

$$[(d^2/ds^2)\rho_g^2(\gamma(s),m)]_{s=0} = [(d^2/ds^2)h(sv,y)]_{s=0} = \Sigma v_i v_j D_i D_j h(0,y). \tag{60}$$

Integrating with respect to Q now yields

$$\sum v_i v_j \Lambda_{ij} = E[(d^2/ds^2)\rho_g^2(\gamma(s), X)]_{s=0}, \quad (X \text{ with distribution } Q). \tag{61}$$

This identifies the matrix Λ from the calculations (56) and (59). To be specific, consider independent observations $X_1, \cdots, X_n$ from Q, and let $Y_j = Exp_{\mu_I}^{-1} X_j (j = 1, \cdots, n)$. In normal coordinates with respect to the above basis of $T_{\mu_I}\Sigma_2^k$, one has the following coordinates of Y_j:

$$(Re(Y_j^1), \cdots, Re(Y_j^{k-2}), Im(Y_j^1), \cdots, Im(Y_j^{k-2})) \in \mathbb{R}^{2k-4}. \tag{62}$$

Writing

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

in blocks of $(k-2) \times (k-2)$ matrices, one arrives at the following expressions of the elements of these matrices, using (59)–(62). Denote $\rho_g^2(\mu_I, X_1) = h(0, Y_1)$ by ρ. Then

$$\begin{aligned}
(\Lambda_{11})_{rs} &= 2E[\rho(\cot\rho)\delta_{rs} - (1/\rho^2)(1 - \rho\cot\rho)(ReY_1^r)(ReY_1^s) \\
&\quad + \rho^{-1}(\tan\rho)(ImY_1^r)(ImY_1^s)]; \\
(\Lambda_{22})_{rs} &= 2E[\rho(\cot\rho)\delta_{rs} - (1/\rho^2)(1 - \rho\cot\rho)(ImY_1^r)(ImY_1^s) \\
&\quad + \rho^{-1}(\tan\rho)(ReY_1^r)(ReY_1^s)]; \\
(\Lambda_{12})_{rs} &= 2E[\rho(\cot\rho)\delta_{rs} - (1/\rho^2)(1 - \rho\cot\rho)(ReY_1^r)(ImY_1^s) \\
&\quad + \rho^{-1}(\tan\rho)(ImY_1^r)(ReY_1^s)]; \\
(\Lambda_{21})_{rs} &= (\Lambda_{12})_{sr}.(r, s = 1, \cdots, k-2).
\end{aligned} \tag{63}$$

One now arrives at the CLT for the intrinsic sample mean μ_n by Theorem 4.1, or the relation (20). A two-sample test for $H_0 : Q_1 = Q_2$, is then provided by (30).

We next turn to extrinsic analysis on Σ_2^k, using the embedding (34). Let μ^J be the mean of $Q \circ J^{-1}$ on $S(k-1, \mathbb{C})$. To compute the projection $P(\mu^J)$, let T be a unitary matrix, $T \in SU(k-1)$ such that $T\mu^J T^* = D = diag(\lambda_1, \cdots, \lambda_{k-1})$, $\lambda_1 \leq \cdots \leq \lambda_{k-2} \leq \lambda_{k-1}$. For $u \in \mathbb{C}S^{k-1}$, $u^*u \in J(\Sigma_2^k)$, write $v = Tu^*$. Then $Tu^*uT^* = vv^*$, and

$$\begin{aligned}
||u^*u - \mu_J||^2 = ||vv^* - D||^2 &= \sum_{i,j} |v_i v_j - \lambda_j \delta_{ij}|^2 \\
&= \sum_j (|v_j|^2 + \lambda_j^2 - 2\lambda_j |v_j|^2) \\
&= \sum_j \lambda_j^2 + 1 - 2\sum_j \lambda_j |v_j|^2,
\end{aligned} \tag{64}$$

which is minimized on $J(\Sigma_2^k)$ by $v = (v^1, \cdots, v^{k-1})$ for which $v^j = 0$ for $j = 1, \cdots, k-2$, and $|v^{k-1}| = 1$. That is, the minimizing u^* in (64) is a unit eigenvector of μ^J with the largest eigenvalue λ_{k-1}, and $P(\mu^J) = u^*u$. This projection is unique if and only if the largest eigenvalue of μ^J is simple, i.e., $\lambda_{k-2} < \lambda_{k-1}$.

Assuming that the largest eigenvalue of μ^J is simple, one may now obtain the asymptotic distribution of the sample extrinsic mean $\mu_{n,E}$, namely, that of $J(\mu_{n,E}) = v_n^* v_n$, where v_n is a unit eigenvector of $\bar{\tilde{X}} = \sum \tilde{X}_j/n$ corresponding to its largest eigenvalue. Here $\tilde{X}_j = J(X_j)$, for i.i.d observations $X_1, \cdots, X_n$ on Σ_2^k. For this purpose, a convenient orthonormal basis (frame) of $T_p S(k-1, \mathbb{C}) \approx S(k-1, \mathbb{C})$ is the following:

$$v_{a,b} = 2^{-1/2}(e_a e_b^t + e_b e_a^t) \text{ for } a < b, v_{a,a} = e_a e_a^t; \tag{65}$$

$$w_{a,b} = i2^{-1/2}(e_a e_b^t - e_b e_a^t) \text{ for } b < a \ (a, b = 1, \cdots, k-1),$$

where e_a is the column vector with all entries zero other than the a-th, and the a-th entry is 1. Let $U_1, \cdots, U_{k-1}$ be orthonormal unit eigenvectors corresponding to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_{k-2} < \lambda_{k-1}$. Then choosing $T = (U_1, \cdots, U_{k-1}) \in SU(k-1)$ $T\mu^J T^* = D = diag(\lambda_1, \cdots, \lambda_{k-1})$, such that the columns of $Tv_{a,b}T^*$ and $Tw_{a,b}T^*$ together constitute an orthonormal basis of $S(k-1, \mathbb{C})$. It is not difficult to check that the differential of the projection operator P satisfies

$$(d_{\mu^J} P)Tv_{a,b}T^* = \begin{cases} 0 & \text{if } 1 \leq a \leq b < k-1, \text{ or } a = b = k-1, \\ (\lambda_{k-1} - \lambda_a)^{-1} Tv_{a,k-1}T^* & \text{if } 1 \leq a < k-1, b = k-1; \end{cases} \tag{66}$$

$$(d_{\mu^J} P)Tw_{a,b}T^* = \begin{cases} 0 & \text{if } 1 \leq a \leq b < k-1, \\ (\lambda_{k-1} - \lambda_a)^{-1} Tw_{a,k-1}T^* & \text{if } 1 \leq a < k-1. \end{cases}$$

To check these, take the projection of a linear curve $c(s)$ in $S(k-1, \mathbb{C})$ such that $\dot{c}(0)$ is one of the basis elements $v_{a,b}$, or $w_{a,b}$, and differentiate the projected curve with respect to s. It follows that $\{Tv_{a,k-1}T^*, Tw_{a,k-1}T^* : a = 1, \cdots, k-2\}$ form an orthonotmal basis of $T_{P(\mu^J)}J(\Sigma_2^k)$. Expressing $\tilde{X}_j - \mu^J$ in the orthonormal basis of $S(k-1, \mathbb{C})$, and $d_{\mu^J} P(\tilde{X}_j - \mu^J)$ with respect to the above basis of $T_{P(\mu^J)}J(\Sigma_2^k)$, one may now apply Theorem 3.3.

For a two-sample test for $H_0 : Q_1 = Q_2$, one may use (22), as explained in Sect. 5.

8 Nonparameric Bayes for Density Estimation and Classification on a Manifold

8.1 Density Estimation

Consider the problem of estimating the density q of a distribution Q on a Riemannian manifold (M, g) with respect to the volume measure λ on M. According to Ferguson [22], given a finite non-zero base measure α on a measurable space $(\mathcal{X}, \Sigma)$, a random probability P on the class $\mathcal{P}$ of all probability measures on $\mathcal{X}$ has the Dirichlet distribution D_α if for every measurable partition $\{B_1, \ldots, B_k\}$ of $\mathcal{X}$, the D_α - distribution of $(P(B_1), \ldots, P(B_k)) = (\theta_1, \ldots, \theta_k)$, say, is Dirichlet with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. Sethuraman [44] gave a very convenient "stick breaking" representation of the random P. To define it, let $u_j (j = 1, \ldots)$ be an i.i.d. sequence of $beta(1, \alpha(\mathcal{X}))$ random variables, independent of a sequence $Y_j (j = 1, \ldots)$ having the distribution $G = \frac{\alpha}{\alpha(\mathcal{X})}$ on $\mathcal{X}$. Sethuraman's representation of the random probability with the Dirichlet prior distribution D_α is

$$P \equiv \sum w_j \delta_{Y_j}, \tag{67}$$

where $w_1 = u_1, w_j = u_j(1 - u_1) \ldots (1 - u_{j-1})(j = 2, \ldots)$, and δ_{Y_j} denotes the Dirac measure at Y_j. As this construction shows, the Dirichlet distribution assigns probability one to the set of all discrete distributions on $\mathcal{X}$, and one cannot retrieve a density estimate from it directly. The Dirichlet priors constitute a conjugate family, i.e., the posterior distribution of a random P with distribution D_α, given observations $X_1, \ldots, X_n$ from P is $D_{\alpha + \sum_{1 \le i \le n} \delta_{X_i}}$. A general method for Bayesian density estimation on a manifold (M, g) may be outlined as follows. Suppose that q is continuous and positive on M. First find a parametric family of densities $m \to K(m; \mu, \tau)$ on M where $\mu \in M$ and $\tau > 0$ are "location" and "scale" parameters, such that K is continuous in its arguments, $K(\cdot; \mu, \tau) d\lambda(\cdot)$ converges to δ_μ as $\tau \downarrow 0$, and the set of all "mixtures" of $K(\cdot; \mu, \tau)$ by distributions on $M \times (0, \infty)$ is dense in the set $C_\lambda(M)$ of all continuous densities on M in the supremum distance, or in $L^1(d\lambda)$. The density q may then be estimated by a suitable mixture. To estimate the mixture, use a prior D_β with full support on the set of all probabilities on the space $M \times (0, \infty)$ of "parameters" (μ, τ). A draw from the prior may be expressed in the form (67), where u_j are i.i.d. $beta(1, b)$ with $b = \beta(M \times (0, \infty))$, independent of $Y_j = (m_j, t_j)$, say, which are i.i.d. $\frac{\beta}{b}$ on $M \times (0, \infty)$. The corresponding random density is then obtained by integrating the kernel K with respect to this random mixture distribution,

$$\sum w_j K(m; m_j, t_j). \tag{68}$$

Given M-valued (Q-distributed) observations $X_1, \ldots X_n$, the posterior distribution of the mixture measure is Dirichlet D_{β_X}, where $\beta_X = \beta + \sum_{1 \le i \le n} \delta_{Z_i}$, with $Z_i = (X_i, 0)$. A draw from the posterior distribution leads to the random density in the form (68), where u_j are i.i.d. $beta(1, b+n)$, independent of (m_j, t_j) which are i.i.d. $\frac{\beta_X}{(b+n)}$. One may also consider using a somewhat different type of priors such as $D_\alpha \times \pi$ where D_α is a Dirichlet prior on M, and π is a prior on $(0, \infty)$, e.g., gamma or Weibull distribution.

Consistency of the posterior is generally established by checking full Kullback-Liebler support of the prior D_β (see [25], pp. 137–139). Strong consistency has been established for the planar shape spaces using the complex Watson family of densities (with respect to the volume measure or the uniform distribution on Σ_2^k) of the form $K([z]; \mu, \tau) = c(\tau) exp \frac{|z * \mu|^2}{\tau}$ in [6, 7], where it has been shown, by simulation from known distributions, that, based on a prior $D_\beta \times \pi$ chosen so as to produce clusters close to the support of the observations, the Bayes estimates of quantiles and other indices far outperform the kernel density estimates of Pelletier [42], and also require much less computational time than the latter. In moderate sample sizes, the nonparametric Bayes estimates perform much better than even the MLE (computed under the true model specification)!

8.2 Classification

Classification of a random observation to one of several groups is one of the most important problems in statistics. This is the objective in medical diagnostics, classification of subspecies and, more generally, this is the target of most image analysis. Suppose there are r groups or populations with a priori given relative sizes or proportions $\pi_i (i = 1, \ldots, r), \sum \pi_i = 1$, and densities $q_i(x)$ (with respect to some sigma-finite measure). Under $0-1$ loss function, the average risk of misclassification (i.e., the Bayes risk) is minimized by the rule: Given a random observation X, classify it to belong to group j if

$\pi_j q_j(X) = max\{\pi_i q_i(X) : i = 1, \ldots, r\}$. Generally, one uses sample estimates of π_i-s and q_i-s, based on random samples from the r groups (training data). Nonparametric Bayes estimates of q_i-s on shapes spaces perform very well in classification of shapes, and occasionally identify outliers and misclassified observations (see, [6, 7]).

9 Examples

In this section we apply the theory to a number of data sets available in the literature.

Example 9.1 (Paleomagnetism). The first statistical confirmation of the shifting of the earth's magnetic poles over geological times, theorized by paleontologists

based on observed fossilised magnetic rock samples, came in a seminal paper by R.A. Fisher [23]. Fisher analyzed two sets of data—one recent (1947–1948) and another old (Quaternary period), using the so-called *von Mises-Fisher model*

$$f(x;\mu,\tau) = c(\tau)\exp\{\tau x'\mu\}(x \in S^2), \tag{69}$$

Here $\mu(\in S^2)$, is the *mean direction*, extrinsic as well as intrinsic ($\mu = \mu_I = \mu_E$), and $\tau > 0$ is the concentration parameter. The maximum likelihood estimate of μ is $\hat{\mu} = \bar{X}/|\bar{X}|$, which is also our sample extrinsic mean. The value of the MLE for the first data set of $n = 9$ observations turned out to be $\hat{\mu} = \hat{\mu}_E = (.2984, .1346, .9449)$, where (0,0,1) is the geographic north pole. Fisher's 95% confidence region for μ is $\{\mu \in S^2 : \rho_g(\hat{\mu}, \mu) \leq 0.1536)\}$. The sample intrinsic mean is $\hat{\mu}_I = (.2990, .1349, .9447)$, which is very close to $\hat{\mu}_E$.The nonparametric confidence region based on $\hat{\mu}_I$, as given by (50), and that based on the extrinsic procedure (53), are nearly the same, and both are about 10% smaller in area than Fisher's region. (See [6], Chap. 2.)

The second data set based on $n = 29$ observations from the Quaternary period that Fisher analyzed, using the same parametric model as above, had the MLE $\hat{\mu} = \bar{X}/|\bar{X}| = (.0172, -.2978, -.9545)$, almost antipodal of that for the first data set, and with a confidence region of geodesic radius .1475 around the MLE. Note that the two confidence regions are not only disjoint, they also lie far away from each other. This provided the first statistical confirmation of the hypothesis of shifts in the earth's magnetic poles, a result hailed by paleontologists (see [30]). Because of difficulty in accessing the second data set, the nonparametric procedures could not be applied to it. But the analysis of another data set dating from the Jurassic period, with $n = 33$, once again yielded nonparametric intrinsic and extrinsic confidence regions very close to each other, and each about 10 % smaller than the region obtained by Fisher's parametric method (see [6], Chap. 5, for details).

Example 9.2 (Brain scan of schizophrenic and normal patients). We consider an example from Bookstein [15] in which 13 landmarks were recorded on a midsagittal two-dimensional slice from magnetic brain scans of each of 14 schizophrenic patients and 14 normal patients. The object is to detect the deformation, if any, in the shape of the k-ad due to the disease, and to use it for diagnostic purposes. The shape space is Σ_2^{13}. The intrinsic two-sample test (22) has an observed value 95.4587 of the asymptotic chisquare statistic with 22 degrees of freedom, and a p-value 3.97×10^{-11}. The extrinsic test based on (24) has an observed value 95.5476 of the chisquare statistic and a p-value 3.8×10^{-11}. The calculations made use of the analytical computations carried out in Example 7.2. It is remarkable, and reassuring, that completely different methodologies of intrinsic and extrinsic inference essentially led to the same values of the corresponding asymptotic chisquare statistics (a phenomenon observed in other examples as well). For details of these calculations and others we refer to [6]. This may also be contrasted with the results of parametric inference in the literature for the same data, as may be found in [19], pp. 146, 162–165. Using a isotropic Normal model for the original landmarks

data, and after removal of "nuisance" parameters for translation, size and rotation, an F-test known as Goodall's F-test (see [26]) gives a p-value .01. A Monte Carlo test based permutation test obtained by 999 random assignments of the data into two groups and computing Goodall's F-statistic, gave a p-value .04. A Hotelling's T^2 test in the tangent space of the pooled sample mean had a p-value .834. A likelihood ratio test based on the isotropic offset Normal distribution on the shape space has the value 43.124 of the chisquare statistic with 22 degrees of freedom, and a p-value .005.

Example 9.3 (Glaucoma detection- a match pair problem in $3D$*).* Our final example is on the $3D$ reflection similarity shape space $R\Sigma_3^k$. To detect shape changes due to glaucoma, data were collected on twelve mature rhesus monkeys.

One of the eyes of each monkey was treated with a chemical agent to temporarily increase the intraocular pressure (IOP). The increase in IOP is known to be a cause of glaucoma. The other eye was left untreated. Measurements were made of five landmarks in each eye, suggested by medical professionals. The data may be found in [11]. The match pair test based on (25) yielded an observed value 36.29 of the asymptotic chisquare statistic with degrees of freedom 8. The corresponding p-value is 1.55×10^{-5} (see [6], Chap. 9). This provides a strong justification for using shape change of the inner eye as a diagnostic tool to detect the onset of glaucoma. An earlier computation using a different nonparametric procedure in [11] provided a p-value .058. Also see [9] where a 95 % confidence region is obtained for the difference between the extrinsic size-and-shape reelection shapes between the treated and untreated eyes.

Acknowledgements The author wishes to thank the referee for helpful suggestions. This research is supported by NSF grant DMS 1107053.

References

1. B. Afsari, Riemannian L^p center of mass: Existence, uniqueness, and convexity. Proc. Am. Math. Soc. **139**, 655–673 (2011)
2. A. Bhattacharya, Nonparametric statistics on manifolds with applications to shape spaces. Ph.D. Thesis, University of Arizona (2008)
3. A. Bhattacharya, Statistical analysis on manifolds: a nonparametric approach for inference on shape spaces. Sankhya, **70**, 43 (2008)
4. A. Bhattacharya, R.N. Bhattacharya, Nonparametric statistics on manifolds with application to shape spaces. Pushing the Limits of Contemporary Statistics: Contributions in honor of J.K. Ghosh. IMS Collections **3**, 282–301 (2008)
5. A. Bhattacharya, R.N. Bhattacharya, Statistics on Riemannian manifolds: asymptotic distribution and curvature. Proc. Am. Math. Soc. **136**, 2957–2967 (2008)
6. A. Bhattacharya, R.N. Bhattacharya, *Nonparametric Inference on Manifolds with Applications to Shape Spaces.* IMS Monograph, No. 2 (Cambridge University Press, Cambridge, 2012)
7. A. Bhattacharya, D. Dunson, Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. Biometrika **97**, 851–865 (2010)

8. A. Bandulasiri, V. Patrangenaru, Algorithms for nonparametric inference on shape manifolds, in *Proceedings of JSM 2005, MN* (2005), pp. 1617–1622
9. A. Bandulasiri, R.N. Bhattacharya, V. Patrangenaru, Nonparametric inference on shape manifolds with applications in medical imaging. J. Multivariate Anal. **100**, 1867–1882 (2009)
10. R.N. Bhattacharya, V. Patrangenaru, Large sample theory of intrinsic and extrinsic sample means on manifolds. Ann. Stat. **31**, 1–29 (2003)
11. R.N. Bhattacharya, V. Patrangenaru, Large sample theory of intrinsic and extrinsic sample means on manifolds-II. Ann. Stat. **33**, 1225–1259 (2005)
12. R.N. Bhattacharya, J. Ramirez, B. Polletta, Manifold structure of the projective shape space. (Unpublished, 2009)
13. F. Bookstein, *The Measurement of Biological Shape and Shape Change*. Lecture Notes in Biomathematics (Springer, Berlin, 1978)
14. F.L. Bookstein, Size and shape spaces of landmark data (with discussion). Stat. Sci. **1**, 181–242 (1986)
15. F.L. Bookstein, *Morphometric Tools for Landmark Data: Geometry and Biology* (Cambridge University Press, Cambridge, 1991)
16. W.M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd edn. (Academic, New York, 1986)
17. M. Do Carmo, *Riemannian Geometry* (Birkhäuser, Boston, 1992)
18. I.L. Dryden, K.V. Mardia, Size and shape analysis of landmark data. Biometrica **79**, 57–68 (1992)
19. I.L. Dryden, K.V. Mardia, *Statistical Shape Analysis* (Wiley, New York, 1998)
20. I.L. Dryden, A. Kume, H. Le, A.T.A. Wood, The MDS model for shape: an alternative approach. *Biometrika* **95**(4), 779–798 (2008)
21. L. Ellingson, F.H. Ruymgaart, V. Patrangenaru, Nonparametric estimation for extrinsic mean shapes of planar contours. (to appear, 2013)
22. T. Ferguson, A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
23. R.A. Fisher, Dispersion on a sphere. Proc. R. Soc. Lond. Ser. A **217**, 295–305 (1953)
24. S. Gallot, D. Hulin, J. Lafontaine, *Riemannian Geometry*. Universitext (Springer, Berlin, 1990)
25. J.K. Ghosh, R.V. Ramamoorthi, *Bayesian Nonparametrics* (Springer, New York, 2003)
26. C.R. Goodall, Procrustes methods in the statistical analysis of shape (with discussion). J. R. Stat. Soc. Ser B. **53**, 285–339 (1991)
27. H. Hendriks, Z. Landsman, Asymptotic tests for mean location on manifolds. C.R. Acad. Sci. Paris Sr. I Math. **322**, 773–778 (1996)
28. H. Hendriks, Z. Landsman, Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions. J. Multivariate Anal. **67**, 227–243 (1998)
29. S. Huckemann, T. Hotz, A. Munk, Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions (with discussions). Stat. Sinica **20**, 1–100 (2010)
30. E. Irving, *Paleomagnetism and Its Application to Geological and Geographical Problems* (Wiley, New York, 1964)
31. H. Karcher, Riemannian center of mass and mollifier smoothing. Comm. Pure Appl. Math. **30**, 509–554 (1977)
32. D.G. Kendall, The diffusion of shape. Adv. Appl. Probab. **9**, 428–430 (1977)
33. D.G. Kendall, Shape manifolds, procrustean metrics, and complex projective spaces. Bull. Lond. Math. Soc. **16**, 81–121 (1984)
34. D.G. Kendall, A survey of the statistical theory of shape. Stat. Sci. **4**, 87–120 (1989)
35. W.S. Kendall, Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. Proc. Lond. Math. Soc **61**, 371–406 (1990)
36. J.T. Kent, New directions in shape analysis. in *The Art of Statistical Science*, ed. by K.V. Mardia. pp. 115–127 (Wiley, New York, 1992)
37. J.T. Kent, The complex Bingham distribution and shape analysis. J. Roy. Stat. Soc. Ser. B **56**, 285–299 (1994)

38. H. Le, Locating Fréchet means with application to shape spaces. Adv. Appl. Prob. **33**, 324–338 (2001)
39. K.V. Mardia, V. Patrangenaru, Directions and projective shapes. Ann. Stat. **33**, 1666–1699 (2005)
40. V. Patrangenaru, Asymptotic statistics on manifolds and their applications. Ph.D. Thesis, Indiana University, Bloomington (1998)
41. V. Patrangenaru, X. Liu, S. Sugathadasa, Nonparametric 3D projective shape estimation from pairs of 2D images-I, in memory of W.P.Dayawansa. J. Multivariate Anal. **101**, 11–31 (2010)
42. B. Pelletier, Kernel density estimation on Riemannian manifolds. Stat. Probab. Lett. **73**(3), 297–304 (2005)
43. M.J. Prentice, K.V. Mardia, Shape changes in the plane for landmark data. Ann. Stat. **23**, 1960–1974 (1995)
44. J. Sethuraman, A constructive definition of Dirichlet priors. Stat. Sinica **4**, 639–650 (1994)
45. G. Sparr, Depth computations from polyhedral images, in *Proceedings of 2nd European Conference on Computer Vision*, ed. by G. Sandini (Springer, New York, 1994), pp. 378–386
46. H. Ziezold, On expected figures and a strong law of large numbers for random elements in quasi-metric spaces, in *Transactions of the Seventh Pragure Conference on Information Theory, Statistical Functions, Random Processes and of the Eighth European Meeting of Statisticians*, vol. A. (Tech. Univ. Prague, Prague (1974), pp. 591–602, Reidel, Dordrecht, 1977

Proportion of Gaps and Fluctuations of the Optimal Score in Random Sequence Comparison

Jüri Lember, Heinrich Matzinger, and Felipe Torres

Abstract We study the asymptotic properties of optimal alignments when aligning two independent i.i.d. sequences over finite alphabet. Such kind of alignment is an important tool in many fields of applications including computational molecular biology. We are particularly interested in the (asymptotic) proportion of gaps of the optimal alignment. We show that when the limit of the average optimal score per letter (rescaled score) is considered as a function of the gap penalty, then given a gap penalty, the proportion of the gaps converges to the derivative of the limit score at that particular penalty. Such an approach, where the gap penalty is allowed to vary, has not been explored before. As an application, we solve the long open problem of the fluctuation of the optimal alignment in the case when the gap penalty is sufficiently large. In particular, we prove that for all scoring functions without a certain symmetry, as long as the gap penalty is large enough, the fluctuations of the optimal alignment score are of order square root of the length of the strings. This order was conjectured by Waterman [Phil. Trans. R. Soc. Lond. B 344(1):383–390, 1994] but disproves the conjecture of Chvatal and Sankoff in [J. Appl. Probab. 12:306–315, 1975].

Keywords Fluctuations • Longest common sequence • McDiarmid's inequality • Random sequence comparison • Waterman conjecture

J. Lember
Tartu University, Institute of Mathematical Statistics, Liivi 2-513 50409, Tartu, Estonia
e-mail: jyril@ut.ee

H. Matzinger (✉)
Georgia Tech, School of Mathematics, Atlanta, GA 30332-0160, USA
e-mail: matzing@math.gatech.edu

F. Torres
Münster University, Institute for Mathematical Statistics, Einsteinstraße 62, 48149 - Münster, Germany
e-mail: ftorrestapia@math.uni-muenster.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_10,

AMS. 60K35, 41A25, 60C05

1 Introduction

1.1 Preliminaries

Throughout this paper $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are two independent sequences of i.i.d. random variables drawn from a finite alphabet $\mathbb{A}$ and having the same distribution. Since we mostly study the finite strings of length n, let $X = (X_1, X_2, \ldots X_n)$ and let $Y = (Y_1, Y_2, \ldots Y_n)$ be the corresponding n-dimensional random vectors. We shall usually refer to X and Y as random sequences.

The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [4, 7, 18, 20, 24] and computational linguistics [13, 16, 17, 25]. In this paper we adopt the same notation as in [11], namely we consider a general scoring scheme, where $S : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ is a *pairwise scoring function* that assigns a score to each couple of letters from $\mathbb{A}$. We assume S to be symmetric, non-constant and we denote by F and E the largest and the second largest possible score, respectively. Formally (recall that S is symmetric and non-constant)

$$F := \max_{(a,b)\in\mathbb{A}\times\mathbb{A}} S(a,b), \quad E := \max_{(a,b):S(a,b)\neq F} S(a,b).$$

An *alignment* is a pair (π, μ) where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ are two increasing sequences of natural numbers, i.e. $1 \leq \pi_1 < \pi_2 < \ldots < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq n$. The integer k is the number of aligned letters, $n - k$ is the number of *gaps* in the alignment and the number

$$q(\pi, \mu) := \frac{n-k}{n} \in [0, 1]$$

is the *proportion of gaps of the alignment* (π, μ). The *average score of aligned letters* is defined by

$$t(\pi, \mu) := \frac{1}{k} \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}).$$

Note that our definition of gap slightly differs from the one that is commonly used in the sequence alignment literature, where a gap consists of maximal number of consecutive *indels* (insertion and deletion) in one side. Our gap actually corresponds to a pair of indels, one in X-side and another in Y-side. Since we consider the sequences of equal length, to every indel in X-side corresponds an indel in Y-side, so considering them pairwise is justified. In other words, the number of gaps in our sense is the number of indels in one sequence. We also consider a *gap price* δ.

Given the pairwise scoring function S and the gap price δ, the score of the alignment (π, μ) when aligning X and Y is defined by

$$U^{\delta}_{(\pi,\mu)}(X, Y) := \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \delta(n - k)$$

which can be written down as the convex combination

$$U^{\delta}_{(\pi,\mu)}(X, Y) = n\left(t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu)\right). \tag{1}$$

In our general scoring scheme δ can also be positive, although usually $\delta \leq 0$ penalizing the mismatch. For negative δ, the quantity $-\delta$ is usually called the *gap penalty*. We naturally assume $\delta \leq F$. The optimal alignment score of X and Y is defined to be

$$L_n(\delta) := \max_{(\pi,\mu)} U^{\delta}_{(\pi,\mu)}(X, Y),$$

where the maximum above is taken over all possible alignments. The alignments achieving the maximum are called *optimal*. For every $\delta \in \mathbb{R}$, let us denote

$$B_n(\delta) := \frac{L_n(\delta)}{n}. \tag{2}$$

Note that to every alignment (π, μ) corresponds an unique pair $(t(\pi, \mu), q(\pi, \mu))$, but different alignments can have the same $t(\pi, \mu)$ and $q(\pi, \mu)$, thus from (1) we get that

$$B_n(\delta) = \max_{(\pi,\mu)} \big(t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu)\big) = \max_{(t,q)} \big(t(1 - q) + \delta q\big), \tag{3}$$

where in the right hand side the maximum is taken over all possible pairs (t, q) corresponding to an alignment of X and Y. In the following, we identify alignments with pairs (t, q), so a pair (t, q) always corresponds to an alignment (π, μ) of X and Y. Let $\mathcal{O}_n(\delta)$ denote the set of optimal pairs, i.e. $(t, q) \in \mathcal{O}_n(\delta)$ if and only if $t(1 - q) + \delta q = B_n(\delta)$. Note that the set $\mathcal{O}_n(\delta)$ is not necessarily a singleton. Let us denote

$$\underline{q}_n(\delta) := \min\{q : (t, q) \in \mathcal{O}_n(\delta)\}$$
$$\overline{q}_n(\delta) := \max\{q : (t, q) \in \mathcal{O}_n(\delta)\}.$$

By Kingman's subadditive ergodic theorem, for any δ there exists a constant $b(\delta)$ so that

$$B_n(\delta) \to b(\delta), \quad \text{a.s..} \tag{4}$$

1.2 The Organization of the Paper and Main Results

In this paper, we use a novel approach regarding the quantities of interest like the proportion of gaps, the rescaled score B_n, etc., as functions of δ. In Sect. 2, we derive some elementary but important properties of $B_n(\delta)$ and we explore the relation between the proportion of gaps of any optimal alignment and the derivatives of $B_n(\delta)$. In particular, we show (Claim 2.2) that for any n and δ,

$$B_n'(\delta_+) = \overline{q}_n(\delta), \quad B_n'(\delta_-) = \underline{q}_n(\delta). \tag{5}$$

In a sense these equalities, which almost trivially follow from the elementary calculus, are the core for the rest of the analysis.

In Sect. 3, we show that when the limit score function b is differentiable at δ, then a.s. $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ both converge to $b'(\delta)$ (by using expression (5)) so that $b'(\delta)$ can be interpreted as the *asymptotic proportion of gaps*. The section ends with an example showing that if b is not differentiable at δ, then the extremal proportions $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ can still a.s. converge to the corresponding one-side derivatives, namely $\underline{q}_n(\delta) \to b'(\delta_+)$ a.s. and $\overline{q}_n(\delta) \to b'(\delta_-)$ a.s.

Section 4 deals with large deviations bounds for the (optimal) proportion of gaps. The main result of this section is Theorem 4.1, which states that for every $\varepsilon > 0$ there exists a $c > 0$ such that for every n big enough the following large deviation inequality holds

$$P\left(b'(\delta_-) - \varepsilon \le \underline{q}_n(\delta) \le \overline{q}_n(\delta) \le b'(\delta_+) + \varepsilon\right) \ge 1 - 4\exp[-c(\varepsilon)n].$$

Combining this last inequality with the result on the speed of convergence proven in [11], we obtain the confidence intervals for the in general unknown quantities $b'(\delta_+)$ and $b'(\delta_-)$ in terms of $B_n(\delta)$ (the inequalities (27) and (27), respectively).

In Sect. 5 we obtain results on the fluctuations of the score of optimal alignments, namely we show that under some asymmetry assumption on the score function there exists a $c > 0$ so that for n large enough $\mathrm{Var}[L_n(\delta)] \ge cn$ provided that the gap penalty $-\delta$ is big enough (Theorem 5.2). This result implies that $\mathrm{Var}[L_n(\delta)] = \Theta(n)$, because as shown by Steele in [21], there exists another constant C such that $\mathrm{Var}[L_n(\delta)] \le Cn$. Our proof is based on the existence of the asymptotic proportion of gaps and, therefore, differs from the previous proofs in the literature.

Finally, Sect. 6 is devoted to the problem of determining the sufficiently large gap penalty δ_o so that the conditions of Theorem 5.2 are fulfilled. We show that when knowing the asymptotic upper bound $\overline{t}(\delta)$ of the average score of aligned letters, then δ_o can be easily found (Claim 6.1). Theorem 6.1 shows how the upper bound $\overline{t}(\delta)$ can be found. The proof of Theorem 6.1 uses similar ideas that the ones used in the proof of Theorem 5.2. The section ends with a practical example (Sect. 6.2).

It is important to notice that we could not find in the literature complete results on the fluctuations of the score in random sequences comparison. Though, a particular model for comparison of random sequences has had an interesting development in

the past 4 decades: the longest common subsequence problem (abbreviated by *LCS problem*). In our setting, the LCS problem corresponds to choose $S(x, y) = 1$ if $x = y$ and $S(x, y) = -\infty$ if $x \neq y$. Already in 1975, Chvatal and Sankoff [5] conjectured that the fluctuations of the length of the LCS is of order $o(n^{2/3})$. But in 1994, Waterman [23] conjectured that those fluctuations should be of order $\Theta(n)$. This last order had been proven by Matzinger et al. [2, 8–10] in a series of relatively recent papers treating extreme models with low entropy. In 2009, the Ph.D. thesis of Torres [14, 15, 22] brought an improvement, proving that the length of the LCS of sequences built by i.i.d. blocks has also fluctuations of order $\Theta(n)$, turning it to be the first time Waterman's conjecture was proven for a model with relatively high entropy. Unfortunately, the block-model of Torres does not have enough ergodicity as to extend the result to the still open original Waterman's conjecture. We believe that the results on the fluctuations of the score of optimal alignments showed in the present paper are an important source of new evidence that Waterman's conjecture might be true, even in more general models of sequence comparison than the LCS problem, provided the score function does not have a certain symmetry.

Note that the LCS problem can be reformulated as a last passage percolation problem with correlated weights [1]. For several last passage percolation models, the order of the fluctuations has been proven to be power $2/3$ of the order of the expectation. But as the previous models and simulations have showed (for simulations, see e.g. [3]), this order seem to be different as the order of the fluctuations of the score in optimal alignments.

2 Basic Properties of B_n

We start by deriving some elementary properties of the function $\delta \mapsto B_n(\delta)$:

Claim 2.1. *For every X and Y, the function $\delta \mapsto B_n(\delta)$ is non-decreasing, piecewise linear and convex.*

Proof. The non-decreasing and piecewise linear properties follow from the definition. For the convexity, with $\lambda \in (0, 1)$ let $\delta = \lambda\delta_1 + (1-\lambda)\delta_2$ and $(t, q) \in \mathcal{O}_n(\delta)$. Note that the pair (t, q) is not necessarily optimal for the proportions δ_1 and δ_2, so that from (3) it follows

$$\begin{aligned} B_n(\lambda\delta_1 + (1-\lambda)\delta_2) &= t(1-q) + (\lambda\delta_1 + (1-\lambda)\delta_2)q \\ &= \lambda\big(t(1-q) + \delta_1 q\big) + (1-\lambda)\big(t(1-q) + \delta_2 q\big) \\ &\leq \lambda B_n(\delta_1) + (1-\lambda)B_n(\delta_2). \end{aligned}$$

□

Claim 2.2. *For any $\delta \in \mathbb{R}$ we have*

$$B_n'(\delta_-) := \lim_{s \searrow 0} \frac{B_n(\delta - s) - B_n(\delta)}{s} = \underline{q}_n(\delta)$$

$$B_n'(\delta_+) := \lim_{s \searrow 0} \frac{B_n(\delta + s) - B_n(\delta)}{s} = \overline{q}_n(\delta).$$

Thus, $\mathcal{O}_n(\delta)$ is singleton if and only if $B_n(\delta)$ is differentiable at δ.

Proof. Fix $\delta \in \mathbb{R}$ and $s > 0$. Let $(t, q) \in \mathcal{O}_n(\delta)$, thus

$$B_n(\delta + s) \geq t(1 - q) + q(\delta + s) = B_n(\delta) + qs$$
$$B_n(\delta - s) \geq t(1 - q) + q(\delta - s) = B_n(\delta) - qs.$$

Hence,

$$\frac{B_n(\delta) - B_n(\delta - s)}{s} \leq q \leq \frac{B_n(\delta + s) - B_n(\delta)}{s}.$$

The inequalities above hold for any optimal (t, q) and for any s, so letting $s \searrow 0$ we have

$$B_n'(\delta_-) \leq \underline{q}_n(\delta) \leq \overline{q}_n(\delta) \leq B_n'(\delta_+). \tag{6}$$

Thus, if B_n is differentiable at δ, then $\underline{q}_n(\delta) = \overline{q}_n(\delta)$ meaning that $\mathcal{O}_n(\delta)$ is a singleton, say $\mathcal{O}_n(\delta) = (t_n(\delta), q_n(\delta))$. To prove that $B_n'(\delta_+) = \overline{q}_n(\delta)$, it is enough to show that there exists a pair $(t, q) \in \mathcal{O}_n(\delta)$ such that $B_n'(\delta_+) = q$. Indeed, since B_n is piecewise linear, for every $\varepsilon > 0$ small enough B_n is differentiable at $\delta + \varepsilon$ and the derivative equals to $B_n'(\delta_+)$. Hence, for every $\varepsilon > 0$ small enough $q := q_n(\delta + \varepsilon) = B_n'(\delta_+)$. Let $t := t_n(\delta + \varepsilon)$. Thus, for every $\varepsilon > 0$ small enough there exists a pair $(t, q) \in \mathcal{O}_n(\delta + \varepsilon)$ such that $q = B_n'(\delta_+)$. This means $t(1-q)+q\delta+q\epsilon = B_n(\delta+\varepsilon)$. Since B_n is continuous, we see that $\lim_{\varepsilon \to 0+} B_n(\delta + \varepsilon) = B_n(\delta) = t(1 - q) + q\delta$, i.e. $(t, q) \in \mathcal{O}_n(\delta)$. With similar arguments one can show that $\underline{q}_n(\delta) = B_n'(\delta_-)$. □

Function $B_n(\delta)$ for large δ. With fairly simple analysis, it is possible to determine $B_n(\delta)$ for large δ. Recall the definition of F. Clearly, when $\delta > F$, the optimal alignment only consists of gaps, namely $\delta \geq F \Rightarrow B_n(\delta) = \delta$. If we decrease the value of δ, say $\delta \in (E, F)$, the optimal alignment tries to align as many pairs of letters which score F as possible, thus minimizing the number of gaps. Formally, such optimal alignment can be obtained by defining a new score function

$$S_1(a, b) := \begin{cases} F & \text{if } S(a, b) = F \\ 0 & \text{if } S(a, b) < F \end{cases}$$

Let $B_n^1(\delta)$ be the corresponding expression (2) for the score function S_1. If (t_n^1, q_n^1) is such that $B_n^1(0) = t_n^1(1 - q_n^1) + 0 \cdot q_n^1$, then $t_n^1 = F$ and $1 - q_n^1$ is the maximal proportion of pairs that score F. Thus (t_n^1, q_n^1) is unique and, therefore, B_n^1 is differentiable at 0. For the original B_n, if $\delta \in [E, F]$, then we have

$$B_n(\delta) = F(1 - q_n^1) + \delta q_n^1 = B_n^1(0) + \delta q_n^1,$$

from where we have

$$B_n(F) = F = B_n^1(0) + F q_n^1. \tag{7}$$

If δ is slightly smaller than E, then the candidate alignments to be optimal alignments are obtained by aligning only those pair of letters that score E or F; amongst such alignments an optimal one will be the one having minimal number of gaps. Formally, we consider the score function

$$S_2(a,b) = \begin{cases} F \text{ if } S(a,b) = F \\ E \text{ if } S(a,b) = E \\ 0 \quad \text{otherwise} \end{cases}$$

Let $B_n^2(\delta)$ be the corresponding expression (2) for the score function S_2. Let (t_n^2, q_n^2) be such that $B_n^2(0) = t_n^2(1 - q_n^2) + 0 \cdot q_n^2$ with the additional property that $q_n^2 \leq q$ for any other optimal pair (t, q) for $B_n^2(0)$. By Claim 2.2, $q_n^2 = (B_n^2)'(0_-)$. Hence, if δ is slightly smaller than E, then $B_n(\delta) = t_n^2(1 - q_n^2) + \delta q_n^2 = B_n^2(0) + \delta q_n^2$.

Hence, we can write down

$$B_n(\delta) = \begin{cases} \delta & \text{if } \delta \geq F \\ F(1 - q_n^1) + \delta q_n^1 & \text{if } E \leq \delta \leq F \\ t_n^2(1 - q_n^2) + \delta q_n^2 & \text{if } E - \varepsilon \leq \delta \leq E \end{cases} \tag{8}$$

for a small $\varepsilon > 0$ which depends on X, Y. Indeed, if δ is much smaller than E but still above the value of the next score, then the optimal alignment (t, q) might align less F-valued letters for in order to achieve less gaps. In other words, the optimal alignment (t, q) can be such that $t(1 - q) < B_n^2(0)$. But it is not so for $\delta = E$ and due to the piecewise linearity of B_n, the $\varepsilon > 0$ described above exists.

By Claim 2.2, for any n we have that

$$\underline{q}_n(F) = q_n^1, \quad \overline{q}_n(F) = 1, \qquad \underline{q}_n(E) = q_n^2, \quad \overline{q}_n(E) = q_n^1. \tag{9}$$

Finally, note that by taking $\delta = E$, we obtain that

$$B_n(E) = B_n^2(0) + E q_n^2 \tag{10}$$

and

$$B_n^1(0) - B_n^2(0) = E(q_n^2 - q_n^1). \tag{11}$$

3 The Asymptotic Proportion of Gaps

¿From the convergence in (4), we see that the limit function $b(\cdot)$ inherits properties from $B_n(\cdot)$. More precisely, the (random) function $B_n(\cdot)$ is convex and non-decreasing, so the same holds for $b(\cdot)$. Moreover, due to the monotonicity, the convergence in (4) is uniform on δ, i.e.

$$\sup_{\delta\in\mathbb{R}} |B_n(\delta) - b(\delta)| \to 0 \quad \text{a.s. as } n \to \infty. \tag{12}$$

But we need to be a bit more careful in deriving properties of the derivative b' from B_n', since the uniform convergence of convex functions implies the convergence of one side derivatives at x only when the limit function is differentiable at x. Indeed, let f_n and f be convex functions that converge pointwise, i.e. $f_n(x) \to f(x)$ as $n \to \infty$, for every x. Then, in general [19] it holds

$$\begin{aligned} f'(x_-) &:= \lim_{s\searrow 0}\lim_{n\to\infty} \frac{f_n(x-s)-f_n(x)}{s} \le \liminf_{n\to\infty}\lim_{s\searrow 0} \frac{f_n(x-s)-f_n(x)}{s} \\ &\le \limsup_{n\to\infty}\lim_{s\searrow 0} \frac{f_n(x+s)-f_n(x)}{s} \le \lim_{s\searrow 0}\lim_{n\to\infty} \frac{f_n(x+s)-f_n(x)}{s} = f'(x_+), \end{aligned}$$

and these inequalities can be strict. In our case these inequalities are

$$b'(\delta_-) \le \liminf_n \underline{q}_n(\delta) \le \limsup_n \overline{q}_n(\delta) \le b'(\delta_+), \quad \text{a.s..} \tag{13}$$

Lemma 3.1. *Let b be differentiable at δ. Then*

$$\underline{q}_n(\delta) \to b'(\delta) \quad \textit{and} \quad \overline{q}_n(\delta) \to b'(\delta) \quad \textit{a.s. as } n \to \infty. \tag{14}$$

Remark 3.1. An interesting question is the following: If b is not differentiable at δ, there exist $\underline{q}, \overline{q} \in (0,1)$ with $\underline{q} \ge b'(\delta_-)$ and $\overline{q} \le b'(\delta_+)$ such that

$$\underline{q}_n(\delta) \to \underline{q} \quad \text{and} \quad \overline{q}_n(\delta) \to \overline{q} \quad \text{a.s. as } n \to \infty? \tag{15}$$

Numerical simulations of the difference $\overline{q}_n - \underline{q}_n$ as $n \to \infty$ do not conclusively show convergence nor boundedness, so perhaps such $\underline{q}, \overline{q}$ do not exist.

Thus, if b is differentiable at δ, the random proportion of gaps of optimal alignments tends to an unique number $q(\delta) := b'(\delta)$ that we can interpreted as the **asymptotic proportion of gaps** at δ. If the function b is not differentiable at δ, then it is not known whether the maximal or minimal proportion of gaps converge. However, as we shall now see this might be the case.

Asymptotic proportion of gaps for large δ. In general, it seems hard to determine where b is not differentiable and the asymptotic proportion of gaps does not exist.

However, based on the elementary properties of B_n and b, we can say something about the differentiability of b for large δ's. Recall B_n^1 and B_n^2, let b^1 and b^2 be the corresponding limits. The following claim shows that the proportions $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ might converge even if b is not differentiable at δ.

Claim 3.1. *The following convergences hold as $n \to \infty$:*

1. $\overline{q}_n(F) \to 1 = b'(F_+)$, *a.s.*
2. $\underline{q}_n(F) \to \frac{F-b^1(0)}{F} = b'(F_-)$, *a.s.*
3. $\overline{q}_n(E) \to \frac{F-b^1(0)}{F} = b'(E_+)$, *a.s.*
4. $\underline{q}_n(E) \to \frac{b(E)-b^2(0)}{E} = \frac{b^1(0)-b^2(0)}{E} + \frac{F-b^1(0)}{F} \geq b'(E_-)$, *a.s.*.

If $b^2(0) > b^1(0) > 0$, then b is not differentiable at $\delta = E$ and $\delta = F$.

Proof. We are going to use the fact that the convergence (12) does not depend on the score function, so there exist constants $b^1(0)$ and $b^2(0)$ such that $B_n^1(0) \to b^1(0)$ and $B_n^2(0) \to b^2(0)$ as $n \to \infty$, a.s.. Hence, from (7)

$$q_n^1 \to \frac{F - b^1(0)}{F}, \quad \text{a.s..}$$

From (10), it follows that

$$q_n^2 \to \frac{b(E) - b^2(0)}{E} = \frac{b^1(0) - b^2(0)}{E} + \frac{F - b^1(0)}{F} \geq b'(E_-), \quad \text{a.s.,} \qquad (16)$$

where the equality comes from (11) and the last inequality comes from (13). So that from (9) the convergences (1)–(4) now follow.

Let us now compare the limits with corresponding derivatives. From (8), we obtain

$$b(\delta) = \begin{cases} \delta & \text{if } \delta \geq F \\ b^1(0) + \delta \frac{F-b^1(0)}{F} & \text{if } E \leq \delta \leq F \end{cases} \qquad (17)$$

Hence, $b'(F_+) = 1$, $b'(F_-) = b'(E_+) = \frac{F-b^1(0)}{F}$. If $b_1(0) > 0$, then $b'(F_+) > b'(F_-)$ so that b is not differentiable at F. When $b^2(0) > b^1(0)$, then

$$b'(E_-) \leq \frac{b^1(0) - b^2(0)}{E} + \frac{F - b^1(0)}{F} < \frac{F - b^1(0)}{F} = b'(E_+),$$

so that b is not differentiable at E. □

We conclude with an important example (see [14, 15, 22]) showing that the case $b^2(0) > b^1(0) > 0$ is realistic.

Example 3.1. Let $m > 0$ be an integer, $\mathbb{A} = \{1, \ldots, m\}$ and $S(a, b) = a \wedge b$. Then $E = m - 1$ and $F = m$. Let every letter in $\mathbb{A}$ having a positive probability. Since $S(a, b) = m$ iff $a = b = m$, obviously $b^1(0) = mP(X_i = m)$ so that

$$b'(F_-) = b'(E_+) = \frac{m - b_1(0)}{m} = 1 - P(X_1 = m) < 1 = b'(F_+).$$

Since $B_n^2(0)$ is bigger than the score of the alignment obtained by aligning as many m-s as possible, thus $B_n^1(0)$, and aligning so many $m-1$'s as possible without disturbing already existing alignment of m's, clearly $b^2(0) > b^1(0)$.

4 Large Deviations

In this section, given $\delta \in \mathbb{R}$, we derive large deviations principle for $B_n'(\delta_+)$ resp. $B_n'(\delta_-)$ by using McDiarmid's inequality. From there, we also derive confidence bounds for $b'(\delta_+)$ resp. $b'(\delta_-)$. Recall that S is symmetric. Let

$$A := \max_{x,y,z \in \mathbb{A}} |S(x,y) - S(x,z)|. \tag{18}$$

For the sake of completeness, let us recall McDiarmid's inequality:

Let $Z_1, \ldots, Z_{2m}$ be independent random variables and $f(Z_1, \ldots, Z_{2m})$ be a function so that changing one variable changes the value at most $K > 0$. Then for any $\sigma > 0$ we have

$$P\Big(f(Z_1, \ldots, Z_{2m}) - Ef(Z_1, \ldots, Z_{2m}) > \sigma\Big) \le \exp\left[-\frac{\sigma^2}{mK^2}\right]. \tag{19}$$

For the proof, we refer to [6]. Another inequality which will be useful later is the so called Höffding's inequality, which is the consequence of McDiarmid's inequality when $f(Z_1, \ldots, Z_m) = \sum_{i=1}^m Z_i$, i.e. for any $\varepsilon > 0$ we have

$$\begin{aligned} P\left(\frac{1}{m}\sum_{i=1}^m Z_i - EZ_1 > \varepsilon\right) &= P\left(\sum_{i=1}^m Z_i - E\big(\sum_{i=1}^m Z_i\big) > \varepsilon m\right) \\ &\le \exp\left[-\frac{(\varepsilon m)^2}{K^2\frac{m}{2}}\right] = \exp\left[-\frac{2\varepsilon^2}{K^2}m\right]. \end{aligned} \tag{20}$$

In our case, for any $\delta \in \mathbb{R}$ changing the value of one of the $2n$ random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ changes the value of $L_n(\delta)$ at most A, hence for every $\varepsilon > 0$ inequality (19) is translated into

$$\begin{aligned} P\big(L_n(\delta) - EL_n(\delta) \ge \varepsilon n\big) &\le \exp\left[-\frac{\varepsilon^2}{A^2}n\right] \\ P\big(L_n(\delta) - EL_n(\delta) \le -\varepsilon n\big) &\le \exp\left[-\frac{\varepsilon^2}{A^2}n\right]. \end{aligned} \tag{21}$$

Let us define $b_n(\delta) := EB_n(\delta)$. For every $\delta \in \mathbb{R}$, by dominated convergence we have $b_n(\delta) \to b(\delta)$ and by monotonicity the convergence is uniform, i.e.

$$\sup_{\delta\in\mathbb{R}} |b_n(\delta) - b(\delta)| \to 0 \qquad \text{as } n \to \infty.$$

Theorem 4.1. *Let $\delta \in \mathbb{R}$. Then, for every $\varepsilon > 0$ there exists $N(\varepsilon) < \infty$ and a constant $c(\varepsilon) > 0$ such that*

$$P\left(b'(\delta_-) - \varepsilon \le \underline{q}_n(\delta) \le \overline{q}_n(\delta) \le b'(\delta_+) + \varepsilon\right) \ge 1 - 4\exp[-c(\varepsilon)n] \tag{22}$$

for every $n > N(\varepsilon)$.

Proof. Given $\delta \in \mathbb{R}$ and $\varepsilon > 0$, we are looking for bounds on $P\left(B_n'(\delta_+) - b'(\delta_+) > \varepsilon\right)$. For any $s > 0$ and any function $\varphi : \mathbb{R} \to \mathbb{R}$ let us define

$$\Delta\varphi := \varphi(\delta + s) - \varphi(\delta). \tag{23}$$

Now, choose a small $1 > s > 0$ depending on ε such that

$$\left|\frac{\Delta b}{s} - b'(\delta_+)\right| \le \frac{\varepsilon}{4}$$

and take n large enough (also depending on ε) such that

$$|\Delta b_n - \Delta b| \le s\frac{\varepsilon}{4}.$$

Thus for those s and n chosen as before we have

$$\begin{aligned}\frac{\Delta B_n}{s} - b'(\delta_+) &= \left(\frac{\Delta B_n}{s} - \frac{\Delta b_n}{s}\right) + \left(\frac{\Delta b_n}{s} - \frac{\Delta b}{s}\right) + \left(\frac{\Delta b}{s} - b'(\delta_+)\right) \\ &\le \left(\frac{\Delta B_n}{s} - \frac{\Delta b_n}{s}\right) + \frac{\varepsilon}{2}.\end{aligned} \tag{24}$$

From (21), it follows

$$\begin{aligned} P\left(B_n(\delta) - b_n(\delta) \le -s\frac{\varepsilon}{4}\right) &\le \exp\left[-\frac{\varepsilon^2 s^2}{16A^2}n\right] = \exp[-c_1(\varepsilon)n] \\ P\left(B_n(\delta + s) - b_n(\delta + s) \ge s\frac{\varepsilon}{4}\right) &\le \exp[-c_1(\varepsilon)n] \end{aligned} \tag{25}$$

where $c_1(\varepsilon) := \varepsilon^2 s^2/(16A^2)$ is a positive constant depending on ε (recall that our s depends on ε). Hence

$$P\left(\frac{\Delta B_n}{s} - \frac{\Delta b_n}{s} \geq \frac{\varepsilon}{2}\right) \leq P\left(B_n(\delta) - b_n(\delta) \leq -s\frac{\varepsilon}{4}\right) + P\left(B_n(\delta+s) - b_n(\delta+s) \geq s\frac{\varepsilon}{4}\right)$$
$$\leq 2\exp[-c_1(\varepsilon)n]. \tag{26}$$

Since B_n is convex, it holds that $B_n'(\delta_+) \leq \Delta B_n/s$ so that for ε and s chosen as before (24) and (26) yield

$$P\left(B_n'(\delta+) - b'(\delta+) \geq \varepsilon\right) \leq P\left(\frac{\Delta B_n}{s} - b'(\delta+) \geq \varepsilon\right) \leq 2\exp[-c_1(\varepsilon)n]. \tag{27}$$

By similar arguments, there exists a positive constant $c_2(\varepsilon) > 0$ so that

$$P\left(B_n'(\delta_-) - b'(\delta_-) \leq \varepsilon\right) \leq 2\exp[-c_2(\varepsilon)n]. \tag{28}$$

for every n big enough. Finally, by taking $c := \min\{c_1, c_2\}$, the inequality (6) implies the inequality (22). □

Note that if b is differentiable at δ, the inequality (22) is satisfied for $q(\delta)$ instead of $b'(\delta_-)$ or $b'(\delta_+)$, namely

Corollary 4.1. *Let b be differentiable at δ. Then, for every $\varepsilon > 0$ there exists $N(\varepsilon) < \infty$ and a constant $c(\varepsilon) > 0$ such that*

$$P\left(q(\delta) - \varepsilon \leq \underline{q}_n(\delta) \leq \overline{q}_n(\delta) \leq q(\delta) + \varepsilon\right) \geq 1 - 4\exp[-c(\varepsilon)n] \tag{29}$$

for every $n > N(\varepsilon)$, where $q(\delta)$ is the unique asymptotic proportion of gaps.

We now derive confidence bounds for $b'(\delta_+)$ resp. $b'(\delta_-)$. Recall the definition $b_n(\delta) = EB_n(\delta) = EL_n(\delta)/n$ and the notation (23). From [11] we have

$$b_n(\delta) \leq b(\delta) \leq b_n(\delta) + v(n)$$

for $n \in \mathbb{N}$ even, where

$$v(n) := A\sqrt{\frac{2}{n-1}\left(\frac{n+1}{n-1} + \ln(n-1)\right)} + \frac{F}{n-1},$$

so it follows

$$\Delta b - v(n) \leq \Delta b_n \leq \Delta b + v(n).$$

Suppose that k samples of $X^i = X_1^i, \ldots, X_n^i$ and $Y^i = Y_1^i, \ldots, Y_n^i$, $i = 1, \ldots, k$ are generated. Let $L_n^i(\delta)$ be the score of the i-th sample. Let

$$\bar{B}_n(\delta) := \frac{1}{kn}\sum_{i=1}^{n} L_n^i(\delta).$$

From (25) we have

$$\begin{aligned} P\big(\Delta\bar{B}_n - \Delta b_n < -c\big) &= P\big(\bar{B}_n(\delta+s) - b_n(\delta+s) + b_n(\delta) - \bar{B}_n(\delta) < -c\big) \\ &\le P\left(\bar{B}_n(\delta+s) - b_n(\delta+s) < -\frac{c}{2}\right) + P\left(b_n(\delta) - \bar{B}_n(\delta) < -\frac{c}{2}\right) \\ &\le 2\exp\left[-\frac{c^2 k}{4A^2}n\right]. \end{aligned}$$

By convexity $sb'(\delta+) \le \Delta b$ for every $s > 0$, so from the last inequality it follows

$$\begin{aligned} P\big(\Delta\bar{B}_n + c + v(n) \ge sb'(\delta+)\big) &\ge P\big(\Delta\bar{B}_n + c + v(n) \ge \Delta b\big) = P\big(\Delta\bar{B}_n + c \ge \Delta b - v(n)\big) \\ &\ge P\big(\Delta\bar{B}_n + c \ge \Delta b_n\big) = P\big(\Delta\bar{B}_n - \Delta b_n \ge -c\big) \\ &\ge 1 - 2\exp\left[-\frac{c^2 k}{4A^2}n\right], \end{aligned}$$

from where we obtain that with probability $1-\varepsilon$

$$b'(\delta+) \le \frac{1}{s}\Big(\bar{B}_n(\delta+s) - \bar{B}_n(\delta) + 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} + v(n)\Big). \tag{30}$$

Since (30) holds for every $s > 0$, we have that with probability $1-\varepsilon$

$$b'(\delta+) \le \min_{s>0}\frac{1}{s}\Big(\bar{B}_n(\delta+s) - \bar{B}_n(\delta) + 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} + v(n)\Big). \tag{31}$$

Similarly, we have that with probability $1-\varepsilon$

$$b'(\delta-) \ge \max_{s>0}\frac{1}{s}\Big(\bar{B}_n(\delta-s) - \bar{B}_n(\delta) - 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} - v(n)\Big). \tag{32}$$

5 Fluctuations of the Score in Optimal Alignments

In this section we prove $\mathrm{Var}[L_n(\delta)] = \Theta(n)$. The $\Theta(n)$ notation means that there exist two constants $0 < c < C < \infty$ such that $cn \le \mathrm{Var}[L_n(\delta)] \le Cn$ for n large enough. The upper bound follows from an Efron-Stein type of inequality proved by Steele in [21], so we aim to provide conditions on the scoring function that guarantee the existence of the lower bound. In this section we show that when $\delta < 0$ and $|\delta|$ is large enough in the sense of Assumption 5.1 (see below), then there exists $c > 0$ so that $\mathrm{Var}(L_n(\delta)) > cn$ for n large enough. In comparison with previous results, here we solve—for the first time—the problem of the fluctuations of the score in optimal alignments for rather realistic high entropy models.

5.1 Order of the Variance

All above mentioned fluctuations results are based in the following strategy: the inequality $\mathrm{Var}[L_n(\delta)] \geq cn$ is satisfied as soon as we are able to establish that changing at random one symbol in the sequences has a biased effect on the optimal alignment score. In details, we choose two letters $a, b \in \mathbb{A}$ and fix a realization of $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_n$. Then among all the a's in X and Y we choose one at random (with equal probability). That chosen letter a is replaced by a letter b. The new sequences thus obtained are denoted by $\tilde{X}$ and $\tilde{Y}$. The optimal score for the strings $\tilde{X}$ and $\tilde{Y}$ is denoted by

$$\tilde{L}_n(\delta) := \max_{(\pi,\mu)} U^{\delta}_{(\pi,\mu)}(\tilde{X}, \tilde{Y}).$$

The following important theorem postulates the mentioned strategy. In full generality, it is proven in [12], for special case of two colors and S corresponding to LCS, see Sect. 3 in [10]; for a special case of $S(a,b) = a \wedge b$ and $\mathbb{A} = \{m-1, m, m+1\}$, see Theorem 2.1 in [14].

Theorem 5.1. *Assume that there exist $\varepsilon > 0$, $d > 0$ and $n_0 < \infty$ such that*

$$P\left(E[\tilde{L}_n(\delta) - L_n(\delta) | X, Y] \geq \varepsilon \right) \geq 1 - e^{-dn} \tag{33}$$

for all $n > n_0$. Then, there exists a constant $c > 0$ not depending on n such that $\mathrm{Var}[L_n(\delta)] \geq cn$ for every n large enough.

Now, our aim is to show that if δ is small enough and the scoring function satisfies some asymmetry property, then there exist letters $a, b \in \mathbb{A}$ so that the condition (33) is fulfilled. Typically, to satisfy the assumptions, δ should be negative so that the main result holds if the gap penalty $|\delta|$ is large enough. Let us introduce our asymmetry assumption on the scoring function:

Assumption 5.1. *Suppose there exist letters $a, b \in \mathbb{A}$ such that*

$$\sum_{c \in \mathbb{A}} P(X_1 = c)\big(S(b,c) - S(a,c)\big) > 0. \tag{34}$$

Remark 5.1. For the alphabet $\mathbb{A} = \{a, b\}$, condition (34) says

$$\big(S(b,a) - S(a,a)\big)P(X_1 = a) + \big(S(b,b) - S(b,a)\big)P(X_1 = b) > 0.$$

Since S is symmetric and one could exchange a and b, the condition (34) actually means

$$\big(S(b,a) - S(a,a)\big)P(X_1 = a) + \big(S(b,b) - S(b,a)\big)P(X_1 = b) \neq 0.$$

When $S(b,b) = S(a,a)$, then Assumption 5.1 is satisfied if and only if $P(X_1 = a) \neq P(X_1 = b)$. For, example when $S(b,b) = S(a,a) > S(b,a)$ (recall that S is assumed to be symmetric and non-constant), then (34) holds if $P(X_1 = a) \neq P(X_1 = b)$.

In the present paper, the main result on the fluctuations of the score in optimal alignments can be formulated as following:

Theorem 5.2. *Suppose Assumption 5.1 holds. Then, there exist constants δ_0 and $c > 0$ not depending on n such that*

$$\mathrm{Var}[L_n(\delta)] \geq cn$$

for all $\delta \leq \delta_0$ and for n large enough.

Before proving the above-stated theorem, we need a preliminary lemma. Suppose Assumption 5.1 holds, then take $a, b \in \mathbb{A}$ satisfying (34) and define the functions $\zeta^x : \mathbb{A} \times \mathbb{A} \mapsto \mathbb{R}$ and $\zeta^y : \mathbb{A} \times \mathbb{A} \mapsto \mathbb{R}$ in the following way:

$$\zeta^x(x,y) = \begin{cases} S(b,y) - S(a,y) & \text{if } x = a \\ 0 & \text{otherwise} \end{cases}$$

$$\zeta^y(x,y) = \begin{cases} S(x,b) - S(x,a) & \text{if } y = a \\ 0 & \text{otherwise.} \end{cases}$$

Note that $S(x,y) = S(y,x)$ implies $\zeta^y(x,y) = \zeta^x(y,x)$. We now define the random variable Z by

$$Z := \zeta^x(X_1,Y_1) + \zeta^y(X_1,Y_1) = \zeta^x(X_1,Y_1) + \zeta^x(Y_1,X_1). \tag{35}$$

Note that (5.1) ensures that Z has strictly positive expectation:

$$\begin{aligned} \rho := EZ &= E\big(\zeta^x(X_1,Y_1) + \zeta^y(X_1,Y_1)\big) = 2E\zeta^x(X_1,Y_1) \\ &= 2E[\zeta^x(a,Y_1)|X_1 = a]P(X_1 = a) = 2\sum_{c\in\mathbb{A}} \zeta^x(a,c)P(Y_1 = c)P(X_1 = a) \\ &= 2P(X_1 = a)\sum_{c\in\mathbb{A}} (S(b,c) - S(a,c))P(X_1 = c) > 0. \end{aligned}$$

Let Λ^* be the Legendre-Fenchel transform of the logarithmic moment generating function of $-Z$, namely

$$\Lambda^*(c) = \sup_{t\in\mathbb{R}} (ct - \ln E[\exp(-Zt)]) \qquad \forall\, c \in \mathbb{R}.$$

It is known that the supremum above can be taken over non-negative t's and, for any $c > E(-Z) = -\rho$, it holds $\Lambda^*(c) > 0$. Since $\rho > 0$, we have for $c = 0$

$$\Lambda^*(0) = -\inf_{t\in\mathbb{R}} \ln E[\exp(-tZ)] = -\inf_{t\geq 0} \ln E[\exp(-tZ)] = -\ln \inf_{t\geq 0} E[\exp(-tZ)] > 0.$$

Let $Z_1, \ldots, Z_k$ be i.i.d. random variables distributed as $-Z$, then for any $c > -\rho$ the following large deviation bound holds

$$P\left(\sum_{i=1}^{k} Z_i > ck\right) \leq \exp[-\Lambda^*(c)k]. \tag{36}$$

Finally, denote $h(q)$ the binary entropy function $h(q) := -q\ln q - (1-q)\ln(1-q)$ and note that the inequality

$$2h(q) < \Lambda^*(0)(1-q)$$

holds when $q > 0$ is small enough, since Λ^* and h are both continuous and $\Lambda^*(0) > 0$.

In what follows, let for any $q \in (0,1)$, $\mathcal{A}^n(q)$ be the set of all alignments with no more than qn gaps, i.e.

$$\mathcal{A}^n(q) := \{(\pi,\mu) : q(\pi,\mu) \leq q\}.$$

We are interested in the event that the sequences X and Y are such that for every alignment (π,μ) with no more than qn gaps we have a biased effect of the random change of at least $\varepsilon > 0$. Let $D_q^n(\varepsilon)$ denote that event i.e.

$$D_q^n(\varepsilon) := \bigcap_{(\pi,\mu)\in\mathcal{A}^n(q)} D_{(\pi,\mu)}^n(\varepsilon) \tag{37}$$

where

$$D_{(\pi,\mu)}^n(\varepsilon) := \left\{ E\left[\sum_{i=1}^{k} \left(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\right) \middle| X, Y\right] \geq \varepsilon \right\}.$$

Now, we are ready to state the key lemma.

Lemma 5.1. *Suppose Assumption 5.1 is fulfilled and take $a, b \in \mathbb{A}$ satisfying (34). Let $q > 0$ small enough such that*

$$2h(q) < \Lambda^*(0)(1-q). \tag{38}$$

Then, there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q, such that

$$P\left((D_q^n(\varepsilon))^c\right) \leq \exp[-\alpha n] \tag{39}$$

for every $n > n_2$.

Proof. Let $x = x_1, \ldots, x_n$ and respectively $y = y_1, \ldots, y_n$ be fixed realizations of X and Y, respectively. Let n_a be the number of a's in both sequences. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ be a fixed alignment of X and Y. Recall that $\tilde{X}$ and $\tilde{Y}$ are obtained by choosing at random one a among all the a's in x and y. Hence, such an a is chosen with probability $1/n_a$. Our further analysis is based on the following observation:

$$E\left[\sum_{i=1}^{k}\left(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\right)\Big| X = x, Y = y\right]$$
$$= \frac{1}{n_a}\sum_{i=1}^{k}\left(\zeta^x(x_{\pi_i}, y_{\mu_i}) + \zeta^y(x_{\pi_i}, y_{\mu_i})\right).$$

Thus, it holds

$$P\left((D^n_{(\pi,\mu)}(\varepsilon))^c\right) = P\left(E\left[\sum_{i=1}^{k}\left(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\right)\Big| X, Y\right] < \varepsilon\right)$$
$$= P\left(\sum_{i=1}^{k} Z_i < N_a\varepsilon\right), \tag{40}$$

where N_a is the (random) number of a's in X and Y and the random variables $Z_1, \ldots, Z_k$ are defined as follows:

$$Z_i := \zeta^x(X_{\pi_i}, Y_{\mu_i}) + \zeta^y(X_{\pi_i}, Y_{\mu_i}) = \zeta^x(X_{\pi_i}, Y_{\mu_i}) + \zeta^x(Y_{\mu_i}, X_{\pi_i})$$

for $i = 1, \ldots, k$. Let us mention again that the random variables Z_i depend on fixed alignment (π, μ) (which is omitted in the notation) and, since $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are i.i.d., so are the random variables $Z_1, \ldots, Z_k$. Clearly, Z_i is distributed as Z defined in (35). Suppose now that the fixed alignment (π, μ) has the proportion of gaps less or equal than q, i.e. $(\pi, \mu) \in \mathcal{A}^n(q)$. Then $\frac{k}{n} \geq 1 - q$ and, since obviously $N_a \leq 2n$, we have

$$\left\{\sum_{i=1}^{k} Z_i < \varepsilon N_a\right\} \subseteq \left\{\sum_{i=1}^{k} Z_i < \varepsilon 2n\right\} = \left\{\sum_{i=1}^{k} Z_i < k2\varepsilon\frac{n}{k}\right\} \subseteq \left\{\sum_{i=1}^{k} Z_i < k\frac{2\varepsilon}{(1-q)}\right\}. \tag{41}$$

Fix now q satisfying (38). Since Λ^* is continuous, there exists ε, depending on the chosen q so that the following two conditions are simultaneously satisfied

$$\frac{-2\varepsilon}{1-q} > -\rho \quad \text{and} \quad -2\alpha := 2h(q) - \Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q) < 0. \tag{42}$$

Using the large deviations bound (36) with $c = \frac{-2\varepsilon}{1-q}$ and the fact that $\frac{k}{n} \geq 1-q$, we have

$$P\left(-\sum_{i=1}^{k} Z_i > -k\frac{2\varepsilon}{(1-q)}\right) \leq \exp\left[-\Lambda^*\left(\frac{-2\varepsilon}{1-q}\right)k\right] \leq \exp\left[-\Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q)n\right]. \quad (43)$$

By (37), (40), (41) and (43), we obtain

$$P\left((D_q^n(\varepsilon))^c\right) \leq |\mathcal{A}^n(q)| \exp\left[-\Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q)n\right]. \quad (44)$$

In order to bound $|\mathcal{A}^n(q)|$, note that the number of different alignment with exactly $(n-k)$ gaps is bounded above by $\binom{n}{n-k}^2$ so that for $q \leq 0.5$ we have

$$|\mathcal{A}^n(q)| \leq \sum_{i \leq qn} \binom{n}{i}^2 \leq \sum_{i \leq qn} \binom{n}{qn}^2 \leq qn\binom{n}{qn}^2 \leq \exp[2h(q)n + \ln(qn)], \quad (45)$$

where $h(q)$ is the binary entropy function. In the second inequality the relation $q \leq 0.5$ was used, while the last inequality is based on the well-known relation $\binom{n}{\gamma n} \leq \exp[h(\gamma)n]$, for any $\gamma \in (0,1)$. Thus, from (42), (44) and (45) we have

$$P\left((D_q^n(\varepsilon))^c\right) \leq \exp\left[\left(2h(q) - \Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q) + \frac{\ln(qn)}{n}\right)n\right] = \exp\left[-\left(2\alpha - \frac{\ln(qn)}{n}\right)n\right]. \quad (46)$$

This implies that there exists n_2 big enough (recall $nq \geq 1$) such that (39) holds. □

Proof of Theorem 5.2. Let $\mathcal{O}(X,Y)$ denote the set of all optimal alignments of (X,Y), i.e.

$$(\pi,\mu) \in \mathcal{O}(X,Y) \Leftrightarrow L_n(\delta) = U^{\delta}_{(\pi,\mu)}(X,Y) = \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \delta q(\pi,\mu)n.$$

Note that the difference $\tilde{L}_n(\delta) - L_n(\delta)$ is bounded from below by

$$\tilde{L}_n(\delta) - L_n(\delta) \geq U^{\delta}_{(\pi,\mu)}(\tilde{X},\tilde{Y}) - U^{\delta}_{(\pi,\mu)}(X,Y) = \sum_{i=1}^{k}\left(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\mu_i})\right).$$

Thus, for every $\varepsilon > 0$ we have

$$\left\{\exists(\pi,\mu) \in \mathcal{O}(X,Y) : E\left[\sum_{i=1}^{k}\left(S(\tilde{X}_{\pi_i},\tilde{Y}_{\mu_i}) - S(X_{\pi_i},Y_{\mu_i})\right)\Big| X,Y\right] \geq \varepsilon\right\}$$
$$\subseteq \left\{E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \varepsilon\right\}. \quad (47)$$

Recall that the event $D_q^n(\varepsilon)$ means that every alignment (π,μ) with no more than qn gaps has a biased effect of the random change at least ε. Now, it is clear that the right side of (47) holds if $D_q^n(\varepsilon)$ holds and there exists an optimal alignment contains no more than qn gaps, i.e. we have the inclusion

$$\left\{\mathcal{O}(X,Y) \subseteq \mathcal{A}^n(q)\right\} \cap D_q^n(\varepsilon) \subseteq \left\{E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \epsilon\right\}. \quad (48)$$

Recall that $b(\delta)$ is convex and increasing, $b'(\delta) = 1$ if δ is big enough and $b'(\delta) = 0$ if δ is small enough. Hence, for every $q \geq 0$ there exists δ so that $b'(\delta+) < q$. Let δ be such and denote $\varepsilon_1 := q - b'(\delta+)$. Then by Theorem 4.1, there exist $c(\varepsilon_1)$ and $n_1(\varepsilon_1)$ such that

$$P\left(\overline{q}_n(\delta) \leq q\right) \geq 1 - 2\exp[c(\varepsilon_1)n] \quad (49)$$

for every $n > n_1$. Therefore we have

$$P\left(\mathcal{O}(X,Y) \subseteq \mathcal{A}^n(q)\right) \geq 1 - 2\exp[c(\varepsilon_1)n] \quad (50)$$

for $n > n_1$. From Lemma 5.1, it follows that if $q > 0$ is small enough to satisfy (38), then there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q so that

$$P\left((D_q^n(\varepsilon))^c\right) \leq \exp[-\alpha n] \quad (51)$$

for every $n > n_2$. To finalize the proof, let us take q satisfying (38) and δ_0 be such that $b'(\delta_{0+}) < q$. Then, there exist $\varepsilon > 0, \alpha > 0$ and $n_0 := \max\{n_2, n_1\}$ so that (50) and (51) hold. Thus, from (48) we have

$$P\left(E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \epsilon\right) \geq 1 - 2\exp[c(\varepsilon_1)n] - \exp[-\alpha n]$$

for every $n > n_0$. Hence, the assumptions of Theorem 5.1 are satisfied. □

An alternative to Lemma 5.1. Recall that δ_0 in Theorem 5.2 was chosen to be such that $b'(\delta_{0+}) < q$, where q satisfies assumptions of Lemma 5.1, namely (38). This assumption comes from the large deviations bound (43). Although, asymptotically it is a sharp inequality, the rate-function Λ^* might not always be easy to compute. Clearly, the statement of Lemma 5.1 holds true for any other type of large deviations inequality giving the same exponential decay. An alternative would be to use Höffding's inequality (20) to get a version of Lemma 5.1 which does not rely on

the computation of Λ^*. The Höffding's inequality gives smaller q, and, therefore, larger δ_0.

Lemma 5.2. *Suppose Assumption 5.1 is fulfilled and take a, b satisfying (34). Let $q > 0$ small enough such that*

$$h(q) < \frac{(1-q)\rho^2}{9A^2}. \tag{52}$$

Then there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q, such that

$$P\left((D_q^n(\varepsilon))^c\right) \le \exp[-\alpha n] \tag{53}$$

for every $n > n_2$.

Proof. Recall the definition of A from (18). Let $q > 0$ be small enough satisfying (52). Then, there exists $\varepsilon > 0$ small enough such that both conditions simultaneously hold:

(1) $2\varepsilon < (1-q)\rho$, which means that $\sigma := \rho - \frac{2\varepsilon}{(1-q)} > 0$;
(2)

$$h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} =: -\alpha(\varepsilon) < 0.$$

Hence, there exists $n_2 < \infty$ such that

$$h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} + \frac{\ln(qn)}{2n} \le -\frac{\alpha}{2} \tag{54}$$

for every $n > n_2$. Recall that $k \ge (1-q)n$. To apply Höffding's inequality, we need to bound the random variable Z. Recall the definition of Z from (35). From the definition, $\zeta^x(x,y)$ and $\zeta^y(x,y)$ are simultaneously non-zero if and only if $x = y = a$, this means that the difference between the maximum and minimum value of Z_i is at most $3A$. For instance, if $S(b,a) < S(a,a)$ then $-2A \le 2(S(b,a) - S(a,a)) \le Z_i \le \max_{c \ne a}(S(b,c) - S(a,c)) \le A$. Then, by using (20), the large deviations bound (43) can be written down as (recall **(1)**)

$$P\left(-\sum_{i=1}^{k} Z_i > -k\frac{2\varepsilon}{(1-q)}\right) = P\left(\frac{1}{k}\sum_{i=1}^{k}(-Z_i) + \rho > \rho - \frac{2\varepsilon}{(1-q)}\right) \le \exp\left[-\frac{2\sigma^2}{(3A)^2}k\right]$$
$$\le \exp\left[-\frac{2\sigma^2(1-q)}{9A^2}n\right] = \exp\left[-\frac{2((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)}n\right].$$

Finally, the inequality (46) can be now written down as

$$P\left(D_q^{nc}(\varepsilon)\right) \le \exp\left[2\left(h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} + \frac{\ln(qn)}{2n}\right)n\right] = \exp[-\alpha n], \tag{55}$$

where the result is proven by using (54). □

6 Determining δ

In the last section, we discuss how to determine δ_0 in Theorem 5.2. Recall, once again, the proof of that theorem: δ_0 is so small that $b'(\delta_{0+}) < q$, where q is small enough to satisfy condition (38). This condition depends on Λ^*, but knowing the distribution of X_1, Λ^* can be found. When Λ^* is unknown, then condition (38) can be substituted by (more restrictive) condition (52). The latter does not depend on Λ^* and can be also used when the distribution of X_1 is unknown. Hence finding q is not a problem. The problem, however, is to determine the function b or its derivatives. In Sect. 4, we found confidence upper bound for $b'(\delta_+)$ (31). That bound is random and holds only with certain probability. In the following, we investigate deterministic ways to estimate $b'(\delta+)$.

Let $(t_n, q_n) \in \mathcal{O}_n(\delta)$ be an optimal pair: $B_n = t_n(1-q_n) + \delta q_n$. Clearly, when (t_n', q_n') is another optimal pair and $q_n' > q_n$, then $t_n' > t_n$. Hence $(\overline{t}_n, \overline{q}_n) \in \mathcal{O}_n(\delta)$, where $\overline{t}_n = \max\{t : (t,q) \in \mathcal{O}_n(\delta)\}$. For every $q \in (0,1)$, let $\overline{t}(q)$ be an asymptotic upper bound for $\overline{t}_n$ in the sense that if $b'(\delta+) < q$ (i.e. $\limsup_n \overline{q}_n(\delta) < q$ almost surely) then

$$P(\text{eventually } \overline{t}_n \le \overline{t}(q)) = 1.$$

Thus, $q \mapsto \overline{t}(q)$ is non-decreasing. In what follows, let $\underline{b}$ be the lower bound of $b(\delta)$ for every δ. Since the asymptotic proportion of gaps goes to zero as $\delta \to -\infty$, $\underline{b}$ can be taken as the limit of gapless alignments. This limit is obviously $ES(X_1, Y_1) =: \gamma$. If the distribution of X_1 is unknown, then $\underline{b}$ can be any lower bound for γ.

Let now $q_o \in (0,1)$ be fixed. We aim to find $\delta_o := \delta(q_o) \ge 0$ such that $b'(\delta_+) < q_o$ for every δ satisfying $-\delta > \delta_o$. The following claim shows that δ_o can be computed as follows:

$$\delta_o = \sup_{q \ge q_o} \frac{\overline{t}(q)(1-q) - \underline{b}}{q}. \tag{56}$$

Claim 6.1. *Let $\delta < 0$ be such that $-\delta > \delta_o$, where δ_o is as in (56). Then $b'(\delta_+) \le q_o$.*

Proof. Take $\delta \le 0$ so small that $-\delta \ge \delta_o$. Without loss of generality, we can assume that b is differentiable at δ. Thus b is differentiable at δ implies that $\overline{q}_n \to b'(\delta) > 0$ a.s. Let now $\varepsilon := |\delta| - \delta_o$ and let $q' > b'(\delta)$ be such that

$$\left|\frac{\overline{t}(q')(1-q') - \underline{b}}{q'} - \frac{\overline{t}(q')(1-b'(\delta)) - \underline{b}}{b'(\delta)}\right| < \varepsilon.$$

Suppose $b'(\delta) \ge q_o$. Then, $q' > q_o$ and by definition of δ_o

$$\delta(q_o) \ge \frac{\overline{t}(q')(1-q') - \gamma}{q'}$$

so that

$$|\delta| = \delta_o + \varepsilon > \frac{\bar{t}\big(q'\big)(1 - b'(\delta)) - \underline{b}}{b'(\delta)} \quad \Leftrightarrow \quad \bar{t}\big(q'\big)(1 - b'(\delta)) - |\delta| b'(\delta) < \underline{b}. \tag{57}$$

Since $b'(\delta) < q'$, then eventually $\bar{t}_n \leq \bar{t}(q')$ a.s. By the convergence $\overline{q}_n \to b'(\delta)$ from the r.h.s. of (57), follows then that eventually

$$B_n = \bar{t}_n(1 - \overline{q}_n) - |\delta| \overline{q}_n < \underline{b}$$

almost surely. We have a contradiction with the almost surely convergence $B_n \to b(\delta) \geq \underline{b}$. □

Remark 6.1. If $\bar{t}(q) \equiv \bar{t}$ is constant, then (56) is

$$\delta(q_o) := \frac{\bar{t}(1 - q_o) - \underline{b}}{q_o}. \tag{58}$$

6.1 Finding $\bar{t}(q)$

For applying (56) the crucial step is to find $\bar{t}$. Since the maximum value of the scoring function is F, a trivial bound is $\bar{t}(q) \equiv F$ and δ_o can be found from (58). However, using the same ideas as in the proof of Theorem 5.2, we could obtain a realistic bound for $\bar{t}(q)$ as follows. In the following theorem, let Λ^* be Legendre-Fenchel transform of $\Lambda(t) := \ln E \exp[tZ]$, where $Z := S(X, Y)$. Clearly Z is a nonnegative random variable $Z \leq F$ and EZ was denoted by γ.

Theorem 6.1. *Let $q_1 \in (0, 1)$ and let $\bar{t}(q_1)$ satisfy one of the following conditions*

$$\frac{2h(q_1 \wedge 0.5)}{1 - q_1} = \Lambda^*(\bar{t}(q_1)). \tag{59}$$

or

$$\bar{t}(q_1) = F \sqrt{\frac{h(q_1 \wedge 0.5)}{1 - q_1}} + \gamma. \tag{60}$$

Then for every δ such that $b'(\delta_+) < q_1$, the following holds

$$P\big(\text{eventually } \bar{t}_n(\delta) \leq \bar{t}(q_1)\big) = 1.$$

Proof. Let $q_1 \in (0, 1)$ be fixed. Let δ be such that $b'(\delta_+) < q_1$. Note that we can find q such that $b'(\delta_+) < q$ and the following conditions both hold

$$\frac{2h(q \wedge 0.5)}{1-q} < \Lambda^*(\bar{t}(q_1)). \tag{61}$$

and

$$\bar{t}(q_1) > F\sqrt{\frac{h(q \wedge 0.5)}{1-q}} + \gamma. \tag{62}$$

Note that (62) implies

$$h(q \wedge 0.5) < \frac{(\bar{t}(q_1) - \gamma)^2}{F^2}(1-q). \tag{63}$$

Let $(\pi_1, \ldots, \pi_k)$ and $(\mu_1, \ldots, \mu_k)$ be a fixed alignment, and let $Z_1, \ldots, Z_k$ be i.i.d. random variables, where $Z_i = S(X_{\pi_i}, Y_{\mu_i})$. Clearly Z_i is distributed as Z defined above. If the alignment (π, μ) is optimal, then

$$t_n = \frac{1}{k}\sum_{i=1}^{k} Z_i.$$

Recall that $\gamma = EZ_i$. Since $\Lambda^*(\gamma) = 0$, the conditions (59) and (60) both guarantee $\bar{t} > \gamma$. Let us define

$$D_q^n(\bar{t}(q_1)) := \bigcap_{(\pi,\mu)\in\mathcal{A}^n(q)} \left\{ \frac{1}{k}\sum_{i=1}^{k} Z_i \le \bar{t}(q_1) \right\}.$$

The event $D_q^n(\bar{t}(q_1))$ states that the average score of aligned letters is smaller than $\bar{t}(q_1)$ for every alignment with proportion of gaps at most q. If all optimal alignments are so, then also $\bar{t}_n(\delta) \le \bar{t}(q_1)$, namely

$$\{\mathcal{O}(X,Y) \in \mathcal{A}^n(q)\} \cap D_q^n(\bar{t}) \subseteq \{\bar{t}_n(\delta) \le \bar{t}(q_1)\}.$$

In order to bound $P(D_q^n(\bar{t}(q_1)))$, we proceed as in Lemma 5.1. Using the large deviations bound

$$P\left(\sum_{i=1}^{k} Z_i > \bar{t}(q_1)k\right) \le \exp[-\Lambda^*(\bar{t}(q_1))k] \le \exp[-\Lambda^*(\bar{t}(q_1))(1-q)n] \tag{64}$$

we obtain the following estimate

$$P\left((D_q^n(\bar{t}))^c\right) \le |\mathcal{A}^n(q)| \exp[-\Lambda^*(\bar{t}(q_1))(1-q)n] \tag{65}$$

For $q \leq 0.5$, we estimate $|\mathcal{A}^n(q)|$ as in Lemma 5.1 by

$$|\mathcal{A}^n(q)| \leq \exp\left[\left(2h(q) + \frac{\ln(qn)}{n}\right)n\right].$$

For $q \in (0.5, 1)$ note that

$$|\mathcal{A}^n(q)| \leq \sum_{i \leq qn} \binom{n}{i}^2 < qn \binom{n}{\frac{1}{2}n}^2 \leq \exp\left[\left(2h(0.5) + \frac{\ln(qn)}{n}\right)n\right].$$

Hence

$$P\left((D_q^n(\bar{t}))^c\right) \leq \exp\left[\left(2h(q \wedge 0.5) + \frac{\ln(qn)}{n} - \Lambda^*(\bar{t}(q_1))(1-q)\right)n\right]. \quad (66)$$

Since (61) holds, then just like in the proof of Theorem 5.2, there exists $\alpha > 0$ and n_o, both depending on $\bar{t}(q_1)$, so that

$$P\left(\bar{t}_n(\delta) > \bar{t}(q_1)\right) \leq \exp[-\alpha n], \quad \forall n > n_o. \quad (67)$$

Thus,by Borel-Cantelli we have

$$P\left(\text{eventually } \bar{t}_n(\delta) \leq \bar{t}(q_1)\right) = 1.$$

With Höffding's inequality the bounds (64) and (66) are

$$P\left(\sum_{i=1}^{k} Z_i > \bar{t}(q_1)k\right) \leq \exp\left[-\frac{2(\bar{t}(q_1)-\gamma)^2}{F^2}k\right] \leq \exp\left[-\frac{2(\bar{t}(q_1)-\gamma)^2}{F^2}(1-q)n\right]$$

$$P\left((D_q^n(\bar{t}))^c\right) \leq \exp\left[2\Big(h(q \wedge 0.5) + \frac{\ln(qn)}{2n} - \frac{(\bar{t}(q_1)-\gamma)^2}{F^2}(1-q)\Big)2n\right].$$

respectively, and the existence of $\alpha > 0$ and n_o comes from (63). □

6.2 Example

Consider a two letter alphabet $\mathbb{A} = \{a, b\}$ with probabilities $P(X_i = b) = P(Y_i = b) = 0.7$, $P(X_i = a) = P(Y_i = a) = 0.3$. Let the scoring function S assign 1 to identical letter pairs and 0 to unequal letters. Then the letters a, b satisfy (5.1). The random variable Z as in (35) is distributed as follows:

$$P(Z = -2) = (0.3)^2 = 0.09, \quad P(Z = 0) = (0.7)^2 = 0.49, \quad P(Z = 1) = 2 \cdot 0.3 \cdot 0.7 = 0.42.$$

Hence $EZ = \rho = 0.24$ and

$$E\exp[-tZ] = (0.09)\exp[2t] + 0.49 + 0.42\exp[-t].$$

This function achieves its minimum at a t^* that is the solution of the equation

$$(2 \cdot 0.09)\exp[2t] = \exp[-t]0.42$$

that is

$$t^* = \frac{1}{3}\ln\frac{42}{18} \approx 0.28.$$

Then

$$-\Lambda^*(0) = \ln\left[(3 \cdot 0.09)\exp[2t^*] + 0.49\right] = -0.03564$$

so that q satisfies (38) if and only if $q < q_o := 0.00255$, because q_o is a solution of the equation

$$h(q_o) = \frac{\Lambda^*(0)}{2}(1 - q_o) \quad \Leftrightarrow \quad h(q_o) = 0.01782(1 - q_o).$$

Let us see, how much q_o changes when we assume the stronger condition (52). Clearly $A = 1$, so to satisfy (52), the proportion of gaps should satisfy the inequality $q < q_o := 0.000784674$, because q_o is the solution of the inequality $9h(q_o) = (1 - q_o)(0.24)^2$ that is

$$h(q_o) = 0.0064(1 - q_o).$$

Determining δ_0. Let us find δ_0 so that $b'(\delta_{0+}) \le q_0 = 0.00255$. In this example $F = 1$, $\gamma = (0.3)^2 + (0.7)^2 = 0.58$. Taking $\bar{t} = 1$, from (58) with $q_o = 0.00255$, we get

$$\delta_o := \frac{(1 - q_o) - 0.58}{q_o} = \frac{(1 - 0.00255) - 0.58}{0.00255} < 164.$$

The inequality (60) is

$$\bar{t}(q) = \sqrt{\frac{h(q \wedge 0.5)}{1 - q}} + \gamma.$$

Thus, from (56), we get

$$\delta_o = \sup_{q \ge q_o} \frac{\bar{t}(q)(1 - q) - \gamma}{q} = \sup_{q \ge q_o} \frac{(\sqrt{\frac{h(q \wedge 0.5)}{1-q}} + \gamma)(1 - q) - \gamma}{q}$$

$$= \sup_{q \ge q_o} \frac{\sqrt{h(q \wedge 0.5)(1 - q)}}{q} - \gamma.$$

Since

$$q \mapsto \frac{\sqrt{h(q \wedge 0.5)(1-q)}}{q}$$

is decreasing, we get a much better bound

$$\delta_o = \frac{\sqrt{h(q_o)(1-q_o)}}{q_o} - \gamma = \frac{\sqrt{h(0.00255)(1-0.00255)}}{0.00255} - 0.58 < 52.$$

The random variable $Z := S(X_1, Y_1)$ has Bernoulli distribution with parameter γ, so it is well known that

$$\Lambda^*(t) = t \ln\left(\frac{t}{\gamma}\right) + (1-t)\ln\left(\frac{1-t}{1-\gamma}\right),$$

provided $t > \gamma$. Therefore, the maximum value of $\Lambda^*(t)$ is achieved for $t = 1$ and that is the solution of (59) for $q_1 = 0.0698$. Hence, (59) has solution $\overline{t}(q_1)$ for every $q_1 \in [0, 0.0698]$ and in the range $[0, 0.0698]$ the function

$$q \mapsto \frac{\overline{t}(q)(1-q) - \gamma}{q}$$

is decreasing. This means that δ_o can be taken as

$$\delta_o = \frac{\overline{t}(q_o)(1-q_o) - \gamma}{q_o}.$$

Since, $\overline{t}(0.00255) = 0.709053$, we thus get

$$\delta_o = \frac{\overline{t}(0.00255)(1-0.00255) - 0.58}{0.00255} = \frac{0.709053(1-0.00255) - 0.58}{0.00255} = 49.9 < 50.$$

Hence in this example, the bound (59) gives only a slight improvement over the bound (60). The reason is that, for Bernoulli random variables with parameter close to 0.5, the Höffding's inequality is almost as good as the one given by the large deviations principle.

Acknowledgements The author "Jüri Lember" was supported by the Estonian Science Foundation Grant nr. 9288 and targeted financing project SF0180015s12. The research work of "Felipe Torres" was supported by the DFG through the SFB 878 at University of Münster.

References

1. K.S. Alexander, The rate of convergence of the mean length of the longest common subsequence. Ann. Appl. Probab. **4**(4), 1074–1082 (1994)
2. F. Bonetto, H. Matzinger, Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. Latin Am. J. Probab. Math. **2**, 195–216 (2006)

3. J. Boutet de Monvel, Extensive simulations for longest common subsequences. Eur. Phys. J. B **7**, 293–308 (1999)
4. N. Christianini, M.W. Hahn, *Introduction to Computational Genomics* (Cambridge University Press, Cambridge, 2007)
5. V. Chvatal, D. Sankoff, Longest common subsequences of two random sequences. J. Appl. Probab. **12**, 306–315 (1975)
6. L. Devroye, G. Lugosi, L. Gyorfi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1996)
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998)
8. C. Durringer, J. Lember, H. Matzinger, Deviation from the mean in sequence comparison with a periodic sequence. ALEA **3**, 1–29 (2007)
9. C. Houdre, H. Matzinger, Fluctuations of the optimal alignment score with and asymmetric scoring function. [arXiv:math/0702036]
10. J. Lember, H. Matzinger, Standard deviation of the longest common subsequence. Ann. Probab. **37**(3), 1192–1235 (2009)
11. J. Lember, H. Matzinger, F. Torres, The rate of the convergence of the mean score in random sequence comparison. Ann. Appl. Probab. **22**(3), 1046–1058 (2012)
12. J. Lember, H. Matzinger, F. Torres, General approach to the fluctuations problem in random sequence comparison. arXiv:1211.5072 (Submitted, 2012).
13. C.-Y. Lin, F.J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain (Association for Computational Linguistics, Stroudsburg 2004), p. 605
14. H. Matzinger, F. Torres, Fluctuation of the longest common subsequence for sequences of independent blocks. [arvix math.PR/1001.1273v3]
15. H. Matzinger, F. Torres, Random modification effect in the size of the fluctuation of the LCS of two sequences of i.i.d. blocks. [arXiv math.PR/1011.2679v2]
16. I.D. Melamed, Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons, in *Proceedings of the Third Workshop on Very Large Corpora* (Massachusetts Institute of Technology, Cambridge, 1995), pp. 184–198. http://books.google.ee/books?id=CHswHQAACAAJ
17. I.D. Melamed, Bitext maps and alignment via pattern recognition. Comput. Linguist. **25**(1), 107–130 (1999)
18. P. Pevzner, *Computational Molecular Biology*. An algorithmic approach, A Bradford Book (MIT, Cambridge, 2000)
19. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
20. T.F. Smith, M.S. Waterman, Identification of common molecular subsequences. J. Mol. Bio. **147**, 195–197 (1981)
21. M.J. Steele, An Efron-Stein inequality for non-symmetric statistics. Ann. Stat. **14**, 753–758 (1986)
22. F. Torres, On the probabilistic longest common subsequence problem for sequences of independent blocks. Ph.D. thesis, Bielefeld University, 2009. Online at http://bieson.ub.uni-bielefeld.de/volltexte/2009/1473/
23. M.S. Waterman, Estimating statistical significance of sequence alignments. Phil. Trans. R. Soc. Lond. B **344**(1), 383–390 (1994)
24. M.S. Waterman, *Introduction to Computational Biology* (Chapman & Hall, London, 1995)
25. K. Wing Li, C.C. Yang, Automatic construction of english/chinese parallel corpora. J. Am. Soc. Inform. Sci. Tech. **54**, 730–742 (2003)

Some Approximation Problems in Statistics and Probability

Yuri V. Prokhorov and Vladimir V. Ulyanov

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract We review the results about the accuracy of approximations for distributions of functionals of sums of independent random elements with values in a Hilbert space. Mainly we consider recent results for quadratic and almost quadratic forms motivated by asymptotic problems in mathematical statistics. Some of the results are optimal and could not be further improved without additional conditions.

Keywords Accuracy of approximations • Eigenvalues of covariance operator • Hilbert space • Power divergence family of statistics • Quadratic forms of random elements • Short asymptotic expansions

2000 *Mathematics Subject Classification.* Primary 62E20, 62H10; Secondary 52A20

Y.V. Prokhorov
Steklov Mathematical Institute, Russian Academy of Sciences, Moscow 119991, Russia
e-mail: tvp@mi.ras.ru

V.V. Ulyanov (✉)
Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow 119991, Russia
e-mail: vulyan@gmail.com

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_11,

1 Quadratic Forms

All three gems in probability theory—the law of large numbers, the central limit theorem and the law of the iterated logarithm—concern the asymptotic behavior of the sums of random variables. It would be natural to extend the results to functionals of the sums, in particular to quadratic forms. Moreover, in mathematical statistics there are numerous asymptotic problems which can be formulated in terms of quadratic or almost quadratic forms. In this article we review the corresponding results with rates of convergence. Some of these results are optimal and could not be further improved without additional conditions. The review does not pretend to completely illuminate the present state of the area under consideration. It reflects mainly the authors interests.

Let $X, X_1, X_2, \ldots$ be independent identically distributed random elements with values in a real separable Hilbert space H. The dimension of H, say $dim(H)$, could be either infinite or finite. Let (x, y) for $x, y \in H$ denote the inner product in H and put $|x| = (x, x)^{1/2}$. We assume that $\mathbf{E}|X_1|^2 < \infty$ and denote by V the covariance operator of X_1:

$$(Vx, y) = \mathbf{E}(X_1 - \mathbf{E}X_1, x)(X_1 - \mathbf{E}X_1, y).$$

Let $\sigma_1^2 \geq \sigma_2^2 \geq \ldots$ be the eigenvalues of V and let $e_1, e_2, \ldots$ be the corresponding eigenvectors which we assume to be orthonormal.

For any integer $k > 0$ we put

$$c_k(V) = \prod_1^k \sigma_i^{-1}, \quad \overline{c}_k(V) = (\prod_1^k \sigma_i^{-1})^{(k-1)/k}. \tag{1}$$

In what follows we use c and $c(\cdot)$, with or without indices, to denote the absolute constants and the constants depending on parameters in brackets. Except for $c_i(V)$ and $\overline{c}_i(V)$ the same symbol may be used for various constants.

We define

$$S_n = n^{-1/2}\sigma^{-1}\sum_{i=1}^n (X_i - \mathbf{E}X_i),$$

where $\sigma^2 = \mathbf{E}|X_1 - \mathbf{E}X_1|^2$. Without loss of generality we may assume that $\mathbf{E}X_1 = 0$ and $\mathbf{E}|X_1|^2 = 1$. The general case can be reduced to this one considering $(X_i - \mathbf{E}X_i)/\sigma$ instead of X_i, $i = 1, 2, \ldots$. Let Y be H-valued Gaussian $(0, V)$ random element. We denote the distributions of S_n and Y by P_n and Q respectively.

The central limit theorem asserts that

$$P_n(B) - Q(B) \to 0$$

for any Borel set B in H provided $Q(\partial B) = 0$, where ∂B is the boundary of B. The estimate of the rate of convergence in the central limit theorem is an estimate of the quantity $\sup_{\mathcal{A}} |P_n(A) - Q(A)|$ for various classes $\mathcal{A}$ of measurable sets A.

The most famous is the Berry-Esseen bound (see [5, 9]) when $H = \mathbf{R}$, i.e. $dim(H) = 1$, and $\mathcal{A} = \mathcal{A}_1 = \{(-\infty, x),\ x \in \mathbb{R}\}$:

$$\sup_{\mathcal{A}_1} |P_n(A) - Q(A)| \le c\, \frac{\mathbf{E}|X_1|^3}{\sqrt{n}}. \tag{2}$$

The bound is optimal with respect to dependence on n and moments of X_1. The lower bound for the constant c in (2) is known (see [11]):

$$c \ge \frac{3 + \sqrt{10}}{6\sqrt{2\pi}} = 0.40\ldots$$

The present upper bounds for c : $c \le 0.47\ldots$ (see [41, 43]) still differ from the lower bound slightly.

In the multidimensional case when $H = \mathbf{R}^d$, i.e. $dim(H) = d > 1$, it is possible to extend the class $\mathcal{A}$ to the class of all convex Borel sets in H and to get a bound (see e.g. [2, 33])

$$\sup_{\mathcal{A}} |P_n(A) - Q(A)| \le c\, \sqrt{d}\, \frac{\mathbf{E}|X_1|^3}{\sigma_d^3 \sqrt{n}}.$$

If we consider an infinite dimensional space H and take $\mathcal{A}$ as the class of all half-spaces in H then one can show (see e.g. pp. 69–70 in [34]) that there exists a distribution of X_1 such that

$$\sup_{\mathcal{A}} |P_n(A) - Q(A)| \ge 1/2. \tag{3}$$

Therefore, in the infinite dimensional case we can construct upper bound for $\sup_{\mathcal{A}} |P_n(A) - Q(A)|$ provided that $\mathcal{A}$ is a relatively narrow class, e.g. the class of all balls $B(a, x) = \{y : y \in H \text{ and } |y - a|^2 \le x\}$ with fixed center a or the class of all balls with fixed bounded radius $\sqrt{x}$. However, the good news are that the numerous asymptotic problems in statistics can be reformulated in terms of these or similar classes (see e.g. Sect. 2).

Put for any $a \in H$

$$F(x) = P_n(B(a, x)),\ \ F_0(x) = Q(B(a, x)),\ \ \delta_n(a) = \sup_x |F(x) - F_0(x)|.$$

According to (3) it is impossible to prove upper bound for $\sup_a \delta_n(a)$ which tends to 0 as $|a| \to \infty$. The upper bound for $\delta_n(a)$ should depend on a and becomes in general bad as $|a|$ grows.

The history of constructing bounds for $\delta_n(a)$ in the infinite dimensional case can be divided roughly into three phases: proving bounds with optimal

- Dependence on n;
- Moment conditions;
- Dependence on the eigenvalues of V.

The first phase started in the middle of 1960s in the twentieth century with bounds of logarithmic order for $\delta_n(a)$ (see [27]) and ended with the result:

$$\delta_n(a) = \mathcal{O}(n^{-1/2}),$$

due to Götze [12], which was based on a Weyl type symmetrization inequality (see Lemma 3.37 (i) in [12]):

Let X, Y, Z be the independent random elements in H. Then

$$\mathbf{E}\exp\{i\tau|X+Y+Z|^2\} \le (\mathbf{E}\exp\{2i\tau(\widetilde{X},\widetilde{Y})\})^{1/4}, \tag{4}$$

where $\widetilde{X}$ is the symmetrization of X, i.e $\widetilde{X} = X - X'$ with independent and identically distributed X and X'. The main point of the inequality is that it enables us to reduce the initial problem with non-linear dependence on X in power of exp to linear one. The inequality since then has been successfully applied and developed by a number of the authors.

The second phase of the history finished with a paper by Yurinskii [48] who proved

$$\delta_n(a) \le \frac{c(V)}{\sqrt{n}}\,(1+|a|^3)\,\mathbf{E}|X_1|^3,$$

where $c(V)$ denotes a constant depending on V only. The Yurinskii result has the optimal dependence on n under minimal moment condition but dependence of $c(V)$ on characteristics of the operator V was still unsatisfactory.

At the end of the third phase it was proved (see [28, 36, 39])

$$\delta_n(a) \le \frac{c\,c_6(V)}{\sqrt{n}}\,(1+|a|^3)\,\mathbf{E}|X_1|^3, \tag{5}$$

where $c_6(V)$ is defined in (1). It is known (see Example 3 in [38]) that for any $c_0 > 0$ and for any given eigenvalues $\sigma_1^2, \dots, \sigma_6^2 > 0$ of a covariance operator V there exist a vector $a \in H = \mathbf{R}^7$, $|a| > c_0$, and a sequence $X_1, X_2, \dots$ of i.i.d. random elements in $H = \mathbf{R}^7$ with zero mean and covariance operator V such that

$$\liminf_{n\to\infty} \sqrt{n}\,\delta_n(a) \ge c\,c_6(V)\,(1+|a|^3)\,\mathbf{E}|X_1|^3. \tag{6}$$

Due to (6) the bound (5) is the best possible in case of the finite third moment of $|X_1|$. For further refinements see e.g. [40]. For the results for the case of non-identically distributed random elements in H see [44].

At the same time better approximations for $F(x)$ are available when we use for approximation an additional term, say $F_1(x)$, of its asymptotic expansion. This term $F_1(x)$ is defined as the unique function satisfying $F_1(-\infty) = 0$ with the Fourier-Stieltjes transform equal to

$$\hat{F}_1(t) = -\frac{2t^2}{3\sqrt{n}}\mathbf{E}e\{t|Y-a|^2\}\left(3(X,Y-a)|X|^2 + 2it(X,Y-a)^3\right). \tag{7}$$

Here and in the following X and Y are independent and we write $e\{x\} = \exp\{ix\}$.

In case $dim(H) < \infty$ the term $F_1(x)$ can be defined in terms of the density function of the normal distribution (see [6]). Let φ denote the standard normal density in $\mathbf{R}^d$. Then the density function $p(y)$ of the normal distribution Q is defined by $p(y) = \varphi(V^{-1/2}y)/\sqrt{\det V}$, $y \in \mathbf{R}^d$. We have

$$F_1(x) = \frac{1}{6\sqrt{n}}\,\chi(A_x), \quad A_x = \{u \in \mathbf{R}^d : |u-a|^2 \le x\}$$

with the signed measure

$$\chi(A) = \int_A \mathbf{E}\, p'''(y)\, X^3\, dy \quad \text{for the Borel sets} \quad A \subset \mathbf{R}^d$$

and

$$p'''(y)\, u^3 = p(y)(3(V^{-1}u,u)(V^{-1}y,u) - (V^{-1}y,u)^3)$$

is the third Frechet derivative of p in the direction u.

Introduce the error

$$\Delta_n(a) = \sup_x |F(x) - F_0(x) - F_1(x)|.$$

Note, that $\hat{F}_1(t) = 0$ and hence $F_1(x) = 0$ when $a = 0$ or X has a symmetric distribution, i.e. when X and $-X$ are identically distributed. Therefore, we get

$$\Delta_n(0) = \delta_n(0).$$

Similar to the developments of the bounds for $\delta_n(a)$ the first task consisted in deriving the bounds for $\Delta_n(a)$ with the optimal dependence on n. Starting with a seminal paper by Esseen [10] for the finite dimensional spaces $H = \mathbf{R}^d$, $d < \infty$, who proved

$$\Delta_n(0) = \mathcal{O}(n^{-d/(d+1)}), \tag{8}$$

a comparable bound

$$\Delta_n(0) = \mathcal{O}(n^{-\gamma})$$

with $\gamma = 1 - \varepsilon$ for any $\varepsilon > 0$ was finally proved in [12, 13], based on the Weyl type inequalities mentioned above. Further refinements and generalizations in the case $a \neq 0$ and $\gamma < 1$ are due to Nagaev and Chebotarev [29], Sazonov et al. [35].

Note however, that the results in the infinite dimensional case did not even yield (8) as corollary when $\sigma_{d+1} = 0$, i.e. $dim(H) = d$. Only 50 years after Esseen's result the optimal bounds (in n) were finally established in [3]

$$\Delta_n(0) \le \frac{c(9,V)}{n}\,\mathbf{E}|X_1|^4, \tag{9}$$

$$\Delta_n(a) \le \frac{c(13,V)}{n}\,(1+|a|^6)\,\mathbf{E}|X_1|^4, \tag{10}$$

where $c(i,V) \le \exp\{c\sigma_i^{-2}\}$, $i = 9, 13$, and in the case of the bound (9) it was additionally assumed that the distribution of X_1 is symmetric. In order to derive these bound new techniques were developed, in particular the so-called multiplicative inequality for the characteristic functions (see Lemma 3.2, Theorem 10.1 and formulas (10.7)–(10.8) in [4]):

Let $\varphi(t)$, $t \ge 0$, denote a continuous function such that $0 \le \varphi \le 1$. Assume that

$$\varphi(t)\,\varphi(t+\tau) \le \theta\mathcal{M}^d(\tau, N)$$

for all $t \ge 0$ and $\tau > 0$ with some $\theta \ge 1$ independent of t and τ, where

$$\mathcal{M}(t,n) = 1/\sqrt{|t|n} + \sqrt{|t|} \ \text{ for } \ |t| > 0.$$

Then for any $0 < B \le 1$ and $N \ge 1$

$$\int_{B/\sqrt{N}}^{1} \frac{\varphi(t)}{t}\,dt \le c(s)\,\theta\,(N^{-1} + (B\sqrt{N})^{-d/2}) \quad \text{for} \quad d > 8. \tag{11}$$

The previous Weyl type inequality (4) gave the bounds for the integrals

$$\int_{D(n,\gamma)} \frac{\mathbf{E}e\{t|S_n - a|^2\}}{|t|}\,dt$$

for the areas $D(n,\gamma) = \{t : n^{1/2} < |t| \le n^{\gamma}\}$ with $\gamma < 1$ only, while (11) enables to extend the areas of integration up to $\gamma = 1$.

The bounds (9) and (10) are optimal with respect to the dependence on n [14] and on the moments. The bound (9) improves as well Esseen's result (8) for the Euclidean spaces $\mathbf{R}^d$ with $d > 8$. However, the dependence on covariance operator V in (9), (10) could be improved. Nagaev and Chebotarev [30] considered the case $a = 0$ and got a bound of type (9) replacing $c(9,V)$ by the following function $c(V)$:

$$c(V) = c\,\left(\overline{c}_{13}(V) + (c_9(V))^{4/9}\sigma_9^{-6}\right),$$

where $\overline{c}_{13}(V)$ and $c_9(V)$ are defined by (1). The general case $a \ne 0$ was considered in [31] (see their Theorem 1.2). The Nagaev and Chebotarev results improve the dependence on the eigenvalues of V (compared to (10)) but still require that $\sigma_{13} > 0$ instead of the weaker condition $\sigma_9 > 0$ in (9). However, it follows from Lemma 2.6 in [17] that for any given eigenvalues $\sigma_1^2, \ldots, \sigma_{12}^2 > 0$ of a covariance operator V

there exist $a \in H = \mathbf{R}^{13}$, $|a| > 1$, and a sequence $X_1, X_2, \ldots$ of i.i.d. random elements in $H = \mathbf{R}^{13}$ with zero mean and covariance operator V such that

$$\liminf_{n\to\infty} n\ \Delta_n(a) \geq c\ c_{12}(V)\ (1 + |a|^6)\ \mathbf{E}|X_1|^4. \tag{12}$$

The bound with dependence on 12 largest eigenvalues of the operator V was obtained only in [46] (for the first version see Corollary 1.3 in [17]). Moreover, in [46] the dependence on the eigenvalues is given in the bound in the explicit form which coincides with the form given by the lower bound (12):

Theorem 1.1. *There exists an absolute constant c such that for any $a \in H$*

$$\Delta_n(a) \leq \frac{c}{n} \cdot c_{12}(V) \cdot \big(\mathbf{E}|X_1|^4 + \mathbf{E}(X_1, a)^4\big) \times\big(1 + (Va, a)\big), \tag{13}$$

where $c_{12}(V)$ is defined in (1).

According to the lower bound (12) the estimate (13) is the best possible in the following sense:

- It is impossible that $\Delta_n(a)$ is of order $\mathcal{O}(n^{-1})$ *uniformly* for all distributions of X_1 with arbitrary eigenvalues $\sigma_1^2, \sigma_2^2, \ldots$;
- The form of the dependence of the right-hand side in (13) on the eigenvalues of V, on n and on $\mathbf{E}|X_1|^4$ coincides with one given in the lower bound.

For earlier versions of this result on the optimality of 12 eigenvalues and a detailed discussion of the connection of the rate problems in the central limit theorem with classical lattice point problems in analytic number theory, see the ICM-1998 Proceedings paper by Götze [14], and also Götze and Ulyanov [17].

Note however, that in the special 'symmetric' cases of the distribution of X_1 or of the center, say a, of the ball, the number of the eigenvalues which are necessary for optimal bounds may well decrease below 12. For example, when $\mathbf{E}(X, b)^3 = 0$ for all $b \in H$, by Corollary 2.7 in [17], for any given eigenvalues $\sigma_1^2, \ldots, \sigma_8^2 > 0$ of a covariance operator V there exists a center $a \in H = \mathbf{R}^9$, $|a| > 1$, and a sequence $X, X_1, X_2, \ldots$ of i.i.d. random elements in $H = \mathbf{R}^9$ with zero mean and the covariance operator V such that

$$\liminf_{n\to\infty} n\ \Delta_n(a) \geq c\ c_8(V)\ (1 + |a|^4)\ \mathbf{E}|X_1|^4.$$

Hence, in this case an upper bound of order $\mathcal{O}(n^{-1})$ for $\Delta_n(a)$ has to involve at least the *eight* largest eigenvalues of V.

Furthermore, lower bounds for $n\Delta_n(a)$ in the case $a = 0$ are not available. A conjecture, see [14], said that in that case the five first eigenvalues of V suffice. That conjecture was confirmed in Theorem 1.1 in [19] with result $\Delta_n(0) = \mathcal{O}(n^{-1})$ provided that $\sigma_5 > 0$ only. Note that for some centered ellipsoids in $\mathbf{R}^d$ with $d \geq 5$

the bounds of order $\mathcal{O}(n^{-1})$ were obtained in [18]. Moreover, it was proved recently (see Corollary 2.4 in [20]) that even for $a \neq 0$ we have $\Delta_n(a) = \mathcal{O}(n^{-1})$ when $H = \mathbf{R}^d$, $5 \leq d < \infty$, and the upper bound for $\Delta_n(a)$ is written in the explicit form and depends on the smallest eigenvalue σ_d (see Theorem 1.4 in [21] as well). It is necessary to emphasize that (13) implies $\Delta_n(a) = \mathcal{O}(n^{-1})$ for general *infinite* dimensional space H with dependence on the first twelve eigenvalues of V only.

The proofs of the recent results due to Götze, Ulyanov and Zaitsev are based on the reduction of the original problem to lattice valued random vectors and on the symmetrization techniques developed in a number of papers, see e.g. Götze [12], Yurinskii [48], Sazonov et al. [35–37], Götze and Ulyanov [17], Bogatyrev et al. [7]. In the proofs we use also the new inequalities obtained in Lemma 6.5 in [20] and in [16] (see Lemma 8.2 in [20]). In fact, the bounds in [20] are constructed for more general quadratic forms of the type $(\mathbb{Q}x, x)$ with non-degenerate linear symmetric bounded operator in $\mathbf{R}^d$.

One of the basic lemma to prove (13) is the following (see Lemma 2.2 in [17]):

Let $T > 0$, $b \in \mathbf{R}^1$, $b \neq 0$, l be an integer, $l \geq 1$, $Y = (Y_1, \ldots, Y_{2l})$ be a Gaussian random vector with values in $\mathbf{R}^{2l}$; $Y_1, \ldots, Y_{2l}$ be independent and $\mathbf{E}Y_i = 0$, $\mathbf{E}Y_i^2 = \sigma_i^2$ for $i = 1, 2, \ldots, 2l$; $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_{2l}^2 > 0$ and $a \in \mathbf{R}^{2l}$. Then there exists a positive constant $c = c(l)$ such that

$$\Big| \int_{-T}^{T} s^{l-1} \mathbf{E} \exp\{is|Y + a|^2\} e^{ibs} ds \Big| \leq c \prod_{j=1}^{2l} \sigma_j^{-1}.$$

For non-uniform bounds with 12 eigenvalues of covariance operator V see [7].

For estimates for the characteristic functions of polynomials (of order higher than 2) of asymptotically normal random variables see [22], for related results see also [23].

2 Applications in Statistics: Almost Quadratic Forms

In this section we consider the accuracy of approximations for the distributions of sums of independent random elements in $k-1$-dimensional Euclidian space. The approximation is considered on the class of sets which are "similar" to ellipsoids. Its appearance is motivated by the study of the asymptotic behavior of the goodness-of-fit test statistics—power divergence family of statistics.

Consider a vector $(Y_1, \ldots, Y_k)^T$ with multinomial distribution $M_k(n, \pi)$, i. e.

$$\Pr(Y_1 = n_1, \ldots, Y_k = n_k) = \begin{cases} n! \prod_{j=1}^{k} (\pi_j^{n_j} / n_j!), & n_j = 0, 1, \ldots, n \ (j = 1, \ldots, k) \\ & \text{and } \sum_{j=1}^{k} n_j = n, \\ 0, & \text{otherwise,} \end{cases}$$

where $\pi = (\pi_1, \ldots, \pi_k)^T, \pi_j > 0, \sum_{j=1}^{k} \pi_j = 1$. From this point on, we will assume the validity of the hypothesis $H_0: \pi = \boldsymbol{p}$. Since the sum of n_i equals n, we can express this multinomial distribution in terms of a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_{k-1})$ and denote its covariance matrix Ω. It is known that so defined Ω equals $(\delta_i^j \; p_i - p_i p_j) \in \mathbf{R}^{(k-1)\times(k-1)}$. The main object of the current study is the power divergence family of goodness-of-fit test statistics:

$$t_\lambda(\boldsymbol{Y}) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} Y_j \left[\left(\frac{Y_j}{np_j} \right)^{\lambda} - 1 \right], \; \lambda \in \mathbf{R},$$

When $\lambda = 0, -1$, this notation should be understood as a result of passage to the limit.

These statistics were first introduced in [8] and [32]. Putting $\lambda = 1, \lambda = -1/2$ and $\lambda = 0$ we can obtain the chi-squared statistic, the Freeman-Tukey statistic, and the log-likelihood ratio statistic respectively.

We consider transformation

$$X_j = (Y_j - np_j)/\sqrt{n}, \; j = 1, \ldots, k, \; r = k-1, \; \boldsymbol{X} = (X_1, \ldots, X_r)^T.$$

Herein the vector $\boldsymbol{X}$ is the vector taking values on the lattice,

$$L = \left\{ \boldsymbol{x} = (x_1, \ldots, x_r)^T; \; \boldsymbol{x} = \frac{\boldsymbol{m} - n\boldsymbol{p}}{\sqrt{n}}, \; \boldsymbol{p} = (p_1, \ldots, p_r)^T, \; \boldsymbol{m} = (n_1, \ldots, n_r)^T \right\},$$

where n_j are non-negative integers.

The statistic $t_\lambda(\boldsymbol{Y})$ can be expressed as a function of $\boldsymbol{X}$ in the form

$$T_\lambda(\boldsymbol{X}) = \frac{2n}{\lambda(\lambda+1)} \left[\sum_{j=1}^{k} p_j \left(\left(1 + \frac{X_j}{\sqrt{n} p_j} \right)^{\lambda+1} - 1 \right) \right], \tag{14}$$

and then, via the Taylor expansion, transformed to the form

$$T_\lambda(\boldsymbol{X}) = \sum_{i=1}^{k} \left(\frac{X_i^2}{p_i} + \frac{(\lambda-1)X_i^3}{3\sqrt{np_i^2}} + \frac{(\lambda-1)(\lambda-2)X_i^4}{12p_i^3 n} + O\left(n^{-3/2}\right) \right).$$

As we see the statistics $T_\lambda(\boldsymbol{X})$ is "close" to quadratic form

$$T_1(\boldsymbol{X}) = \sum_{i=1}^{k} \frac{X_i^2}{p_i},$$

considered in Sect. 1.

We call a set $B \subset \mathbf{R}^r$ *an extended convex set*, if for for all $l = 1, \ldots, r$ it can be expressed in the form:

$$B = \{\boldsymbol{x} = (x_1, \ldots, x_r)^T : \lambda_l(x^*) < x_l < \theta_l(x^*) \text{ and } x^* = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_r)^T \in B_l\},$$

where B_l is some subset of $\mathbf{R}^{r-1}$ and $\lambda_l(x^*), \theta_l(x^*)$ are continuous functions on $\mathbf{R}^{r-1}$. Additionally, we introduce the following notation

$$[h(\boldsymbol{x})]_{\lambda_l(x^*)}^{\theta_l(x^*)} = h(x_1, \ldots, x_{l-1}, \theta_l(x^*), x_{l+1}, \ldots, x_r) - h(x_1, \ldots, x_{l-1}, \lambda_l(x^*), x_{l+1}, \ldots, x_r).$$

It is known that the distributions of all statistics in the family converge to chi-squared distribution with $k-1$ degrees of freedom (see e.g. [8], p. 443). However, more intriguing is the problem to find the rate of convergence to the limiting distribution.

For any bounded extended convex set B in [47] it was obtained an asymptotic expansion, which in [42] was converted to

$$\Pr(\boldsymbol{X} \in B) = J_1 + J_2 + O(n^{-1}). \tag{15}$$

with

$$J_1 = \int \cdots \int_B \phi(\boldsymbol{x}) \left\{1 + \frac{1}{\sqrt{n}} h_1(\boldsymbol{x}) + \frac{1}{n} h_2(\boldsymbol{x})\right\} dx, \text{ where}$$

$$h_1(\boldsymbol{x}) = -\frac{1}{2} \sum_{j=1}^{k} \frac{x_j}{p_j} + \frac{1}{6} \sum_{j=1}^{k} x_j \left(\frac{x_j}{p_j}\right)^2,$$

$$h_2(\boldsymbol{x}) = \frac{1}{2} h_1(\boldsymbol{x})^2 + \frac{1}{12}\left(1 - \sum_{j=1}^{k} \frac{1}{p_j}\right) + \frac{1}{4} \sum_{j=1}^{k} \left(\frac{x_j}{p_j}\right)^2 - \frac{1}{12} \sum_{j=1}^{k} x_j \left(\frac{x_j}{p_j}\right)^3;$$

$$J_2 = -\frac{1}{\sqrt{n}} \sum\nolimits_{l=1}^{r} n^{-(r-l)/2} \sum\nolimits_{x_{l+1} \in L_{l+1}} \cdots \sum\nolimits_{x_r \in L_r} \left[\int \cdots \int_{B_l} [S_1(\sqrt{n} x_l + n p_l) \phi(\boldsymbol{x})]_{\lambda_l(x^*)}^{\theta_l(x^*)} dx_1, \cdots, dx_{l-1}\right]; \tag{16}$$

$$L_j = \left\{ \boldsymbol{x} \colon x_j = \frac{n_j - np_j}{\sqrt{n}},\ n_j \text{ and } p_j \text{ defined as before} \right\};$$

$$S_1(x) = x - \lfloor x \rfloor - 1/2,\ \lfloor x \rfloor \text{ is the integer part of } x;$$

$$\phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{r/2}|\Omega|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \Omega^{-1} \boldsymbol{x} \right).$$

In [47] it was shown that $J_2 = O(n^{-1/2})$.

Using elementary transformations it can be easily shown that the determinant of the matrix Ω equals $\prod_{i=1}^{k} p_i$.

In [47] it was also examined the expansion for the most known power divergence statistic, which is the chi-squared statistic. Put $B^\lambda = \{\boldsymbol{x} \mid T_\lambda(\boldsymbol{x}) < c\}$. It is easy to show that B^1 is an ellipsoid, which is a particular case of a bounded extended convex set. Yarnold managed to simplify the item (16) in this simple case and converted the expansion (15) to

$$\Pr(\boldsymbol{X} \in B^1) = G_r(c) + (N^1 - n^{r/2} V^1) e^{-c/2} \Big/ \left((2\pi n)^r \prod_{j=1}^{k} p_j \right)^{1/2} + O(n^{-1}), \tag{17}$$

where $G_r(c)$ is the chi-squared distribution function with r degrees of freedom; N^1 is the number of points of the lattice L in B^1; V^1 is the volume of B^1. Using the result from Esseen [10], Yarnold obtained an estimate of the second item in (17) in the form $O(n^{-(k-1)/k})$. If we estimate the second term in (17) taking the result from Götze [15] instead of Esseen's one from Esseen [10] we get (see [18]) in the case of the Pearson chi-squared statistics, i.e. when $\lambda = 1$, that for $r \geq 5$

$$\Pr(\boldsymbol{X} \in B^1) = G_r(c) + O(n^{-1}).$$

In [42] it was shown that, when $\lambda = 0, \lambda = -1/2$, we have

$$J_1 = G_r(c) + O(n^{-1})$$

$$J_2 = (N^\lambda - n^{r/2} V^\lambda)\, e^{-c/2} \Big/ \left((2\pi n)^r \prod_{j=1}^{k} p_j \right)^{1/2} + o(1), \tag{18}$$

$$V^\lambda = V^1 + O(n^{-1}).$$

These results were expanded by Read to the case $\lambda \in \mathbf{R}$. In particular, Theorem 3.1 in [32] implies

$$\Pr\left(T_\lambda < c\right) = \Pr\left(\chi_r^2 < c\right) + J_2 + O\left(n^{-1}\right). \tag{19}$$

This reduces the problem to the estimation of the order of J_2.

It is worth mentioning that in [42] and in [32] there is no estimate for the residual in (18). Consequently, it is impossible to construct estimates of the rate of convergence of the statistics T_λ to the limiting distribution, based on the simple representation for J_2 initially suggested by Yarnold.

In [45] and in [1] the rate of convergence for the residual in (18) was obtained for any power divergence statistic. Then we constructed an estimate for J_2 based on the fundamental number theory results of Hlawka [25] and Huxley [26] about an approximation of a number of the integer points in the convex sets (more general than ellipsoids) by the Lebesgue measure of the set.

Therefore, one of the main point is to investigate the applicability of the aforementioned theorems from number theory to the set B^λ.

In [45] it is shown that $B^\lambda = \{\boldsymbol{x} \mid T_\lambda(\boldsymbol{x}) < c\}$ is a bounded extended convex (strictly convex) set. As it has been already mentioned, in accordance with the results of Yarnold [47]

$$J_2 = O\left(n^{-1/2}\right).$$

For the specific case of $r = 2$ this estimate has been considerably refined in [1]:

$$J_2 = O\left(n^{-3/4+\varepsilon}(\log n)^{315/146}\right) \tag{20}$$

with $\varepsilon = 3/4 - 50/73 < 0,0651$. As it follows from (18), the rate of convergence of J_2 to 0 cannot be better than the results in the lattice point problem for the ellipsoids in number theory, where for the case $r = 2$ we have the lower bound of the order $O\left(n^{-3/4}\log\log n\right)$ (see [24]). Therefore, the relation (20) gives for J_2 the order that is not far from the optimal one.

In [1] it was used the following theorem from Huxley [26]:

Theorem 2.2. *Let D be a two-dimensional convex set with area A, bounded by a simple closed curve C, divided into a finite number of pieces each of those being 3 times continuously differentiable in the following sense. Namely, on each piece C_i the radius of curvature ρ is positive (and not infinite), continuous, and continuously differentiable with respect to the angle of contingence ψ. Then in a set that is obtained from D by translation and linear expansion of order M, the number of integer points equals*

$$N = AM^2 + O\left(IM^K(\log M)^\Lambda\right)$$

$$K = \frac{46}{73},\ \Lambda = \frac{315}{146},$$

where I is a number depending only on the properties of the curve C, but not on the parameters M or A.

In [45] the results from Asylbekov et al. [1] were generalized to any dimension. The main reason why two cases when $r = 2$ and $r \geq 3$ are considered separately consists in the fact that for $r \geq 3$ it is much more difficult than for $r = 2$ to check

the applicability of the number theory results to B^λ. In [45] we used the following result from Hlawka [25]:

Theorem 2.3. *Let D be a compact convex set in $\mathbf{R}^m$ with the origin as its inner point. We denote the volume of this set by A. Assume that the boundary C of this set is an $(m-1)$-dimensional surface of class $\mathbf{C}^\infty$, the Gaussian curvature being non-zero and finite everywhere on the surface. Also assume that a specially defined "canonical" map from the unit sphere to D is one-to-one and belongs to the class $\mathbf{C}^\infty$. Then in the set that is obtained from the initial one by translation along an arbitrary vector and by linear expansion with the factor M the number of integer points is*

$$N = AM^m + O\left(IM^{m-2+\frac{2}{m+1}}\right)$$

where the constant I is a number dependent only on the properties of the surface C, but not on the parameters M or A.

Providing that $m = 2$, the statement of Theorem 2.3 is weaker than the result of Huxley.

The above theorem is applicable in [45] with $M = \sqrt{n}$. Therefore, for any fixed λ we have to deal not with a single set, but rather with a sequence of sets $B^\lambda(n)$ which are, however, "close" to the limiting set B^1 for all sufficiently large n (see the representation for $T_\lambda(\boldsymbol{X})$ after (14)). It is necessary to emphasize that the constant I in our case, generally speaking, is $I(n)$, i.e. it depends on n. Only having ascertained the fulfillment of the inequality

$$|I(n)| \leqslant C_0,$$

where C_0 is an absolute constant, we are able to apply Theorem 2.3 without a change of the overall order of the error with respect to n.

In [45] we prove the following estimate of J_2 in the space of any fixed dimension $r \geqslant 3$.

Theorem 2.4. *For the term J_2 from the decomposition (19) the following estimate holds*

$$J_2 = O\left(n^{-r/(r+1)}\right),\ r \geqslant 3,$$

The Theorem implies that for the statistics $t_\lambda(\boldsymbol{Y})$ and $T_\lambda(\boldsymbol{X})$ (see formula (14)) it holds that

$$\Pr(t_\lambda(\boldsymbol{Y}) < c) = \Pr(T_\lambda(\boldsymbol{X}) < c) = G_r(c) + O\left(n^{-1+\frac{1}{r+1}}\right),\ r \geqslant 3.$$

Acknowledgements The authors are partly supported by RFBR grants, No. 11-01-00515 and No. 11-01-12104. The second author is partly supported as well by CRC 701 at Bielefeld University.

References

1. Zh.A. Asylbekov, V.N. Zubov, V.V. Ulyanov, On approximating some statistics of goodness-of-fit tests in the case of three-dimensional discrete data. Siberian Math. J. **52**(4), 571–584 (2011)
2. V. Bentkus, On dependence of Berry–Esseen bounds on dimensionality. Lithuanian Math. J. **26**, 205–210 (1986)
3. V. Bentkus, F. Götze, Uniform rates of convergence in the CLT for quadratic forms in multidimensional spaces. Probab. Theor. Relat. Fields **109**, 367–416 (1997)
4. V. Bentkus, F. Goẗze, Optimal bounds in non-Gaussian limit theorems for UU-statistics. Ann. Probab. **27**(1), 454–521 (1999)
5. A.C. Berry, The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Am. Math. Soc. **49**, 122–136 (1941)
6. R.N. Bhattacharya, R. Ranga Rao, *Normal Approximation and Asymptotic Expansions* (Robert E. Krieger Publishing Co., Inc., Melbourne, 1986), pp. xiv+291. ISBN: 0-89874-690-6
7. S.A. Bogatyrev, F. Götze, V.V. Ulyanov, Non-uniform bounds for short asymptotic expansions in the CLT for balls in a Hilbert space. J. Multivariate Anal. **97**(9), 2041–2056 (2006)
8. N.A.C. Cressie, T.R.C. Read, Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B, **46**, 440–464 (1984)
9. C.G. Esseen, On the Liapounoff limit of error in the theory of probability. Ark. Mat. Astr. Fys. **28A**(9), 19 (1942)
10. C.G. Esseen, Fourier analysis of distribution functions. Acta Math. **77**, 1–125 (1945)
11. C.G. Esseen, A moment inequality with an application to the central limit theorem. Skand. Aktuarietidskr. **39**, 160–170 (1956)
12. F. Götze, Asymptotic expansion for bivariate von Mises functionals. Z. Wahrsch. Verw. Gebiete **50**, 333–355 (1979)
13. F. Götze, Expansions for von Mises functionals. Z. Wahrsch. Verw. Gebiete **65**, 599–625 (1984)
14. F. Götze, Lattice point problems and the central limit theorem in Euclidean spaces. Doc. Math. J. DMV, Extra vol. **ICM III**, 245–255 (1998)
15. F. Götze, Lattice point problems and values of quadratic forms. Inventiones mathematicae **157**, 195–226 (2004)
16. F. Götze, G.A. Margulis, Distribution of values of quadratic forms at integral points. Preprint. http://arxiv.org/abs/1004.5123 (2010)
17. F. Götze, V.V. Ulyanov, Uniform approximations in the CLT for balls in Euclidian spaces. Preprint 00-034 SFB 343, Univ.Bielefeld (2000)
18. F. Götze, V.V. Ulyanov, Asymptotic disrtribution of χ^2-type statistics. Preprint 03-033, Research group "Spectral analysis, asymptotic distributions and stochastic dynamics" (2003)
19. F. Götze, A.Yu. Zaitsev, Uniform rates of convergence in the CLT for quadratic forms. Preprint 08119. SFB 701, Univ.Bielefeld (2008)
20. F. Götze, A.Yu. Zaitsev, Explicit rates of approximation in the CLT for quadratic forms. http://arxiv.org/pdf/1104.0519.pdf (2011)
21. F. Götze, A.Yu. Zaitsev, Uniform rates of approximation by short asymptotic expansions in the CLT for quadratic forms of sums of i.i.d. random vectors. Preprint 09073 SFB 701, Univ. Bielefeld, Bielefeld, (2009); published in J.Math.Sci. (N.Y.) **176**(2), 162–189 (2011)
22. F. Götze, Yu.V. Prokhorov, V.V. Ulyanov, Estimates for the characteristic functions of polynomials of asymptotically normal random variables. (Russian) Uspekhi Mat. Nauk **51** 2(308), 3–26 (1996); translation in Russ. Math. Surv. **51**(2), 181–204 (1996)
23. F. Götze, Yu.V. Prokhorov, V.V. Ulyanov, On the smooth behavior of probability distributions under polynomial mappings. (Russian) Teor. Veroyatnost. i Primenen. **42**(1), 51–62 (1997); translation in Theor. Probab. Appl. **42**(1), 28–38 (1998)
24. G. Hardy, On Dirichlet's divisor problem. Proc. Lond. Math. Soc. **15**, 1–25 (1916)
25. E. Hlawka, Über integrale auf konvexen körpern I. Mh. Math. **54**, 1–36 (1950)

26. M.N. Huxley, Exponential sums and lattice points II. Proc. Lond. Math. Soc. **66**, 279–301 (1993)
27. N.P. Kandelaki, On limit theorem in Hilbert space. Trudy Vychisl. Centra Akad. Nauk Gruzin. SSR **11**, 46–55 (1965)
28. S.V. Nagaev, On new approach to study of distribution of a norm of a random element in a Hilbert space. Fifth Vilnius conference on probability theory and mathematical statistics. Abstracts **4**, 77–78 (1989)
29. S.V. Nagaev, V.I. Chebotarev, A refinement of the error estimate of the normal approximation in a Hilbert space. Siberian Math. J. **27**, 434–450 (1986)
30. S.V. Nagaev, V.I. Chebotarev, On the accuracy of Gaussian approximation in Hilbert space. Acta Applicandae Mathematicae **58**, 189–215 (1999)
31. S.V. Nagaev, V.I. Chebotarev, On the accuracy of Gaussian approximation in a Hilbert space. (Russian) Mat. Tr. **7**(1), 91–152 (2004); translated in Siberian Adv. Math. **15**(1), 11–73 (2005)
32. T.R.C. Read, Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics. Ann. Math. Stat. Part A **36**, 59–69 (1984)
33. V.V. Sazonov, On the multi-dimensional central limit theorem. Sankhya Ser. A **30**(2), 181–204 (1968)
34. V.V. Sazonov, *Normal Approximation – Some Recent Advances*. Lecture Notes in Mathematics, vol. 879 (Springer, Berlin, 1981)
35. V.V. Sazonov, V.V. Ulyanov, B.A. Zalesskii, Normal approximation in a Hilbert space. I, II. Theor. Probab. Appl. **33**, 207–227, 473–483 (1988)
36. V.V. Sazonov, V.V. Ulyanov, B.A. Zalesskii, A sharp estimate for the accuracy of the normal aapproximation in a Hilbert space. Theor. Probab. Appl. **33**, 700–701 (1988)
37. V.V. Sazonov, V.V. Ulyanov, B.A. Zalesskii, A precise estimate of the rate of covergence in the CLT in Hilbert space. Mat.USSR Sbornik **68**, 453–482 (1991)
38. V.V. Senatov, Four examples of lower bounds in the multidimensional central limit theorem. Theor. Probab. Appl. **30**, 797–805 (1985)
39. V.V. Senatov, On rate of convergence in the central limit theorem in a Hilbert space. Fifth Vilnius conference on probability theory and mathematical statistics. Abstracts **4**, 222 (1989)
40. V.V. Senatov, Qualitive effects in the estimates of convergence rate in the central limit theorem in multidimensional spaces, in *Proceedings of the Steklov Institute of Mathematics*, vol. 215, Moscow, Nauka (1996)
41. I.G. Shevtsova, On the absolute constants in the BerryEsseen type inequalities for identically distributed summands. Preprint http://arxiv.org/pdf/1111.6554.pdf (2011)
42. M. Siotani, Y. Fujikoshi, Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics. Hiroshima Math. J. **14**, 115–124 (1984)
43. I.S. Tyurin, Sharpening the upper bounds for constants in Lyapunov's theorem. (Russian) Uspekhi Mat. Nauk **65** 3(393), 201–201 (2010); translation in Russ. Math. Surv. **65**(3), 586–588 (2010)
44. V.V. Ulyanov, Normal approximation for sums of nonidentically distributed random variables in Hilbert spaces. Acta Sci. Math. (Szeged) **50**(3–4), 411–419 (1986)
45. V.V. Ulyanov, V.N. Zubov, Refinement on the convergence of one family of goodness-of-fit statistics to chi-squared distribution. Hiroshima Math. J. **39**(1), 133–161 (2009)
46. V.V. Ulyanov, F. Götze, Short asymptotic expansions in the CLT in Euclidian spaces: a sharp estimate for its accuracy. *Proceedings 2011 World Congress on Engineering and Technology*, vol. 1, 28 Oct–2 Nov 2011, Shanghai, China (IEEE, New York, 2011), pp. 260–262
47. J.K. Yarnold, Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. Ann. Math. Stat. **43**, 1566–1580 (1972)
48. V.V. Yurinskii, On the accuracy of normal approximation of the probability of hitting a ball. Theor. Probab. Appl. **27**, 280–289 (1982)

Part IV
Random Matrices

Moderate Deviations for the Determinant of Wigner Matrices

Hanna Döring and Peter Eichelsbacher

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract We establish a moderate deviations principle (MDP) for the log-determinant $\log|\det(M_n)|$ of a Wigner matrix M_n matching four moments with either the GUE or GOE ensemble. Further we establish Cramér-type moderate deviations and Berry-Esseen bounds for the log-determinant for the GUE and GOE ensembles as well as for non-symmetric and non-Hermitian Gaussian random matrices (Ginibre ensembles), respectively.

Keywords Cumulants • Determinant • Four Moment Theorem • Gaussian ensembles • Ginibre ensembles • Moderate deviations • Wigner random matrices

2010 *Mathematics Subject Classification.* Primary 60B20; Secondary 60F10, 15A18

1 Introduction

In random matrix theory, the *determinant* is naturally an important functional. The study of determinants of random matrices has a long history. The earlier papers focused on the determinant $\det A_n$ of a non-Hermitian iid matrix A_n, where the

H. Döring
Ruhr-Universität Bochum, Fakultät für Mathematik, NA 3/68, D-44780 Bochum, Germany
e-mail: hanna.doering@ruhr-uni-bochum.de

P. Eichelsbacher (✉)
Ruhr-Universität Bochum, Fakultät für Mathematik, NA 3/66, D-44780 Bochum, Germany
e-mail: peter.eichelsbacher@ruhr-uni-bochum.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_12,

entries of the matrix were independent random variables with mean 0 and variance 1. Szekeres and Turán [23] studied an extremal problem. Later, in a series of papers moments of the determinants were computed, see [20] and [4] and references therein. In [24], Tao and Vu proved for Bernoulli random matrices, that with probability tending to one as n tends to infinity

$$\sqrt{n!}\exp(-c\sqrt{n\log n}) \leq |\det A_n| \leq \sqrt{n!}\ \omega(n) \tag{1}$$

for any function $\omega(n)$ tending to infinity with n. This shows that almost surely, $\log|\det A_n|$ is $(\frac{1}{2}+o(1))n\ \log n$. In [11], Goodman considered the random Gaussian case, where the entries of A_n are iid standard real Gaussian variables. Here the square of the determinant can be expressed as a product of independent chi-square variables and it was proved that

$$\frac{\log(|\det A_n|) - \frac{1}{2}\log n! + \frac{1}{2}\log n}{\sqrt{\frac{1}{2}\log n}} \to N(0,1)_{\mathbb{R}}, \tag{2}$$

where $N(0,1)_{\mathbb{R}}$ denotes the real standard Gaussian (convergence in distribution). A similar analysis also works for complex Gaussian matrices, in which the entries remain jointly independent but now have the distribution of the complex Gaussian $N(0,1)_{\mathbb{C}}$. In this case a slightly different law holds true:

$$\frac{\log(|\det A_n|) - \frac{1}{2}\log n! + \frac{1}{4}\log n}{\sqrt{\frac{1}{4}\log n}} \to N(0,1)_{\mathbb{R}}. \tag{3}$$

Girko [9] stated that (2) holds for real iid matrices under the assumption that the fourth moment of the atom variables is 3. In [10] he claimed the same result under the assumption that the atom variables have bounded $(4+\delta)$-th moment. Recently, Nguyen and Vu [19] gave a proof for (2) under an exponential decay hypothesis on the entries. They also present an estimate for the rate of convergence, which is that the Kolmogorov distance of the distribution of the left hand side of (2) and the standard real Gaussian can be bounded by $\log^{-\frac{1}{3}+o(1)} n$. In our paper we will be able to improve the bound to $\log^{-\frac{1}{2}} n$ in the Gaussian case.

In the non-Hermitian iid model A_n it is a crucial fact that the rows of the matrix are jointly independent. This independence no longer holds true for *Hermitian* random matrices, which makes the analysis of determinants of Hermitian random matrices more challenging. The analogue of (1) for Hermitian random matrices was first proved in [25, Theorem 31] as a consequence of the famous Four Moment Theorem. Even in the Gaussian case, it is not simple to prove an analogue of the Central Limit Theorem (CLT) (3). The observations in [11] do not apply due to the dependence between the rows. In [18] and in [15], the authors computed the moment generating function of the log-determinant for the Gaussian unitary and Gaussian

orthogonal ensembles, respectively, and discussed the central limit theorem via the method of cumulants (see [15, (40) and Appendix D]): consider a Hermitian $n \times n$ matrix X_n in which the atom distribution ζ_{ij} are given by the complex Gaussian $N(0,1)_{\mathbb{C}}$ for $i < j$ and the real Gaussian $N(0,1)_{\mathbb{R}}$ for $i = j$ (which is called the Gaussian Unitary Ensemble (GUE)). The calculations in [15] should imply a Central Limit Theorem (see Remark 2.4 in our paper):

$$\frac{\log(|\det X_n|) - \frac{1}{2}\log n! + \frac{1}{4}\log n}{\sqrt{\frac{1}{2}\log n}} \to N(0,1)_{\mathbb{R}}, \tag{4}$$

Recently, Tao and Vu [26] presented a different approach to prove this result approximating the log-determinant as a sum of weakly dependent terms, based on analyzing a tridiagonal form of the GUE due to Trotter [27]. They have to apply stochastic calculus and a martingale central limit theorem to get their result. This method is quite different and also quite involved. More important for us, the techniques due to Tao and Vu seem not to be applicable to get finer asymptotics like Cramér-type moderate deviations, Berry-Esseen bounds and moderate deviations principles. The reason for this is the quality of the approximation by a sum of weakly dependent terms they have chosen is not sharp enough. Let us emphasize that Tao and Vu proved the CLT (4) for certain Wigner matrices, generating a Four Moment Theorem for determinants.

The aim of our paper is to use a closed formula for the moments of the determinant of a GUE matrix, giving at the same time a closed formula for the cumulant generating function of the log-determinant. We will be able to present good bounds for all cumulants. As a consequence we will obtain Cramér-type moderate deviations, Berry-Esseen bounds and moderate deviation principle (for definitions see Sect. 2) for the log-determinant of the GUE, improving results in [15] and [26]. Moreover we will obtain similar results for the GOE ensemble. Good estimates on the cumulants imply such results. To do so we apply a celebrated lemma of the theory of large deviations probabilities due to Rudzkis et al. [21, 22] as well as results on moderate deviation principles via cumulants due to the authors [6]. Applying the recent Four Moment theorem for determinants due to Tao and Vu [26], we are able to prove the moderate deviation principle and Berry-Esseen bounds for the log-determinant for Wigner matrices matching four moments with either the GUE or GOE ensemble. Moreover we will be able to prove moderate deviations results and will improve the Berry-Esseen type bounds in [19] in the cases of non-symmetric and non-Hermitian Gaussian random matrices, called Ginibre ensembles.

Remark that the first universal result of a moderate deviations principle was proved in [7] and [8] for the number of eigenvalues of a Wigner matrix, based on fine asymptotics of the variance of the eigenvalue counting function of GUE matrices, on the Four Moment theorem and on localization results.

2 Gaussian Ensembles and Wigner Matrices

Among the ensembles of $n \times n$ random matrices X_n, Gaussian orthogonal and unitary ensembles have been studied extensively and are still being investigated. Their probability densities are proportional to $\exp(-\operatorname{tr}(X_n^2))$, where tr denotes the trace. Matrices are real symmetric for the Gaussian orthogonal ensemble (GOE) and Hermitian for the Gaussian unitary ensemble (GUE). The joint distributions of eigenvalues for the Gaussian ensembles are ([1, Theorem 2.5.2], [17, Chap. 3])

$$P_{n,\beta}(\lambda_1, \ldots, \lambda_n) := \frac{1}{Z_{n,\beta}} \exp\left(-\frac{\beta}{4} \sum_{i=1}^{n} \lambda_i^2\right) \prod_{1 \le j < k \le n} |\lambda_j - \lambda_k|^\beta, \tag{5}$$

where $\beta = 1, 2$ for the orthogonal and unitary ensembles, respectively, and $Z_{n,\beta}$ is the normalizing constant, sometimes called the Mehta integral (see [1, Theorem 2.5.2, formula (2.5.4), and Corollay 2.5.9, Selberg's integral formula]).

Let us denote by X_n^β the random matrices of the two Gaussian ensembles. We are interested in the moments of $|\det X_n^\beta|$ for these ensembles, that is

$$M_{n,\beta}(s) := \langle |\det X_n^\beta|^s \rangle_\beta := \int_{\mathbb{R}^n} P_{n,\beta}(\lambda_1, \ldots, \lambda_n) \prod_{i=1}^{n} |\lambda_i|^s \, d\lambda_i .$$

All information about the distribution of $\log|\det X_n^\beta|$ can be obtained from the generating function $M_{n,\beta}(s)$. The moments of $\log|\det X_n^\beta|$ may be obtained from the coefficients in the Taylor expansion of $M_{n,\beta}$ evaluated at $s = 0$,

$$M_{n,\beta}(s) = \sum_{j \ge 0} \frac{\langle (\log|\det X_n^\beta|)^j \rangle_\beta}{j!} s^j ,$$

the corresponding cumulants $\Gamma_j(n,\beta) := (-i)^j \frac{d^j}{dt^j} \log \mathbb{E}\big[e^{it \log|\det X_n^\beta|}\big]\big|_{t=0}$ are related to the Taylor coefficients of $\log M_{n,\beta}$ via

$$\log M_{n,\beta}(s) = \sum_{j \ge 0} \frac{\Gamma_j(n,\beta)}{j!} s^j .$$

In the literature the *Mellin transform* of the probability density of $|\det X_n^\beta|$ was calculated for the Gaussian ensembles, giving an explicit formula for $M_{n,\beta}(s)$. To be more precise, if $g_{n,\beta}(\cdot)$ denotes the probability density of the determinant of a GOE or a GUE matrix and $g_{n,\beta}^+(y) := \frac{1}{2}(g_{n,\beta}(y) + g_{n,\beta}(-y))$ be the even part, the Mellin transform of $g_{n,\beta}^+$ is defined by

$$\mathcal{M}_{n,\beta}(s) := \int_0^\infty y^{s-1} g^+_{n,\beta}(y) dy.$$

For the GOE and GUE ensembles we obtain

$$\mathcal{M}_{n,\beta}(s) = \frac{1}{2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P_{n,\beta}(\lambda_1, \ldots, \lambda_n) |\lambda_1 \cdots \lambda_n|^{s-1} d\lambda_1 \cdots d\lambda_n$$

and an obvious consequence is the relation

$$M_{n,\beta}(s) = 2\mathcal{M}_{n,\beta}(s+1). \tag{6}$$

It is quite involved to calculate the Mellin transform even for the Gaussian ensembles. The case $\beta = 1$ was calculated in [15, formulas (31), (19) and (26)] (see also [17, Chap. 26.5]). Here the Mellin transform is a Pfaffian of an anti-symmetric matrix applying the method of (skew) orthogonal polynomials. With (6), for $n = 2p + 1$ one obtains

$$M_{2p+1,1}(s) = 4^{ns/2} \prod_{m=1}^{n} \frac{\Gamma(\frac{s}{2} + \frac{1}{2} + b_m^1)}{\Gamma(\frac{1}{2} + b_m^1)} \tag{7}$$

with $b_m^1 := \frac{1}{2} \lfloor \frac{m-1}{2} \rfloor + \frac{1}{4}$. If $n = 2p$ one obtains

$$M_{2p,1}(s) = 2^{\frac{(n+1)s}{2}} F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right) \frac{\Gamma((s+1)/2)\Gamma((n+1)/2)}{\Gamma(\frac{1}{2})\Gamma((n+1+s)/2)} \prod_{m=1}^{p} \frac{\Gamma(s+m+\frac{1}{2})}{\Gamma(m+\frac{1}{2})}, \tag{8}$$

where F is the (Gauß) hypergeometric function

$$F(a, b; c; z) := \sum_{m=0}^{\infty} \frac{(a)^{(m)}(b)^{(m)}}{(c)^{(m)}} \frac{z^m}{m!} \tag{9}$$

with $(x)^{(m)} := x(x+1)(x+2)\cdots(x+m-1)$ denoting the Pochhammer symbol. F is convergent for arbitrary a, b, c and for real $-1 < z < 1$. In [3], an alternative derivation for (7) and (8) is presented using terminating hypergeometric series. The case $\beta = 2$ was calculated in [18, Sect. 2]. Here a knowledge of determinants and orthogonal polynomials is needed. One obtains

$$M_{n,2}(s) = 2^{ns/2} \prod_{m=1}^{n} \frac{\Gamma(\frac{s}{2} + \frac{1}{2} + b_m^2)}{\Gamma(\frac{1}{2} + b_m^2)} \tag{10}$$

with $b_m^2 = \lfloor \frac{m}{2} \rfloor$. As a consequence of (10) we obtain the following results for the cumulants $\Gamma_j(n, 2)$ of $\log |\det X_n^2|$:

Lemma 2.1 (Bounds for the cumulants of $\log|\det X_n^2|$, GUE). *For the Gaussian unitary ensemble $\beta = 2$ we obtain*

$$\Gamma_1(n,2) = -\frac{n}{2}(1+\log 2) + \frac{n}{2}\log\big(2\lfloor n/2\rfloor\big) + \text{const} + O(1/n)$$

and

$$\sigma_2^2 := \Gamma_2(n,2) = \frac{1}{2}\log\big(2\lfloor n/2\rfloor\big) + \frac{1}{2}(\gamma + \log 2 + 1) + O(1/n),$$

where γ denotes the Euler-Mascheroni constant. Moreover for any $j \geq 3$ we have

$$\big|\Gamma_j(n,2)\big| \leq \text{const}\ j!. \tag{11}$$

Proof. Let us remark that some of our calculations can be found in [15]. We work out all the details to get good bounds on the cumulants, which is not the aim in [15]. With $\psi(x) := \frac{d}{dx}\log\Gamma(x)$ we denote the *digamma function*. From (10) we obtain

$$\Gamma_1(n,2) = \frac{d}{ds}\log M_{n,2}(s)\Big|_{s=0} = \frac{n}{2}\log 2 + \frac{1}{2}\sum_{i=1}^{n}\psi(1/2+b_i^2). \tag{12}$$

For any $n = 2k+1$ we obtain $\frac{1}{2}\sum_{i=1}^{n}\psi(1/2+b_i^2) = \sum_{j=1}^{k}\psi(1/2+j) + \frac{1}{2}\psi(\frac{1}{2})$ and for $n = 2k$ we have $\frac{1}{2}\sum_{i=1}^{n}\psi(1/2+b_i^2) = \sum_{j=1}^{k}\psi(1/2+j) + \frac{1}{2}\psi(1/2) - \frac{1}{2}\psi(\frac{n+1}{2})$. With $\Gamma(1+x) = x\Gamma(x)$ it follows that $\psi(1+x) = \psi(x) + \frac{1}{x}$ and therefore recursively $\psi(1/2+j) = \psi(1/2) + 2\Big(\sum_{l=1}^{j}\frac{1}{2l-1}\Big)$, see [14, Sect. 1.3, (1.3.9)]. Using

$$2\sum_{j=1}^{k}\sum_{l=1}^{j}\frac{1}{2l-1} = 2(k+1)\sum_{l=1}^{k}\frac{1}{2l-1} - \sum_{l=1}^{k}\frac{2l}{2l-1} = (2k+1)\Big(\sum_{l=1}^{2k}\frac{1}{l} - \sum_{l=1}^{k}\frac{1}{2l}\Big) - k$$

we obtain $\sum_{j=1}^{k}\psi(1/2+j) = k\psi(1/2) - k + (2k+1)\Big(\sum_{l=1}^{2k}\frac{1}{l} - \sum_{l=1}^{k}\frac{1}{2l}\Big)$. Applying

$$\sum_{l=1}^{n}\frac{1}{l} = \gamma + \log n + \frac{1}{2n} + O(\frac{1}{n^2}), \tag{13}$$

it follows that $(2k+1)\Big(\sum_{l=1}^{2k}\frac{1}{l} - \sum_{l=1}^{k}\frac{1}{2l}\Big) = (2k+1)\frac{1}{2}(\gamma + 2\log 2) + (2k+1)\frac{1}{2}\log k + O(\frac{1}{k})$. With $\psi(1/2) = -2\log 2 - \gamma$ we have

$$\sum_{j=1}^{k}\psi(1/2+j) + \frac{1}{2}\psi(1/2) = -k + (k+\frac{1}{2})\log k + O(\frac{1}{k}). \tag{14}$$

In the case $n = 2k$ we have to consider in addition the term $\frac{1}{2}\psi(1/2+k) = \frac{1}{2}\log k + O(\frac{1}{k})$. Summarizing we obtain for every n:

$$\Gamma_1(n,2) = -\frac{n}{2}\big(\log 2 + 1\big) + \frac{n}{2}\log(2k) + \text{const} + O(1/n).$$

From (10) and (12) we obtain for $n = 2k+1$

$$\Gamma_j(n,2) = \frac{d^j}{ds^j}\log M_{n,2}(s)\Big|_{s=0} = \frac{1}{2^j}\psi^{(j-1)}(1/2) + \frac{1}{2^{j-1}}\sum_{i=1}^{k}\psi^{(j-1)}(1/2+i) \tag{15}$$

with the *polygamma function* $\psi^{(k)}(x) := \frac{d^k}{dx^k}\log\Gamma(x)$. For $n = 2k$ one has to subtract from the right hand side the term $\frac{1}{2^j}\psi^{(j-1)}(\frac{n+1}{2})$. We remind the representation of $\Gamma(x)^{-1}$ due to Weierstrass (see for example [14, Sect. 1.3, (1.3.17)]): $\frac{1}{\Gamma(x)} = xe^{\gamma x}\prod_{k=1}^{\infty}(1+\frac{x}{k})e^{-\frac{x}{k}}$. Differentiating $-\log\Gamma(x)$ leads to

$$\psi(x) = -\gamma - \frac{1}{x} + \sum_{k=1}^{\infty}\left(\frac{1}{k} - \frac{1}{x+k}\right) = -\gamma + \sum_{n=0}^{\infty}\left(\frac{1}{n+1} - \frac{1}{x+n}\right).$$

Therefore one obtains

$$\psi^{(k)}(x) = (-1)^{k+1}\,k!\sum_{n=0}^{\infty}\frac{1}{(x+n)^{k+1}}. \tag{16}$$

It follows that

$$\begin{aligned}\sum_{i=1}^{k}\psi^{(j-1)}(1/2+i) &= (-1)^j\,(j-1)!\,2^j\sum_{i=1}^{k}\sum_{m=i}^{\infty}\frac{1}{(2m+1)^j}\\ &= (-1)^j\,(j-1)!\,2^{j-1}\left(2\sum_{i=1}^{k}\sum_{m=i}^{k}\frac{1}{(2m+1)^j} + 2\sum_{i=1}^{k}\sum_{m=k+1}^{\infty}\frac{1}{(2m+1)^j}\right)\\ &=: T_1 + T_2.\end{aligned}$$

With $2\sum_{i=1}^{k}\sum_{m=i}^{k}\frac{1}{(2m+1)^j} = \sum_{m=1}^{k}\frac{1}{(2m+1)^{j-1}} - \sum_{m=1}^{k}\frac{1}{(2m+1)^j}$ we obtain

$$\begin{aligned}T_1 = (-1)^j\,(j-1)!\,2^{j-1}\sum_{m=0}^{k}\frac{1}{(2m+1)^{j-1}} - (-1)^j\,(j-1)!\,2^{j-1}\\ -(-1)^j\,(j-1)!\,2^{j-1}\sum_{m=1}^{k}\frac{1}{(2m+1)^j}.\end{aligned}$$

Further we get

$$T_2 = (-1)^j\,(j-1)!\,2^{j-1}\,2k\sum_{m=k+1}^{\infty}\frac{1}{(2m+1)^j}.$$

Hence using (16) for $\psi^{(j-1)}$ we obtain

$$\begin{aligned}\sum_{i=1}^{k}\psi^{(j-1)}(1/2+i) &= (-1)^j\,(j-1)!\,2^{j-1}\sum_{m=0}^{k}\frac{1}{(2m+1)^{j-1}} - \frac{1}{2}\psi^{(j-1)}\Big(\frac{1}{2}\Big)\\ &\quad + (-1)^j\,(j-1)!\,2^{j-1}(2k+1)\sum_{m=k+1}^{\infty}\frac{1}{(2m+1)^j}.\end{aligned} \tag{17}$$

In particular for $j=2$, we have

$$\begin{aligned}\sum_{i=1}^{k}\psi^{(1)}(1/2+i) &= 2\Big(\frac{1}{2}\log(k)+\frac{1}{2}(\gamma+2\log(2))\Big) - \frac{1}{2}\psi^{(1)}(1/2)\\ &\quad +\frac{1}{2}(2k+1)\psi^{(1)}\Big(k+\frac{3}{2}\Big)\\ &= \log(k)+\gamma+2\log(2) - \frac{1}{2}\psi^{(1)}(1/2)+1+O\Big(\frac{1}{n}\Big).\end{aligned} \tag{18}$$

With (15) we obtain for $n=2k+1$ that

$$\Gamma_j(n,2) = (-1)^j(j-1)!\sum_{m=0}^{k}\frac{1}{(2m+1)^{j-1}} + (-1)^j(j-1)!(2k+1)\sum_{m=k+1}^{\infty}\frac{1}{(2m+1)^j}.$$

The first term is $-2^{1-j}(j-1)\psi^{(j-2)}(\frac{1}{2})+O(1/k)$. The second term is $2^{-j}(2k+1)\psi^{(j-1)}(\frac{1}{2}+k+1)$. For $n=2k$ we have to subtract $2^{-j}\psi^{(j-1)}(\frac{1}{2}+k)$. Finally we will apply some bounds for the polygamma functions $\psi^{(j)}$. Therefore we will apply the following integral-representation (see for example [14, Sect. 1.4, (1.4.12)]):

$$\begin{aligned}\psi(x) &= \log(x) - \int_0^{\infty} e^{-tx}\left(tf(t)+\frac{1}{2}\right)dt \quad \text{with}\\ f(t) &:= \left(\frac{1}{2}-\frac{1}{t}+\frac{1}{e^t-1}\right)\frac{1}{t},\ t\geq 0.\end{aligned} \tag{19}$$

Differentiating we see that for $j \geq 1$:

$$\psi^{(j)}(x) = (-1)^{j-1} j!x^{-j} + (-1)^{j-1} \int_0^\infty e^{-tx} t^j \left(tf(t) + \frac{1}{2} \right) dt. \tag{20}$$

Notice that $0 < \left(tf(t) + \frac{1}{2}\right) < 1$ for every $t \geq 0$; hence we obtain for every $x \geq 0$ and every $j \geq 1$:

$$|\psi^{(j)}(x)| \leq j!x^{-j} + j!x^{-j-1}. \tag{21}$$

Let us consider the variance $\sigma_2^2 = \Gamma_2(n, 2)$. With (21) we have $|\psi^{(1)}(1/2 + k)| \leq (\frac{1}{2} + k)^{-1} + (\frac{1}{2} + k)^{-2}$. Hence we have $\sigma_2^2 = \frac{1}{2} \sum_{i=1}^k \psi^{(1)}(1/2 + i) + \frac{1}{2}\psi(1/2) + O(1/k)$ and with (18) we obtain

$$\sigma_2^2 = \frac{1}{2} \log k + \frac{1}{2}(\gamma + 2 \log 2 + 1) + O(1/k).$$

For $j \geq 3$ the cumulants can be bounded by: With (21) we obtain

$$\begin{aligned} |\Gamma_j(n, 2)| &\leq \left| 2^{1-j}(j-1)\psi^{(j-2)}(1/2) \right| + \left| 2^{-j}(2k+1)\psi^{(j-1)}(1/2 + k + 1) \right| \\ &\quad + \left| 2^{-j}\psi^{(j-1)}(1/2 + k) \right| + O(1/k) \\ &\leq 6(j-1)! + \text{const} \left(\frac{(j-1)!}{2^{j-1}} \frac{1}{k^{j-2}} + \frac{(j-1)!}{2^{j-1}} \frac{1}{k^{j-1}} \right) \leq \text{const}(j-1)!. \end{aligned}$$

Therefore the cumulants satisfy the stated bounds. □

With some more technical effort we obtain similar results for the Gaussian orthogonal ensembles:

Lemma 2.2 (Bounds for the cumulants of $\log |\det X_n^1|$**, GOE).** *For the orthogonal Gaussian ensemble ($\beta = 1$) we obtain*

$$\Gamma_1(n, 1) = \frac{n}{2} \log\left(2\lfloor n/2 \rfloor\right) - \frac{n}{2} + \text{const} + O(1/n)$$

and

$$\sigma_1^2 := \Gamma_2(n, 1) = \log\left(2\lfloor n/2 \rfloor\right) + \frac{\gamma}{2} + 1 - 2K + \frac{\pi^2}{4} + O(1/n),$$

where K denotes Catalan's constant $K = \sum_{m=0}^\infty \frac{(-1)^m}{(2m+1)^2}$, and for any $j \geq 3$

$$\left|\Gamma_j(n, 1)\right| \leq \text{const } j!.$$

Proof. For $\beta = 1$ and $n = 2k+1$, formula (7) for the Mellin transform implies

$$\begin{aligned}\Gamma_1(n,1) &= \frac{d}{ds}\log M_{n,s}(s)\big|_{s=0} = \frac{n}{2}\log(4) + \frac{1}{2}\sum_{i=1}^{n}\psi\big(\frac{1}{2}+\frac{1}{2}\lfloor\frac{i-1}{2}\rfloor+\frac{1}{4}\big)\\ &= n\log(2) + \sum_{i=0}^{k-1}\psi\big(\frac{3}{4}+\frac{i}{2}\big) + \frac{1}{2}\psi\big(\frac{3}{4}+\frac{k}{2}\big)\\ &= n\log(2) + \frac{1}{2}\psi\big(\frac{3}{4}\big) + \sum_{i=1}^{k}\Big(\frac{1}{2}\psi\big(\frac{3}{4}+\frac{i-1}{2}\big) + \frac{1}{2}\psi\big(\frac{3}{4}+\frac{i}{2}\big)\Big).\end{aligned}$$

The last transformation is useful since we are now able to apply Legendre's duplication formula $\Gamma(z)\Gamma(z+1/2) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$ (see for example [14, Sect. 1.2]). This implies

$$\frac{1}{2}\psi(z) + \frac{1}{2}\psi\big(z+\frac{1}{2}\big) = \psi(2z) - \log(2). \tag{22}$$

With $z = 3/4 + i/2 - 1/2$ we obtain

$$\Gamma_1(n,1) = n\log(2) + \frac{1}{2}\psi\big(\frac{3}{4}\big) + \sum_{i=1}^{k}\psi\big(1/2+i\big) - k\log(2). \tag{23}$$

The summand $\frac{1}{2}\psi\big(\frac{3}{4}\big)$ equals via the same identity $\psi\big(\frac{1}{2}\big) - \log(2) - \frac{1}{2}\psi\big(\frac{1}{4}\big) = \psi\big(\frac{1}{2}\big) - \log(2) + \frac{\pi}{4} + \frac{3}{2}\log(2) + \frac{1}{2}\gamma = \frac{\pi}{4} - \frac{3}{2}\log(2) - \frac{1}{2}\gamma$. As in the GUE case, we have $\sum_{i=1}^{k}\psi\big(1/2+i\big) = -\frac{1}{2}\psi\big(\frac{1}{2}\big) - k + \big(k+\frac{1}{2}\big)\log(k) + O\big(\frac{1}{k}\big)$, see (14). Now (23) implies that

$$\Gamma_1(n,1) = \frac{n}{2}\log(n-1) - \frac{n}{2} + \frac{\pi+2}{4} + O\big(\frac{1}{n}\big).$$

The jth cumulant, $j \geq 2$, is given by

$$\begin{aligned}\Gamma_j(n,1) &= \frac{d^j}{ds^j}\log M_{n,s}(s)\big|_{s=0} = \frac{1}{2^j}\sum_{i=1}^{n}\psi^{(j-1)}\big(\frac{1}{2}+\frac{1}{2}\lfloor\frac{i-1}{2}\rfloor+\frac{1}{4}\big)\\ &= \frac{1}{2^{j-1}}\sum_{i=0}^{k-1}\psi^{(j-1)}\big(\frac{3}{4}+\frac{i}{2}\big) + \frac{1}{2^j}\psi^{(j-1)}\big(\frac{3}{4}+\frac{k}{2}\big).\end{aligned}$$

Differentiating (22) implies $\psi^{(j-1)}(2z) = \frac{1}{2^j}\psi^{(j-1)}(z) + \frac{1}{2^j}\psi^{(j-1)}\big(z+\frac{1}{2}\big)$ and therefore

$$\Gamma_j(n,1) = \frac{1}{2^j}\psi^{(j-1)}\big(\frac{3}{4}\big) + \sum_{i=1}^{k}\psi^{(j-1)}\big(1/2+i\big) \tag{24}$$

hold. The duplicity formula for $z = \frac{1}{4}$ implies $\frac{1}{4}\psi^{(1)}\big(\frac{3}{4}\big) = \psi^{(1)}\big(\frac{1}{2}\big) - \frac{1}{4}\psi^{(1)}\big(\frac{1}{4}\big)$, where $\psi^{(1)}\big(\frac{1}{4}\big) = 16\sum_{m=0}^{\infty}\frac{1}{(4m+1)^2} = 8\sum_{m=0}^{\infty}\big(\frac{1}{(2m+1)^2} + \frac{(-1)^m}{(2m+1)^2}\big) = 2\sum_{m=0}^{\infty}\frac{1}{(m+\frac{1}{2})^2} + 8\sum_{m=0}^{\infty}\frac{(-1)}{(2m+1)^2} = 2\psi^{(1)}\big(\frac{1}{2}\big) + 8K$ with Catalan's constant K, resulting in $\frac{1}{4}\psi^{(1)}\big(\frac{3}{4}\big) = \frac{\pi^2}{4} - 2K$. With (24) and (18) we can conclude

$$\begin{aligned}\Gamma_2(n,1) &= \frac{1}{4}\psi^{(1)}\big(\frac{3}{4}\big) + \sum_{i=1}^{k}\psi^{(1)}\big(1/2+i\big)\\ &= \frac{\pi^2}{4} - 2K + \log(k) + \frac{\gamma}{2} + \log(2) + 1 + O\big(\frac{1}{n}\big).\end{aligned} \tag{25}$$

For every $j \geq 3$, the first summand can be bounded using (21)

$$\big|\frac{1}{2^j}\psi^{(j-1)}\big(\frac{3}{4}\big)\big| \leq (j-1)!\big(\frac{2}{3}\big)^{j-1} + (j-1)!2\big(\frac{2}{3}\big)^{j} = (j-1)!\frac{7}{3}\big(\frac{2}{3}\big)^{j-1},$$

and the remaining sum in (24) is the same as in the GUE case: With (17) we have

$$\begin{aligned}&\sum_{i=1}^{k}\psi^{(j-1)}\big(1/2+i\big) + \frac{1}{2}\psi^{(j-1)}\big(\frac{1}{2}\big)\\ &\quad= -2(j-1)\psi^{(j-2)}\big(\frac{1}{2}\big) + (2k+1)\psi^{(j-1)}\big(1/2+k+1\big) + O\big(\frac{1}{k}\big).\end{aligned}$$

Applying (21) we obtain $\left|\sum_{i=1}^{k}\psi^{(j-1)}(1/2+i) + \frac{1}{2}\psi^{(j-1)}\big(\frac{1}{2}\big)\right| \leq \text{const}(j-1)!$, which implies the bound for the jth cumulant, $j \geq 3$.

In the case of $n = 2k$ even, we have to study the asymptotic behaviour of the hypergeometric function (see (9)): $F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right) := 1 + \sum_{m=1}^{\infty} x_m$, denoting $\frac{\big(\frac{1+s}{2}\big)^{(m)}\big(-\frac{s}{2}\big)^{(m)}}{\big(\frac{n+1+s}{2}\big)^{(m)}}\frac{1}{2^m m!}$ by x_m. Each x_m is of order $O(n^{-m})$ and, for $s \in [0,2)$ and n large enough, the hypergeometric function takes values in the interval $(-1,1)$. Therefore we can study the power series of the logarithm and get

$$\begin{aligned}\log F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right) &= \log\left(1 + \sum_{m=1}^{\infty} x_m\right)\\ &= \sum_{m=1}^{\infty} x_m + \sum_{l=2}^{\infty}(-1)^l\frac{1}{l}\big(\sum_{m=1}^{\infty} x_m\big)^l.\end{aligned}$$

We differentiate each x_m via the quotient rule and the product rule in the enumerator. Setting $s = 0$, the only remaining term in the enumerator is the one where we differentiate the factor $-\frac{s}{2}$. Thus the square of the denominator cancels out. The derivative of x_m equals a constant times $\frac{1}{2^m m!} \frac{1}{\left(\frac{n+1}{2}\right)^{(m)}}$. It follows that the sum over l is of order $O(n^{-1})$, too. Similarly we obtain that for every $j \geq 1$

$$\frac{d^j}{ds^j} \log F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right)\Big|_{s=0} = O\big(1/n\big).$$

Thus with (8) and (14) it follows that

$$\begin{aligned}
\Gamma_1(n,1) &= \frac{n+1}{2}\log(2) + \frac{d}{ds}\log F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right)\Big|_{s=0} \\
&\quad + \frac{1}{2}\psi\big(\frac{1}{2}\big) - \frac{1}{2}\psi\big(\frac{n+1}{2}\big) + \sum_{m=1}^{k} \psi\big(1/2+m\big) \\
&= \frac{n+1}{2}\log(2) + O\big(\frac{1}{n}\big) + \frac{1}{2}\psi\big(\frac{1}{2}\big) - \frac{1}{2}\psi\big(\frac{n+1}{2}\big) \\
&\quad - \frac{1}{2}\psi\big(\frac{1}{2}\big) - \frac{n}{2} + \frac{n+1}{2}\log\big(\frac{n}{2}\big) \\
&= \frac{n}{2}\log(n) - \frac{n}{2} + \frac{1}{2}\log(2) + O\big(1/n\big)
\end{aligned}$$

and by (17)

$$\begin{aligned}
\Gamma_j(n,1) &= \frac{d^j}{ds^j}\log F\left(\frac{s+1}{2}, -\frac{s}{2}; \frac{n+1+s}{2}; \frac{1}{2}\right)\Big|_{s=0} + \frac{1}{2^j}\psi^{(j-1)}\big(\frac{1}{2}\big) \\
&\quad - \frac{1}{2^j}\psi^{(j-1)}\big(\frac{n+1}{2}\big) + \sum_{m=1}^{k} \psi^{(j-1)}\big(1/2+m\big) \\
&= \frac{1}{2^j}\psi^{(j-1)}\big(\frac{1}{2}\big) - \frac{1}{2^j}\psi^{(j-1)}\big(\frac{n+1}{2}\big) + \sum_{m=1}^{k} \psi^{(j-1)}\big(1/2+m\big) + O\big(1/n\big).
\end{aligned}$$

Note that the only difference to the case $n = 2k+1$, see (24), is the summand $\frac{1}{2^j}\psi^{(j-1)}\big(\frac{n+1}{2}\big)$, which is of order $O\big(1/n\big)$. Therefore the second and higher cumulant satisfy the stated bounds for all n. □

Good estimates on cumulants imply asymptotic results for the log-determinant of GUE and GOE ensembles, respectively. Before we state our results, we remind the reader on Cramér-type moderate deviations and a moderate deviation principle. The classical result due to Cramér is the following. For independent and identically distributed random variables $X_1, \ldots, X_n$ with $\mathbb{E}(X_1) = 0$ and $\mathbb{E}(X_1^2) = 1$ such that

$\mathbb{E}e^{t_0|X_1|} \leq c < \infty$ for some $t_0 > 0$, the following expansion for tail probabilities can be proved:

$$\frac{P(W_n > x)}{1 - \Phi(x)} = 1 + O(1)(1 + x^3)/\sqrt{n}$$

for $0 \leq x \leq n^{1/6}$ with $W_n := (X_1 + \cdots + X_n)/\sqrt{n}$, Φ the standard normal distribution function, and $O(1)$ depends on c and t_0. This result is sometimes called a *large deviations relation*. Let us recall the definition of a large deviation principle (LDP) due to Varadhan, see for example [5]. A sequence of probability measures $\{(\mu_n), n \in \mathbb{N}\}$ on a topological space $\mathcal{X}$ equipped with a σ-field $\mathcal{B}$ is said to satisfy the LDP with speed $s_n \nearrow \infty$ and good rate function $I(\cdot)$ if the level sets $\{x : I(x) \leq \alpha\}$ are compact for all $\alpha \in [0, \infty)$ and for all $\Gamma \in \mathcal{B}$ the lower bound

$$\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(\Gamma) \geq - \inf_{x \in \mathrm{int}(\Gamma)} I(x)$$

and the upper bound

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(\Gamma) \leq - \inf_{x \in \mathrm{cl}(\Gamma)} I(x)$$

hold. Here $\mathrm{int}(\Gamma)$ and $\mathrm{cl}(\Gamma)$ denote the interior and closure of Γ respectively. We say a sequence of random variables satisfies the LDP when the sequence of measures induced by these variables satisfies the LDP. Formally a moderate deviation principle is nothing else but the LDP. However, we will speak about a moderate deviation principle (MDP) for a sequence of random variables, whenever the scaling of the corresponding random variables is between that of an ordinary Law of Large Numbers and that of a Central Limit Theorem.

We consider

$$W_{n,\beta} := \frac{\log|\det X_n^\beta| - \Gamma_1(n,\beta)}{\sigma_\beta} \quad \text{for} \quad \beta = 1, 2 \tag{26}$$

as well as

$$\widetilde{W}_{n,\beta} := \frac{\log|\det X_n^\beta| - \frac{n}{2}\log n + \frac{n}{2}}{\sqrt{\frac{1}{\beta}\log n}} \quad \text{for} \quad \beta = 1, 2. \tag{27}$$

Theorem 2.3. *For $\beta = 1, 2$ we can prove:*

(1) Cramér-type moderate deviations: There exist two constants C_1 and C_2 depending on β, such that the following inequalities hold true:

$$\left| \log \frac{P(W_{n,\beta} \geq x)}{1 - \Phi(x)} \right| \leq C_2 \frac{1 + x^3}{\sigma_\beta}$$

and

$$\left|\log\frac{P(W_{n,\beta}\le -x)}{\Phi(-x)}\right|\le C_2\frac{1+x^3}{\sigma_\beta}$$

for all $0\le x\le C_1\sigma_\beta$. *On all cases* σ_β *is of order* $\sqrt{\log n}$.

(2) *Berry-Esseen bounds: We obtain the following bounds:*

$$\sup_{x\in\mathbb{R}}\left|P(W_{n,\beta}\le x)-\Phi(x)\right|\le C(\beta)(\log n)^{-1/2},$$

$$\sup_{x\in\mathbb{R}}\left|P(\widetilde{W}_{n,\beta}\le x)-\Phi(x)\right|\le C(\beta)(\log n)^{-1/2}.$$

(3) *Moderate deviations principle: For any sequence* $(a_n)_n$ *of real numbers such that* $1\ll a_n\ll\sigma_\beta$ *the sequences* $\left(\frac{1}{a_n}W_{n,\beta}\right)_n$ *and* $\left(\frac{1}{a_n}\widetilde{W}_{n,\beta}\right)_n$ *satisfy a MDP with speed* a_n^2 *and rate function* $I(x)=\frac{x^2}{2}$, *respectively.*

Remark 2.4. The Berry-Esseen bound implies the Central Limit Theorem stated in (4). The statement of the central limit theorem in [15] was given differently. In section 3, they considered a variance of order $2\sigma^2=\frac{1}{\beta n}$, meaning that the spectrum of the GUE model is concentrated on a finite interval (the support of the semicircular law). Then the D is the determinant of the rescaled (!) GUE model, given a $\frac{n}{2}\log n+n\log 2$ summand in addition to the expectation $-n(\frac{1}{2}+\log 2)+O(\frac{1}{n})$ they stated in [15, (43)]. This is actually the expectation in (4). Choosing the variance $\sigma^2=\frac{1}{4n}$ in the case $\beta=2$ implies that we have to rescale each matrix-entry ζ_{ij} by $\zeta_{ij}/(2\sqrt{n})$ and hence the determinant of the rescaled matrix is $2^n n^{n/2}$ times the determinant of the matrix X_n^2.

Proof. With the bound on the cumulants (11) we obtain that $\left|\Gamma_j(W_{n,2})\right|\le 7\frac{j!}{\sigma_2^j}$. With $\sigma_2^2\ge\frac{1}{2}(\gamma+2\log 2+1)$ we get

$$\left|\Gamma_j(W_{n,2})\right|\le j!\frac{1}{\sigma_2^{j-2}}\frac{7\cdot 2}{(\gamma+2\log 2+1)}\le j!\frac{1}{\sigma_2^{j-2}}5\le j!\Big(\frac{5}{\sigma_2}\Big)^{j-2}\le\frac{j!}{\Delta^{j-2}}$$

with $\Delta=\sigma_2/5$ for all $n\ge 2$. With Lemma 2.3 in [22] one obtains

$$\frac{P\big(W_{n,2}\ge x\big)}{1-\Phi(x)}=\exp(L(x))\left(1+q_1\phi(x)\frac{x+1}{\Delta_1}\right)$$

and

$$\frac{P\big(W_{n,2}\le -x\big)}{\Phi(-x)}=\exp(L(-x))\left(1+q_2\phi(x)\frac{x+1}{\sqrt{2}\Delta_1}\right)$$

for $0\le x\le\Delta_1$, where $\Delta_1=\sqrt{2}\Delta/36$,

$$\phi(x)=\frac{60\big(1+10\Delta_1^2\exp\big(-(1-x/\Delta_1)\sqrt{\Delta_1}\big)\big)}{1-x/\Delta_1},$$

q_1, q_2 are constants in the interval $[-1, 1]$ and L is a function defined in [22, Lemma 2.3, (2.8)] satisfying $\left|L(x)\right| \leq \frac{|x|^3}{3\Delta_1}$ for all x with $|x| \leq \Delta_1$. The Cramér-type moderate deviations follow applying [7, Lemma 6.2]. The Berry-Esseen bound follows from [22, Lemma 2.1] which is

$$\sup_{x\in\mathbb{R}} \left| P\left(W_{n,2} \leq x\right) - \Phi(x)\right| \leq \frac{18}{\Delta_1} = \text{const}\, \frac{1}{(\log n)^{1/2}}.$$

The same Berry-Esseen bound follows using the asymptotic behavior of the first two moments. Finally the MDP follows from [6, Theorem 1.1] which is a MDP for $\left(\frac{1}{a_n} W_{n,2}\right)_n$ for any sequence $(a_n)_n$ of real numbers growing to infinity slow enough such that $a_n/\Delta \to 0$ as $n \to \infty$. Moreover $\left(\frac{1}{a_n} W_{n,2}\right)_n$ and $\left(\frac{1}{a_n} \widetilde{W}_{n,2}\right)_n$ are exponentially equivalent in the sense of [5, Definition 4.2.10]: with $\hat{W}_{n,2} := \frac{\log|\det X_n^2| - \frac{n}{2}\log n + \frac{n}{2}}{\sigma_2}$ we have that $|W_{n,2} - \hat{W}_{n,2}| \to 0$ as $n \to \infty$, and it follows that $\left(\frac{1}{a_n} \hat{W}_{n,2}\right)_n$ and $\left(\frac{1}{a_n} W_{n,2}\right)_n$ are exponentially equivalent. By Taylor we have $\left|\frac{1}{a_n}(\hat{W}_{n,2} - \widetilde{W}_{n,2})\right| = o(1)\, \hat{W}_{n,2}$ and hence the result follows with [5, Theorem 4.2.13]. □

Next we will consider the following class of random matrices. Consider two independent families of i.i.d. random variables $(Z_{i,j})_{1\leq i<j}$ (complex-valued) and $(Y_i)_{1\leq i}$ (real-valued), zero mean, such that $\mathbb{E}Z_{1,2}^2 = 0, \mathbb{E}|Z_{1,2}|^2 = 1$ and $\mathbb{E}Y_1^2 = 1$. Consider the (Hermitian) $n\times n$ matrix M_n with entries $M_n^*(j,i) = M_n(i,j) = Z_{i,j}$ for $i < j$ and $M_n^*(i,i) = M_n(i,i) = Y_i$. Such a matrix is called *Hermitian Wigner matrix*. The GUE matrices are the special case with complex Gaussian random variables $N(0,1)_{\mathbb{C}}$ in the upper triangular and real Gaussian random variables $N(0,1)_{\mathbb{R}}$ on the diagonal.

We say that a Wigner Hermitian matrix obeys Condition (C_1) for some constant C if one has

$$\mathbb{E}|Z_{i,j}|^C \leq C_1 \quad \text{and} \quad \mathbb{E}|Y_i|^C \leq C_1 \tag{28}$$

for some constant C_1 independent on n. Two Wigner Hermitian matrices $M_n = (\zeta_{i,j})_{1\leq i,j\leq n}$ and $M_n' = (\zeta_{i,j}')_{1\leq i,j\leq n}$ match to order m off the diagonal and to order k on the diagonal if one has

$$\mathbb{E}((\mathrm{Re}(\zeta_{i,j}))^a(\mathrm{Im}(\zeta_{i,j}))^b) = \mathbb{E}((\mathrm{Re}(\zeta_{i,j}'))^a(\mathrm{Im}(\zeta_{i,j}'))^b)$$

for all $1 \leq i \leq j \leq n$ and natural numbers $a, b \geq 0$ with $a + b \leq m$ for $i < j$ and $a + b \leq k$ for $i = j$.

Applying [26, Theorem 5], the Four Moment Theorem for the determinant, we are able to prove an MDP for the log-determinant even for a class of Wigner Hermitian matrices. For any Wigner Hermitian matrix M_n consider

$$W_n := \frac{\log|\det M_n| - \frac{1}{2}\log n! + \frac{1}{4}\log n}{\sqrt{\frac{1}{2}\log n}}.$$

Theorem 2.5 (Universal moderate deviations principle). *Let M_n be a Wigner Hermitian matrix whose atom distributions are independent of n, have real and imaginary parts that are independent and match GUE to fourth order and obey Condition (C_1), (28), for some sufficiently large C, then for any sequence $(a_n)_n$ of real numbers such that $1 \ll a_n \ll \sqrt{\log n}$ the sequence $\left(\frac{1}{a_n} W_n\right)_n$ satisfies a MDP with speed a_n^2 and rate function $I(x) = \frac{x^2}{2}$. If M_n matches GOE instead of GUE, then one instead has that $\left(\frac{\sqrt{\frac{1}{2}\log n}}{a_n \sqrt{\log n}} W_n\right)_n$ satisfies the MDP with same speed and rate function.*

Proof. Let M_n be the Wigner Hermitian matrix whose entries satisfy the conditions of the theorem and M_n' denotes the GUE matrix. Then [26, Theorem 5] says that there exists a small $c_0 > 0$ such that for all $G : \mathbb{R} \to \mathbb{R}_+$ with $\left|\frac{d^j}{dx^j} G(x)\right| = O(n^{c_0})$ for $j = 0, \ldots, 5$, we have

$$\left|\mathbb{E}\big(G(\log|\det(M_n)|)\big) - \mathbb{E}\big(G(\log|\det(M_n')|)\big)\right| \leq n^{-c_0}$$

We consider for any $b, c \in \mathbb{R}$ the interval $I_n := [b_n, c_n]$ with

$$b_n := b\, a_n \sqrt{\frac{1}{2}\log n} + \frac{1}{2}\log n! - \frac{1}{4}\log n \text{ and } c_n := c\, a_n \sqrt{\frac{1}{2}\log n} + \frac{1}{2}\log n! - \frac{1}{4}\log n$$

With $I_n^+ := [b_n - n^{-c_0/10}, c_n + n^{-c_0/10}]$ and $I_n^- := [b_n + n^{-c_0/10}, c_n - n^{-c_0/10}]$ we construct a bump function $G_n : \mathbb{R} \to \mathbb{R}_+$ which is equal to one on the smaller interval I_n^- and vanishes outside the larger interval I_n^+. It follows that $P(\log|\det(M_n)| \in I_n) \leq \mathbb{E}G_n(\log|\det(M_n)|)$ and $\mathbb{E}G_n(\log|\det(M_n')|) \leq P(\log|\det(M_n')| \in I_n^+)$. One can choose G_n to satisfy the condition $\left|\frac{d^j}{dx^j} G_n(x)\right| = O(n^{c_0})$ for $j = 0, \ldots, 5$ and hence

$$P(\log|\det(M_n)| \in I_n) \leq P(\log|\det(M_n')| \in I_n^+) + n^{-c_0}. \tag{29}$$

By the same argument we get

$$P(\log|\det(M_n')| \in I_n^-) - n^{-c_0} \leq P(\log|\det(M_n)| \in I_n). \tag{30}$$

With $P\left(\frac{1}{a_n} W_n \in [b, c]\right) = P\big(\log|\det(M_n)| \in I_n\big)$. With (29) and [5, Lemma 1.2.15] we see that

$$\limsup_{n\to\infty} \frac{1}{a_n^2} \log P\left(W_n/a_n \in [b, c]\right)$$
$$\leq \max\left(\limsup_{n\to\infty} \frac{1}{a_n^2} \log P(\log|\det(M_n')| \in I_n^+); \limsup_{n\to\infty} \frac{1}{a_n^2} \log n^{-c_0}\right).$$

For the first object we have

$$\limsup_{n\to\infty} \frac{1}{a_n^2} \log P(\log|\det(M_n')| \in I_n^+)$$
$$= \limsup_{n\to\infty} \frac{1}{a_n^2} \log P\left(\frac{1}{a_n}\widetilde{W}_{n,2} \in [b-\eta(n), c+\eta(n)]\right)$$

with $\eta(n) := n^{-c_0/10}\big(a_n\sqrt{\frac{1}{2}\log n}\big)^{-1} \to 0$ as $n \to \infty$. Since $c_0 > 0$ and $\log n/a_n^2 \to \infty$ for $n \to \infty$ by assumption, applying Theorem 2.3 we have

$$\limsup_{n\to\infty} \frac{1}{a_n^2} \log P\big(W_n/a_n \in [b,c]\big) \le -\inf_{x\in[b,c]} \frac{x^2}{2}.$$

Applying (30) we obtain in the same manner that

$$\limsup_{n\to\infty} \frac{1}{a_n^2} \log P\big(W_n/a_n \in [b,c]\big) \ge -\inf_{x\in[b,c]} \frac{x^2}{2}.$$

The conclusion follows applying [5, Theorem 4.1.11 and Lemma 1.2.18]. □

Remark 2.6. The bump function G_n in the proof of Theorem 2.5 can be chosen to fulfill $\big|\frac{d^j}{dx^j}G_n(x)\big| = O(n^{c_0})$ for $j = 0,\dots,5$ uniformly in the endpoints of the interval $[b,c]$. Hence the Berry-Esseen bound in Theorem 2.3 can be obtained for Wigner matrices considered in Theorem 2.5:

$$\sup_{x\in\mathbb{R}} \big|P(W_n \le x) - \Phi(x)\big| \le \text{const}\,\big((\log n)^{-1/2} + n^{-c_0}\big).$$

We omit the details.

3 Non-symmetric and Non-Hermitian Gaussian Random Matrices

As already mentioned, recently Nguyen and Vu proved in [19], that for A_n be an $n \times n$ matrix whose entries are independent real random variables with mean zero and variance one, the Berry-Esseen bound

$$\sup_{x\in\mathbb{R}} \big|P(W_n \le x) - \Phi(x)\big| \le \log^{-1/3+o(1)} n$$

with

$$W_n := \frac{\log(|\det A_n|) - \frac{1}{2}\log(n-1)!}{\sqrt{\frac{1}{2}\log n}} \tag{31}$$

holds true. We will prove good bounds for the cumulants of W_n in the case where the entries are Gaussian random variables. Therefore we will be able to prove Cramér-type moderate deviations and an MDP as well as a Berry-Esseen bound of order $(\log n)^{-1/2}$ (and it seems that one cannot have a rate of convergence better than this). In the Gaussian case, again the calculation of the Mellin transform is the main tool. Fortunately, the transform can be calculated much easier.

Let A_n be an $n \times n$ matrix whose entries are independent real or complex Gaussian random variables with mean zero and variance one. Denote by $A_n^\dagger$ the transpose or Hermitian conjugate of A_n according as A_n is real or complex. Then $A_n A_n^\dagger$ is positive semi-definite and its eigenvalues are real and non-negative. The positive square roots of the eigenvalues of $A_n A_n^\dagger$ are known as the singular values of A_n. One has that

$$\prod_{i=1}^{n} \lambda_i^2 = \det(A_n A_n^\dagger) = |\det A_n|^2 = \prod_{i=1}^{n} |x_i|^2,$$

where λ_i are the singular values and x_i are the eigenvalues of A_n. Now $A_n A_n^\dagger$ is called *Wishart matrix*. For the real case we consider independent $N(0,1)_{\mathbb{R}}$ distributed entries, for the complex case we assume that the real and imaginary parts are independent and $N(0,1)_{\mathbb{R}}$ distributed entries. These ensembles are called *Ginibre ensembles*. One obtains for the joint probability distribution of the eigenvalues of $A_n A_n^\dagger$ on $\mathbb{R}_+^n$ the density

$$\frac{1}{\tilde{Z}_{n,\beta}} \exp\Big(-\frac{\beta}{2}\sum_{i=1}^{n} y_i\Big) \prod_{i=1}^{n} y_i^{\beta/2-1} \prod_{i<j} |y_i - y_j|^\beta$$

with $\beta = 1$ for the real and $\beta = 2$ for the complex case and $\tilde{Z}_{n,\beta}$ being the normalizing constant (see for example [2, Chap. 7]). As a result the Gaussian joint probability density for the singular values λ_i gets transformed to

$$Q_{n,\beta}(\lambda_1, \ldots, \lambda_n) := \frac{1}{Z_{n,\beta}(n)} \exp\Big(-\frac{\beta}{2}\sum_{i=1}^{n} \lambda_i^2\Big) \prod_{i=1}^{n} \lambda_i^{\beta-1} \prod_{i<j} |\lambda_i^2 - \lambda_j^2|^\beta$$

with

$$Z_{n,\beta}(p) := \int \cdots \int \exp\Big(-\frac{\beta}{2}\sum_{i=1}^{n} \lambda_i^2\Big) \prod_{i=1}^{n} \lambda_i^{(p-n)+\beta-1} \prod_{i<j} |\lambda_i^2 - \lambda_j^2|^\beta \prod_{i=1}^{n} d\lambda_i \tag{32}$$

Now the Mellin transform of the probability density of the determinant of A_n is given by

$$\mathcal{M}_{n,\beta}(s) = \int_0^\infty \cdots \int_0^\infty |\lambda_1 \cdots \lambda_n|^{s-1} Q_{n,\beta}(\lambda_1, \ldots, \lambda_n) \prod_{i=1}^n d\lambda_i = \frac{Z_{n,\beta}(n+s-1)}{Z_{n,\beta}(n)}.$$

But using the Selberg identity of the Laguerre form, [17, formula 17.6.5], we obtain for the moment generating function $M_{n,\beta}(s) = \mathcal{M}_{n,\beta}(s-1)$:

$$M_{n,\beta}(s) = \left(\frac{2}{\beta}\right)^{ns/2} \prod_{i=1}^n \frac{\Gamma\big((s+i\,\beta)/2\big)}{\Gamma\big((i\,\beta)/2\big)}. \tag{33}$$

This formula makes even sense for $\beta = 4$, where A_n is a quaternion matrix and $A_n^\dagger$ denotes the dual of A_n (see [17, Sect. 15.4] for a discussion of the definition of a determinant in this case). We will concentrate on the real case $\beta = 1$. The results of the following theorem can be stated and proved similarly in the two other cases $\beta = 2, 4$. We omit the details. We consider W_n as in (31) and

$$\widetilde{W}_n := \frac{\log|\det A_n| - \mathbb{E}(\log|\det A_n|)}{\mathbb{V}(\log|\det A_n|)^{1/2}}. \tag{34}$$

Theorem 3.1. *Let A_n be an $n \times n$ matrix whose entries are independent real $N(0,1)_{\mathbb{R}}$ random variables. Then we have:*

(1) Cramér-type moderate deviations: There exists two constants C_1 and C_2 depending on β, such that the following inequalities hold true:

$$\left|\log \frac{P(\widetilde{W}_n \geq x)}{1-\Phi(x)}\right| \leq C_2 \frac{1+x^3}{\sigma_\beta}$$

and

$$\left|\log \frac{P(\widetilde{W}_n \leq -x)}{\Phi(-x)}\right| \leq C_2 \frac{1+x^3}{\sigma_\beta}$$

for all $0 \leq x \leq C_1 \mathbb{V}(\log|\det A_n|)^{1/2}$.

(2) Berry-Esseen bounds: We obtain the following bounds:

$$\sup_{x\in\mathbb{R}} \big|P(W_n \leq x) - \Phi(x)\big| \leq C(\beta)(\log n)^{-1/2},$$

$$\sup_{x\in\mathbb{R}} \big|P(\widetilde{W}_n \leq x) - \Phi(x)\big| \leq C(\beta)(\log n)^{-1/2}.$$

(3) Moderate deviations principle: For any sequence $(a_n)_n$ of real numbers such that $1 \ll a_n \ll \sigma_\beta$ the sequences $\left(\frac{1}{a_n} W_n\right)_n$ and $\left(\frac{1}{a_n} \widetilde{W}_n\right)_n$ satisfies a MDP with speed a_n^2 and rate function $I(x) = \frac{x^2}{2}$, respectively.

Proof. With (33) we are able to estimate the cumulants $\Gamma_j(n)$ of $\log|\det A_n|$. The calculations will benefit from a few results presented in the proofs of Lemmas 2.1

and 2.2. Therefore we restrict ourselves to the major steps of the proof. We denote by ψ the digamma function and by $\psi^{(k)}$, $k \in \mathbb{N}$, the polygamma function (see Lemma 2.1). With (33) we have

$$\Gamma_1(n) = \frac{n}{2}\log n + \frac{1}{2}\sum_{i=1}^{n}\psi(i/2) \quad \text{and} \quad \Gamma_j(n) = \frac{1}{2^j}\sum_{i=1}^{n}\psi^{(j-1)}(i/2) \text{ for } j \geq 2.$$

For $n = 2k+1$ we have $\frac{1}{2}\sum_{i=1}^{n}\psi(i/2) = \frac{1}{2}\big(\sum_{i=0}^{k}\psi(1/2+i) + \sum_{i=1}^{k}\psi(i)\big)$. Using (14) the first summand is equal to $-\frac{k}{2} + \frac{k}{2}\log k + \frac{1}{4}\log k + \frac{1}{4}\psi(1/2) + O(1/k)$. With $\psi(1+x) = \psi(x) + \frac{1}{x}$ (see Lemma 2.1) one obtains that $\psi(i) = \psi(1) + \sum_{j=1}^{i-1}\frac{1}{j}$. Thus applying (13) we have $\frac{1}{2}\sum_{i=1}^{k}\psi(i) = \frac{k}{2}\log(k-1) - \frac{k}{2} + \text{const} + O(1/k)$. Summarizing we get

$$\begin{aligned}\Gamma_1(2k+1) &= -k + k\log k + \frac{1}{4}\log k + \text{const} + O(1/k)\\ &= -\frac{n}{2}(1+\log 2) + \frac{n}{2}\log(n-1) - \frac{1}{4}\log(n-1) + \text{const} + O(1/n).\end{aligned}$$

Therefore the leading term of the expectation of $\log|\det A_n|$ is $\log\big((n-1)!\big)$. In the case $n = 2k$ one obtains the same order. For $\Gamma_j(2k+1)$ with $j \geq 2$ we proceed as following:

$$\begin{aligned}\Gamma_j(2k+1) &= \frac{1}{2^j}\sum_{i=1}^{2k+1}\psi^{(j-1)}(i/2)\\ &= \frac{1}{2^j}\left(\psi^{(j-1)}(1/2) + \sum_{i=1}^{k}\psi^{(j-1)}(1/2+i) + \sum_{i=1}^{k}\psi^{(j-1)}(i)\right).\end{aligned}$$

Take the representation (16) to see that $\psi^{(j-1)}(i) = (-1)^j(j-1)!\sum_{m=i}^{\infty}\frac{1}{m^j}$, such that

$$\begin{aligned}\sum_{i=1}^{k}\psi^{(j-1)}(i) &= (-1)^j(j-1)!\left(\sum_{m=1}^{k}\frac{1}{m^{j-1}} + k\sum_{m=k+1}^{\infty}\frac{1}{m^j}\right)\\ &= -(j-1)\psi^{(j-2)}(1) + O(1/k) + k\psi^{(j-1)}(k+1).\end{aligned}$$

With the help of (17) we obtain for $j \geq 3$ that

$$\begin{aligned}\Gamma_j(n) = {}& \frac{1}{2^{j+1}}\psi^{(j-1)}(1/2) - \frac{1}{2^j}(j-1)\big(\psi^{(j-2)}(1/2) + \psi^{(j-2)}(1)\big)\\ &+ \frac{1}{2^{j+1}}(2k+1)\psi^{(j-1)}(1/2+k+1) + \frac{1}{2^j}\,k\,\psi^{(j-1)}(k+1) + O(1/k).\end{aligned}$$

With (21) we are able to bound the cumulants in a similar way as in the proof of Lemma 2.1 and obtain $|\Gamma_j(n)| \leq \text{const}\, j!$. Moreover with (18) we obtain for the variance

$$\Gamma_2(n) = \frac{1}{2}\log n + \frac{1}{2}\big(\gamma + 1 + \frac{\pi^2}{8}\big) + O(1/n).$$

Therefore the leading term of the variance of $\log|\det A_n|$ is $\frac{1}{2}\log n$. Now the theorem follows exactly as in the proof of Theorem 2.3. □

Remark 3.2. Let A_n be an $n \times n$ matrix whose entries are independent complex and quaternion, respectively. Then W_n and $\widetilde{W}_n$ as defined before satisfy Cramér-type moderate deviations, Berry-Esseen bounds and a moderate deviations principle. This can easily be checked noting that, for $\beta = 1, 2, 4$,

$$\Gamma_j^{(\beta)}(n) = \frac{n}{2}\log\big(\frac{2}{\beta}\big)\delta_{\{j=1\}} + \frac{1}{2^j}\sum_{i=1}^{n}\psi^{(j-1)}\left(\frac{i\beta}{2}\right)$$

is of order $\frac{1}{2\beta}\log(n)$: For $\beta = 2$ we have already bounded these summands in the proof above. In the case $\beta = 4$ use (22) and its derivatives to see, that the cumulant can be represented via sums of $\psi^{(j-1)}(i)$ and $\psi^{(j-1)}(i + 1/2)$.

Remark 3.3 (Trace-fixed ensembles). In [16], the authors considered fixed-trace Gaussian random matrix ensembles (real-symmetric and Hermitian ones). Here the trace of the matrix is kept constant with no other restriction on the matrix elements. These ensembles are shown to be equivalent as far as finite moments of the matrix elements are concerned. Especially, the Mellin transform of the fixed-trace Gaussian matrices can be deduced from the Mellin transform of the Gaussian orthogonal and unitary ensemble, respectively, see [16, formulas (17), (20) and (22)]. Hence it is expected that the distribution of the log-determinant of these ensembles is asymptotically Gaussian with a variance of order $\log n$. We would be able to deduce the results in Theorem 3.1 for the Gaussian trace-fixed ensembles by the same technique. We omit the details. Remark, that universal limits for the eigenvalue correlation functions in the bulk of the spectrum for fixed trace matrix ensembles are considered in [12, 13]. In this case, the class of matrices are of nondeterminantal structure.

Acknowledgements The second author has been supported by Deutsche Forschungsgemeinschaft via SFB/TR 12.

References

1. G.W. Anderson, A. Guionnet, O. Zeitouni, *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics, vol. 118 (Cambridge University press, Cambridge, 2010)
2. T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics (Wiley, New York, 1984) MR 771294 (86b:62079)

3. G.E. Andrews, I.P. Goulden, D.M. Jackson, Determinants of random matrices and Jack polynomials of rectangular shape. Stud. Appl. Math. **110**(4), 377–390 (2003). MR 1971134 (2005g:15014)
4. A. Dembo, On random determinants. Q. Appl. Math. **47**(2), 185–195 (1989). MR 998095 (91a:62125)
5. A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications* (Springer, New York, 1998)
6. H. Döring, P. Eichelsbacher, Moderate deviations via cumulants. J. Theor. Probab. 1–26 (2012). doi: 10.1007/s10959-012-0437-0
7. H. Döring, P. Eichelsbacher, Moderate deviations for the eigenvalue counting function of Wigner matrices. Lat. Am. J. Probab. Math. Stat. Vol. X, pp. 27–44 (2013)
8. H. Döring, P. Eichelsbacher, in *Edge Fluctuations of Eigenvalues for Wigner Matrices*. High Dimensional Probability VI: The Banff volume. Progress in Probability (Springer, Berlin, 2013)
9. V.L. Girko, A central limit theorem for random determinants. Teor. Veroyatnost. i Primenen. **24**(4), 728–740 (1979). MR 550529 (82g:60035)
10. V.L. Girko, A refinement of the central limit theorem for random determinants. Teor. Veroyatnost. i Primenen. **42**(1), 63–73 (1997). MR 1453330 (98k:60034)
11. N.R. Goodman, The distribution of the determinant of a complex Wishart distributed matrix. Ann. Math. Statist. **34**, 178–180 (1963). MR 0145619 (26 #3148b)
12. F. Götze, M. Gordin, Limit correlation functions for fixed trace random matrix ensembles. Comm. Math. Phys. **281**(1), 203–229 (2008). MR 2403608 (2009d:82069)
13. F. Götze, M.I. Gordin, A. Levina, The limit behavior at zero of correlation functions of random matrices with a fixed trace. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **341**, no. Veroyatn. i Stat. 11, 68–80, 230 (2007). MR 2363585 (2009g:62077)
14. N.N. Lebedev, *Special Functions and Their Applications*, Revised English edition. Translated and edited by Richard A. Silverman (Prentice-Hall, Englewood Cliffs, 1965). MR 0174795 (30 #4988)
15. G. Le Caër, R. Delannay, Distribution of the determinant of a random real-symmetric matrix from the Gaussian orthogonal ensemble. Phys. Rev. E (3) **62**(2), part A, 1526–1536 (2000). MR 1797664 (2001m:82039)
16. G. Le Caër, R. Delannay, The distributions of the determinant of fixed-trace ensembles of real-symmetric and of Hermitian random matrices. J. Phys. A **36**(38), 9885–9898 (2003). MR 2006448 (2004h:15008)
17. M.L. Mehta, *Random Matrices*, 3rd edn. Pure and Applied Mathematics (Amsterdam), vol. 142 (Elsevier/Academic, Amsterdam, 2004)
18. M.L. Mehta, J.-M. Normand, Probability density of the determinant of a random Hermitian matrix. J. Phys. A **31**(23), 5377–5391 (1998). MR 1634820 (2000b:82018)
19. H.H. Nguyen, V. Vu, Random matrices: law of the determinant. Ann. Probab. (2013) (to appear)
20. A. Prékopa, On random determinants. I. Studia Sci. Math. Hungar. **2**, 125–132 (1967). MR 0211439 (35 #2319)
21. R. Rudzkis, L. Saulis, V. Statuljavičus, A general lemma on probabilities of large deviations. Litovsk. Mat. Sb. **18**(2), 99–116, 217 (1978). MR 0501287 (58 #18681)
22. L. Saulis, V. A. Statulevičius, *Limit Theorems for Large Deviations*. Mathematics and Its Applications (Soviet Series), vol. 73 (Kluwer, Dordrecht, 1991). Translated and revised from the 1989 Russian original. MR 1171883 (93e:60055b)
23. G. Szekeres, P. Turán, On an extremal problem in the theory of determinants. Math. Naturwiss. Am. Ungar. Akad. Wiss. **56**, 796–806 (1937)
24. T. Tao, V. Vu, On random ± 1 matrices: singularity and determinant. Random Struct. Algorithms **28**(1), 1–23 (2006). MR 2187480 (2006g:15048)

25. T. Tao, V. Vu, Random matrices: universality of local eigenvalue statistics. Acta Math. **206**, 127–204 (2011)
26. T. Tao, V. Vu, A central limit theorem for the determinant of a Wigner matrix. Adv. Math. **231**(1), 74 – 101 (2012)
27. H.F. Trotter, Eigenvalue distributions of large Hermitian matrices; Wigner's semicircle law and a theorem of Kac, Murdock, and Szegö. Adv. Math. **54**(1), 67–82 (1984). MR 761763 (86c:60055)

The Semicircle Law for Matrices with Dependent Entries

Olga Friesen and Matthias Löwe

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract We investigate the spectral distribution of random matrix ensembles with correlated entries. The matrices considered are symmetric, have real-valued entries and stochastically independent diagonals. Along the diagonals the entries may be correlated. We show that under sufficiently nice moment conditions and sufficiently strong decay of correlations the empirical eigenvalue distribution converges almost surely weakly to the semi-circle law. The present note improves an earlier result (see [Friesen and Löwe, J. Theor. Probab., 2011]) by the authors using similar techniques.

Keywords Random matrices • Dependent entries • Wigner semicircle law

2010 *Mathematics Subject Classification.* 60F05.

1 Introduction

Large-dimensional random matrices were first considered in the context of application in statistics and in theoretical physics, among others, in particular they served as a model when studying the properties of atoms with heavy nuclei.

O. Friesen · M. Löwe (✉)
Westfälische Wilhelms-Universität Münster, Fachbereich Mathematik, Einsteinstraße 62, 48149 Münster, Germany
e-mail: olga.friesen@uni-muenster.de; maloewe@math.uni-muenster.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_13,

However, nowadays the field of random matrices is considered to be interesting in its own rights, since it gave rise to many interesting results, such as Wigner's semi-circle law for the limiting spectral distribution of a symmetric or Hermitian random matrix or the Tracy-Widom distribution as the limiting distribution of the (appropriately scaled) largest eigenvalue of a random Hermitian matrix. Moreover, there are connections to many questions in pure mathematics as well as applications in a multitude of areas outside of mathematics, e.g. telecommunications.

As indicated by the results mentioned above, one of the most interesting and best studied problems, has been to investigate the properties of the eigenvalues of random matrices. The most prominent example is Wigner's semi-circle law. Wigner, in his seminal paper [17] showed that the spectral distribution of symmetric Bernoulli matrices under appropriate scaling converges to the semi-circle law. In [18] he remarked that this result may be generalized to the spectral distribution of a wide class of random matrices, among them symmetric or Hermitian random matrices with independent Gaussian entries, otherwise.

A result in this spirit was proved by Arnold [3] in the situation of symmetric or Hermitian random matrices filled with independent and identically distributed (i.i.d.) random variables with sufficiently many moments. Other generalizations of Wigner's semi-circle law concern matrix ensembles with entries drawn according to weighted Haar measures on classical (e.g., orthogonal, unitary, symplectic) groups. Such results are particularly interesting, since such random matrices also play a major role in non-commutative probability (see e.g. [10]); other applications are in graph theory, combinatorics or algebra. For a broad overview the interested reader is referred to Mehta's classical textbook [14] or to the rather recent work by Anderson et al. [2].

This note addresses a question that is much in the spirit of Arnold's generalization of the semi-circle law. Even though a couple of random matrix models include situations with stochastically correlated entries (see especially [6], where the case of random Toeplitz and Hankel matrices is treated), the dependencies are not very natural from a stochastic point of view. A generic way to construct random matrices with dependent entries could be to consider a two dimensional (stationary) random field indexed by $\mathbb{Z}^2$ with correlations that decay with the distance of the indices and to take an $n \times n$ block as entries for a random $n \times n$ matrix.

This setup would, of course, be very general, and the present note is just a first step to study the asymptotic eigenvalue distribution of such matrix ensembles. Here we will deviate from the independence assumption by considering (real) random fields with entries that may be dependent on each diagonal, but with stochastically independent diagonals. For such matrices we will prove a semi-circle law under a sufficiently fast decay of correlations along the diagonals. It should be noted, that a similar result under an arbitrarily slow decay of correlations cannot be expected, since in such a situation one should already get into the realm of Toeplitz matrices as treated in [6].

The setup of the present note may look at first glance a bit more artificial than a situation where the matrices are filled with row- or columnwise independent random variables (e.g. with row- or columnwise independent Markov chains).

Note, however, that in order to guarantee for real eigenvalues we will need to restrict ourselves to symmetric random matrices. This would imply that a matrix with rowwise independent entries above the diagonal has columnwise independent entries below it. Not only is this a rather strange setup, also can one see from simulations that their asymptotic eigenvalue distribution is probably not the semi-circle law.

It should also be remarked that the conditions in this note are weaker than those in earlier article (see [8]) and therefore the results are more general then our previous ones. Example 4.3 below shows that these weaker assumptions extend the validity of Theorem 2.2 to a set of rather natural examples that could not be treated by means of the main theorem in [8]. Moreover, we also find an indication that the rather strong moment conditions we impose (see (C1) below) may be relaxed. Theorem 2.4 below is a first step into this direction.

It also should be mentioned that a similar situation has been studied by Khorunzhy and Pastur in [12]. They consider the eigenvalue distribution of so called deformed Wigner ensembles that consist of matrices which can be written as a sum of Wigner matrix (a symmetric matrix with independent entries above the diagonal) and a deterministic matrix. It is proven that in this situation the empirical eigenvalue density converges in probability to a non-random limit. This setup, yet similar, is different from ours.

The question, whether stochastically dependent entries could be allowed in order for the semi-circle law to hold, is not new. For example, Bai [4] p. 626 raises the question of whether Wigner's theorem is still holding true when the independence condition in the Wigner matrix is weakened. Also Götze and Tikhomirov [9], Hofmann-Credner and Stolz [11], and Schenker and Schulz-Baldes [16] consider a situation where the entries of a random matrix are in a natural way stochastically dependent. However, their conditions do not cover our situation.

On the other hand, some extra conditions apart from a weak dependence structure are necessary. Indeed, Anderson and Zeitouni [1] show that convergence to the semicircle law does not hold in general under finite range of dependence.

The rest of the note is organized as follows. In the second section we will formalize the situation we want to consider and state our main result. Section 3 is devoted to the proof, that is based on a moment method. These naturally lead to some combinatorial problems, that need to be solved. Section 4 contains some examples. In particular we consider Gaussian random fields as well as Markov chains on a finite state space.

2 The Main Result

In this section we will state our main theorem, a semi-circle law for symmetric random matrices with independent diagonals (for a precise formulation see Theorem 2.2 below). These random matrices are constructed as follows:

Let $\{a_n(p,q), 1 \le p \le q < \infty\}$ be a real valued random field. For any $n \in \mathbb{N}$, define the symmetric random $n \times n$ matrix $\mathbf{X}_n$ by

$$\mathbf{X}_n(q,p) = \mathbf{X}_n(p,q) = \frac{1}{\sqrt{n}} a_n(p,q), \qquad 1 \le p \le q \le n,$$

We will have to impose the conditions on $\mathbf{X}_n$. To be able to formulate them introduce

$$m_k := \sup_{n\in\mathbb{N}} \max_{1\le p\le q\le n} \mathbb{E}\left[|a_n(p,q)|^k\right] \quad k \in \mathbb{N}. \tag{1}$$

We will assume the following.

(C1) $\mathbb{E}[a_n(p,q)] = 0, \mathbb{E}\left[a_n(p,q)^2\right] = 1$ and

$$m_k < \infty, \quad \forall k \in \mathbb{N}. \tag{2}$$

(C2) The diagonals of $\mathbf{X}_n$, i.e. the families $\{a_n(p,p+r), p \in \mathbb{N}\}$, $r \in \mathbb{N}_0$, are independent,

(C3) The covariance of two entries on the same diagonal can be bounded by some constant depending only on their distance, i.e. for any $\tau \in \mathbb{N}_0$ there is a constant $c(\tau) \ge 0$ such that

$$|\mathrm{Cov}(a_n(p,q), a_n(p+\tau, q+\tau))| \le c(\tau), \qquad p,q \in \mathbb{N},$$

(C4) The entries on the diagonals have a quickly decaying dependency structure, which will be expressed in terms of the condition

$$\lim_{\tau\to\infty} c(\tau)\ \tau^{\varepsilon} = 0,$$

for some $\varepsilon > 0$.

Remarks 2.1. (i) The choice of the first two moments in (C1) is just for standardization, while (2) is a condition one might eventually want to relax. This, however, not that easy in general. We will see that for a weaker form of our main result the weaker condition

(C1') $\mathbb{E}[a_n(p,q)] = 0, \mathbb{E}\left[a_n(p,q)^2\right] = 1$ and

$$m_2 < \infty \tag{3}$$

suffices.

(ii) Note that condition (C4) implies that $\sum_{\tau=1}^{n} c(\tau) = o(n)$ since

$$\frac{1}{n}\sum_{\tau=1}^{n} c(\tau) \le \frac{1}{n^{\varepsilon}} \sum_{\tau=1}^{n} \frac{c(\tau)\ \tau^{\varepsilon}}{\tau} \to 0, \quad \text{as } n \to \infty.$$

(iii) In particular (C4) is an improvement over the condition of summable correlations

$$\sum_{\tau=0}^{\infty} |c(\tau)| < \infty$$

we imposed in [8].

As noted, we will show that the empirical eigenvalue distribution of the (appropriately scaled) random matrices introduced above converges to a limit law μ, the so-called semi-circle distribution. We choose its density to be concentrated on the interval $[-2, 2]$. Then its density is given by

$$\frac{d\mu}{dx} = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2} & \text{if } -2 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

To state our main theorems denote the ordered (real) eigenvalues of $\mathbf{X}_n$ by

$$\lambda_1^{(n)} \le \lambda_2^{(n)} \le \dots \lambda_n^{(n)}.$$

Let μ_n be the empirical eigenvalue distribution, i.e.

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k^{(n)}}.$$

With these notations we are able to formulate the central result of this note.

Theorem 2.2. *Assume that the symmetric random matrix $\mathbf{X}_n$ as defined above satisfies the conditions (C1), (C2), (C3), and (C4). Then, with probability* 1*, the empirical spectral distribution of $\mathbf{X}_n$ converges weakly to the standard semi-circle distribution, i.e.*

$$\mu_n \Rightarrow \mu \qquad \text{as } n \to \infty$$

both, in expectation and $\mathbb{P}$ − almost surely. Here "$\Rightarrow$" denotes weak convergence.

Remark 2.3. As stated in the introduction for the semi-circle law to hold, it is not possible to renounce condition (C4) without any replacement. To understand this, consider for example a Toeplitz matrix, that is a Hermitian matrix with identical entries on each diagonal. If the variance of the entries is positive, we clearly have

$$c(\tau) = \mathcal{O}(1).$$

Indeed, it was shown in [6] that the empirical distribution of a sequence of Toeplitz matrices tends with probability 1 to a nonrandom probability measure with unbounded support.

If we assume (C1') instead of (C1) we still obtain a convergence result.

Theorem 2.4. *Assume that the symmetric random matrix $\mathbf{X}_n$ as defined above satisfies the conditions (C1'), (C2), (C3), and (C4). Then, the empirical spectral distribution of $\mathbf{X}_n$ converges weakly in probability to the semi-circle distribution.*

3 Proof of Theorems 2.2 and 2.4

The basic tool in our proofs will be the method of moments. Naively, one could expect an approach using Stieltjes transforms to work as well (and in this case one would probably be able to work with weaker moment conditions than those imposed in (C1)). This method is rather standard in random matrix theory (see e.g. Chap. 2.4 in [2]). However, if the entries of a random matrix are not independent, this method seems to have serious technical difficulties.

The method of moments, on the other hand, also is traditional in the proof of a semicircle law. Among others, Wigner used this method to derive the semicircle distribution as the limiting spectral distribution for scaled Bernoulli matrices, see [17]. Also Arnold's generalization of the semicircle law [3] relies on the same technique. A recent example for matrices with dependent entries is [16]. Our proof is particularly inspired by the latter of these paper. An excellent reference to this method is also provided by Bose and Sen, cf. [5].

A key observation for the following is that the techniques in [8] suffice to prove Theorem 2.2 also under the weaker assumption (C4). For the reader's convenience we will give this proof next (following the lines in [8], of course).

To this end, let Y be distributed according to the semi-circle distribution. For the proof of the theorem it will be important to notice that the moments of Y are given by

$$\mathbb{E}(Y^k) = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ C_{\frac{k}{2}}, & \text{if } k \text{ is even,} \end{cases} \tag{4}$$

where

$$C_{\frac{k}{2}} = \frac{k!}{\frac{k}{2}!\left(\frac{k}{2}+1\right)!}$$

denote the Catalan numbers. Note that these moments determine the semicircle distribution uniquely.

This implies, that the weak convergence of the expected empirical distribution will follow from the convergence of the empirical moments, i.e. from the relation

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ C_{\frac{k}{2}}, & \text{if } k \text{ is even,} \end{cases}$$

where $\operatorname{tr}(\cdot)$ denotes the trace operator. The first part of the proof is to verify this convergence.

To start with, consider the set $\mathcal{T}_n(k)$ of k-tuples of consistent pairs, that is elements of the form $(P_1, \ldots, P_k)$ with $P_j = (p_j, q_j) \in \{1, \ldots, n\}^2$ satisfying $q_j = p_{j+1}$ for any $j = 1, \ldots, k$, where $k+1$ is identified with 1. Then, we have

$$\frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{X}_n^k\right)\right] = \frac{1}{n^{1+\frac{k}{2}}} \sum_{(P_1,\ldots,P_k)\in\mathcal{T}_n(k)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right].$$

Further, define $\mathcal{P}(k)$ to be the set of all partitions π of $\{1, \ldots, k\}$. Any partition π induces an equivalence relation $\sim_\pi$ on $\{1, \ldots, k\}$ by

$$i \sim_\pi j \quad :\Longleftrightarrow \quad i \text{ and } j \text{ belong to the same set of the partition } \pi.$$

We say that an element $(P_1, \ldots, P_k) \in \mathcal{T}_n(k)$ is a π-consistent sequence if

$$|p_i - q_i| = \left|p_j - q_j\right| \quad \Longleftrightarrow \quad i \sim_\pi j.$$

At this stage, the independence of the diagonals enters crucially. Indeed, due to condition (C2), this implies that $a_n(P_{i_1}), \ldots, a_n(P_{i_l})$ are stochastically independent if $i_1, \ldots, i_l$ belong to l different blocks of π. The set of all π-consistent sequences $(P_1, \ldots, P_k) \in \mathcal{T}_n(k)$ is denoted by $S_n(\pi)$. Thus, we can write

$$\frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{X}_n^k\right)\right] = \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi\in\mathcal{P}(k)} \sum_{(P_1,\ldots,P_k)\in S_n(\pi)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right].$$

Now fix a $k \in \mathbb{N}$. For any $\pi \in \mathcal{P}(k)$ let $\#\pi$ denote the number of equivalence classes of π.

We distinguish different cases. This will show that eventually only those π satisfying $\#\pi = \frac{k}{2}$ count.

First case: $\#\pi > \frac{k}{2}$.

Since π is a partition of $\{1, \ldots, k\}$, there is at least one equivalence class with a single element l. Consequently, for any sequence $(P_1, \ldots, P_k) \in S_n(\pi)$ we have

$$\mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right] = \mathbb{E}\Big[\prod_{i\neq l} a_n(P_i)\Big]\cdot\mathbb{E}\left[a_n(P_l)\right] = 0,$$

due to the independence of elements in different equivalence classes.

Hence, we obtain

$$\frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{X}_n^k\right)\right] = \frac{1}{n^{1+\frac{k}{2}}} \sum_{\substack{\pi\in\mathcal{P}(k),\\ \#\pi\leq\frac{k}{2}}} \sum_{(P_1,\ldots,P_k)\in S_n(\pi)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right].$$

Second case: $r := \#\pi < \frac{k}{2}$.

We need to calculate $\#S_n(\pi)$. To fix an element $(P_1, \dots, P_k) \in S_n(\pi)$, we first choose the pair $P_1 = (p_1, q_1)$. There are at most n possibilities to assign a value to p_1 and another n possibilities for q_1. To fix $P_2 = (p_2, q_2)$, note that the consistency of the pairs implies $p_2 = q_1$. If now $1 \sim_\pi 2$, the condition $|p_1 - q_1| = |p_2 - q_2|$ allows at most two choices for q_2. Otherwise, if $1 \not\sim_\pi 2$, we have at most n possibilities. We now proceed sequentially to determine the remaining pairs. When arriving at some index i, we check whether i is equivalent to any preceding index $1, \dots, i-1$. If this is the case, then we have at most two choices for P_i and otherwise, we have n. Since there are exactly r different equivalence classes, we can conclude that

$$\#S_n(\pi) \leq n^2 \cdot n^{r-1} \cdot 2^{k-r} \leq C \cdot n^{r+1}$$

with a constant $C = C(r,k)$ depending on r and k.

Now the uniform boundedness of the moments and the Hölder inequality together imply that for any sequence $(P_1, \dots, P_k)$,

$$|\mathbb{E}\left[a_n(P_1) \cdot \ldots \cdot a_n(P_k)\right]| \leq \left[\mathbb{E}\left|a_n(P_1)\right|^k\right]^{\frac{1}{k}} \cdot \ldots \cdot \left[\mathbb{E}\left|a_n(P_k)\right|^k\right]^{\frac{1}{k}} \leq m_k. \quad (5)$$

Consequently, taking account of the relation $r < \frac{k}{2}$, we get

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\substack{\pi \in \mathcal{P}(k),\\ \#\pi < \frac{k}{2}}} \sum_{(P_1,\dots,P_k) \in S_n(\pi)} |\mathbb{E}\left[a_n(P_1) \cdot \ldots \cdot a_n(P_k)\right]| \leq C \cdot \frac{1}{n^{1+\frac{k}{2}}} \cdot n^{r+1} = o(1).$$

Combining the calculations in the first and the second case, we can conclude that

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\substack{\pi \in \mathcal{P}(k),\\ \#\pi = \frac{k}{2}}} \sum_{(P_1,\dots,P_k) \in S_n(\pi)} \mathbb{E}\left[a_n(P_1) \cdot \ldots \cdot a_n(P_k)\right],$$

if the limits exist.

Now consider the case where k is *odd*. Since then the condition $\#\pi = \frac{k}{2}$ cannot be satisfied, the considerations above immediately yield

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = 0.$$

It remains to cope with *even* k. Denote by $\mathcal{PP}(k) \subset \mathcal{P}(k)$ the set of all pair partitions of $\{1, \dots, k\}$. In particular, $\#\pi = \frac{k}{2}$ for any $\pi \in \mathcal{PP}(k)$. On the other hand, if $\#\pi = \frac{k}{2}$ but $\pi \notin \mathcal{PP}(k)$, we can conclude that π has

at least one equivalence class with a single element and hence, as in the first case, the expectation corresponding to the π-consistent sequences will become zero. Consequently,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi\in\mathcal{PP}(k)} \sum_{(P_1,\ldots,P_k)\in S_n(\pi)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right],$$

if the limits exist. We have now reduced the original set $\mathcal{P}(k)$ to the subset $\mathcal{PP}(k)$. Next we want to fix a $\pi \in \mathcal{PP}(k)$ and cope with the set $S_n(\pi)$.

Lemma 3.1 (cf. [6], Proposition 4.4.). *Let $S_n^*(\pi) \subseteq S_n(\pi)$ denote the set of π-consistent sequences $(P_1, \ldots, P_k)$ satisfying*

$$i \sim_\pi j \quad \Longrightarrow \quad q_i - p_i = p_j - q_j$$

for all $i \neq j$. Then, we have

$$\#\left(S_n(\pi)\backslash S_n^*(\pi)\right) = o\left(n^{1+\frac{k}{2}}\right).$$

Proof. We call a pair (P_i, P_j) with $i \sim_\pi j, i \neq j$, positive if $q_i - p_i = q_j - p_j > 0$ and negative if $q_i - p_i = q_j - p_j < 0$. Since $\sum_{i=1}^k q_i - p_i = 0$ by consistency, the existence of a negative pair implies the existence of a positive one. Thus, we can assume that any sequence $(P_1, \ldots, P_k) \in S_n(\pi)\backslash S_n^*(\pi)$ contains a positive pair (P_l, P_m). To fix such a sequence, we first determine the positions of l and m and then, we fix the signs of the remaining differences $q_i - p_i$. The number of possibilities to accomplish that depends only on k and not on n. Now we choose one of n possible values for p_l. In a next step, we fix the values of the differences $|q_i - p_i|$ for all P_i except for P_l and P_m. We have $n^{\frac{k}{2}-1}$ possibilities for that, since π is a pair partition and for any $i \sim_\pi j$ it holds that $|p_i - q_i| = |p_j - q_j|$ by definition. Then, $\sum_{i=1}^k q_i - p_i = 0$ implies that

$$0 < 2(q_l - p_l) = q_l - p_l + q_m - p_m = \sum_{\substack{i=1,\\ i\neq l,m}}^{k} p_i - q_i.$$

Since we have already chosen the signs of the differences as well as their absolute values, we know the value of the sum on the right hand side. Hence, the difference $q_l - p_l = q_m - p_m$ is fixed. We now have the index p_l, all differences $|q_i - p_i|, i \in \{1, \ldots, k\}$, and their signs. Thus, we can start at P_l and

go systematically through the whole sequence $(P_1, \dots, P_k)$ to see that it is uniquely determined. Consequently, our considerations lead to

$$\#\left(S_n(\pi) \backslash S_n^*(\pi)\right) \leq C \cdot n^{\frac{k}{2}} = o\left(n^{1+\frac{k}{2}}\right).$$

□

As a consequence of Lemma 3.1 and relation (5), we obtain

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi \in \mathcal{PP}(k)} \sum_{(P_1,\dots,P_k)\in S_n^*(\pi)} \mathbb{E}\left[a_n(P_1) \cdot \ldots \cdot a_n(P_k)\right],$$

if the limits exist.

We call a pair partition $\pi \in \mathcal{PP}(k)$ *crossing* if there are indices $i < j < l < m$ with $i \sim_\pi l$ and $j \sim_\pi m$. Otherwise, we call π *non-crossing*. The set of all non-crossing pair partitions is denoted by $\mathcal{NPP}(k)$.

Lemma 3.2. *For any crossing $\pi \in \mathcal{PP}(k) \backslash \mathcal{NPP}(k)$, we have*

$$\sum_{(P_1,\dots,P_k)\in S_n^*(\pi)} \mathbb{E}\left[a_n(P_1) \cdot \ldots \cdot a_n(P_k)\right] = o\left(n^{\frac{k}{2}+1}\right).$$

Proof. Let π be crossing and consider a sequence $(P_1, \dots, P_k) \in S_n^*(\pi)$. Note that if there is an $l \in \{1, \dots, k\}$ with $l \sim_\pi l+1$, where $k+1$ is identified with 1, we immediately have

$$a_n(P_l) = a_n(P_{l+1}),$$

since $q_l = p_{l+1}$ by consistency and then $p_l = q_{l+1}$ by definition of $S_n^*(\pi)$. In particular,

$$\mathbb{E}\left[a_n(P_l) \cdot a_n(P_{l+1})\right] = 1.$$

The sequence $(P_1, \dots, P_{l-1}, P_{l+2}, \dots, P_k)$ is still consistent because of the relation $q_{l-1} = p_l = q_{l+1} = p_{l+2}$. Since there are at most n choices for $q_l = p_{l+1}$, it follows

$$\#S_n^*(\pi) \leq n \cdot \#S_n^*(\pi^{(1)}),$$

where $\pi^{(1)} \in \mathcal{PP}(k-2) \backslash \mathcal{NPP}(k-2)$ is the pair partition induced by π after eliminating the indices l and $l+1$. Let r denote the maximum number of pairs of indices that can be eliminated in this way. Since π is crossing, there are at least two pairs left and hence, $r \leq \frac{k}{2} - 2$. By induction, we conclude that

$$\#S_n^*(\pi) \leq n^r \cdot \#S_n^*(\pi^{(r)}),$$

where now $\pi^{(r)} \in \mathcal{PP}(k-2r) \backslash \mathcal{NPP}(k-2r)$ is the still crossing pair partition induced by π. Thus, we so far have

$$\sum_{(P_1,\ldots,P_k)\in S_n^*(\pi)} |\mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right]| \leq n^r \sum_{(P_1^{(r)},\ldots,P_{k-2r}^{(r)})\in S_n^*(\pi^{(r)})} \left|\mathbb{E}\left[a_n(P_1^{(r)})\cdot\ldots\cdot a_n(P_k^{(r)})\right]\right|. \tag{6}$$

Choose $i \sim_{\pi^{(r)}} i+j$ such that j is minimal. We want to count the number of sequences $(P_1^{(r)},\ldots,P_{k-2r}^{(r)}) \in S_n^*(\pi^{(r)})$ given that $p_i^{(r)}$ and $q_{i+j}^{(r)}$ are fixed. Therefore, we start with choosing one of n possible values for $q_i^{(r)}$. But then, we can also deduce the value of

$$p_{i+j}^{(r)} = q_i^{(r)} - p_i^{(r)} + q_{i+j}^{(r)}.$$

Since j is minimal, any element in $\{i+1,\ldots,i+j-1\}$ is equivalent to some element outside the set $\{i,\ldots,i+j\}$. There are n possibilities to fix $P_{i+1}^{(r)}$ as $p_{i+1}^{(r)} = q_i^{(r)}$ is already fixed. Proceeding sequentially, we have n possibilities for the choice of any pair $P_l^{(r)}$ with $l \in \{i+2,\ldots,i+j-2\}$ and there is only one choice for $P_{i+j-1}^{(r)}$ since $q_{i+j-1}^{(r)} = p_{i+j}^{(r)}$ is already chosen. For any other pair that has not yet been fixed, there are at most n possibilities if it is not equivalent to one pair that has already been chosen. Otherwise, there is only one possibility. Hence, assuming that the elements $p_i^{(r)}$ and $q_{i+j}^{(r)}$ are fixed, we have at most

$$n \cdot n^{j-2} \cdot n^{\frac{k}{2}-r-j} = n^{\frac{k}{2}-r-1}$$

possibilities to choose the rest of the sequence $(P_1^{(r)},\ldots,P_{k-2r}^{(r)}) \in S_n^*(\pi^{(r)})$. Consequently, estimating the term in (6) further, we obtain

$$\begin{aligned}
&\sum_{(P_1,\ldots,P_k)\in S_n^*(\pi)} |\mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right]| \\
&\quad\leq n^r \sum_{(P_1^{(r)},\ldots,P_{k-2r}^{(r)})\in S_n^*(\pi^{(r)})} \left|\mathbb{E}\left[a_n(P_i^{(r)})\, a_n(P_{i+j}^{(r)})\right]\right| \\
&\quad\leq n^{\frac{k}{2}-1} \sum_{p_i^{(r)},q_{i+j}^{(r)}=1}^{n} c(|q_{i+j}^{(r)} - p_i^{(r)}|) \\
&\quad\leq C\cdot n^{\frac{k}{2}} \sum_{\tau=0}^{n-1} c(\tau) = o\left(n^{1+\frac{k}{2}}\right),
\end{aligned}$$

since $\sum_{\tau=0}^{n-1} c(\tau) = o(n)$ by condition (C4) and Remark 2.1 (ii). □

Lemma 3.2 now guarantees that we need to consider only non-crossing pair partitions, that is

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi\in\mathcal{NPP}(k)} \sum_{(P_1,\ldots,P_k)\in S_n^*(\pi)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right],$$

if the limits exist.

Lemma 3.3. *Let $\pi \in \mathcal{NPP}(k)$. For any $(P_1,\ldots,P_k) \in S_n^*(\pi)$, we have*

$$\mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right] = 1.$$

Proof. Let $l < m$ with $m \sim_\pi l$. Since π is non-crossing, the number $l-m-1$ of elements between l and m must be even. In particular, there is $l \le i < j \le m$ with $i \sim_\pi j$ and $j = i+1$. By the properties of S_n^*, we have $a_n(P_i) = a_n(P_j)$, and the sequence $(P_1,\ldots,P_l,\ldots,P_{i-1},P_{i+2},\ldots,P_m,\ldots,P_k)$ is still consistent. Applying this argument successively, all pairs between l and m vanish and we see that the sequence $(P_1,\ldots,P_l,P_m,\ldots,P_k)$ is consistent, that is $q_l = p_m$. Then, the identity $p_l = q_m$ also holds. In particular, $a_n(P_l) = a_n(P_m)$. Since l,m have been chosen arbitrarily, we obtain

$$\mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right] = \prod_{\substack{l<m\\ l\sim_\pi m}} \mathbb{E}\left[a_n(P_l)\cdot a_n(P_m)\right] = 1.$$

□

It remains to verify

Lemma 3.4. *For any $\pi \in \mathcal{NPP}(k)$, we have*

$$\lim_{n\to\infty} \frac{\#S_n^*(\pi)}{n^{\frac{k}{2}+1}} = 1.$$

Proof. To calculate the number of elements in $S_n^*(\pi)$, first choose P_1. There are n^2 possibilities for that choice. If $1 \sim_\pi 2$, then P_2 is uniquely determined since $p_2 = q_1$ and by definition of $S_n^*(\pi)$, $q_2 = p_1$. If $1 \not\sim_\pi 2$, then there are $n-1$ possibilities to fix P_2. Proceeding in the same way, we see that if $i \in \{2,\ldots,k\}$ is equivalent to some element in $\{1,\ldots,i-1\}$, there is always only one value P_i

can take. Otherwise there are asymptotically n choices. The latter case will occur exactly $\frac{k}{2}-1$ times. In conclusion,

$$\#S_n^*(\pi) \sim n^2 \cdot n^{\frac{k}{2}-1} = n^{1+\frac{k}{2}}.$$

□

Lemmas 3.3 and 3.4 now provide that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right] \\
&\qquad = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi\in\mathcal{NPP}(k)} \sum_{(P_1,\ldots,P_k)\in S_n^*(\pi)} \mathbb{E}\left[a_n(P_1)\cdot\ldots\cdot a_n(P_k)\right] \\
&\qquad = \lim_{n\to\infty} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\pi\in\mathcal{NPP}(k)} \#S_n^*(\pi) = \#\mathcal{NPP}(k).
\end{aligned}$$

Since the number of non-crossing pair partitions $\#\mathcal{NPP}(k)$ equals exactly the Catalan number $C_{\frac{k}{2}}$, we can conclude that the expected empirical spectral distribution of $\mathbf{X}_n$ tends to the semi-circle law. This is the asserted convergence in expectation.

It remains to deduce almost sure convergence. Therefore, we want to follow the ideas of [6]. To this end, we need

Lemma 3.5. *Suppose the conditions of Theorem 2.2 hold. Then, for any $k, n \in \mathbb{N}$,*

$$\mathbb{E}\left[\left(\operatorname{tr}\left(\mathbf{X}_n^k\right) - \mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right]\right)^4\right] \leq C \cdot n^2.$$

Proof. Fix $k, n \in \mathbb{N}$. Using the notation

$$P = (P_1,\ldots,P_k) = ((p_1,q_1),\ldots,(p_k,q_k)), \qquad a_n(P) = a_n(P_1)\cdot\ldots\cdot a_n(P_k),$$

we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\operatorname{tr}\left(\mathbf{X}_n^k\right) - \mathbb{E}\left[\operatorname{tr}\left(\mathbf{X}_n^k\right)\right]\right)^4\right] \\
&= \frac{1}{n^{2k}} \sum_{\pi^{(1)},\ldots,\pi^{(4)}\in\mathcal{P}(k)} \sum_{P^{(i)}\in S_n\left(\pi^{(i)}\right),i=1,\ldots,4} \mathbb{E}\Big[\prod_{j=1}^{4}\left(a_n(P^{(j)}) - \mathbb{E}\left[a_n(P^{(j)})\right]\right)\Big]. \quad (7)
\end{aligned}$$

Now consider a partition $\boldsymbol{\pi}$ of $\{1,\ldots,4k\}$. We say that a sequence $(P^{(1)},\ldots,P^{(4)})$ is $\boldsymbol{\pi}$-consistent if each $P^{(i)}, i=1,\ldots,4$, is a consistent sequence and

$$\left|q_l^{(i)} - p_l^{(i)}\right| \;=\; \left|q_m^{(j)} - p_m^{(j)}\right| \quad\Longleftrightarrow\quad l + (i-1)k \;\sim_{\boldsymbol{\pi}}\; m + (j-1)k.$$

Let $\mathcal{S}_n(\boldsymbol{\pi})$ denote the set of $\boldsymbol{\pi}$-consistent sequences with entries in $\{1,\dots,n\}$. Then, (7) becomes

$$\mathbb{E}\left[\left(\mathrm{tr}\mathbf{X}_n^k - \mathbb{E}\left[\mathrm{tr}\mathbf{X}_n^k\right]\right)^4\right] = \frac{1}{n^{2k}} \sum_{\boldsymbol{\pi}\in\mathcal{P}(4k)} \sum_{(P^{(1)},\dots,P^{(4)})\in\mathcal{S}_n(\boldsymbol{\pi})} \mathbb{E}\Big[\prod_{j=1}^{4}\left(a_n(P^{(j)}) - \mathbb{E}\left[a_n(P^{(j)})\right]\right)\Big]. \quad (8)$$

We want to analyze the expectation on the right hand side. Therefore, fix a $\boldsymbol{\pi} \in \mathcal{P}(4k)$. We call $\boldsymbol{\pi}$ a matched partition if

(i) Any equivalence class of $\boldsymbol{\pi}$ contains at least two elements and
(ii) For any $i \in \{1,\dots,4\}$ there is a $j \neq i$ and $l, m \in \{1,\dots,k\}$ with

$$l + (i-1)k \;\sim_{\boldsymbol{\pi}}\; m + (j-1)k.$$

In case $\boldsymbol{\pi}$ is not matched, we can conclude that

$$\sum_{(P^{(1)},\dots,P^{(4)})\in\mathcal{S}_n(\boldsymbol{\pi})} \mathbb{E}\Big[\prod_{j=1}^{4}\left(a_n(P^{(j)}) - \mathbb{E}\left[a_n(P^{(j)})\right]\right)\Big] = 0.$$

Thus, we only have to consider matched partitions to evaluate the sum in (8). Let $\boldsymbol{\pi}$ be such a partition and denote by $r = \#\boldsymbol{\pi}$ the number of equivalence classes of $\boldsymbol{\pi}$. Note that condition (i) implies $r \leq 2k$. To count all $\boldsymbol{\pi}$-consistent sequences $(P^{(1)},\dots,P^{(4)})$, we first choose one of at most n^r possibilities to fix the r different equivalence classes. Afterwards, we fix the elements $p_1^{(1)},\dots,p_1^{(4)}$, which can be done in n^4 ways. Since now the differences $|q_l^{(i)} - p_l^{(i)}|$ are uniquely determined by the choice of the corresponding equivalence classes, we can proceed sequentially to see that there are at most two choices left for any pair $P_l^{(i)}$. To sum up, we have at most

$$2^{4k}\cdot n^4 \cdot n^r = C\cdot n^{r+4}$$

possibilities to choose $(P^{(1)},\dots,P^{(4)})$. If now $r \leq 2k-2$, we can conclude that

$$\#\mathcal{S}_n(\boldsymbol{\pi}) \leq C\cdot n^{2k+2}. \quad (9)$$

Hence, it remains to consider the case where $r = 2k-1$ and $r = 2k$, respectively.

To begin with, let $r = 2k-1$. Then, we have either two equivalence classes with three elements or one equivalence class with four. Since $\boldsymbol{\pi}$ is matched, there must exist an $i \in \{1,\dots,4\}$ and an $l \in \{1,\dots,k\}$ such that $P_l^{(i)}$ is not equivalent to any other pair in the sequence $P^{(i)}$. Without loss of generality, we can assume

that $i = 1$. In contrast to the construction of $(P^{(1)}, \dots, P^{(4)})$ as above, we now alter our procedure as follows: We fix all equivalence classes except of that $P_l^{(1)}$ belongs to. There are n^{r-1} possibilities to accomplish that. Now we choose again one of n^4 possible values for $p_1^{(1)}, \dots, p_1^{(4)}$. Hereafter, we fix $q_m^{(1)}$, $m = 1, \dots, l-1$, and then start from $q_k^{(1)} = p_1^{(1)}$ to go backwards and obtain the values of $p_k^{(1)}, \dots, p_{l+1}^{(1)}$. Each of these steps leaves at most two choices to us, that is 2^{k-1} choices in total. But now, $P_l^{(1)}$ is uniquely determined since $p_l^{(1)} = q_{l-1}^{(1)}$ and $q_l^{(1)} = p_{l+1}^{(1)}$ by consistency. Thus, we had to make one choice less than before, implying (9).

Now, let $r = 2k$. In this case, each equivalence class has exactly two elements. Since we consider a matched partition, we can find here as well an $l \in \{1, \dots, k\}$ such that $P_l^{(1)}$ is not equivalent to any other pair in the sequence $P^{(1)}$. But in addition to that, we also have an $m \in \{1, \dots, k\}$ such that, possibly after relabeling, $P_m^{(2)}$ is neither equivalent to any element in $P^{(1)}$ nor to any other element in $P^{(2)}$. Thus, we can use the same argument as before to see that this time, we can reduce the number of choices to at most $C \cdot n^{r+2} = C \cdot n^{2k+2}$. In conclusion, (9) holds for any matched partition $\boldsymbol{\pi}$. To sum up our results, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathrm{tr}\mathbf{X}_n^k - \mathbb{E}\left[\mathrm{tr}\mathbf{X}_n^k\right]\right)^4\right] \\
&\quad = \frac{1}{n^{2k}} \sum_{\substack{\boldsymbol{\pi} \in \mathcal{P}(4k), \\ \boldsymbol{\pi} \text{ matched}}} \sum_{(P^{(1)}, \dots, P^{(4)}) \in \mathcal{S}_n(\boldsymbol{\pi})} \mathbb{E}\Big[\prod_{j=1}^{4} \left(a_n(P^{(j)}) - \mathbb{E}\left[a_n(P^{(j)})\right]\right)\Big] \\
&\quad \leq C \cdot n^2,
\end{aligned}$$

which is the statement of Lemma 3.5. □

From Lemma 3.5 and Chebyshev's inequality, we can now conclude that for any $\varepsilon > 0$ and any $k, n \in \mathbb{N}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\mathrm{tr}\mathbf{X}_n^k - \mathbb{E}\left[\frac{1}{n}\mathrm{tr}\mathbf{X}_n^k\right]\right| > \varepsilon\right) \leq \frac{C}{\varepsilon^4 n^2}.$$

Hence, the convergence in expectation part of Theorem 2.2 together with the Borel-Cantelli lemma yield that

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{tr}\mathbf{X}_n^k = \mathbb{E}\left[Y^k\right] \qquad \text{almost surely,}$$

where Y is distributed according to the standard semi-circle law. In particular, we have that, with probability 1, the empirical spectral distribution of $\mathbf{X}_n$ converges weakly to the semi-circle law. This finishes the proof of Theorem 2.2.

Theorem 2.4 can be proved along the lines of the proof of Theorem 2.1.21 in [2]. Indeed, by a result of Hoffman and Wielandt for any two symmetric $n \times n$

matrices $\mathbf{A}$ and $\mathbf{B}$ and their ordered eigenvalues $\lambda_1^{\mathbf{A}} \leq \ldots \leq \lambda_n^{\mathbf{A}}$ and $\lambda_1^{\mathbf{B}} \leq \ldots \leq \lambda_n^{\mathbf{B}}$, respectively, it holds

$$\sum_{i=1}^{n} |\lambda_i^{\mathbf{A}} - \lambda_i^{\mathbf{B}}| \leq \mathrm{tr}(\mathbf{A} - \mathbf{B})^2.$$

Since for random matrices the right hand side just depends on the second moments of the entries of $\mathbf{A}$ and $\mathbf{B}$, this bound can be used to establish a truncation technique and first show a semicircle law in probability for the truncated matrices (as above) and then extend it to matrices satisfying (C1'). For details we refer to the proof of Theorem 2.1.21 in [2].

4 Examples

4.1 Gaussian Processes

Let $\{a(p, p+r), p \in \mathbb{N}\}$, $r \in \mathbb{N}_0$, be independent families of stationary Gaussian Markov processes with mean 0 and variance 1. In addition to this, we assume that the processes are non-degenerate in the sense that $\mathbb{E}[a(p, p+r)|a(q, q+r), q \leq p-1] \neq a(p, p+r)$. In this case, the conditions of Theorem 2.2 are satisfied. Indeed, for fixed $r \in \mathbb{N}_0$ and any $p \in \mathbb{N}$, we can represent $a_p := a(p, p+r)$ as

$$a_p = x_p \sum_{j=1}^{p} y_j \xi_j,$$

where $\{\xi_j\}$ is a family of independent standard Gaussian variables and $x_p, y_1, \ldots, y_p \in \mathbb{R} \setminus \{0\}$. Then, we obtain

$$c(\tau) := |\mathrm{Cov}(a_p, a_{p+\tau})| = \left| \frac{x_{p+\tau}}{x_p} \right|,$$

implying $c(\tau) = c(1)^\tau$ for any $\tau \in \mathbb{N}_0$. By calculating the second moment of $a_2 = x_2 y_2 \xi_2 + \mathrm{Cov}(a_1, a_2) a_1$, we can conclude that $c(1) < 1$. Thus, condition (C4) is satisfied for any $\varepsilon > 0$.

4.2 Markov Chains with Finite State Space

We want to verify that condition (C4) holds for stationary N-state Markov chains which are aperiodic and irreducible. Let $\{X_k, k \in \mathbb{N}\}$ be such a Markov chain on the

state space $S = \{s_1, \ldots, s_N\}$ with mean 0 and variance 1. Denote by π its stationary distribution. In [13], Theorem 4.9, it is stated that for some constant $C > 0$ and some $\alpha \in (0, 1)$,

$$\max_{i,j \in \{1,\ldots,N\}} |\mathbb{P}(X_k = s_i \mid X_1 = s_j) - \pi(i)| \leq C\alpha^{k-1}, \quad k \in \mathbb{N}.$$

In particular, we obtain

$$|\mathrm{Cov}(X_k, X_1)| = \left| \sum_{i,j=1}^{N} s_i s_j \left(\mathbb{P}(X_k = s_i \mid X_1 = s_j) - \pi(i) \right) \pi(j) \right| \leq C\alpha^{k-1}.$$

Thus $\mathrm{Cov}(X_k, X_1)$ decays exponentially to 0 as $k \to \infty$ and condition (C4) is satisfied.

4.3 Fractional Brownian Motion

We want to consider a stochastic process that exhibits long-range dependence but satisfies condition (C4) nevertheless. To this end, consider fractional Brownian motion $(B_t^H)_t$ with Hurst parameter (or index) H. Recall that B_t^H is a continuous stochastic process, starting in 0, obeying $\mathbb{E}B_t^H = 0$ for all t and

$$\mathrm{Cov}(B_t^H, B_s^H) = \frac{1}{2}(t^{2H} + s^{2H} - |t - s|^{2H}).$$

In particular,

$$\mathbb{V}(B_t^H) = t^{2H}$$

implying that for Hurst parameter $H = \frac{1}{2}$ the process is a Brownian motion. A standard reference for fractional Brownian motion is the textbook by Samorodnitsky and Taqqu [15], Chap. 7.

Now for each diagonal, we take independent fractional Brownian motions with index $H \in (0, 1)$ and for the entries $\{X_k, k \in \mathbb{N}\}$ on a (fixed) diagonal, we take the integer times of fractional Gaussian noise, i.e. $X_k = B_k^H - B_{k-1}^H$. Thus the $\{X_k, k \in \mathbb{N}\}$ are stationary with mean zero and variance one. Using the above covariance formula it can be further shown that for $H \neq 1/2$,

$$\mathrm{Cov}(X_{k+1}, X_1) = \frac{1}{2}\left((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\right) \sim H(2H-1)k^{2H-2},$$

as $k \to \infty$ (cf. [7], Proposition 3.1). Hence, condition (C4) is satisfied. In particular, the sum $\sum_{k=0}^{\infty} |\mathrm{Cov}(X_{k+1}, X_1)|$ diverges if $1/2 < H < 1$ implying that we have a long-range dependence.

References

1. G.W. Anderson, O. Zeitouni, A law of large numbers for finite-range dependent random matrices. Comm. Pure Appl. Math. **61**(8), 1118–1154 (2008)
2. G.W. Anderson, A. Guionnet, O. Zeitouni, *An Introduction to Random Matrices.* Cambridge Studies in Advanced Mathematics, vol. 118 (Cambridge University Press, Cambridge, 2010)
3. L. Arnold, On Wigner's semicircle law for the eigenvalues of random matrices. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **19**, 191–198 (1971)
4. Z.D. Bai, Methodologies in spectral analysis of large-dimensional random matrices, a review. Stat. Sinica **9**(3), 611–677 (1999)
5. A. Bose, A. Sen, nother look at the moment method for large dimensional random matrices. Electron. J. Probab. **13**(21), 588–628 (2008)
6. W. Bryc, A. Dembo, T. Jiang, Spectral measure of large random Hankel, Markov and Toeplitz matrices. Ann. Probab. **34**(1), 1–38 (2006)
7. P. Doukhan, G. Oppenheim, M.S. Taqqu, *Theory and Applications of Long-Range Dependence* (Birkhäuser, Basel, 2003)
8. O. Friesen, M. Löwe, The semicircle law for matrices with independent diagonals. Preprint, J. Theor. Probab. (2013)
9. F. Götze, A.N. Tikhomirov, Limit theorems for spectra of random matrices with martingale structure, in *Stein's Method and Applications*. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., vol. 5 (Singapore University Press, Singapore, 2005), pp. 181–193
10. A. Guionnet, *Large Random Matrices: Lectures on Macroscopic Asymptotics.* Lecture Notes in Mathematics, vol. 1957. Lectures from the 36th Probability Summer School held in Saint-Flour, 2006 (Springer, Berlin, 2009)
11. K. Hofmann-Credner, M. Stolz, Wigner theorems for random matrices with dependent entries: ensembles associated to symmetric spaces and sample covariance matrices. Electron. Comm. Probab. **13**, 401–414 (2008)
12. A.M. Khorunzhy, L. Pastur, On the eigenvalue distributon of the deformed Wigner ensemble of random matrices. Adv. Sov. Math. **19**, 97–127 (1994)
13. D.A. Levin, Y. Peres, E.L. Wilmer, *Markov Chains and Mixing Times* (American Mathematical Society, Providence, 2006)
14. M.L. Mehta, *Random Matrices.* Pure and Applied Mathematics (Amsterdam), vol. 142 (Elsevier/Academic, Amsterdam, 2004)
15. G. Samorodnitsky, M.S. Taqqu, *Stable Non-Gaussian Random Processes* (Chapman & Hall, London, 1994)
16. J. Schenker, H. Schulz-Baldes, Semicircle law and freeness for random matrices with symmetries or correlations. Math. Res. Lett. **12**, 531–542 (2005)
17. E. Wigner, Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. Second Ser. **62**, 548–564 (1955)
18. E. Wigner, On the distribution of the roots of certain symmetric matrices. Ann. Math. **67**, 325–328 (1958)

Limit Theorems for Random Matrices

Alexander Tikhomirov

Dedicated to Friedrich Götze on the occasion of his sixtieth birthday

Abstract This note gives a survey of some results on limit theorems for random matrices that have been obtained during the last 10 years in the joint research of the author and F. Götze. We consider the rate of convergence to the semi-circle law and Marchenko–Pastur law, Stein's method for random matrices, the proof of the circular law, and some limit theorems for powers and products of random matrices.

Keywords Limit theorems • Number theory • Probability • Statistics • Random Matrices • Semi-Circle Law • Marchenko-Pastur Law • Circular Law

2010 *Mathematics Subject Classification.* 60F05.

1 Introduction

In this note we describe some results obtained jointly with F. Götze in the last 10 years. We consider limit theorems for Wigner matrices (Wigner matrix means symmetric real matrix or Hermitian matrix with independent entries up to symmetry), limit theorems for sample covariance matrices and limit theorems for powers and products of independent Ginibre-Girko matrices (that means matrices with all independent entries without any symmetry). We consider results about

A. Tikhomirov (✉)
Department of Mathematics, Komi Research Center Ural Branch of Russian Academy of Sciences, Syktyvkar State University, Chernova 3a, 197001 Syktyvkar, Russia
e-mail: tichomir@math.uni-bielefeld.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_14,

the rate of convergence to the semi-circular law and Marchenko–Pastur law for the empirical spectral distribution function of Wigner matrices and of sample covariance matrices, respectively. For Girko–Ginibre matrices and their powers and products we discuss the results on convergence to the limit distributions. We consider also random matrices with dependent entries and describe Stein's method for random matrices with some martingale structure of dependence of entries.

2 Wigner Matrices

Let X_{jk} $(1 \le j \le k \le n)$ be independent random variables (possibly complex) with $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}|X_{jk}|^2 = 1$, defined on the same probability space $\{\Omega, \mathfrak{R}, \mathbb{P}\}$. We define the Hermitian (symmetric in real case) matrix $\mathbf{X}$ with entries $[\mathbf{X}]_{jk} = \frac{1}{\sqrt{n}} X_{jk}$ for $1 \le j \le k \le n$. Consider the eigenvalues of the matrix $\mathbf{X}$ denoted in non-increasing order by $\lambda_1 \ge \cdots \ge \lambda_n$ and define the empirical spectral distribution function of this matrix as

$$\mathcal{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{\lambda_j \le x\},$$

where $\mathbb{I}\{A\}$ denotes indicator of the event A. Introduce also the expected spectral distribution function $F_n(x) := \mathbb{E}\mathcal{F}_n(x)$ of matrix $\mathbf{X}$. Wigner [39] considered the symmetric random matrix $\mathbf{X}$ with entries $X_{jk} = \pm 1$ with probability $\frac{1}{2}$ and proved that

$$\Delta_n := \sup_x |F_n(x) - G(x)| \to 0, \quad \text{as} \quad n \to \infty, \tag{1}$$

where $G(x)$ is the distribution function of the semi-circular law with the density $G'(x) = \frac{1}{2\pi}\sqrt{4 - x^2}\mathbb{I}\{|x| \le 2\}$. This problem has been studied by several authors. Wigner's result [39] was extended later to different classes of distributions of random variables X_{jk}. In particular, Wigner in [40] proved that (1) holds for symmetric random matrices with sub-Gaussian entries. (A random variable ξ is called subgaussian random variable if there exists a positive constant $\beta > 0$ such that $\mathbb{P}\{|\xi| > x\} \le \exp\{-\beta x^2\}$ for any $x > 0$.) Later it was shown that the semi-circular law (the statement (1)) holds under the assumption of Lindeberg condition for the distributions of matrix entries, i e.,

$$L_n(\tau) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=j}^{n} \mathbb{E}|X_{jk}|^2 \mathbb{I}\{|X_{jk}| \ge \tau\sqrt{n}\} \to 0 \quad \text{as } n \to \infty, \tag{2}$$

for any $\tau > 0$ (see, e.g., [16]). It was shown also that under the same assumptions

$$\Delta_n^* := \sup_x |\mathcal{F}_n(x) - G(x)| \to 0 \quad \text{in probability as} \quad n \to \infty. \tag{3}$$

We have investigated the rate of convergence in (1) and (3). This problem has been studied by several authors. In particular, Bai [5] proved that $\Delta_n = O(n^{-\frac{1}{4}})$ assuming that $\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^4 \le M_4 < \infty$. Later Bai et al. [9] under the condition that $\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^8 \le M_8 < \infty$ proved that $\Delta_n = O(n^{-\frac{1}{2}})$ and $\mathbb{E}\Delta_n^* = O(n^{-\frac{2}{5}})$. Girko [18] proved that $\Delta_n = O(n^{-\frac{1}{2}})$ assuming $\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^4 \le M_4 < \infty$. A very interesting result was obtained recently by Erdös et al. in [13]. It follows from their results that for random matrices whose entries have distributions with exponential tails, i.e., $\mathbb{P}\{|X_{jk}| > t\} \le A\exp\{-t^\varkappa\}$ for some $A, \varkappa > 0$, the following holds

$$\mathbb{P}\left\{\Delta_n^* \le Cn^{-1}(\log n)^{C\ln\ln n}\right\} \ge 1 - C\exp\{-(\log n)^{c\ln\ln n}\} \tag{4}$$

with some positive constants C and c depending on $A, \varkappa$ only.

We state the results obtained jointly with F. Götze in several theorems below.

Theorem 2.1 (Götze and Tikhomirov [20]). *Let* $\mathbb{E}X_{jk} = 0$ *and* $\mathbb{E}|X_{jk}|^2 = 1$. *Let*

$$\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^4 \le M_4 < \infty. \tag{5}$$

Then there exist a numerical constant $C > 0$ *such that*

$$\Delta_n \le CM_4^{\frac{1}{2}}n^{-\frac{1}{2}}. \tag{6}$$

If in addition

$$\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^{12} \le M_{12} < \infty,$$

then

$$\mathbb{E}\Delta_n^* \le CM_{12}^{\frac{1}{6}}n^{-\frac{1}{2}}. \tag{7}$$

Assuming instead of (5) the condition (8) below, we have obtained the following result.

Theorem 2.2 (Tikhomirov [37]). *Let* X_{jk} *be independent random variables with* $\mathbb{E}X_{jk} = 0$ *and* $\mathbb{E}|X_{jk}|^2 = 1$. *Assume that for some* $0 < \delta \le 2$ *the following relation holds*

$$\sup_{1\le j\le k}\mathbb{E}|X_{jk}|^{2+\delta} =: M_{2+\delta} < \infty. \tag{8}$$

Then there exists a numerical $C > 0$ *such that*

$$\Delta_n \le C\left(\frac{M_{2+\delta}^{\frac{\delta}{2+\delta}}}{n^{\frac{\delta}{2+\delta}}}\right)^{1-\frac{(1-\delta)_+}{2}},$$

where $(1-\delta)_+ = \max\{1-\delta, 0\}$.

Under stronger assumptions on the distribution of X_{jk} we get bounds for Δ_n of order $O(n^{-\frac{1}{2}-\gamma})$ with some positive $\gamma > 0$. In particular, in the paper of Bobkov et al. [13] we consider random variables X_{jk} with distributions satisfying a Poincaré-type inequality. Let us recall that a probability measure μ on $\mathbb{R}^d$ is said to satisfy a Poincaré-type, $PI(\sigma^2)$, or a spectral gap inequality with constant σ^2 if for any bounded smooth function g on $\mathbb{R}^d$ with gradient ∇g

$$\mathrm{Var}(g) \le \sigma^2 \int_{\mathbb{R}^d} |\nabla g|^2 d\mu, \tag{9}$$

where $\mathrm{Var}(g) = \int_{\mathbb{R}^d} g^2 d\mu - \left(\int_{\mathbb{R}^d} g d\mu \right)^2$.

Theorem 2.3 (Bobkov et al. [13]). *If the distributions of X_{jk}'s satisfy the Poincaré-type inequality $PI(\sigma^2)$ on the real line, then*

$$\Delta_n \le C n^{-2/3},$$

where the constant C depends on σ only. Moreover,

$$\mathbb{E}\Delta_n^* \le C n^{-2/3} \log^2(n+1).$$

For any positive constants $\alpha > 0$ and $\varkappa > 0$ define the quantities

$$l_{n,\alpha} := \log n \, (\log\log n)^{\alpha} \quad \text{and} \quad \beta_n := (l_{n,\alpha})^{\frac{1}{\varkappa}+\frac{1}{2}}. \tag{10}$$

The best known result for the rate of convergence in probability to the semi-circular law is the following:

Theorem 2.4 (Götze and Tikhomirov [28]). *Let $\mathbb{E}X_{jk} = 0$, $\mathbb{E}X_{jk}^2 = 1$. Assume that there exist constants A and $\varkappa > 0$ such that*

$$\mathbb{P}\{|X_{jk}| \ge t\} \le A \exp\{-t^{\varkappa}\}, \tag{11}$$

for any $1 \le j \le k \le n$ and any $t \ge 1$. Then, for any positive $\alpha > 0$ there exist positive constants C and c depending on A and $\varkappa$ and α only, such that

$$\mathbb{P}\Big\{ \sup_x |\mathcal{F}_n(x) - G(x)| > n^{-1}\beta_n^2 \Big\} \le C \, \exp\{-c \, l_{n,\alpha}\}.$$

Remark 2.5. In the result of (4) [13] $\Delta_n^* = O_P(n^{-1}(\log n)^{O(\log\log n)})$. In our result $\Delta_n^* = O_P(n^{-1}(\log n)^{O(1)})$.

Remark 2.6. If $\mathbf{X}$ belongs to Gaussian Unitary Ensemble (GUE) [23] or Gaussian Orthogonal Ensemble (GOE) [38] then there exists an absolute constant $C > 0$ such that

$$\Delta_n \le C n^{-1}. \tag{12}$$

3 Sample Covariance Matrices

In this section we consider the so called sample covariance matrices and their generalization. Let $\mathbf{X}$ be rectangular matrices of order $[n \times p]$ with independent entries (possible complex) X_{jk}, $j = 1, \dots, n$; $k = 1 \dots, p$. We shall assume that $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}|X_{jk}|^2 = 1$. Consider the matrix $\mathbf{W} = \frac{1}{p}\mathbf{X}\mathbf{X}^*$. Such matrices are called sample covariance matrices and they were first considered in 1928 by Wishart [41]. He obtained the joint distribution of entries of the matrix $\mathbf{W}$ as X_{jk} are standard Gaussian random variables. We shall be interested in the asymptotic distribution of the spectrum of the matrix $\mathbf{W}$. Note that the matrix $\mathbf{W}$ is semi-positive definite and its eigenvalues are non-negative. Denote the eigenvalues of the matrix $\mathbf{W}$ in decreasing order by $s_1^2 \geq \dots \geq s_n^2 \geq 0$. (Note that the numbers $s_1, \dots, s_n$ are called singular values of matrix $\frac{1}{\sqrt{p}}\mathbf{X}$.) Define the empirical spectral distribution function of the matrix $\mathbf{W}$ by the equality

$$\mathcal{H}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{s_j^2 \leq x\}. \tag{13}$$

Let $H_y(x)$ be the distribution function with the density

$$H_y'(x) = \frac{\sqrt{(b-x)(x-a)}}{2\pi x y} + (1 - \frac{1}{y})_+ \delta_0(x), \tag{14}$$

where $a = (1 - \sqrt{y})^2$, $b = (1 + \sqrt{y})^2$, and $\delta_0(x)$ denotes Dirac δ-function, $a_+ = \max\{a, 0\}$ for any real a. This distribution is called Marchenko–Pastur distribution with parameter y. Assuming that $p = p(n)$ where $\lim_{n\to\infty} \frac{n}{p} = y$, and assuming the moment condition (5), Marchenko and Pastur [29] have shown that there exists

$$\lim_{n\to\infty} \mathbb{E}\mathcal{H}_n(x) = H_y(x), \tag{15}$$

The result of Marchenko–Pastur was improved by many authors. As a final result we have the following Theorem.

Theorem 3.1. *Let the random variables X_{jk}, $1 \leq j \leq n$, $1 \leq k \leq p$ be independent for any $n \geq 1$ and have zero mean and unit variance. Assume that $p = p(n)$ such that $\lim_{n\to\infty} \frac{n}{p} = y$. Further suppose that the Lindeberg condition holds, i.e.,*

$$L_n(\tau) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{p} \mathbb{E}|X_{jk}|^2 \mathbb{I}\{|X_{jk}| \geq \tau\sqrt{n}\} \to 0,$$

for any $\tau > 0$. Then

$$\sup_x |\mathbb{E}\mathcal{H}_n(x) - H_y(x)| \to 0, \ \textit{as} \quad n \to \infty.$$

Moreover, $\mathcal{H}_n(x)$ converges to the Marchenko–Pastur distribution in probability.

The proof of this result may be found in [8]. We have investigated the rate of convergence of the expected and empirical spectral distribution function of sample covariance matrix to the Marchenko–Pastur law. This question was considered also in the papers of Bai [6], and in Bai and co-authors [10]. Bai et al. in [10] and independently Götze and Tikhomirov in [21] established the bound of the rate of convergence in Kolmogorov distance $\Delta_n = \sup_x |\mathbb{E}\mathcal{H}_n(x) - H_y(x)| = O(n^{-\frac{1}{2}})$, assuming that

$$\max_{j,k\geq 1} \mathbb{E}|X_{jk}|^8 \leq C \tag{16}$$

for some positive constant $C > 0$ independent of n. Götze and and Tikhomirov [21] proved as well that $\Delta_n^* = \mathbb{E}\sup_x |\mathcal{H}_n(x) - H_y(x)| = O(n^{-\frac{1}{2}})$, assuming

$$\max_{j,k\geq 1} \mathbb{E}|X_{jk}|^{12} \leq C \tag{17}$$

for some positive constant $C > 0$ independent of n. Somewhat later these bounds were improved in the paper of Götze and Tikhomirov [26] and in the paper of Tikhomirov [36]. We formulate the following result.

Theorem 3.2. *Let the random variables X_{jk}, $1 \leq j \leq n$, $1 \leq k \leq p$ be independent for any fixed $n \geq 1$ and have zero mean and unit variance. Assume that $p = p(n)$, where $\frac{n}{p} = y \leq 1$. Let for some $0 < \delta \leq 2$*

$$M_{2+\delta} := \sup_{j,k,n} \mathbb{E}|X_{jk}|^{2+\delta} < \infty.$$

Then there exist a positive constant $C = C(\delta)$, depending on δ only , such that

$$\Delta_n \leq C M_{2+\delta}^{\frac{\delta}{2+\delta}} n^{-\frac{\delta}{2+\delta}}. \tag{18}$$

The bound (18) for $\delta = 2$ was obtained in [26], the bound for the case $0 < \delta < 2$ in [36]. The question about optimality of the above mentioned bounds is still open. But assuming that the random variables X_{jk} are independent standard complex Gaussian random variables (so-called Laguerre unitary ensemble) the optimal bound of the rate of convergence of the expected spectral distribution of the matrix $\mathbf{W}$ was obtained. It turns out that $\Delta_n = O(n^{-1})$, which was proved by Götze and Tikhomirov [23]. Recall that the distribution of a random variable X has so-called exponential tail means that there exist constants $A > 0$ and $\varkappa > 0$ such that

$$\mathbb{P}\{|X| \geq t\} \leq A\exp\{-t^{\varkappa}\}. \tag{19}$$

Assuming that the entries of the matrix $\mathbf{X}$ have distribution with exponential tails, Götze and Tikhomirov have proved in [27] that

$$\mathbb{P}\Big\{\sup_x |\mathcal{H}_n(x) - H_y(x)| > n^{-1}\beta_n^2\Big\} \le C\ \exp\{-c\, l_{n,\alpha}\}, \tag{20}$$

for any $\alpha > 0$. Here β_n and $l_{n,\alpha}$ were defined in (10). The constants $C > 0$ and $c > 0$ depend on $A, \varkappa$ and α only. It would be interesting to extend the results about sample covariance matrices to more general situations. First we consider the singular values of powers of random matrices. And then we consider the asymptotic distribution of singular values of products of independent random matrices.

3.1 Powers of Random Matrices

Let $\mathbf{X} = (X_{jk})_{j,k=1}^n$ be a square random matrix of order n with independent entries such that $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}|X_{jk}|^2 = 1$. In this section we shall investigate the asymptotic distribution of the singular values of the matrix $\mathbf{W} = n^{-\frac{m}{2}}\mathbf{X}^m$ or the eigenvalues of the matrix $\mathbf{V} = \mathbf{W}\mathbf{W}^*$. For $m = 1$ it is the case of sample covariance matrix with parameter $y = 1$. Denote by $s_1^2 \ge \ldots \ge s_n^2$ the eigenvalues of the matrix $\mathbf{V}$. (Note that $s_1 \ge \ldots \ge s_n$ are the singular values of the matrix $\mathbf{W}$.) Let

$$\mathcal{H}_n^{(m)}(x) = \frac{1}{n}\sum_{j=1}^m \mathbb{I}\{s_j^2 \le x\} \tag{21}$$

denote the empirical spectral distribution function of the matrix $\mathbf{V}$. Let $FC(k,m) = \frac{1}{mk+1}\binom{mk+k}{k}$ denote the kth Fuss–Catalan number with parameter m, for $k \ge 1$. These numbers are the moments of some distribution which we denote by $H^{(m)}(x)$. It is well known that the Stieltjes transform of this distribution, $s^{(m)}(z) = \int \frac{1}{x-z} dH^{(m)}(x)$, satisfies the equation

$$1 + zs^{(m)}(z) + (-1)^{m+1} z^{m+1} (s^{(m)}(z))^{m+1} = 0,$$

In the joint papers of Alexeev et al. [2] and [1] the following was proved:

Theorem 3.3. *Let random variables X_{jk} be independent for any fixed $n \ge 1$ and for any $1 \le j,k \le n$. Assume that $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}|X_{jk}|^2 = 1$ for any $j,k \ge 1$ and*

$$\sup_{j,k\ge 1} \mathbb{E}|X_{jk}|^4 \le C, \tag{22}$$

for some positive constant $C > 0$. Assume also that for any $\tau > 0$

$$L_n(\tau) = \frac{1}{n^2}\sum_{j,k=1}^n \mathbb{E}|X_{jk}|^4 \mathbb{I}\{|X_{jk}| \ge \tau\sqrt{n}\} \to 0 \quad \text{as } n \to \infty. \tag{23}$$

Then

$$\lim_{n\to\infty} \sup_x |\mathbb{E}\mathcal{H}_n^{(m)}(x) - H^{(m)}(x)| = 0. \tag{24}$$

For the proof of this result we use the method of moments, see [2]. The proof of Theorem 3.2 by the method of Stieltjes transform is given in [4].

3.2 *Product of Random Matrices*

Let $m \geq 1$ be a fixed integer. For any $n \geq 1$ consider an $(m+1)$-tuple of integers $(p_0, \ldots, p_m)$ with $p_0 = n$ and $p_\nu = p_\nu(n)$ for $\nu = 1, \ldots, m$, such that

$$\lim_{n\to\infty} \frac{n}{p_\nu(n)} = y_\nu \in (0, 1]. \tag{25}$$

Furthermore, we consider an array of independent complex random variables $X_{jk}^{(\nu)}$, $1 \leq j \leq p_{\nu-1}, 1 \leq k \leq p_\nu$, $\nu = 1, \ldots, m$ defined on a common probability space $\{\Omega_n, \mathbb{F}_n, \mathbb{P}\}$ with $\mathbb{E}X_{jk}^{(\nu)} = 0$ and $\mathbb{E}|X_{jk}^{(\nu)}|^2 = 1$. Let $\mathbf{X}^{(\nu)}$ denote the $p_{\nu-1} \times p_\nu$ matrix with entries $[\mathbf{X}^{(\nu)}]_{jk} = \frac{1}{\sqrt{p_\nu}} X_{jk}^{(\nu)}$, for $1 \leq j \leq p_{\nu-1}, 1 \leq k \leq p_\nu$. The random variables $X_{jk}^{(\nu)}$ may depend on n but for simplicity we shall not make this explicit in our notations. Denote by $s_1 \geq \ldots \geq s_n$ the singular values of the random matrix $\mathbf{W} := \prod_{\nu=1}^{m} \mathbf{X}^{(\nu)}$ and define the empirical distribution of its squared singular values by

$$\mathcal{H}_n^{(m)}(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{{s_k}^2 \leq x\}.$$

We shall investigate the approximation of the expected spectral distribution $H_n^{(m)}(x) = \mathbb{E}\mathcal{H}_n^{(m)}(x)$ by the distribution function $H_{\mathbf{y}}(x)$ which is defined by its Stieltjes transform $s_{\mathbf{y}}(z)$ in the following way:

$$1 + z s_{\mathbf{y}}(z) - s_{\mathbf{y}}(z) \prod_{l=1}^{m} (1 - y_l - z y_l s_{\mathbf{y}}(z)) = 0, \tag{26}$$

where $0 \leq y_l \leq 1$.

Remark 3.4. In the case $y_1 = y_2 = \cdots = y_m = 1$ the distribution $H_{\mathbf{y}}$ has moments $M(k, m) = FC(k, m)$. The Stieltjes transform of the distribution $H_{\mathbf{y}}(x)$ satisfies in this case the equation

$$1 + zs(z) + (-1)^{m+1} z^m s(z)^{m+1} = 0.$$

The main result of this subsection.

Theorem 3.5. *Assume that condition (25) holds. Let* $\mathbb{E}X_{jk}^{(\nu)} = 0$, $\mathbb{E}|X_{jk}^{(\nu)}|^2 = 1$. *Suppose that the Lindeberg condition holds, i.e.,*

$$L_n(\tau) := \max_{\nu=1,\dots,m} \frac{1}{p_{\nu-1}\, p_\nu} \sum_{j=1}^{p_{\nu-1}} \sum_{k=1}^{p_\nu} \mathbb{E}|X_{jk}^{(\nu)}|^2 I_{\{|X_{jk}^{(\nu)}| \geq \tau\sqrt{n}\}} \to 0 \quad as\ n \to \infty, \quad (27)$$

for any $\tau > 0$. *Then*

$$\lim_{n\to\infty} \sup_x |H_n^{(m)}(x) - H_{\mathbf{y}}(x)| = 0.$$

Remark 1. For $m = 1$ we get the well-known result of Marchenko-Pastur for sample covariance matrices [29].

Remark 2. We see that the limit distribution for the distribution of singular values of product of independent square random matrices is the same as for powers of random matrices with independent entries, see [2].

The statement of Theorem 3.5 was published in [1] and a proof of this result is given in [3].

4 Circular Law and Its Generalization

4.1 Circular Law

Let $X_{jk}, 1 \leq j,k < \infty$, be complex random variables with $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}|X_{jk}|^2 = 1$. For a fixed $n \geq 1$, denote by $\lambda_1, \dots, \lambda_n$ the eigenvalues of the $n \times n$ matrix

$$\mathbf{X} = (X_n(j,k))_{j,k=1}^n, \quad X_n(j,k) = \frac{1}{\sqrt{n}} X_{jk} \text{ for } 1 \leq j,k \leq n, \quad (28)$$

and define its empirical spectral distribution function by

$$G_n(x,y) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{\mathrm{Re}\{\lambda_j\} \leq x,\ \mathrm{Im}\{\lambda_j\} \leq y\}. \quad (29)$$

We investigate the convergence of the expected spectral distribution function $\mathbb{E}G_n(x,y)$ to the distribution function $G(x,y)$ of the uniform distribution in the unit disc in $\mathbb{R}^2$.

The main results which was obtained in [19] is the following.

Theorem 4.1. *Let* X_{jk} *be independent random variables with*

$$\mathbb{E}X_{jk} = 0, \qquad \mathbb{E}|X_{jk}|^2 = 1 \quad and \quad \mathbb{E}|X_{jk}|^2 \varphi(X_{jk}) \leq \varkappa,$$

where $\varphi(x) = (\ln(1+|x|))^{19+\eta}$ *for some* $\eta > 0$. *Then* $\mathbb{E}G_n(x,y)$ *converges weakly to the distribution function* $G(x,y)$ *as* $n \to \infty$.

We shall prove the same result for the following class of sparse matrices. Let ε_{jk}, $j,k = 1,\dots,n$ denote Bernoulli random variables which are independent in aggregate and independent of $(X_{jk})_{j,k=1}^n$ with success probability $p_n := \mathbb{P}\{\varepsilon_{jk} = 1\}$. Consider the matrix $\mathbf{X}^{(\varepsilon)} = \frac{1}{\sqrt{np_n}}(\varepsilon_{jk}X_{jk})_{j,k=1}^n$. Let $\lambda_1^{(\varepsilon)},\dots,\lambda_n^{(\varepsilon)}$ denote the (complex) eigenvalues of the matrix $\mathbf{X}^{(\varepsilon)}$ and denote by $G_n^{(\varepsilon)}(x,y)$ the empirical spectral distribution function of the matrix $\mathbf{X}^{(\varepsilon)}$, i. e.

$$G_n^{(\varepsilon)}(x,y) := \frac{1}{n}\sum_{j=1}^n \mathbb{I}\{\mathrm{Re}\,\{\lambda_j^{(\varepsilon)}\} \le x,\ \mathrm{Im}\,\{\lambda_j^{(\varepsilon)}\} \le y\}. \tag{30}$$

Theorem 4.2. *Let* X_{jk} *be independent random variables with*

$$\mathbb{E}X_{jk} = 0, \qquad \mathbb{E}|X_{jk}|^2 = 1 \quad \textit{and} \quad \mathbb{E}|X_{jk}|^2\varphi(X_{jk}) \le \varkappa,$$

where $\varphi(x) = (\ln(1+|x|))^{19+\eta}$ *for some* $\eta > 0$. *Assume that* $p_n^{-1} = \mathcal{O}(n^{1-\theta})$ *for some* $1 \ge \theta > 0$. *Then* $\mathbb{E}G_n^{(\varepsilon)}(x,y)$ *converges weakly to the distribution function* $G(x,y)$ *as* $n \to \infty$.

Remark 4.3. The crucial problem of the proofs of Theorems 4.1 and 4.2 is to find bounds for the smallest singular values $s_n(z)$ respectively $s_n^{(\varepsilon)}(z)$ of the shifted matrices $\mathbf{X}-z\mathbf{I}$ respectively $\mathbf{X}^{(\varepsilon)}-z\mathbf{I}$. These bounds are based on the results obtained by Rudelson and Vershynin in [32]. In the version of paper [25] we have used the corresponding results of Rudelson [31] proving the circular law in the case of i.i.d. sub-Gaussian random variables. In fact, the results in [25] actually imply the circular law for i.i.d. random variables with $\mathbb{E}|X_{jk}|^4 \le \varkappa_4 < \infty$ in view of the fact (explicitly stated by Rudelson in [31]) that in his results the sub-Gaussian condition is needed for the proof of $\mathbb{P}\{\|\mathbf{X}\| > K\} \le C\exp\{-cn\}$ only. This result was written by Pan and Zhou in [30].

The strong circular law assuming moment condition of order larger than 2 and comparable sparsity assumptions was proved by Tao and Vu in [33] based on their results in [34] in connection with the multivariate Littlewood Offord problem. In [35] Tao and Vu proved the circular law without sparsity assuming a moment condition of order 2 only.

The investigation of the convergence of the spectral distribution functions of real or complex (non-symmetric and non-Hermitian) random matrices with independent entries has a long history. Ginibre in 1965, [14], studied the real, complex and quaternion matrices with i. i. d. Gaussian entries. He derived the joint density for the distribution of eigenvalues of matrix and determined the density of the expected spectral distribution function of random matrix with Gaussian entries with independent real and imaginary parts and deduced the circle law. Using the Ginibre

results, Edelman in 1997, [12], proved the circular law for matrices with i. i. d. Gaussian real entries. Girko in 1984, [15], investigated the circular law for general matrices with independent entries assuming that the distributions of the entries have densities. As pointed out by Bai [7], Girko's proof had serious gaps. Bai in [7] gave a proof of the circular law for random matrices with independent entries assuming that the entries have bounded densities and finite sixth moments. It would be interesting to consider the following generalization of the circular law.

4.2 Asymptotic Spectrum of the Product of Independent Random Matrices

Let $m \geq 1$ be a fixed integer. For any $n \geq 1$ consider mutually independent identically distributed (i.i.d.) omplex random variables $X_{jk}^{(\nu)}$, $1 \leq j,k \leq n$, $\nu = 1,\dots,m$, with $\mathbb{E}X_{jk}^{(\nu)} = 0$ and $\mathbb{E}|X_{jk}^{(\nu)}|^2 = 1$ defined on a common probability space $(\Omega_n, \mathbb{F}_n, \mathbb{P})$. Let $\mathbf{X}^{(\nu)}$ denote the $n \times n$ matrix with entries $[\mathbf{X}^{(\nu)}]_{jk} = \frac{1}{\sqrt{n}} X_{jk}^{(\nu)}$, for $1 \leq j,k \leq n$. Denote by $\lambda_1,\dots,\lambda_n$ the eigenvalues of the random matrix $\mathbf{W} := \prod_{\nu=1}^{m} \mathbf{X}^{(\nu)}$ and define its empirical spectral distribution function by

$$\mathcal{F}_n(x,y) := \mathcal{F}_n^{(m)}(x,y) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{\mathrm{Re}\,\lambda_k \leq x, \mathrm{Im}\,\lambda_k \leq y\},$$

where $\mathbb{I}\{B\}$ denotes the indicator of an event B. We shall investigate the convergence of the expected spectral distribution $F_n(x,y) = \mathbb{E}\mathcal{F}_n(x,y)$ to the distribution function $F(x,y)$ corresponding to the m-th power of the uniform distribution on the unit disc in the plane $\mathbb{R}^2$ with Lebesgue-density

$$f(x,y) = \frac{1}{\pi m (x^2+y^2)^{\frac{m-1}{m}}} \mathbb{I}\{x^2+y^2 \leq 1\}.$$

We consider the Kolmogorov distance between the distributions $F_n(x,y)$ and $F(x,y)$,

$$\Delta_n := \sup_{x,y} |F_n(x,y) - F(x,y)|.$$

We have proved the following.

Theorem 4.4. *Let $\mathbb{E}X_{jk}^{(\nu)} = 0$, $\mathbb{E}|X_{jk}^{(\nu)}|^2 = 1$. Then, for any fixed $m \geq 1$,*

$$\lim_{n\to\infty} \Delta_n = 0.$$

The result holds in the non-i.i.d. case as well.

Theorem 4.5. *Let* $\mathbb{E}X_{jk}^{(\nu)} = 0$, $\mathbb{E}|X_{jk}^{(\nu)}|^2 = 1$, $\nu = 1,\dots,m$, $j,k = 1,\dots,n$. *Assume that the random variables* $X_{jk}^{(\nu)}$ *have uniformly integrable second moments, i. e.*

$$\max_{\nu,j,k,n} \mathbb{E}|X_{jk}^{(\nu)}|^2 \mathbb{I}\{|X_{jk}^{(\nu)}| > M\} \to 0 \quad as \quad M \to \infty. \tag{31}$$

Then for any fixed $m \geq 1$,

$$\lim_{n\to\infty} \Delta_n = 0.$$

Definition 4.6. Let $\mu_n(\cdot)$ denote the empirical spectral measure of an $n \times n$ random matrix $\mathbf{X}$ (uniform distribution on the eigenvalues of matrix $\mathbf{X}$) and let $\mu(\cdot)$ denote the uniform distribution on the unit disc in the complex plane $\mathbb{C}$. We say that the circular law holds for the random matrices $\mathbf{X}$ if $\mathbb{E}\mu_n(\cdot)$ converges weakly to the measure $\mu(\cdot)$ in the complex plane $\mathbb{C}$.

Remark 4.7. For $m = 1$ we recover the well-known circular law for random matrices [19, 35].

5 Bounds on Levy and Kolmogorov Distance in Terms of Stieltjes Transform

One of the first bounds on the Kolmogorov distance between distribution functions via their Stieltjes transforms was obtained by Girko in [17]. Bai in [5] proved a new inequality bounding the Kolmogorov distance of distribution functions by their Stieltjes transforms. The proofs of Theorems 2.1–3.3 are based on a smoothing inequality for the Kolmogorov distance between distribution functions in terms of their Stielties transform. Recall that the Stieltjes transform $S_F(z)$ of a distribution function $F(x)$ is defined by the equality

$$S_F(z) := \int_{\mathbb{R}} \frac{1}{z-x} dF(x),$$

for all $z = u + iv$ with $u \in \mathbb{R}$ and $v > 0$. For any distribution functions F and G define the Levy distance as

$$L(F,G) := \inf\{\delta > 0 : \ F(x-\delta) - \delta \leq G(x) \leq F(x+\delta) + \delta, \quad \text{for all } x \in \mathbb{R}\}. \tag{32}$$

In [13] the following result was proved.

Theorem 5.1. *Let* F *and* G *be distribution functions. Given* $v > 0$, *let an interval* $[\alpha,\beta] \subset \mathbb{R}$ *be chosen to satisfy* $G(\alpha) < v$ *and* $1 - G(\beta) < v$. *Then*

$$L(F,G) \le \sup_{x\in[\alpha-2v,\beta+2v]} \left| \int_{-\infty}^{x} (S_F(x+iv) - S_G(x+iv))dx \right| + 4v + 50\mathrm{Im}\, S_G(x+iv). \tag{33}$$

The following corollaries are very important for applications.

Corollary 5.2 (Bobkov et al. [13]). *Let F and G be arbitrary distribution functions. With some universal constant $c > 0$, for any $v_1 > v_0 > 0$,*

$$cL(F,G) \le v_0 + v_0 \sup_{x\in\mathbb{R}} \mathrm{Im}\, S_G(x+iv_0) + \int_{\mathbb{R}} |S_G(u+iv_1) - S_F(u+iv_1)|du + \sup_{x\in[\alpha-2v,\beta+2v]} \int_{v_0}^{v_1} |S_G(x+iv) - S_F(x+iv)|dv, \tag{34}$$

where $\alpha < \beta$ are chosen to satisfy $G(\alpha) < v_0$, and $1 - G(\beta) < v_0$.

Corollary 5.3. *If G is the distribution function of the standard semi-circular law, and F is any distribution function, we have for all $v_1 > v_0 > 0$, up to some universal constant $c > 0$,*

$$c\|F-G\| := c \sup_{x\in\mathbb{R}} |F(x) - G(x)| \le v_0 + \int_{\mathbb{R}} |S_G(u+iv_1) - S_F(u+iv_1)|du + \sup_{x\in[\alpha-2v,\beta+2v]} \int_{v_0}^{v_1} |S_G(x+iv) - S_F(x+iv)|dv \tag{35}$$

This result improved a similar inequality (2.4) in [20]. The main idea of such type of inequalities belongs to F. Götze. We consider the first integral in the right hand side of (35) ("horizontal") far from the real line. A distance from a real line in the second integral ("vertical") has an order $O(n^{-1}\log n^b)$ in one point only. To obtain a bound of order $O(n^{-1}\log n^b)$ for Δ_n we need some modification of the last result. Let $\gamma = \sqrt{4-x^2}$.

Theorem 5.4. *Let $v > 0$ and a and $\varepsilon > 0$ be positive numbers such that*

$$\alpha = \frac{1}{\pi} \int_{|u|\le a} \frac{1}{u^2+1} du = \frac{3}{4}, \tag{36}$$

and

$$2va \le \varepsilon\sqrt{\gamma}. \tag{37}$$

If G denotes the distribution function of the standard semi-circular law, and F is any distribution function, there exists some absolute constant $c > 0$, such that

$$c\|F-G\| \le \int_{-\infty}^{\infty} |S_F(u+iV)-S_G(u+iV)|du + v + \varepsilon^{\frac{3}{2}} + \sup_{x\in[\alpha-2v,\beta+2v]} \int_{v'}^{V} |S_F(x+iu)-S_G(x+iu)|du, \qquad (38)$$

where α, β are defined in Theorem 5.2 and $v' = v/\sqrt{\gamma}$.

In the inequality (38) the right hand side is "sensitive" to the closeness of the point x to the end points of the support of semi-circular distribution function.

6 Stein's Method for Random Matrices

One of the more interesting direction of our joint work with F. Götze was a development of Stein's method for random matrices. This idea belongs exclusively to F. Götze. The obtained results were published in several papers [22, 24, 25], and we give a short review of them in this section. The goal of this review is to illustrate the possibilities of Stein's method for the investigation of the convergence of the empirical spectral distribution function of random matrices. We consider two ensembles of random matrices: real symmetric matrices and sample covariance matrices of real observations. We give a simple characterization of both semicircle and Marchenko-Pastur distributions via linear differential equations. Using conjugate differential operators, we give a simple criterion for convergence to these distributions. We state also the general sufficient conditions for the convergence of the expected spectral distribution functions of random matrices.

6.1 Real Symmetric Matrices

Let $X_{jk}, 1 \le j \le k < \infty$, be a triangular array of random variables with $\mathbb{E}X_{jk} = 0$ and $\mathbb{E}X_{jk}^2 = \sigma_{jk}^2$, and let $X_{kj} = X_{jk}$, for $1 \le j < k < \infty$. For a fixed $n \ge 1$, denote by $\lambda_1 \le \ldots \le \lambda_n$ the eigenvalues of a symmetric $n \times n$ matrix

$$\mathbf{W}_n = (W_n(jk))_{j,k=1}^n, \quad W_n(jk) = \frac{1}{\sqrt{n}} X_{jk}, \text{ for } 1 \le j \le k \le n, \qquad (39)$$

and define its empirical spectral distribution function by

$$\mathcal{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{\lambda_j \le x\}. \qquad (40)$$

We investigate the convergence of the expected spectral distribution function, $F_n(x) := \mathbb{E}\mathcal{F}_n(x)$, to the distribution function of Wigner's semicircle law.

Let $g(x)$ and $G(x)$ denote the density and the distribution function of the standard semicircle law, i.e.

$$g(x) = \frac{1}{2\pi}\sqrt{4-x^2}\,\mathbb{I}\{|x| \le 2\}. \quad G(x) = \int_{-\infty}^{x} g(u)du. \tag{41}$$

6.2 Stein's Equation for the Semicircle Law

Introduce a class of functions

$$\mathbb{C}^1_{\{-2,2\}} = \{f : \mathbb{R} \to \mathbb{R} : f \in \mathbb{C}^1(\mathbb{R} \setminus \{-2,2\});$$
$$\overline{\lim}_{|y|\to\infty}|yf(y)| < \infty;\ \limsup_{y\to\pm 2}|4-y^2||f'(y)| < C\}.$$

By $\mathbb{C}(\mathbb{R})$ we denote the class of continuous functions on $\mathbb{R}$, by $\mathbb{C}^1(B)$, $B \subset \mathbb{R}$, we denote the class of all functions $f : \mathbb{R} \to \mathbb{R}$ differentiable on B with bounded derivative on all compact subsets of B. We state the following

Lemma 6.1. *Assume that a bounded function $\varphi(x)$ without discontinuity of second order satisfies the following conditions*

$$\varphi(x) \text{ is continuous at the points } x = \pm 2 \tag{42}$$

and

$$\int_{-2}^{2} \varphi(u)\sqrt{4-u^2}du = 0. \tag{43}$$

Then there exists a function $f \in \mathbb{C}^1_{\{-2,2\}}$ such that, for any $x \neq \pm 2$,

$$(4-x^2)f'(x) - 3xf(x) = \varphi(x). \tag{44}$$

If $\varphi(\pm 2) = 0$ then there exists a continuous solution of (44).

As a simple implication of this Lemma we get

Proposition 6.2. *The random variable ξ has distribution function $G(x)$ if and only if the following equality holds, for any function $f \in \mathbb{C}^1_{\{-2,2\}}$,*

$$\mathbb{E}\left((4-\xi^2)f'(\xi) - 3\xi f(\xi)\right) = 0. \tag{45}$$

6.3 Stein Criterion for Random Matrices

Let $\mathbf{W}$ denote a symmetric random matrix with eigenvalues $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$. If $\mathbf{W} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix, one defines $f(\mathbf{W}) = \mathbf{U}^{-1} f(\mathbf{\Lambda})\mathbf{U}$, where $f(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))$.

We can now formulate the convergence to the semicircle law for the spectral distribution function of random matrices.

Theorem 6.3. *Let $\mathbf{W}_n$ denote a sequence of random matrices of order $n \times n$ such that, for any function $f \in \mathbb{C}^1_{\{-2,2\}}$*

$$\frac{1}{n}\mathbb{E}\mathrm{Tr}(4\mathbf{I}_n - \mathbf{W}_n^2) f'(\mathbf{W}_n) - \frac{3}{n}\mathbb{E}\mathrm{Tr}\mathbf{W}_n f(\mathbf{W}_n) \to 0, \quad \text{as } n \to \infty. \tag{46}$$

Then

$$\Delta_n := \sup_x |\mathbb{E}F_n(x) - G(x)| \to 0, \text{ as } n \to \infty. \tag{47}$$

6.4 Resolvent Criterion for the Spectral Distribution Function of a Random Matrix

We introduce the resolvent matrix for a symmetric matrix $\mathbf{W}$ and any non-real z,

$$\mathbf{R}(z) = (\mathbf{W} - z\mathbf{I})^{-1}, \tag{48}$$

where $\mathbf{I}$ denotes the identity matrix of order $n \times n$.

Proposition 6.4. *Assume that, for any $v \neq 0$,*

$$\mathcal{R}_n(\mathbf{W})(z) := \frac{1}{n}\mathbb{E}\mathrm{Tr}(4\mathbf{I} - \mathbf{W}^2)\mathbf{R}^2(z) + \frac{3}{n}\mathbb{E}\mathrm{Tr}\mathbf{W}\mathbf{R}(z) \to 0, \text{ as } n \to \infty \tag{49}$$

uniformly on compact sets in $\mathbb{C}\backslash\mathbb{R}$. Then

$$\Delta_n \to 0, \text{ as } n \to \infty. \tag{50}$$

6.5 General Conditions of the Convergence of the Expected Distribution Function of Random Matrices to the Semi Circular Law

We shall assume that $\mathbb{E}X_{jl} = 0$ and $\sigma_{jl}^2 := \mathbb{E}X_{jl}^2$, for $1 \le j \le l \le n$. Introduce σ-algebras $\mathcal{F}^{jl} = \sigma\{X_{km} : 1 \le k \le m \le n,\ \{k,m\} \neq \{j,l\}\}$, $1 \le j \le l \le n$, and

$\mathcal{F}^j = \sigma\{X_{km} : 1 \le k \le m \le n,\ k \ne j \text{ and } m \ne j\}$, $1 \le j \le n$. We introduce as well Lindeberg's ratio for random matrices, that is for any $\tau > 0$,

$$L_n(\tau) := \frac{1}{n^2} \sum_{j,l=1}^{n} \mathbb{E} X_{jl}^2 I_{\{|X_{jl}| > \tau\sqrt{n}\}}. \tag{51}$$

Theorem 6.5. *Assume that the random variables $X_{jl}, 1 \le j \le l \le n,\ n \ge 1$ satisfy the following conditions*

$$\mathbb{E}\{X_{jl} | \mathcal{F}^{jl}\} = 0, \tag{52}$$

$$\varepsilon_n^{(1)} := \frac{1}{n^2} \sum_{1 \le j \le l \le n} \mathbb{E} |\mathbb{E}\{X_{jl}^2 | \mathcal{F}^j\} - \sigma_{jl}^2| \to 0 \quad as \quad n \to \infty, \tag{53}$$

there exists $\sigma^2 > 0$, such that

$$\varepsilon_n^{(2)} := \frac{1}{n^2} \sum_{1 \le j \le l \le n} |\sigma_{jl}^2 - \sigma^2| \to 0 \quad as \quad n \to \infty, \tag{54}$$

and

for any fixed $\tau > 0$,

$$L_n(\tau) \to 0 \quad as \quad n \to \infty. \tag{55}$$

Then

$$\Delta_n := \sup_x |\mathbb{E} F_n(x) - G(x\sigma^{-1})| \to 0 \quad as \quad n \to \infty. \tag{56}$$

Corollary 6.6. *Let $X_{lj}^{(n)}$, $1 \le l \le j \le n$ be distributed uniformly in the ball of the radius $\sqrt{N}$ in $\mathbb{R}^N$ with $N = \frac{n(n+1)}{2}$, for any $n \ge 1$. Then*

$$\Delta_n \to 0, \quad as \quad n \to \infty. \tag{57}$$

6.6 Sample Covariance Matrices

Let $X_{jk}, 1 \le j, k < \infty$, be random variables with $\mathbb{E} X_{jk} = 0$ and $\mathbb{E} X_{jk}^2 = \sigma_{jk}^2$. For fixed $n \ge 1$ and $m \ge 1$, we introduce a matrix $n \times m$

$$\mathbf{X} = \left(X_{lj}\right)_{1 \le l \le n,\ 1 \le j \le m}. \tag{58}$$

Denote by $\lambda_1 \leq \ldots \leq \lambda_n$ the eigenvalues of the symmetric $n \times n$ matrix

$$\mathbf{W}_n = \frac{1}{p}\mathbf{X}\mathbf{X}^T, \tag{59}$$

and define its empirical spectral distribution function by

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{\lambda_j \leq x\}. \tag{60}$$

We investigate the convergence of the expected spectral distribution function $\mathbb{E}F_n(x)$ to the distribution function of the Marchenko-Pastur law.

Let $g_\alpha(x)$ and $G_\alpha(x)$ denote the density and the distribution function of the Marchenko-Pastur law with parameter $\alpha \in (0, \infty)$, that is

$$g_\alpha(x) = \frac{1}{x\pi}\sqrt{(x-a)(b-x)}\, I_{\{x\in[a,b]\}}, \quad G_\alpha(x) = \int_{-\infty}^{x} g_\alpha(u)du, \tag{61}$$

where $a = (1-\sqrt{\alpha})^2$, $b = (1+\sqrt{\alpha})^2$.

6.7 Stein's Equation for the Marchenko-Pastur Law

Introduce a class of functions

$$\mathbb{C}^1_{\{a,b\}} = \{f : \mathbb{R} \to \mathbb{R} : f \in \mathbb{C}^1(\mathbb{R} \setminus \{a,b\});$$

$$\overline{\lim}_{|y|\to\infty}|yf(y)| < \infty; \quad \limsup_{y \to \frac{a-b}{2} \pm \frac{a+b}{2}} |(\frac{(a-b)^2}{4} - (y - \frac{a+b}{2})^2||f'(y)| < C\}.$$

At first we state the following

Lemma 6.7. *Let $\alpha \neq 1$. Assume that a bounded function $\varphi(x)$ without discontinuity of second order satisfies the following conditions*

$$\varphi(x) \text{ is continuous in the points } x = a, x = b \tag{62}$$

and

$$\int_a^b \varphi(u)g_\alpha(u)du = 0. \tag{63}$$

Then there exists a function $f \in \mathbb{C}^1_{\{a,b\}}$ such that, for any $x \neq a$ or $x \neq b$,

$$(x-a)(b-x)xf'(x) - 3x(x-\frac{a+b}{2})f(x) = \varphi(x). \tag{64}$$

If $\varphi(a) = 0$ ($\varphi(b) = 0$) then there exists a continuous solution of the equation (64).

Proposition 6.8. *The random variable ξ has distribution function $G_\alpha(x)$ if and only if the following equality holds, for any function $f \in \mathbb{C}^1_{\{a,b\}}$,*

$$\mathbb{E}\left((\xi-a)(b-\xi)\xi f'(\xi) - 3\xi(\xi-\frac{a+b}{2})f(\xi)\right) = 0. \tag{65}$$

6.8 Stein's Criterion for Sample Covariance Matrices

Let $\mathbf{W}$ denote a sample covariance matrix with eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$. If $\mathbf{W} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{U}$ is an orthogonal and $\mathbf{\Lambda}$ a diagonal matrix, one defines $f(\mathbf{W}) = \mathbf{U}^{-1}f(\mathbf{\Lambda})\mathbf{U}$, where $f(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))$.

We can now formulate the convergence to the Marchenko-Pastur law for the spectral distribution function of random matrices.

Theorem 6.9. *Let $\mathbf{W}_n$ denote a sequence of sample covariance matrices of order $n \times n$ such that, for any function $f \in \mathbb{C}^1_{\{a,b\}}$*

$$\frac{1}{n}\mathbb{E}\mathrm{Tr}(\mathbf{W}_n - a\mathbf{I}_n)(b\mathbf{I}_n - \mathbf{W}_n)\mathbf{W}_n f'(\mathbf{W}_n) - \frac{3}{n}\mathbb{E}\mathrm{Tr}\mathbf{W}_n(\mathbf{W}_n - \frac{a+b}{2}\mathbf{I}_n)f(\mathbf{W}_n) \to 0, \quad as\ n \to \infty. \tag{66}$$

Then

$$\Delta_n := \sup_x |\mathbb{E}F_n(x) - G_\alpha(x)| \to 0,\ as\ n \to \infty. \tag{67}$$

6.9 Resolvent Criterion for Sample Covariance Matrices

Denote by $\mathbf{R}(z)$ the resolvent matrix for the sample covariance matrix $\mathbf{W}$.

Proposition 6.10. *Assume that, for any $v \neq 0$,*

$$\mathcal{R}_n(\mathbf{W})(z) := \frac{1}{n}\mathbb{E}\mathrm{Tr}\mathbf{W}(\mathbf{W} - a\mathbf{I})(b\mathbf{I} - \mathbf{W})\mathbf{R}^2(z) + \frac{3}{n}\mathbb{E}\mathrm{Tr}(\mathbf{W} - \frac{a+b}{2}\mathbf{I})\mathbf{W}\mathbf{R}(z) \to 0,\ as\ n \to \infty \tag{68}$$

uniformly on compacts sets in $\mathbb{C}\backslash\mathbb{R}$. *Then*

$$\Delta_n \to 0, \text{ as } n \to \infty. \tag{69}$$

6.10 Convergence to the Marchenko-Pastur Distribution

We shall assume that $\mathbb{E}X_{jl} = 0$ and $\sigma_{jl}^2 := \mathbb{E}X_{jl}^2$, for $1 \le j \le n$ and $1 \le l \le m$. Introduce σ-algebras $\mathcal{F}^{jl} = \sigma\{X_{kq} : 1 \le k \le n, 1 \le q \le m, \{k,q\} \ne \{j,l\}\}$, $1 \le j \le n, 1 \le l \le m$, and $\mathcal{F}^{l} = \sigma\{X_{js} : 1 \le j \le n, 1 \le s \le m, s \ne l\}$, $1 \le l \le m$. We introduce as well Lindeberg's ratio for random matrices, that is for any $\tau > 0$,

$$L_n(\tau) = \frac{1}{nm}\sum_{j=1}^{n}\sum_{l=1}^{m}\mathbb{E}X_{jl}^2\mathbb{I}\{|X_{jl}| > \tau\sqrt{n}\}, \tag{70}$$

as well as the notation $X_{jl}^{(\tau)} := X_{jl}\mathbb{I}\{|X_{jl}| \le \tau\sqrt{n}\} - \mathbb{E}\mathbb{X}_{jl}\mathbb{I}\{|X_{jl}| \le \tau\sqrt{n}\}$, $\xi_{jl}^{(\tau)} := \mathbb{E}\left\{X_{jl}^{(\tau)}\middle|\mathcal{F}^{(jl)}\right\}$. Introduce also the vectors $\mathbf{X}_l^{(\tau)} = (X_{1,l}^{(\tau)}, \dots, X_{n,l}^{(\tau)})^T$ and $\xi_{,l}^{(\tau)} = (\xi_{1,l}^{(\tau)}, \dots, \xi_{n,l}^{(\tau)})^T$.

Theorem 6.11. *Let* $m = m(n)$ *depend on* n, *such that*

$$\frac{m(n)}{n} \to \alpha \in (0,1), \text{ as } n \to \infty. \tag{71}$$

Assume that the random variables $X_{jl}, 1 \le j \le n, 1 \le l \le m$, $n, m \ge 1$ *satisfy the following conditions*

$$\mathbb{E}\{X_{jl}|\mathcal{F}^{jl}\} = 0, \tag{72}$$

$$\varepsilon_n^{(1)} := \frac{1}{nm}\sum_{j=1}^{n}\sum_{l=1}^{m}\mathbb{E}|\mathbb{E}\{X_{jl}^2|\mathcal{F}^{l}\} - \sigma_{jl}^2| \to 0 \quad as \quad n \to \infty, \tag{73}$$

there exists $\sigma^2 > 0$, *such that*

$$\varepsilon_n^{(2)} := \frac{1}{nm}\sum_{j=1}^{n}\sum_{l=1}^{m}|\sigma_{jl}^2 - \sigma^2| \to 0, \tag{74}$$

$$\varepsilon_n^{(3)} := \frac{1}{nm^2}\sum_{l=1}^{m}\sum_{j,k=1}^{n}\mathbb{E}\left|\mathbb{E}\{((X_{jl}^{(\tau)})^2 - \mathbb{E}(X_{jl}^{(\tau)})^2)((X_{kl}^{(\tau)})^2 - \mathbb{E}(X_{kl}^{(\tau)})^2)\middle|\mathcal{F}^{l}\}\right| \to 0, \tag{75}$$

$$\varepsilon_n^{(4)} := \frac{1}{nm^2}\sum_{l=1}^{m}\sum_{1\le j\ne k\le n}\mathbb{E}\left|(\mathbb{E}\{((\xi_{jl}^{(\tau)})^2-\mathbb{E}(\xi_{jl}^{(\tau)})^2)\right.$$
$$\left.\times\,((\xi_{kl}^{(\tau)})^2-\mathbb{E}(\xi_{kl}^{(\tau)})^2)\Big|\mathcal{F}^l\}\right|\to 0,\quad as\ n\to\infty, \tag{76}$$

and

$$L_n(\tau)\to 0,\quad for\ any\ fixed\ \tau>0,\quad as\quad n\to\infty. \tag{77}$$

Then

$$\Delta_n := \sup_x|\mathbb{E}F_n(x)-G_\alpha(x\sigma^{-1})|\to 0\quad as\quad n\to\infty. \tag{78}$$

Remark 6.12. Note that condition (74) implies that

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\mathrm{Tr}\,\mathbf{W}_n=\sigma^2<\infty. \tag{79}$$

Corollary 6.13. *Assume (71). Let, for any $n,m\ge 1$, X_{jl}, $1\le j\le n, 1\le l\le m$, be independent and $\mathbb{E}X_{jl}=0$, $\mathbb{E}X_{jl}^2=\sigma^2$. Suppose that, for any fixed $\tau>0$,*

$$L_n(\tau)\to 0,\quad as\quad n\to\infty. \tag{80}$$

Then the expected spectral distribution function of the sample covariance matrix W converges to the Marchenko-Pastur distribution,

$$\Delta_n := \sup_x|\mathbb{E}F_n(x)-G_\alpha(x\sigma^{-1})|\to 0,\quad as\quad n\to\infty. \tag{81}$$

Acknowledgements Alexander Tikhomirov partially was supported by SFB 701, by grants of RFBR N 11-01-00310-a, N 11-01-122104-ofi-m-2011, and by Program of Basic Research of Urals Division of RAS.

References

1. N. Alexeev, F. Götze, A.N. Tikhomirov, On the singular spectrum of powers and products of random matrices. Dokl. Math. **82**(1), 505–507 (2010)
2. N. Alexeev, F. Götze, A.N. Tikhomirov, Asymptotic distribution of singular values of powers of random matrices. Lithuanian math. J. **50**(2), 121–132 (2010)
3. N. Alexeev, F. Götze, A.N. Tikhomirov, On the asymptotic distribution of singular values of products of large rectangular random matrices. J. Math. Sci. **408**, 9–43 (2012 (Zapiski nauchnyh seminarov POMI (in Russia)). arXiv:1012.2586
4. N. Alexeev, F. Götze, A.N. Tikhomirov, On the asymptotic distribution of the singular values of powers of random matrices. Preprint, arXiv:1012.2743
5. Z.D. Bai, Convergence rate of expected spectral distributions of large random matrices. I. Wigner matrices. Ann. Probab. **21**(2), 625–648 (1993)
6. Z.D. Bai, Convergence rate of expected spectral distributions of large random matrices. II. Sample covariance matrices. Ann. Probab. **21**(2), 649–672 (1993)

7. Z.D. Bai, Circular law. Ann. Probab. **25**, 494–529 (1997)
8. Z.D. Bai, Methodologies in spectral analysis of large-dimensional random matrices, a review. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author. Statist. Sinica **9**(3), 611–677 (1999)
9. Z.D. Bai, B. Miao, J. Tsay, Convergence rates of the spectral distributions of large Wigner matrices. Int. Math. J. **1**(1), 65–90 (2002)
10. Z.D. Bai, B. Miao, J.-F. Yao, Convergence rates of spectral distributions of large sample covariance matrices. SIAM J. Matrix Anal. Appl. **25**(1), 105–127 (2003)
11. S.G. Bobkov, F. Götze, A.N. Tikhomirov, On concentration of empirical measures and convergence to the semi-circle law. J. Theoret. Probab. **23**(3), 792–823 (2010)
12. A. Edelman, The probability that a random real Gaussian matrix has k real eigenvalues, related distributions, and circular law. J. Mult. Anal. **60**, 203–232 (1997)
13. L. Erdös, H.-T. Yau, J. Yin, Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. **229**(3), 1435–1515 (2012) arXiv:1007.4652.
14. J. Ginibre, Statistical ensembles of complex, quaternion, and real matrices. J. Math. Phys. **6**, 440–449 (1965)
15. V.L. Girko, The circular law. (Russian) Teor. Veroyatnost. i Primenen. **29**(4), 669–679 (1984)
16. V.L. Girko, Spectral theory of random matrices. (Russian) Uspekhi Mat. Nauk **40** 1(241), 67–106 (1985)
17. V.L. Girko, Asymptotics of the distribution of the spectrum of random matrices. (Russian) Uspekhi Mat. Nauk **44** 4(268), 7–34, 256 (1989); translation in Russ. Math. Surv. **44**(4), 3–36 (1989)
18. V.L. Girko, Extended proof of the statement: convergence rate of expected spectral functions of the sample covariance matrix $\hat{R}_{m_n}(n)$ is equal to $O(n^{-1/2})$ under the condition $\frac{m_n}{n} \leq c < 1$ and the method of critical steepest descent. Random Oper. Stoch. Equat. **10**(4), 351–405 (2002)
19. F. Götze, A. Tikhomirov, The circular law for random matrices. Ann. Probab. **38**(4), 1444–1491 (2010)
20. F. Götze, A.N. Tikhomirov, Rate of convergence to the semi-circular law. Probab. Theor. Relat. Fields **127**, 228–276 (2003)
21. F. Götze, A.N. Tikhomirov, Rate of convergence in probability to the Marchenko-Pastur law. Bernoulli **10**(3), 503–548 (2004)
22. F. Götze, A.N. Tikhomirov, *Limit Theorems for Spectra of Random Matrices with Martingale Structure. Stein's Method and Applications.* Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., vol. 5 (Singapore University Press, Singapore, 2005), pp. 181–193
23. F. Götze, A.N. Tikhomirov, The rate of convergence for spectra of GUE and LUE matrix ensembles. Cent. Eur. J. Math. **3**(4), 666–704 (2005)
24. F. Götze, A. Tikhomirov, Limit theorems for spectra of positive random matrices under dependence. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **311**, Veroyatn. i Stat. 7, 92–123, 299 (2004); translation in J. Math. Sci. (N. Y.) **133**(3), 1257–1276 (2006)
25. F. Götze, A.N. Tikhomirov, Limit theorems for spectra of random matrices with martingale structure. Teor. Veroyatn. Primen. **51**(1), 171–192 (2006); translation in Theor. Probab. Appl. **51**(1), 42–64 (2007)
26. F. Götze, A.N. Tikhomirov, The rate of convergence of spectra of sample covariance matrices. Teor. Veroyatn. Primen. **54**(1), 196–206 (2009); translation in Theor. Probab. Appl. **54**(1), 129–140 (2010)
27. F. Götze, A.N. Tikhomirov, On the Rate of Convergence to the Marchenko–Pastur Distribution. Preprint, arXiv:1110.1284
28. F. Götze, A.N. Tikhomirov, The rate of convergence to the semi-circular law, in *High Dimensional Probability, VI*. Progress in Probability, vol. 66 (Birkhaeuser, Basel, 2013), pp. 141–167. arXiv:1109.0611
29. V. Marchenko, L. Pastur, The eigenvalue distribution in some ensembles of random matrices. Math. USSR Sbornik **1**, 457–483 (1967)
30. G. Pan, W. Zhou, Circular law, extreme singular values and potential theory. J. Multivariate Anal. **101**(3), 645–656 (2010). arXiv:0705.3773

31. M. Rudelson, Invertibility of random matrices: norm of the inverse. Ann. Math. (2) **168**(2), 575–600 (2008). http://arXiv:math/0507024.
32. M. Rudelson, R. Vershynin, The Littlewood–Offord problem and invertibility of random matrices. Adv. Math. **218**(2), 600–633 (2008). http://arXiv.org/abs/math.PR/0703307.
33. T. Tao, V. Vu, Random matrices: the circular law. Comm. Contemp. Math. **10**(2), 261–307 (2008)
34. T. Tao, V. Vu, Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. Ann. Math. **169**(2), 595–632 (2009). http://arXiv:math/0511215.
35. T. Tao, V. Vu, Random matrices: universality of ESDs and the circular law. With an appendix by Manjunath Krishnapur. Ann. Probab. **38**(5), 2023–2065 (2010)
36. A.N. Tikhomirov, On the rate of convergence of the expected spectral distribution function of a Wigner matrix to the semi-circular law. Siberian Adv. Math. **19**(3), 211–223 (2009)
37. A.N. Tikhomirov, The rate of convergence of the expected spectral distribution function of a sample covariance matrix to the Marchenko-Pastur distribution. Siberian Adv. Math. **19**(4), 277–286 (2009)
38. D.A. Timushev, A.N. Tikhomirov, A.A. Kholopov, On the accuracy of the approximation of the GOE spectrum by the semi-circular law. (Russian) Teor. Veroyatn. Primen. **52**(1), 180–185 (2007); translation in Theor. Probab. Appl. **52**(1), 171–177 (2008)
39. E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. (2) **62**, 548–564 (1955)
40. E.P. Wigner, On the distribution of the roots of certain symmetric matrices. Ann. Math. (2) **67**, 325–327 (1958)
41. J. Wishart, The generalised product moment distribution in samples from a normal multivariate population. Biometrika **20A**(1–2), 32–52 (1928)