Additive Number Theory

FESTSCHRIFT IN HONOR OF THE SIXTIETH BIRTHDAY OF MELVYN B. NATHANSON

David Chudnovsky
Gregory Chudnovsky
(Editors)

Springer

Additive Number Theory

David Chudnovsky • Gregory Chudnovsky
Editors

Additive Number Theory

Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson

Springer

Editors
David Chudnovsky
Polytechnic Institute of NYU
IMAS
6 MetroTech Center
Brooklyn, NY 11201, USA
david@imas.poly.edu

Gregory Chudnovsky
Polytechnic Institute of NYU
IMAS
6 MetroTech Center
Brooklyn, NY 11201, USA
gregory@imas.poly.edu

ISBN 978-0-387-37029-3 e-ISBN 978-0-387-68361-4
DOI 10.1007/978-0-387-68361-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010929948

Mathematics Subject Classification (2010): 11P32, 11P70, 11P82, 11P99, 11B13, 11B25, 11B30, 11B83, 11K38

Preface

This Festschrift is dedicated to Melvyn B. Nathanson by his colleagues, friends, and students. This volume celebrates his many contributions to various areas of number theory. Mel's outstanding career as a mathematician and a public figure resulted in many achievements both in science and in the public arena.

It is appropriate to quote here the tribute of the great I.M. Gelfand to Mel:

I remember Melvyn as a young man attending my seminar in Moscow. He participated in my Rutgers seminar as well and taught us a lot of number theory. I enjoy his love of mathematics and the way he thinks about it. I wish him all the best and expect new wonderful results from him.

We thank Jean Bourgain, M.-C. Chang, Javier Cilleruelo, Shalom Eliahou, Christian Elsholtz, Ron Graham, Ben Green, Yahya O. Hamidoune, Peter Hegarty, Alex Iosevich, Sergei V. Konyagin, D. Labrousse, Cédric Lecouvey, Vsevolod F. Lev, Máté Matolcsi, Steven J. Miller, Tom Morgan, Marina Nechayeva, Lan Nguyen, Kevin O'Bryant, J.L. Ramírez Alfonsín, Burton Randol, Øystein J. Rødseth, Svetlana Roudenko, Imre Z. Ruzsa, Ilda da Silva, Jonathan Sondow, Daniel Scheinerman, Oriol Serra, Zhi-Wei Sun, Julia Wolf, and Michael E. Zieve for their contributions to this volume.

David Chudnovsky
Gregory Chudnovsky
Editors

Photo by Alex Nathanson

Contents

Addictive Number Theory

Melvyn B. Nathanson

A True Story

In 1996, just after Springer-Verlag published my books *Additive Number Theory: The Classical Bases* [34] and *Additive Number Theory: Inverse Problems and the Geometry of Sumsets* [35], I went into my local Barnes and Noble superstore on Route 22 in Springfield, New Jersey, and looked for them on the shelves. Suburban bookstores do not usually stock technical mathematical books, and, of course, the books were not there. As an experiment, I asked if they could be ordered. The person at the information desk typed in the titles, and told me that his computer search reported that the books did not exist. However, when I gave him the ISBN numbers, he did find them in the Barnes and Noble database. The problem was that the book titles had been cataloged incorrectly. The data entry person had written *Addictive Number Theory*.[1]

I have always found it addictive to think about mathematics. Of course, as many have observed, it is better for one's career to think about fashionable things, or about things that appeal to fashionable people. To me, fashionable is boring, and I prefer to think about problems that interest almost no one. Of course, if what appeals to you is what is already popular, then that is what you should study. We mathematicians are free to investigate whatever we like.

In the preface to the first volume, *The Classical Bases*, I wrote

> Additive number theory is a deep and beautiful part of mathematics, but for too long it has been obscure and mysterious, the domain of a small number of specialists, who have often been specialists only in their own small part of additive number theory. This is the first

[1] I have told this story many times, and like every good story, it has acquired an independent existence. I have heard others tell variations on the tale, always with the same additive–addictive punch line.

M.B. Nathanson
Department of Mathematics, Lehman College (CUNY), Bronx, New York 10468
and
CUNY Graduate Center, New York, New York 10016
e-mail: melvyn.nathanson@lehman.cuny.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_1,

> of several books on additive number theory. I hope that these books will demonstrate the richness and coherence of the subject and that they will encourage renewed interest in the field.

The results have far exceeded my expectations. The second volume, *Inverse Problems*, has developed into a major field of mathematics, sometimes called "additive combinatorics," and has, *mirabile dictu*, become fashionable. The central result in this book is an extraordinary "inverse theorem" of Gregory Freiman about the structure of a finite set A of integers whose sumset $A + A$ is small. I had been interested in this result for a long time, and, when Freiman emigrated from the former Soviet Union and was invited to the Institute for Advanced Study, I visited him and discussed it with him. He was astonished, and years later remarked, "No one mentioned my theorem for decades until you asked me about it in Princeton." A few years later, after the publication of *Inverse Theorems*, the British mathematician Tim Gowers used Freiman's theorem in his work on effective bounds for Szemerédi's theorem on long arithmetic progressions in dense sets of integers. I met Gowers for the first time also at the Institute for Advanced Study, and he told the following story, which he recounted in a recent email:

> I had got to the stage of understanding that Freiman's theorem would be useful ... but I couldn't understand Freiman's proof, and Ruzsa's was spread over more than one paper and published in obscure journals so I couldn't piece that together either. And then I found myself browsing in the mathematics section of Blackwell's in Oxford (even though I myself am and was at Cambridge), and saw your book. The title was promising, and to my great delight I saw that it contained a full account of Ruzsa's proof. This was a great stroke of luck: your book gave me exactly the help I needed at exactly the right time.

Gowers received a Fields Medal in large part for his work on Szemerédi's theorem.

To veterans in combinatorial and additive number theory, who are used to at best benign neglect and at worst scorn and ridicule, this is an astounding transformation. Paul Erdős, one of the great figures in 20th century mathematics, was not highly regarded by the mathematical mafiosi. Combinatorial and additive number theory have only recently come into fashion, but even now, attention is paid to only a small part of the subject, the part connected with harmonic analysis and ergodic theory. This is because there have been and continue to be remarkable theorems arising from the union of analysis and combinatorial number theory, and everyone focuses on (that is, the herd stampedes toward) the successful. In the next few years I plan to complete at least two more volumes on additive number theory, with an emphasis on other strange and beautiful but still not well known results. It will be curious to see if suddenly they, too, become hot topics.

Remarks on Some of My Articles

The editors of this volume have asked me to comment on some of my articles that I particularly like. The first, of course, are juvenilia: Articles that I wrote while I was a graduate student in mathematics at the University of Rochester from 1966 to 1971.

(My mathematical life started rather late: I studied philosophy as an undergraduate at the University of Pennsylvania, and then spent a year at Harvard as a graduate student in biophysics before switching to math.) My Rochester advisor was Sanford L. Segal [53, 54], an erudite and charming analytic number theorist and historian of mathematics under the Nazis. Our work did not intersect, but many years later I wrote a short article on functional equations [31] that Sandy generalized [52].

My Rochester articles were on a variety of topics, for example, an exponential diophantine equation [24, 56], the greatest order of an element from the symmetric group [26] (I subsequently learned that I did not invent this problem, and that Edmund Landau [19] had used prime number theory to determine the asymptotics), complementing sets of lattice points [23], the fundamental domain of a discrete group [28], and a result, sometimes called the "fundamental theorem of additive number theory," about the structure of the iterated sumsets hA of a finite set of integers [27]. Many years later, my student Sandie Han, Christoph Kirfel, and I extended this to linear forms [12], and I later generalized a related result of Khovanskii [15, 16] to linear forms in abelian semigroups [37]. The latter result used some commutative algebra, specifically, the Hilbert polynomial in several variables for finitely generated algebras. Ruzsa and I have published a purely combinatorial proof [49]. My student Jaewoo Lee has studied a related problem [20].

In 1970 I spent the Lent and Easter terms as a visiting research student at the University of Cambridge in DPPMS, the Department of Pure Mathematics and Mathematical Statistics, in its former building at 16 Mill Lane. One of the reasons I went to Cambridge was to talk to Cassels, who had written two beautiful articles [3, 4] on the Catalan conjecture ("8 and 9 are the only consecutive powers"). Another forgotten bit of juvenilia is my proof that the analog of the Catalan conjecture is true in any field of rational functions [29]. A friend at Cambridge was Béla Bollobás, and it may have been Béla who first introduced me to Erdős.

My plan was to stay in Europe for the summer and attend the International Congress of Mathematicians in Nice in September. At the end of the academic year I travelled to Russia, and then to Hungary and Israel, where I wanted to find a university where I could work on my own. I showed up unannounced at the Weizmann Institute of Science in Rehovot, and told someone that I was looking for a place to study. I was sent to a math professor there, Shlomo Sternberg, who asked what I was interested in. I told him about additive number theory. "No one in Israel is interested in that," he said, "so you might as well stay here." Weizmann gave me an office and library access, and found a place for me to live. Browsing in the journals in the library, I learned about an idea of Milnor to define a "random" binary sequence, and wrote my first articles, "Derivatives of binary sequences" [22] and "Integrals of binary sequences" [25], which were published in the *SIAM Journal of Applied Mathematics*.

The Weizmann Institute library had a copy of Halberstam and Roth's book *Sequences, Vol. I* [11], which I carefully studied. (I gave a lecture at Weizmann in 2001, and looked for the book in the library. It was still on the shelf. No one had signed it out since I did in 1970.) I became and am still fascinated by the Erdős-Turán conjecture that the representation function of an asymptotic basis for the

nonnegative integers of order two must be unbounded. In the process of trying to construct a counterexample, I invented the concept of a *minimal asymptotic basis*, which is a set A of nonnegative integers with the property that the sumset $A + A$ contains all sufficiently large integers, but, for every element $a^* \in A$, there are infinitely many positive integers that cannot be represented as the sum of two elements from the set $A \setminus \{a^*\}$. I constructed explicit examples of minimal asymptotic bases. This was my first original idea about additive bases. Later I learned that minimal bases had been previously defined by Stöhr [55], and that Härtter [13] had proved their existence, but that I had constructed the first nontrivial examples. Many years later I realized that the opposite of the Erdős-Turán conjecture holds for bases for the additive group of all integers, and that every function $f : \mathbf{Z} \to \mathbf{N}_0 \cup \{\infty\}$ with only finitely many zeros is the representation function of an asymptotic basis for $\mathbf{Z}$ [5, 39, 41–43]. This is essentially what distinguishes a group and a semigroup.

In September, 1971, I began my first job, as an instructor at Southern Ilinois University in Carbondale. There were two other number theorists there, Lauwerens Kuipers and Harald Niederreiter, who were completing their monograph *Uniform Distribution of Sequences* [18]. SIU had a Ph.D. program in mathematics, an excellent library, and an atmosphere that was, for me, conducive to research. I continued to think about minimal bases. Driving home to Philadelphia from Carbondale for Thanksgiving, I realized that the set B of nonnegative even integers has the property that infinitely many positive integers (i.e. the odd numbers) cannot be represented as the sum of two elements of B, but that, if b^* is any nonnegative integer not in B (i.e. any odd positive integer) then the set $B \cup \{b^*\}$ is an asymptotic basis of order 2. Thus, B can reasonably be called a *maximal asymptotic nonbasis*, which is the natural dual of a minimal asymptotic basis. I was able to describe all maximal asymptotic nonbases consisting of unions of congruence classes, and also to construct examples of other types of maximal asymptotic nonbases.

I combined my various results in the article "Minimal bases and maximal nonbases in additive number theory," which appeared in the *Journal of Number Theory* [30]. The article contained a list of unsolved problems. I had mailed a preprint to Erdős in Budapest. In a short time I received a letter from him with a presumptive solution to one of the problems. I found his proof difficult, and worked hard to understand it. Finally I understood the idea of the proof, but I also realized that the proof was wrong, and that, modifying the argument, I could prove exactly the opposite of what Erdős had claimed was true. This did answer my question, but with a "change of sign." We published this in "Maximal asymptotic nonbases" [6], the first of nearly 20 articles that Erdős and I wrote together. My two favorite articles with Erdős are on oscillations of bases [7] and on representation functions of minimal bases [8].

Although I was on the faculty of SIU from 1971 to 1981, I was actually on leave for four of my first 7 years. I received an IREX fellowship for the academic year 1972–1973 to study with Gel'fand at Moscow State University in the USSR. One result was the article "Classification problems in K-categories" [33]. In 1974–1975 I was appointed Assistant to André Weil at the Institute for Advanced Study. I arrived in Princeton in the summer, when Weil was in Paris. When he returned in

the fall, I asked him, "As your Assistant, what do I have to do for you?" He replied, "Nothing, and conversely." A few weeks later, however, he asked if I would take notes of his lectures on the history of number theory, which became Weil's book *Elliptic Functions according to Eisenstein and Kronecker* [57]. I spent 1975–1976 at Rockefeller University and Brooklyn College (CUNY), and 1977–1978 at Harvard University. In addition to my appointment in mathematics at Harvard, I was also a member of the nuclear nonproliferation working group of the Program for Science and International Affairs (now the Belfer Center for Science and International Affairs in the Kennedy School of Government), and we wrote a book, *Nuclear Nonproliferation: The Spent Fuel Problem* [10]. About this time I also wrote another nonmathematical book, *Komar-Melamid: Two Soviet Dissident Artists* [32].

From 1981–1986 I was Dean of the Graduate School of Rutgers-Newark and on the doctoral mathematics faculty at Rutgers-New Brunswick. My Rutgers Ph. D. student John C. M. Nash and I wrote "Cofinite subsets of asymptotic bases for the positive integers" [21]. Since 1986 I have been Professor of Mathematics at Lehman College (CUNY) and the CUNY Graduate Center. For the first 5 years (1986–1991) I was also Provost at Lehman. During 10 years of administrative duty I was, to Erdős' satisfaction, still able to find the time to prove and conjecture, and published many articles. With my CUNY Ph.D. student Xing-De Jia I wrote several articles, including a new construction of thin minimal asymptotic bases [14].

For many years I was also an adjunct member of the faculty of Rockefeller University in the laboratory of Morris Schreiber. At Rockefeller in 1976, I organized my first number theory conference. Erdős gave a lecture in which he discussed the following problem about the number of sums and products of a finite set of positive integers: Prove that for every $\varepsilon > 0$ there exists a number $K(\varepsilon)$ such that, if A is a set of k positive integers and $k \geq K(\varepsilon)$ then there are at least $k^{2-\varepsilon}$ integers that can be represented in the form $a + a'$ or aa' with $a, a' \in A$. At the time, there were no results on this problem, but in 1983 Erdős and Szemerédi [9] proved that there exists a $\delta > 0$ such that the number of sums and products is at least $k^{1+\delta}$. Eventually, I was able to obtain an explicit value for δ (Nathanson [36]), and the sum-product problem has become another hot topic in number theory.

A more recent subject is work with my students Brooke Orosz and Manuel Silva, together with Kevin O'Bryant and Imre Ruzsa, on the comparative theory of binary linear forms evaluated at finite sets of integers [48]. There is much more to be done in this area.

Finally, I would like to mention three other very new topics of research. In work with Blair Sullivan on the Caccetta-Haggkvist conjecture in graph theory [44,50], a new definition of the height of a subspace in a finite projective space was introduced. This height function has been further studied by O'Bryant [51] and Batson [1].

In a different direction, I have studied multiplicative functional equations satisfied by formal power series that look like quantum integers (for example, [2,38,40]), and, with Alex Kontorovich, their additive analogs [17].

At the Institute for Advanced Study in 1974–1975, I noticed some articles of Jack Milnor and Joe Wolf about the growth of finitely generated groups, and thought that this work that should be investigated as a kind of "nonabelian additive number

theory." Thirty six years later, I have finally started to think about this subject, now called "geometric group theory" and "metric geometry," and have obtained some new results [45–47].

Acknowledgements I want to thank David and Gregory Chudnovsky for organizing and editing this volume. Back in 1982, the Chudnovskys and I, together with Harvey Cohn, created the New York Number Theory Seminar at the CUNY Graduate Center, and we have been running this weekly seminar together for more than a quarter century. It has been a pleasure to know them and work with them.

Most of all, I acknowledge the love and support of my wife Marjorie and children Becky and Alex.

References

1. J. Batson, *Nathanson heights in finite vector spaces*, J. Number Theory **128** (2008), no. 9, 2616–2633.
2. A. Borisov, M. B. Nathanson, and Y. Wang, *Quantum integers and cyclotomy*, J. Number Theory **109** (2004), no. 1, 120–135.
3. J. W. S. Cassels, *On the equation* $a^x - b^y = 1$, Am. J. Math. **75** (1953), 159–162.
4. J. W. S. Cassels, *On the equation* $a^x - b^y = 1$. *II*, Proc. Cambridge Philos. Soc. **56** (1960), 97–103.
5. J. Cilleruelo and M. B. Nathanson, *Dense sets of integers with prescribed representation functions*, preprint, 2007.
6. P. Erdős and M. B. Nathanson, *Maximal asymptotic nonbases*, Proc. Am. Math. Soc. **48** (1975), 57–60.
7. P. Erdős and M. B. Nathanson, *Partitions of the natural numbers into infinitely oscillating bases and nonbases*, Comment. Math. Helv. **51** (1976), no. 2, 171–182.
8. P. Erdős and M. B. Nathanson, *Systems of distinct representatives and minimal bases in additive number theory*, Number theory, Carbondale 1979 (Proceedings of the Southern Illinois Conference, Southern Illinois University, Carbondale, Ill., 1979), Lecture Notes in Mathematics, vol. 751, Springer, Berlin, 1979, pp. 89–107.
9. P. Erdős and E. Szemerédi, *On sums and products of integers*, Studies in Pure Mathematics, To the Memory of Paul Turán (P. Erdős, L. Alpár, G. Halász, and A. Sárközy, eds.), Birkhäuser Verlag, Basel, 1983, pp. 213–218.
10. Harvard University Nuclear Nonproliferation Study Group, *Nuclear nonproliferation: the spent fuel problem*, Pergamon policy studies on energy and environment, Pergamon Press, New York, 1979.
11. H. Halberstam and K. F. Roth, *Sequences*, Vol. 1, Oxford University Press, Oxford, 1966, Reprinted by Springer-Verlag, Heidelberg, in 1983.
12. S.-P. Han, C. Kirfel, and M. B. Nathanson, *Linear forms in finite sets of integers*, Ramanujan J. **2** (1998), no. 1–2, 271–281.
13. E. Härtter, *Ein Beitrag zur Theorie der Minimalbasen*, J. Reine Angew. Math. **196** (1956), 170–204.
14. X.-D. Jia and M. B. Nathanson, *A simple construction of minimal asymptotic bases*, Acta Arith. **52** (1989), no. 2, 95–101.
15. A. G. Khovanskiĭ, *The Newton polytope, the Hilbert polynomial and sums of finite sets*, Funktsional. Anal. i Prilozhen. **26** (1992), no. 4, 57–63, 96.
16. A. G. Khovanskiĭ, *Sums of finite sets, orbits of commutative semigroups and Hilbert functions*, Funktsional. Anal. i Prilozhen. **29** (1995), no. 2, 36–50, 95.
17. A. V. Kontorovich and M. B. Nathanson, *Quadratic addition rules for quantum integers*, J. Number Theory **117** (2006), no. 1, 1–13.

18. L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*, Wiley, New York, 1974, reprinted by Dover Publications in 2006.
19. E. Landau, *Handbuch der Lehre von der verteilung der primzahlen*, Chelsea Publishing Company, New York, 1909, reprinted by Chelsea in 1974.
20. J. Lee, *Geometric structure of sumsets*, preprint, 2007.
21. J. C. M. Nash and M. B. Nathanson, *Cofinite subsets of asymptotic bases for the positive integers*, J. Number Theory **20** (1985), no. 3, 363–372.
22. M. B. Nathanson, *Derivatives of binary sequences*, SIAM J. Appl. Math. **21** (1971), 407–412.
23. M. B. Nathanson, *Complementing sets of n-tuples of integers*, Proc. Am. Math. Soc. **34** (1972), 71–72.
24. M. B. Nathanson, *An exponential congruence of Mahler*, Am. Math. Monthly **79** (1972), 55–57.
25. M. B. Nathanson, *Integrals of binary sequences*, SIAM J. Appl. Math. **23** (1972), 84–86.
26. M. B. Nathanson, *On the greatest order of an element of the symmetric group*, Am. Math. Monthly **79** (1972), 500–501.
27. M. B. Nathanson, *Sums of finite sets of integers*, Am. Math. Monthly **79** (1972), 1010–1012.
28. M. B. Nathanson, *On the fundamental domain of a discrete group*, Proc. Am. Math. Soc. **41** (1973), 629–630.
29. M. B. Nathanson, *Catalan's equation in $K(t)$*, Am. Math. Monthly **81** (1974), 371–373.
30. M. B. Nathanson, *Minimal bases and maximal nonbases in additive number theory*, J. Number Theory **6** (1974), 324–333.
31. M. B. Nathanson, *Multiplication rules for polynomials*, Proc. Am. Math. Soc. **69** (1978), no. 2, 210–212. MR MR0466087 (57 #5970)
32. M. B. Nathanson, *Komar-Melamid: Two Soviet Dissident Artists*, Southern Illinois University Press, Carbondale, IL, 1979.
33. M. B. Nathanson, *Classification problems in K-categories*, Fund. Math. **105** (1979/80), no. 3, 187–197.
34. M. B. Nathanson, *Additive number theory: the classical bases*, Graduate Texts in Mathematics, vol. 164, Springer-Verlag, New York, 1996.
35. M. B. Nathanson, *Additive number theory: inverse problems and the geometry of sumsets*, Graduate Texts in Mathematics, vol. 165, Springer-Verlag, New York, 1996.
36. M. B. Nathanson, *On sums and products of integers*, Proc. Am. Math. Soc. **125** (1997), no. 1, 9–16.
37. M. B. Nathanson, *Growth of sumsets in abelian semigroups*, Semigroup Forum **61** (2000), no. 1, 149–153.
38. M. B. Nathanson, *A functional equation arising from multiplication of quantum integers*, J. Number Theory **103** (2003), no. 2, 214–233.
39. M. B. Nathanson, *Unique representation bases for the integers*, Acta Arith. **108** (2003), no. 1, 1–8.
40. M. B. Nathanson, *Formal power series arising from multiplication of quantum integers*, Unusual applications of number theory, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 64, American Mathematical Society, Providence, RI, 2004, pp. 145–167.
41. M. B. Nathanson, *The inverse problem for representation functions of additive bases*, Number theory (New York, 2003), Springer, New York, 2004, pp. 253–262.
42. M. B. Nathanson, *Representation functions of additive bases for abelian semigroups*, Int. J. Math. Math. Sci. (2004), no. 29–32, 1589–1597.
43. M. B. Nathanson, *Every function is the representation function of an additive basis for the integers*, Port. Math. (N.S.) **62** (2005), no. 1, 55–72.
44. M. B. Nathanson, *Heights on the finite projective line*, Intern. J. Number Theory **5** (2009), 55–65.
45. M. B. Nathanson, *Bi-Lipschitz equivalent metrics on groups, and a problem in additive number theory*, preprint, 2009.
46. M. B. Nathanson, *Phase transitions in infinitely generated groups, and a problem in additive number theory*, Integers, to appear.

47. M. B. Nathanson, *Nets in groups, minimum length g-adic representations, and minimal additive complements*, preprint, 2009.
48. M. B. Nathanson, K. O'Bryant, B. Orosz, I. Ruzsa, and M. Silva, *Binary linear forms over finite sets of integers*, Acta Arith. **129** (2007), 341–361.
49. M. B. Nathanson and I. Z. Ruzsa, *Polynomial growth of sumsets in abelian semigroups*, J. Théor. Nombres Bordeaux **14** (2002), no. 2, 553–560.
50. M. B. Nathanson and B. D. Sullivan, *Heights in finite projective space, and a problem on directed graphs*, Integers **8** (2008), A13, 9.
51. K. O'Bryant, *Gaps in the spectrum of Nathanson heights of projective points*, Integers **7** (2007), A38, 7 pp. (electronic).
52. S. L. Segal, *On Nathanson's functional equation*, Aequationes Math. **28** (1985), no. 1–2, 114–123.
53. S. L. Segal, *Mathematicians under the Nazis*, Princeton University Press, Princeton, NJ, 2003.
54. S. L. Segal, *Nine introductions in complex analysis*, revised ed., North-Holland Mathematics Studies, vol. 208, Elsevier Science B.V., Amsterdam, 2008.
55. A. Stöhr, *Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe. I, II*, J. Reine Angew. Math. **194** (1955), 40–65, 111–140.
56. S. S. Wagstaff, *Solution of Nathanson's exponential congruence*, Math. Comp. **33** (1979), 1097–1100.
57. A. Weil, *Elliptic Functions according to Eisenstein and Kronecker*, Springer-Verlag, Berlin, 1976.

Sum-Product Theorems and Applications

Jean Bourgain

(To M. Nathanson)

Summary This is a brief account of recent developments in the theory of exponential sums and on methods from Arithmetic Combinatorics.

Keywords Exponential sum · Sum-product

Mathematics Subject Classifications (2010). Primary: 11L07, Secondary: 11T23

Introduction

These Notes originate from some lectures given by the author in the Fall of 2007 at IAS during the program on Arithmetic Combinatorics. Their purpose was twofold. The first was to illustrate the interplay between Additive Number Theory and problems on exponential sums, by reviewing various recent contributions in this general area and how they relate to several classical problems. The second was to present a proof of the Gauss sum estimate

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| < C|H|^{1-\delta}$$

for subgroups $H < \mathbb{F}_p^*$, $|H| > p^\varepsilon$ ($\varepsilon > 0$ fixed and arbitrary), which is a typical sample of those developments. My intent here was to make the argument as elementary and self-contained as possible (which it is, up to the Plunnecke–Ruzsa theory of set addition).

Therefore, what follows is not written in a homogeneous style. The first three sections are indeed presented in great detail, while the remainder is rather a survey with

J. Bourgain
Institute for Advanced Study, Princeton, NJ 08540, USA
e-mail: bourgain@ias.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_2,

only statements of the results. Note that this presentation is mostly geared toward the author's own research and is certainly far from complete, either from mathematical or historical perspective (the interested reader may wish to consult books such as [K-S] or [T-V] for background material). The reference list only serves this exposé and a more complete bibliography may be found in [K-S] and [T-V].

0 Sum-Product Theorem in $\mathbb{F}_p$

Theorem 1 ([B-K-T] and [B-G-K]).
Given $\varepsilon > 0$, there is $\delta > 0$ such that if $A \subset \mathbb{F}_p$ and $1 < |A| < p^{1-\varepsilon}$, then

$$|A + A| + |A.A| > c|A|^{1+\delta}.$$

There is the following quantitative statement.

Theorem 2 ([Ga] and [Ka-S]).

$$|A + A| + |A.A| > c \min\left(|A|^{\frac{14}{13}}, p^{\frac{1}{12}}|A|^{\frac{11}{12}}\right).$$

Denote

$$E_+(A, B) = |\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 | x_1 + y_1 = x_2 + y_2\}|$$

(additive energy)

$$E_\times(A, B) = |\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 | x_1 y_1 = x_2 y_2\}|$$

(multiplicative energy).

The Sum-Product theorem follows then from:

Proposition 1.

$$E_\times(A, A)^4 \ll |A + A|^9 |A|^2 + \frac{1}{p}|A + A|^8 |A|^5$$

using the inequality

$$|A.A| \geq \frac{|A|^4}{E_\times(A)}.$$

1 Preliminaries from Additive Combinatorics

(Plünnecke–Ruzsa Theory).
We consider subsets of an additive group $G, +$.

Lemma 1 (triangle inequality).

$$|A-B| \le \frac{|A-C|\,|B-C|}{|C|}.$$

Theorem 1 ([P-R]). *Let $X, A_1, \dots, A_k \subset G$ satisfy*

$$|X + A_i| \le \alpha_i |X| \qquad (1 \le i \le k).$$

Then there is $X_1 \subset X$ with

$$|X_1 + A_1 + \cdots + A_k| \le \alpha_1\alpha_2\cdots\alpha_k |X_1|.$$

Corollary 1.

$$|A_1 + \cdots + A_k| \le \frac{|A_1 + X|\cdots|A_k + X|}{|X|^{k-1}}.$$

Corollary 2 ([Ka-S]). *There exists $X' \subset X, |X'| > \frac{1}{2}|X|$ with*

$$|X' + A_1 + \cdots + A_k| \lesssim \frac{|A_1 + X|\cdots|A_k + X|}{|X|^{k-1}}.$$

Proof. If $Y \subset X, |Y| \ge \frac{1}{2}|X|$, then

$$\frac{|A_i + Y|}{|Y|} \le 2\frac{|A_i + X|}{|X|} = 2\alpha_i. \qquad (*)$$

Use [P-R] iteratively.
Construct disjoint set $X_s \subset X$ s.t.

$$|X_s + A_1 + \cdots + A_k| \le 2^k \alpha_1 \cdots \alpha_k |X_s|. \qquad (**)$$

Assume $X_1, \dots, X_s$ obtained. Let $Y = X\backslash(X_1 \cup \dots \cup X_s)$. If $|Y| < \frac{1}{2}|X|$, set $X' = X_1 \cup \dots \cup X_s$. From $(**)$

$$|X' + A_1 + \cdots + A_k| \le \sum_{s' \le s} |X_{s'} + A_1 + \cdots + A_k| \le 2^k \alpha_1 \cdots \alpha_k |X'|.$$

If $|Y| \ge \frac{1}{2}|X|$, then $(*)$. Apply [P-R] to $Y \Rightarrow X_{s+1} \subset Y$ such that

$$|X_{s+1} + A_1 + \cdots + A_k| \le (2\alpha_1)\cdots(2\alpha_k)|X_{s+1}|.$$

□

Proof of Proposition.

$$E_\times(A) = \sum_{a,b\in A} |aA \cap bA|.$$

Hence, there is $b_0 \in A$ and $A_1 \subset A$, $1 \leq N \leq |A|$ with

$$|aA \cap b_0 A| \sim N \text{ if } a \in A_1$$

and

$$|A_1|N \gg \frac{E_\times(A)}{|A|}. \qquad (*)$$

Case 1.

$$\frac{A_1 - A_1}{A_1 - A_1} = \mathbb{F}_p.$$

Then, there is $\xi = \frac{a_1-a_2}{a_3-a_4} (a_i \in A_i)$ s.t.

$$\left| \left\{ (x_1, x_2, x_3, x_4) \in A_1^4 | \xi = \frac{x_1 = x_2}{x_3 - x_4} \right\} \right| \leq \frac{|A_1|^4}{p}.$$

Hence

$$|(a_1 - a_2)A_1 + (a_3 - a_4)A_1| = |\xi A_1 + A_1| \geq \frac{|A_1|^2 |\xi A_1|^2}{E_+(\xi A_1, A_1)} \geq p.$$

Estimate

$$\begin{aligned} |(a_1 - a_2)A_1 + (a_3 - a_4)A_1| &\leq |a_1 A_1 - a_2 A_1 + a_3 A_1 - a_4 A_1| \\ &\overset{[P-R]}{\leq} |A|^{-3} \prod_{i=1}^{4} |a_i A \pm b_0 A|. \end{aligned}$$

From triangle inequality

$$\begin{aligned} |a_i A \pm b_0 A| &\leq \frac{|a_i A + (a_i A \cap b_0 A)| \, |b_0 A + (a_i A \cap b_0 A)|}{|a_i A \cap b_0 A|} \\ &< \frac{|A + A|^2}{N} \text{ (since } a_i \in A_1). \end{aligned}$$

Hence,

$$p \lesssim |A|^{-3} \left(\frac{|A + A|^2}{N} \right)^4 \lesssim |A|^{-3} |A + A|^8 |A|^8 E_\times(A, A)^{-4}$$

since N satisfies $(*)$

$$E_\times(A)^4 \gg \frac{1}{p}|A+A|^8\,|A|^5.$$

Case 2.

$$\frac{A_1-A_1}{A_1-A_1} \neq \mathbb{F}_p.$$

Hence,

$$\frac{A_1-A_1}{A_1-A_1} \not\supset \frac{A_1-A_1}{A_1-A_1}+1$$

and there is $\xi = \frac{a_1-a_2}{a_3-a_4} + 1(a_i \in A_1)$ s.t.

$$\xi \notin \frac{A_1-A_1}{A_1-A_1}.$$

Therefore, for any subset $A' \subset A_1$

$$\begin{aligned}|A'|^2 = |A'+\xi A'| &= |(a_1-a_2)A' + (a_1-a_2+a_3-a_4)A'| \\ &\leq |(a_1-a_2)A' + (a_1-a_2)A_1 + (a_3-a_4)A_1|.\end{aligned}$$

Using the Corollary to [P-R], take A' s.t. $X' = (a_1-a_2)A'$ satisfies $|X'| = |A'| > \frac{1}{2}|A_1|$ and

$$|X' + (a_1-a_2)A_1 + (a_3-a_4)A_1| \lesssim \frac{|(a_1-a_2)A_1+X|\,|(a_3-a_4)A_1+X|}{|X|}$$

where $X = (a_1-a_2)A_1$.

Hence,

$$|A_1|^2 \sim |A'|^2 \lesssim \frac{|A_1+A_1|.|(a_3-a_4)A_1+(a_1-a_2)A_1|}{|A_1|}$$

and

$$|A_1|^3 \lesssim |A+A|\,|a_1A_1 - a_2A_1 + a_3A_1 - a_4A_1|.$$

As before, since $a_i \in A_1$

$$|a_1A - a_2A + a_3A - a_4A| \ll |A|^{-3}\frac{|A+A|^8}{N^4}.$$

Therefore,

$$|A|^{-3}|A+A|^9 \gtrsim N^4|A_1|^3 \geq \frac{(N.|A_1|)^4}{|A|} \underset{(*)}{\gg} \frac{E_x(A)^4}{|A|^5}$$

and

$$E_\times(A)^4 \gg |A+A|^9|A|^2.$$

□

2 Some Tools from Graph Theory: The Balog–Szemerédi–Gowers Theorem

Statement. *Let $G, +$ be an additive group. There is an absolute constant C such that the following holds. Let $A \subset G$ be a finite set and $K \in \mathbb{R}_+$ such that*

$$E_+(A, A) > \frac{1}{K}|A|^3.$$

Then there is a subset $A' \subset A$ such that

$$|A'| > K^{-C}|A|$$
$$|A' \pm A'| < K^C|A'|.$$

Remark. Underlying Balog–Szemerédi–Gowers is in fact a result from graph theory, which will be implicit in the argument.

Also, Balog–Szemerédi–Gowers is not restricted to an Abelian setting and there are variants for general groups, both in discrete and continuous settings, using similar proofs (see the book [T-V]).

Sketch of the Proof.

Main idea. We construct a large subset $A' \subset A$, such that whenever $x, x' \in A'$, then there are at least $K^{-C}|A|^7$ representations

$$x - x' = x_1 - x_2 + x_3 - x_4 + x_5 - x_6 + x_7 - x_8 \text{ with } x_i \in A.$$

Hence

$$|A' - A'| \leq \frac{|A|^8}{K^{-C}|A|^7}.$$

The construction.

Let $\omega(x) = |\{(x_1, x_2) \in A^2 | x = x_1 - x_2\}|$ for $x \in G$.
Hence,

$$\sum_{x \in G} \omega(x) = |A|^2$$
$$\sum \omega(x)^2 = E_+(A).$$

Define

$$D = \left\{ z \in G | \omega(z) > \frac{1}{2K}|A| \right\}$$

(the 'popular' differences).

Then

$$\frac{1}{K}|A|^3 < \sum_{z \in D} \omega(z)^2 + \left(\frac{1}{2K}|A|\right)|A|^2$$

and

$$\sum_{z\in D}\omega(z)^2 > \frac{1}{2K}|A|^3.$$

Define the following (directed) graph $R \subset A \times A$

$$(x, y) \in R \Leftrightarrow x - y \in D.$$

Hence,

$$|R| = \sum_{z\in D}\omega(z) > \frac{1}{2K}|A|^2.$$

Denote R_x, R_y the sections of R. Thus,

$$\frac{1}{2K}|A|^2 < \sum_{y\in A}|R_y| \le |A|^{\frac{1}{2}}\left(\sum_{y\in A}|R_y|^2\right)^{\frac{1}{2}}$$

and

$$\sum_{y\in A}|R_y|^2 > \frac{1}{4K^2}|A|^3. \tag{1}$$

Define

$$Y = \{(x, x') \in A \times A|\ |R_x \cap R_{x'}| < \theta|A|\}$$

where we take

$$\theta = 10^{-3}K^{-2}.$$

Then,

$$\sum_{y\in A}|(R_y \times R_y) \cap Y| = \sum_{(x,x')\in Y}|R_x \cap R_{x'}| < \theta|A|^3 \tag{2}$$

and from (1), (2)

$$\sum_{y\in A}|R_y|^2 > \frac{1}{8K^2}|A|^3 + \frac{1}{8K^2\theta}\sum_{y\in A}|(R_y \times R_y) \cap Y|.$$

Therefore, there is $y_0 \in A$ with

$$|R_{y_0}|^2 > \frac{1}{8K^2}|A|^2 + 10|(R_{y_0} \times R_{y_0}) \cap Y|$$
$$\Rightarrow |R_{y_0}| > \frac{1}{3K}|A|.$$

The set A' is defined by

$$A' = \left\{x \in R_{y_0}|\ |(\{x\} \times R_{y_0}) \cap Y| < \frac{1}{3}|R_{y_0}|\right\}.$$

Since

$$\frac{1}{3}|R_{y_0}\setminus A'|\,|R_{y_0}| \le |(R_{y_0}\times R_{y_0})\cap Y| < \frac{1}{10}|R_{y_0}|^2$$

we have

$$|A'| > \frac{1}{2}|R_{y_0}| > \frac{1}{6K}|A|.$$

Take any $x_1, x_2 \in A'$. Then,

$$|\{x \in R_{y_0} | (x_1, x) \notin Y \text{ and } (x_2, x) \notin Y\}| > \left(1 - \frac{2}{3}\right)|R_{y_0}|$$

and

$$|R_{x_1} \cap R_x| > \theta|A|, |R_{x_2} \cap R_x| > \theta|A|$$

for at least $\frac{1}{3}|R_{y_0}|$ elements $x \in R_{y_0}$.

Write

$$\begin{aligned} x_1 - x_2 &= (x_1 - x) - (x_2 - x) \\ &= (x_1 - y_1) - (x - y_1) - (x_2 - y_2) + (x - y_2) \\ &\quad \text{where } y_i \in R_{x_i} \cap R_x \quad (i = 1, 2). \end{aligned}$$

Since $x_1 - y_1, x - y_1, x_2 - y_2, x - y_2 \in D$, each difference has at least $\frac{1}{2K}|A|$ representations in $A - A$. Hence, there are at least

$$\frac{1}{3}|R_{y_0}|.(\theta.|A|)^2.\left(\frac{1}{2K}|A|\right)^4 \gtrsim K^{-9}|A|^7$$

representations

$$x_1 - x_2 = z_1 - z_2 + z_3 - z_4 + z_5 - z_6 + z_7 - z_8$$

with $z_i \in A$, as claimed.

This proves the Balog–Szemerédi–Gowers theorem.

3 Exponential Sum Estimate

We will establish the following estimate on Gauss sums.

Theorem 1. *Let H be a multiplicative subgroup of $\mathbb{F}_p^*$ and $|H| > p^\varepsilon$ for some $\varepsilon > 0$. Then,*

$$\max_{(a,p)=1}\left|\sum_{x\in H} e_p(ax)\right| < C|H|^{1-\delta} \text{ where } \delta = \delta(\varepsilon) > 0.$$

Denote

$$\hat{f}(k) = \sum_{x \in \mathbb{F}_p} e_p(kx) f(x) \qquad (k \in \mathbb{F}_p)$$

the Fourier transform of $f : \mathbb{F}_p \to \mathbb{C}$.

Lemma 2 (harmonic analysis). *Let $\mu : \mathbb{F}_p \to [0,1]$ be a probability measure ($\sum \mu(x) = 1$).*

Denote for $\delta > 0$

$$\Lambda_\delta = \{k \in \mathbb{F}_p | \, |\hat{\mu}(k)| > p^{-\delta}\}.$$

Then,

$$|\{(k_1, k_2) \in \Lambda_\delta | k_1 - k_2 \in \Lambda_{2\delta}\}| > p^{-2\delta} |\Lambda_\delta|^2.$$

Proof. Let $|\hat{\mu}(k)| = c_k \hat{\mu}(k)$ with $c_k \in \mathbb{C}, |c_k| = 1$. We have

$$|\Lambda_\delta| . p^{-\delta} < \sum_{k \in \Lambda_\delta} c_k \hat{\mu}(k) = \sum_{x \in \mathbb{F}_p} \left[\sum_{k \in \Lambda_\delta} c_k e_p(kx) \right] \mu(x)$$

and

$$|\Lambda_\delta|^2 p^{-2\delta} < \sum_{x \in \mathbb{F}_p} \left| \sum_{k \in \Lambda_\delta} c_k e_p(kx) \right|^2 \mu(x) \leq \sum_{k_1, k_2 \in \Lambda_\delta} |\hat{\mu}(k_1 - k_2)|.$$

□

Corollary 3.

$$E_+(\Lambda_\delta, \Lambda_\delta) > p^{-4\delta} \frac{|\Lambda_\delta|^4}{|\Lambda_{2\delta}|}.$$

Corollary 4. *There is the following dichotomy. Let $\kappa > \delta > 0$.*

Either

$$|\Lambda_{2\delta}| > p^\kappa |\Lambda_\delta|$$

or there is $\Lambda \subset \Lambda_\delta$ such that

$$|\Lambda| > p^{-C\kappa} |\Lambda_\delta|$$
$$|\Lambda + \Lambda| < p^{C\kappa} |\Lambda|.$$

Proof. Corollary 1+ Balog–Szemerédi–Gowers. □

Let $H < \mathbb{F}_p^*, |H| = p^\alpha$ for some $\alpha > 0$.

Definition. A probability measure μ on $\mathbb{F}_p$ is H-invariant provided

$$\hat{\mu}(k) = \hat{\mu}(hk) \text{ for all } k \in \mathbb{F}_p, h \in H.$$

Example.

$$\mu(x) = \begin{cases} \frac{1}{|H|} \text{ if } x \in H \\ 0 \text{ if } x \notin H. \end{cases} .$$

Main Proposition.

For all $\rho < 1$ and $\delta > 0$, there is $\delta' = \delta'(\alpha, \rho, \delta) > 0$ such that

$$\Lambda_{\delta'} \neq \{0\} \Rightarrow |\Lambda_\delta| > p^\rho.$$

Here, $\Lambda_\delta = \Lambda_\delta(\mu)$, where μ is an arbitrary H-invariant measure.

The argument gives $\delta'(\alpha, \rho, \delta) = \frac{\delta}{\exp(\frac{1}{\alpha(1-\rho)})^C}$.

The limitation of the method: $|H| = p^\alpha$ with $\alpha \sim \frac{1}{\log\log p}$ (see [B1]).

Proof of Theorem using Proposition.

Take $\rho = 1 - \frac{\alpha}{3}, \delta = \frac{\alpha}{4} \Rightarrow \delta'$, according to the Proposition.

Apply the Proposition with $\mu = \frac{1}{|H|} 1_H$.

Assume $|\hat{\mu}(a)| > p^{-\delta'}$ for some $a \in \mathbb{F}_p^* \Rightarrow \Lambda_{\delta'} \neq \{0\}$.

Hence, $|\Lambda_\delta| > p^\rho \Rightarrow$

$$p^{\rho-2\delta} < \sum_{k \in \mathbb{F}_p} |\hat{\mu}(k)|^2 = p \sum_{x \in \mathbb{F}_p} \mu(x)^2 = \frac{p}{|H|} = p^{1-\alpha}$$

(contradiction).

Proof of the Main Proposition.

By H-invariance: $\Lambda_\delta = H.\Lambda_\delta$.

Hence,

$$\Lambda_\delta \neq \{0\} \Rightarrow |\Lambda_\delta| \geq p^\alpha.$$

Thus, the statement certainly holds for $\rho = \alpha$.

Assume now we established the statement for some $\rho < 1$. Thus

$$(*)\ \forall \delta > 0, \exists \delta' > 0 \text{ such that } \Lambda_{\delta'} \neq \{0\} \Rightarrow |\Lambda_\delta| > p^\rho$$

for arbitrary H-invariant μ.

We will then derive the statement for $\rho_1 = \rho + c \min(\rho, 1 - \rho)$.

Take $\delta > 0$ (small enough) $\underset{(*)}{\Rightarrow} \delta' < \delta$ and $\frac{1}{2}\delta' \underset{(*)}{\Rightarrow} \delta''$.

Assume $\Lambda_{\delta''} \neq \{0\} \Rightarrow |\Lambda_{\frac{1}{2}\delta'}| > p^\rho$.

Apply dichotomy from Corollary 2. Fix $\delta < \kappa < \rho$ to be specified. There are 2 cases

$$|\Lambda_{\delta'}| > p^{\kappa}|\Lambda_{\frac{1}{2}\delta'}| \Rightarrow |\Lambda_{\delta}| > p^{\rho+\kappa} \text{ and we are done}$$

$$\nearrow$$

$$\searrow$$

$$\exists \Lambda \subset \Lambda_{\frac{1}{2}\delta'} \cap \mathbb{F}_p^* \text{ with } \begin{cases} |\Lambda| > p^{-C\kappa}|\Lambda_{\frac{1}{2}\delta'}| > p^{\rho - C\kappa} \\ |\Lambda + \Lambda| < p^{C\kappa}|\Lambda| \end{cases}$$

where C is the constant from Corollary 2.

Since $|\Lambda + \Lambda| < p^{C\kappa}|\Lambda|$, the sum-product proposition implies

$$E_{\times}(\Lambda)^4 \ll |\Lambda + \Lambda|^9|\Lambda|^2 + \frac{1}{p}|\Lambda + \Lambda|^8|\Lambda|^5$$

hence,

$$E_{\times}(\Lambda) \ll p^{3C\kappa}\left(|\Lambda|^{\frac{11}{4}} + p^{-\frac{1}{4}}|\Lambda|^{\frac{13}{4}}\right). \qquad (**)$$

Denote $\mu_-(x) = \mu(-x)$ and

$$(\mu * \mu_-)(x) = \sum_{y \in \mathbb{F}_p} \mu(x-y)\mu_-(y)$$

the (additive) convolution of μ and μ_-. Hence,

$$\widehat{\mu * \mu_-}(k) = |\hat{\mu}(k)|^2.$$

Define a new probability measure η on $\mathbb{F}_p$ by

$$\eta(x) = \frac{1}{|\Lambda|}\sum_{y \in \Lambda}(\mu * \mu_-)\left(\frac{x}{y}\right).$$

Hence,

$$\hat{\eta}(k) = \frac{1}{|\Lambda|}\sum_{y \in \Lambda}|\hat{\mu}(yk)|^2.$$

Since $\Lambda \subset \Lambda_{\frac{\delta'}{2}}$, it follows

$$\begin{aligned} \hat{\eta}(1) = \frac{1}{|\Lambda|}\sum_{y \in \Lambda}|\hat{\mu}(y)|^2 &> p^{-\delta'} \\ &\Rightarrow \Lambda_{\delta'}(\eta) \neq \{0\} \\ &\underset{(*)}{\Rightarrow} |\Lambda_{\delta}(\eta)| > p^{\rho}. \end{aligned}$$

Denote

$$\tilde{\Lambda} = \Lambda_\delta(\eta), \text{ hence } |\tilde{\Lambda}| > p^\rho.$$

Then,

$$\sum_{k\in\tilde{\Lambda}} \hat{\eta}(k) > p^{-\delta}|\tilde{\Lambda}|$$
$$\Rightarrow \sum_{y\in\Lambda, k\in\tilde{\Lambda}} |\hat{\mu}(yk)|^2 > p^{-\delta}|\Lambda|.|\tilde{\Lambda}|.$$

Denote

$$\omega(z) = |\{(y,k) \in \Lambda \times \tilde{\Lambda} | yk = z\}|.$$

Thus,

$$\sum \omega(z)|\hat{\mu}(z)|^2 > p^{-\delta}|\Lambda|.|\tilde{\Lambda}|$$
$$\Rightarrow \omega(\Lambda_\delta) > \frac{1}{2}p^{-\delta}|\Lambda|.|\tilde{\Lambda}|.$$

Also

$$\begin{aligned}\omega(\Lambda_\delta) &= |\{(y,k) \in \Lambda \times \tilde{\Lambda} | yk \in \Lambda_\delta\}| \\ &\le |\Lambda_\delta|^{\frac{1}{2}} E_\times(\Lambda, \tilde{\Lambda})^{\frac{1}{2}} \\ &\le |\Lambda_\delta|^{\frac{1}{2}} E_\times(\tilde{\Lambda})^{\frac{1}{4}} E_\times(\Lambda)^{\frac{1}{4}} \\ &\le |\Lambda_\delta|^{\frac{1}{2}} |\tilde{\Lambda}|^{\frac{3}{4}} E_\times(\Lambda)^{\frac{1}{4}}.\end{aligned}$$

We use here the inequalities

$$E(A,B) \le E(A,A)^{\frac{1}{2}}.E(B,B)^{\frac{1}{2}}$$

and

$$E(A,A) \le |A|^3.$$

Therefore,

$$p^{-\delta}|\Lambda|\,|\tilde{\Lambda}|^{\frac{1}{4}} \le |\Lambda_\delta|^{\frac{1}{2}} E_\times(\Lambda)^{\frac{1}{4}}$$

and bounding $E_\times(\Lambda)$ using $(**)$ gives

$$p^{-\delta}|\Lambda|\,|\tilde{\Lambda}|^{\frac{1}{4}} \le p^{C\kappa}|\Lambda_\delta|^{\frac{1}{2}}(|\Lambda|^{\frac{11}{16}} + p^{-\frac{1}{16}}|\Lambda|^{\frac{13}{16}}).$$

Recalling that

$$|\Lambda| > p^{\rho - C\kappa}$$
$$|\tilde{\Lambda}| > p^\rho$$

it follows that either

$$|\Lambda_\delta| > p^{\frac{9}{8}\rho - 8C\kappa} > p^{\frac{10}{9}\rho}$$

or

$$|\Lambda_\delta| > p^{\frac{1}{8}(1+7\rho)-8C\kappa} > p^{\frac{1+9\delta}{10}}$$

(letting $\kappa \sim \min(\rho, 1-\rho)$ be small enough).

This completes the proof. □

4 Additive Relations in Multiplicative Groups

Obtaining nontrivial bounds on Gauss sums is essentially equivalent with estimates on the number of additive relations.

For $A \subset \mathbb{F}_p$ and $k \in \mathbb{Z}_+$, define

$$E_{(k)}(A) = |\{(x_1, \dots, x_{2k}) \in A^{2k} | x_1 + \cdots + x_k = x_{k+1} + \cdots + x_{2k}\}|.$$

Thus,

$$E_{(2)}(A) = E_+(A, A).$$

Lemma 1. *Assume A satisfies an exponential sum bound*

$$\max_{a \in \mathbb{F}_p^*} \left| \sum_{x \in A} e_p(ax) \right| < p^{-\varepsilon}|A| \text{ for some } \varepsilon > 0.$$

Then

$$E_{(k)}(A) = \left(\frac{1}{p} + 0(p^{-2k\varepsilon}) \right) |A|^{2k}.$$

In particular,

$$E_{(k)}(A) < \frac{2}{p}|A|^{2k} \quad \text{for } k > \frac{1}{2\varepsilon}.$$

Proof. Use the circle method. Thus

$$E_{(k)}(A) = \frac{1}{p} \sum_{a=0}^{p-1} \left| \sum_{x \in A} e_p(ax) \right|^{2k}$$

and isolating the contribution of $a = 0$, we get

$$\left| E_{(k)}(A) - \frac{|A|^{2k}}{p} \right| < \max_{a \in \mathbb{F}_p^*} \left| \sum_{x \in A} e_p(ax) \right|^{2k} < p^{-2k\varepsilon}|A|^{2k}.$$

□

Conversely, we have the following:

Lemma 2. *Let $H < \mathbb{F}_p^*$ and assume*

$$E_{(2k)}(H) < p^{-\frac{1}{2}-\delta}|H|^{4k}$$

for some $k \in \mathbb{Z}_+$ and $\delta > 0$. Then,

$$\max_{a \in \mathbb{F}_p^*} \left| \sum_{x \in H} e_p(ax) \right| \leq p^{-\frac{\delta}{4k^2}} |H|.$$

Proof. Write

$$\left| \sum_{x \in H} e_p(ax) \right|^{2k} = \sum_{x \in \mathbb{F}_p} \omega(z) e_p(az)$$

with

$$\omega(z) = |\{(x_1, \ldots, x_{2k}) \in H^{2k} | x_1 + \cdots + x_k - x_{k+1} - \cdots - x_{2k} = z\}|.$$

Since H is a multiplicative group

$$\omega(z) = \omega(xz) \text{ if } x \in H, z \in \mathbb{F}_p.$$

Also

$$\sum_z \omega(z)^2 = \|\omega\|_2^2 = E_{2k}(H).$$

□

Next

$$\begin{aligned}
\left| \sum_{x \in H} e_p(ax) \right|^{4k^2} &= \left| \frac{1}{|H|} \left[\sum_{\substack{z \in \mathbb{F}_p \\ x \in H}} \omega(z) e_p(axz) \right] \right|^{2k} \\
&\leq \frac{1}{|H|^{2k}} \left(\sum_{z \in \mathbb{F}_p} \omega(z) \left| \sum_{x \in H} e_p(axz) \right| \right)^{2k} \\
&\overset{\text{(Hölder)}}{\leq} \frac{1}{|H|^{2k}} \left(\sum \omega(z) \right)^{2k-1} \left[\sum_z \omega(z) \left| \sum_{x \in H} e_p(axz) \right|^{2k} \right] \\
&= |H|^{4k(k-1)} \left| \sum_{z,z'} \omega(z) \omega(z') e_p(azz') \right| \\
&\overset{\text{(Hadamard)}}{\leq} |H|^{4k(k-1)} \|\omega\|_2^2 \sqrt{p} < |H|^{4k^2 - \delta}
\end{aligned}$$

proving Lemma 2.

Recall the classical bound

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| \le \sqrt{p}|H|,$$

which is nontrivial for $|H| > \sqrt{p}$.

Prior to [B-G-K], completely explicit Gauss-sum estimates (with power-saving) for smaller groups (up to $|H| > p^{\frac{1}{4}+\varepsilon}$) had been obtained using variants of Stepanov's method (Garcia–Voloch, Shparlinski, Heath–Brown, Heath-Brown-Konyagin, Konyagin). In particular, one has

Proposition 1. *Let $H < \mathbb{F}_p^*$ and $|H| < p^{2/3}$. Then,*

$$E_{(2)}(H) \ll |H|^{\frac{5}{2}}$$

(see [K-S]). The proof is a variant of Stepanov's method.

Corollary 1. *If $H < \mathbb{F}_p^*$ and $|H| > p^{\frac{1}{3}+\varepsilon}$. Then*

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| \ll p^{-\frac{3}{8}\varepsilon}|H|.$$

Proof. By Proposition 1

$$E_{(2)}(H) \ll |H|^4 p^{-\frac{3}{2}(\frac{1}{3}+\varepsilon)}$$

and Lemma 2 applies with $k = 1, \delta = \frac{3}{2}\varepsilon$. □

Proposition 2 ([Kon]). *Let $H < \mathbb{F}_p^*$, $|H| < p^{\frac{1}{2}}$. Then for $k \in \mathbb{Z}_+$*

$$E_{(k)}(H) \ll |H|^{2k-2+2^{1-k}}.$$

This allows us to replace the $\frac{1}{3}$-exponent by $\frac{1}{4}$.

Corollary 2 ([Kon]). *Let $H < \mathbb{F}_p^*$ and $p^{\frac{1}{4}+\varepsilon} < |H| < p^{\frac{1}{2}}$. Then*

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| < p^{-\varepsilon_1}|H|$$

with

$$\varepsilon_1 = \max_{k\in\mathbb{Z}_+} \frac{1}{4k^2}\left(2\varepsilon - \left(\frac{1}{2}+2\varepsilon\right)4^{-k}\right).$$

The [B-G-K] exponents are still explicit but rather poor.

Combining Proposition 2 with the sum-product techniques, one gets for instance.

Proposition 3 ([B-G]). *Let $H < \mathbb{F}_p^*$ and $|H| > p^{\frac{1}{4}}$. Then,*

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| < |H|^{0,999984073+o(1)}.$$

For even smaller groups, one has:

Proposition 4. *Let $H < \mathbb{F}_p^*$ and $|H| > p^{\alpha}$. Then*

$$\max_{a\in\mathbb{F}_p^*}\left|\sum_{x\in H} e_p(ax)\right| \ll p^{-e^{-5k}}|H|$$

if $k \geq 4$ is a power of 2 satisfying

$$\alpha k > \left(\frac{k}{2}\right)^{0,968}.$$

Present technology requires

$$\frac{\log|H|}{\log p} > \frac{C}{(\log\log p)}$$

for some constant C. See [B1].

On the other hand, there is the following conjecture due to Montgomery, Vaughan, and Wooley (partly based on numerics).

Conjecture ([M-V-W]).

$$\max_{(a,p)=1}\left|\sum_{x\in H} e_p(ax)\right| < \min\left(p^{\frac{1}{2}}, C(\log p)^{\frac{1}{2}}|H|^{\frac{1}{2}}\right).$$

According to this conjecture, $\frac{|H|}{\log p} \to \infty$ would imply equidistribution of H (mod p).

Problem. (related to Furstenberg's conjecture $\times 2, \times 3$).

Let G_p be the group generated by 2 and 3 in $\mathbb{F}_p^*$, thus,

$$G_p = \langle 2, 3\rangle < \mathbb{F}_p^*.$$

Does G_p become equidistributed for $p \to \infty$?

Remark.

$$\frac{\mathrm{ord}_p(2) + \mathrm{ord}_p(3)}{\log p} \to \infty \text{ for } p \to \infty$$

(Corvaja–Zannier, based on the subspace theorem).

See also [B-L-M-V] for recent developments.

5 Multilinear Exponential Sums

The [B-G-K] argument provides in fact more general results for products of arbitrary sets $A_j \subset \mathbb{F}_p^*$.

Theorem 1 ([B-G-K], [B-G]). *Let $k \geq 4$ be a power of 2 and $A_1, \ldots, A_k \subset \mathbb{F}_p^*$ satisfy*

$$|A_1| \ldots |A_k| > p^{(\frac{k}{2})^{0,968}}.$$

Then

$$\left| \sum_{x_1 \in A_1, \ldots, x_k \in A_k} e_p(x_1 \ldots x_k) \right| \ll p^{-e^{-5k}} |A_1| \cdots |A_k|.$$

Recall the classical (Hadamard) inequality for $k = 2$

$$\left| \sum_{x \in A, y \in B} e_p(xy) \right| \leq |A|^{\frac{1}{2}} |B|^{\frac{1}{2}} p^{\frac{1}{2}}.$$

This inequality is nontrivial provided $|A|.|B| > p$.

A multilinear analogue with sharp entropy requirement on the sources is given by the following:

Theorem 2 ([B1]). *Assume $0 < \delta < \delta_0 < \frac{1}{4}$ and $k \geq 3$.*

There is

$$\delta' > \left(\frac{\delta}{k} e^{-1/\delta_0} \right)^{Ck}$$

such that if $A_1, \ldots, A_k \subset \mathbb{F}_p$ satisfy

(i) $|A_i| > p^{\delta_0}$ *for* $i = 1, \ldots, k$

(ii) $|A_1| \cdots |A_k| > p^{1+\delta}$.

Then

$$\left| \sum_{x_1 \in A_1, \ldots, x_k \in A_k} e_p(x_1 \ldots x_k) \right| < p^{-\delta'} |A_1| \cdots |A_k|.$$

6 Extensions to 'Almost Groups'

New estimates on exponential sums involving exponential functions may be obtained as well.

Theorem 1. *For all $\delta > 0$, there is $\delta' > 0$ such that if $\theta \in \mathbb{Z}_+$ satisfies*

$(\theta, p) = 1$ *and* $O_p(\theta) \geq t > p^\delta$
$(O_p(\theta) =$ *multiplicative order of* θ *mod* $p)$

then

$$\max_{(a,p)=1} \left| \sum_{s=1}^{t} e_p(a\theta^s) \right| < tp^{-\delta'}$$

Thus the sum may be incomplete. This result has many applications (see [K-S]), in particular.

Number fields.
Minimum norm representatives in residue classes and the Euclidean division algorithm in algebraic number fields (Egami's problem).

Coding theory.
The Odlyzko–Stanley enumeration problem.

Hyperelliptic curves.
Supersingularity of mod p reduction (Kodama's problem).
See [K-S] for details.

7 Sum-Product Theorem and Gauss Sums in Arbitrary Finite Fields

The results for prime fields generalize as follows.

Theorem 1 ([B-K-T], explicit exponents in [K-S]). *Assume $S \subset \mathbb{F}_q, |S| > q^\delta$ ($\delta > 0$ arbitrary) and*

$$|S + S| + |S.S| < K|S|.$$

Then there is a subfield G of $\mathbb{F}_q$ and $\xi \in \mathbb{F}_q^*$ such that

$$|G| < K^C|S|$$

and

$$|S \backslash \xi G| < K^C$$

where $C = C(\delta)$.

Exponential sum bounds in $\mathbb{F}_q$.

$$q = p^m \qquad Tr(x) = x + x^p + \cdots + p^{m-1}$$
$$\psi(x) = e_p(Tr(x)) \qquad \text{additive character.}$$

Theorem 2 ([B-C]). *Let $g \in \mathbb{F}_q^*$ of order t and*

$$t \geq t_1 > q^{\varepsilon}$$
$$\max_{\substack{1 \leq \nu < m \\ \nu | m}} gcd\,(p^\nu - 1, t) < q^{-\varepsilon} t$$

($\varepsilon > 0$ arbitrary).
Then

$$\max_{a \in \mathbb{F}_q^*} \left| \sum_{j \leq t_1} \psi(a g^j) \right| < C q^{-\delta} t_1$$

where $\delta = \delta(\varepsilon) > 0$.

8 The Case of General Polynomial (mod p)

We first recall Weil's estimate.

Theorem 1 (Weil). *Let $f(x) \in \mathbb{F}_p[X]$ of degree d. Then*

$$\left| \sum_{1 \leq x \leq p} e_p(f(x)) \right| \leq d\sqrt{p}.$$

Problem. Obtain non-trivial estimates for $d \geq \sqrt{p}$.

Sum-product technology enables one to obtain such results for special (sparse) polynomials (as considered by Mordell, cf. [Mor]).

Theorem 2 ([B2]). *Let*

$$f(x) = \sum_{i=1}^{r} a_i x^{k_i} \in \mathbb{Z}[X] \text{ and } (a_i, p) = 1$$

such that

$$(k_i, p-1) < p^{1-\delta} \qquad (1 \leq i \leq r)$$
$$(k_i - k_j, p-1) < p^{1-\delta} \qquad (1 \leq i \neq j < r).$$

Then

$$\left| \sum_{x=1}^{p} e_p\left(f(x)\right) \right| < Cp^{1-\delta'}$$

where $\delta' = \delta'(r, \delta) > 0 \quad (\delta > 0$ *arbitrary).*

The following example shows that the second condition is necessary.

Example (Cochrane–Pinner). Let

$$f(x) = x^{\frac{p-1}{2}+1} - x.$$

Then

$$\begin{aligned} \sum e_p(f(x)) &= \frac{p-1}{2} + \sum_{(\frac{x}{p})=-1} e_p(-2x) \\ &= \frac{p-1}{2} + 0(\sqrt{p}). \end{aligned}$$

Theorem 3. *Let* $\theta_1, \ldots, \theta_r \in \mathbb{F}_p^*$ *satisfy* ($\delta > 0$ *arbitrary)*

$$\begin{aligned} 0(\theta_i) &> p^{\delta} \qquad (1 \le i \le r) \\ 0(\theta_i \theta_j^{-1}) &> p^{\delta} \qquad (1 \le i \ne j \le r). \end{aligned}$$

Then, for $t > p^{\delta}$

$$\max_{a_i \in \mathbb{F}_p^*} \left| \sum_{s=1}^{t} e_p \left(\sum_{i=1}^{r} a_i \theta_i^s \right) \right| < Cp^{-\delta'} t$$

where $\delta' = \delta'(r, \delta)$.

Applications to cryptography and distributional properties of Diffie–Hellman triples $\{\theta^x, \theta^y, \theta^{xy}\}$.

Power generators $u_{n+1} = u_n^e$.

Blum-Blum-Shub generator ($e = 2$).

Theorems 2 and 3 rely on an extension of the sum-product theorem to Cartesian products.

Theorem 4. *Fix* $\varepsilon > 0$. *There is* $\delta'(\delta) \overset{\delta \to 0}{\longrightarrow} 0$ *such that if*

$$A \subset \mathbb{F}_p \times \mathbb{F}_p \quad (p^{\epsilon} < |A| < p^{2-\varepsilon})$$

and

$$|A + A| + |A.A| < p^{\delta}|A|$$

then

$$p^{1-\delta'} < |A| < p^{1+\delta'}$$

and there is a line $L \subset \mathbb{F}_p \times \mathbb{F}_p$ *of the form*

$$L = \{a\} \times \mathbb{F}_p, L = \mathbb{F}_p \times \{a\},$$

$$L = \{(x, ax) | x \in \mathbb{F}_p\}$$

such that

$$|A \cap L| > p^{1-\delta'}.$$

9 The Sum-Product in $\mathbb{Z}_q = \mathbb{Z}/q\mathbb{Z}$

Because of the presence of subrings when q is composite, additional restrictions on $A \subset \mathbb{Z}_q$ are needed.

The following gives a uniform statement in the modulus q.

Theorem 1 ([B3]). *Given* $0 < \delta_1 < 1$, $0 < \delta_2 < 1$, *there are* $\varepsilon = \varepsilon(\delta_1, \delta_2) > 0$ *and* $\gamma = \gamma(\delta_1, \delta_2) > 0$ *such that the following holds.*

Let $A \subset \mathbb{Z}_q$ *(q arbitrary and large enough) satisfy:*

(i) $|A| < q^{1-\delta_1}$

(ii) $|\pi_{q_1}(A)| > q_1^{\delta_2}$ *for all* $q_1 | q$ *with* $q_1 > q^{\varepsilon}$ *where* $\pi_{q_1} : \mathbb{Z}_q \to \mathbb{Z}_{q_1}$ *is the quotient map.*

Then

$$|A + A| + |A.A| > q^{\gamma}|A|.$$

Remark. Let $q = p_1^{m_1} p_2^{m_2} \ldots$ be the prime factorization. Then

$$\mathbb{Z}_q \simeq \mathbb{Z}_{p_1^{m_1}} \times \mathbb{Z}_{p_2^{m_2}} \times \ldots$$

We first establish the theorem for $q = p^m$ a prime power (with uniformity in p and m) and then recombine the factors.

Gauss sums (mod q).

Theorem 2 ([B4]). *For all* $\varepsilon > 0$, *there is* $\delta = \delta(\varepsilon)$ *such that if* $H < \mathbb{Z}_q^*$ *(q arbitrary) satisfies*

$$|H| > q^{\varepsilon}.$$

Then

$$\max_{\xi \in \mathbb{Z}_q^*} \left| \sum_{x \in H} e_q(\xi x) \right| < q^{-\delta}|H|.$$

Corresponding statement for incomplete sums is more restrictive.

Example. $q = p^2$.

Take $g = 1 + p$. Then

$$g^s \equiv 1 + sp \pmod{q}.$$

Hence

$$\left| \sum_{1 \le s < \frac{p}{5}} e_q(g^s) \right| \sim p.$$

Theorem 3 ([B3]). *Given $\varepsilon > 0$, there is $\delta = \delta(\varepsilon) > 0$ such that the following holds.*

Let $q \in \mathbb{Z}_+$ be large enough and $g \in \mathbb{Z}_q^*$ satisfy

$$ord_{q_1}(g) > q_1^{\varepsilon} \text{ whenever } q_1 | q, q_1 > q^{\delta}.$$

Then for $t > q^{\varepsilon}$

$$\max_{\xi \in \mathbb{Z}_q^*} \left| \sum_{1 \le s \le t} e_q(\xi g^s) \right| < t^{1-\delta}.$$

Special Case. $q = p^m$ (p fixed) and g fixed ($m \to \infty$).

Remark. $\operatorname{ord}_{p^m}(g) \sim p^m$.

There is the following more precise result due to Korobov.

Theorem 4 ([Kor]). *Given p and $g \in \mathbb{Z}_+, g \ge 2$, there is a constant $\rho = \rho(p, g)$ such that for $m \to \infty$, $t < q = p^m$*

$$\max_{(\xi, p)=1} \left| \sum_{s=1}^{t} e_{p^m}(\xi g^s) \right| \ll t . \exp\left(-\rho \frac{(\log t)^3}{(\log q)^2} \right).$$

This estimate is non-trivial for $\log t \gg (\log q)^{\frac{2}{3}}$.

Sketch of the Proof.

Fix d and take $m_1 \in \mathbb{Z}_+$ with $d m_1 < m \le (d+1) m_1$.

Take $s_1 < p^{m_1}$ such that $g^{s_1} \equiv 1 (\operatorname{mod} p^{m_1})$. Hence $g^{s_1} = 1 + b p^{m_1}$.

Apply the binomial formula

$$g^{k s_1} = (1 + b p^{m_1})^k \equiv \sum_{j \le d} \binom{k}{j} b^j p^{j m_1} \pmod{p^m}.$$

This is a polynomial in k of degree $d < \frac{m}{m_1}$.

Take m_1 with $p^{m_1} < t^{\frac{1}{4}}$ and $N \sim t^{\frac{1}{4}}$. Write

$$\sum_{1\le x,y\le N} e_{p^m}(\xi g^{s_1 xy}) = \sum_{1\le x,y\le N} e\left(F(x,y)\right)$$

where

$$F(x,y) = \sum_{j\le d} \binom{xy}{j} b^j \frac{1}{p^{m-jm_1}}.$$

Apply then Vinogradov exponential sum bound.
$\Rightarrow$ nontrivial estimate provided $d^2 = o(\log N)$, hence for

$$m^2 = o\left((\log t)^3\right).$$

There is the following general multilinear bound for composite modulus.

Theorem 5 ([B3]). *Given $\varepsilon > 0$, there are $\delta > 0$ and $k \in \mathbb{Z}_+$ such that if $A_1, \dots, A_k \subset \mathbb{Z}_q$ (q arbitrary) satisfy*

$$|A_i| > q^{\varepsilon} \qquad (1 \le i \le k)$$

and also

$$\max_{\xi\in\mathbb{Z}_{q_1}} \left|A_i \cap \pi_{q_1}^{-1}(\xi)\right| < q_1^{-\varepsilon}|A_i| \text{ whenever } q_1|q, q_1 > q^{\delta}.$$

Then

$$\left|\sum_{x_1\in A_1,\dots,x_k\in A_k} e_q(x_1\dots x_k)\right| < q^{-\delta}|A_1|\cdots|A_k|.$$

10 Exponential Sums in Finite Commutative Rings

Let R be a finite commutative ring with unit and assume

$$|R| = q \text{ with no small prime divisors.}$$

Denote $R^* =$ invertible elements.

Theorem 1 ([B5]). *Let $H < R^*, |H| > q^{\delta}$ (δ arbitrary).*

For all $\varepsilon > 0$, there is $\varepsilon' = \varepsilon'(\varepsilon) \to 0$ such that one of the following alternatives holds

(1) $\max_{\mathcal{X}\neq\mathcal{X}_0} \left|\sum_{x\in H} \mathcal{X}(x)\right| < |H|^{1-\varepsilon}$

($\mathcal{X} =$ *additive character of* R).

(2) There is nontrivial ideal I in R with

$$|H \cap (1+I)| > |H|^{1-\varepsilon'}.$$

(3) There is a nontrivial subring R_1 of R, such that $1 \in R_1$ and

$$|H \cap R_1| > |H|^{1-\varepsilon'}.$$

Application to Heilbronn sums (dv. [Od]).

Theorem 2 ([B4]). *For given $m \in \mathbb{Z}_+, m \geq 2$ and $r \geq 1$, there is $\delta = \delta(m,r) > 0$ such that*

$$\left| \sum_{x=1}^{p} e_{p^m} \left(\sum_{s=1}^{r} a_s x^{sp^{m-1}} \right) \right| < p^{1-\delta}$$

for p (prime) sufficiently large and $(a_1, \ldots, a_r) \in \mathbb{Z}_{p^m}^r \setminus \{0\}$.

Remark. Since $(x+py)^{p^{m-1}} \equiv x^{p^{m-1}} \pmod{p^m}$, the sum is complete.

Earlier results.

Heath–Brown,(1995), $\delta(2,1) = \frac{1}{12}$ ([H-B]).
Heath–Brown, Konyagin, (2000), $\delta(2,1) = \frac{1}{8}$ ([H-K]).
Malykhin-Bourgain-Chang, (2005), $\delta(m,1)$ for general m. (see [B-C3]).

11 Euclidean Algorithm in Algebraic Number Fields

Let K be a given algebraic number field and O its maximal order.
Let I be an integral ideal and $\alpha \in O/I$. Define

$$N_I(\alpha) = \min_{x \in \alpha} |N(x)|$$

(the minimal norm (taken in $\mathbb{Q}$) of all elements of α.)
Define further

$$L(K,I) = \max_{\alpha \in (O/I)^*} N_I(\alpha).$$

One has always the inequality

$$L(K,I) \leq N(I) = |O/I|.$$

K is an Euclidean field if

$$L(K,I) < N(I)$$

for all principal ideals I.

Only a few examples are known. For instance (cf. [K-S]), the only Euclidean quadratic fields $\mathbb{Q}(\sqrt{d})$ are obtained for

$$d \in \{-11, -7, -3, -2, -1, 2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73\}.$$

However, if K has an infinite group $U(K)$ of units or equivalently, if

$$r_1 + r_2 - 1 > 0$$

where $[K : \mathbb{Q}] = n = r_1 + 2r_2$, with r_1 (resp. r_2) the number of real (resp. complex) embeddings of K in $\mathbb{C}$, then K is 'almost Euclidean' in the sense that

$$L(K, I) = o(N(I))$$

for almost all ideals I (Egami's problem).
(results by Konyagin–Shparlinski, Bourgain–Chang, ...).
More precise statements obtained in [B-C2].

Prime ideal case.

Denote $\pi_K(T)$ the number of prime ideals of norm $\leq T$. Then

$$\pi_K(T) = (1 + o(1)) \frac{T}{\log T}.$$

Theorem 1 ([B-C2]). *For all $\varepsilon > 0$, there is $\delta = \delta(\varepsilon, K) > 0$ such that for $T \to \infty$*

$$|\{\mathcal{P} | \mathcal{P} \text{ prime ideal with } N(\mathcal{P}) \leq T, L(K, \mathcal{P}) > N(\mathcal{P})^{1-\delta}\}| < T^{\varepsilon}.$$

General integral ideals.

Denote $M(T)$ the number of ideals I in O of norm $N(I) \leq T$. Then

$$M(T) = (h(K)\kappa + o(1))\, T$$

where

$$\begin{aligned} h(K) &= \text{ class number} \\ \kappa = \kappa(K) &= 2^{r_1}(2\pi)^{r_2} R(K) |d\,(K)|^{-\frac{1}{2}}\, w(K)^{-1} \\ R(K) &= \text{ regulator} \\ d(K) &= \text{discriminant} \\ w(K) &= |E(K)|. \end{aligned}$$

Theorem 2 ([B-C2]). *There is $\delta' = \delta'(\delta) \to 0$ with $\delta \to 0$ such that*

$$L(K, I) < N(I)^{1-\delta}$$

for ideals I outside a sequence of asymptotic density at most δ'.

Main idea.

Consider the quotient map $\varphi : O \to O/I$.
If

$$I = \prod \mathcal{P}^{a(\mathcal{P})} \quad \text{(prime ideal factorization)}$$

then

$$O/I = \prod O/\mathcal{P}^{a(\mathcal{P})}.$$

Let $U = U(K)$ be the group of units and consider

$$G = \varphi(U) < (O/I)^*.$$

The main issue is to establish equidistribution results of G in O/I.

12 Application to QUE

We refer the reader to [K-R] for background material.

The Quantum Cat Map.

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$ and consider the toral automorphism $\mathbb{T}^2 \to \mathbb{T}^2$: $x \mapsto Ax$.

The classical evolution on $\mathcal{C}^\infty(\mathbb{T}^2)$ is defined by

$$f \to f \circ A.$$

We describe next the quantization of Hannay and Berry. Assume

$$ab \equiv cd \equiv 0 \pmod 2.$$

Let $N \in \mathbb{Z}_+$ and consider the Hilbert space

$$\mathcal{H}_N = L^2(\mathbb{Z}_N) \qquad \mathbb{Z}_N = \mathbb{Z}/N\mathbb{Z}$$

with inner product

$$\langle \phi, \psi \rangle = \frac{1}{N} \sum_{x \in \mathbb{Z}_N} \phi(x)\, \overline{\psi(x)}.$$

We associate to $f = \sum_{n \in \mathbb{Z}^2} \hat{f}(n) e^{2\pi i n x}$ in $C^\infty(\mathbb{T}^2)$ its 'quantization' $Op_N(f)$ acting on $\mathcal{H}_N$ and defined by

$$Op_N(f) = \sum_{n \in \mathbb{Z}^2} \hat{f}(n) T_N(n)$$

where

$$T_N(n)\phi(x) = e^{i\pi\frac{n_1 n_2}{N}} e^{2\pi i \frac{n_2 x}{N}} \phi(x + n_1).$$

One may then assign to A a unitary operator $U_N(A)$ called 'quantum propagator' satisfying the 'exact' Egorov theorem

$$U_N(A)^* Op_N(f) U_N(A) = Op_N(f \circ A).$$

We are concerned with the eigenfunctions ψ of $U_N(A)$.

In the context of cat maps, Schnirelman's general theorem takes the following form.

Let $f \in \mathcal{C}^\infty(\mathbb{T}^2)$ and let for each $N \in \mathbb{Z}_+$, $\{\psi_j\}_{1 \le j \le N}$ be an orthonormal basis of $\mathcal{H}_N$ of eigenfunctions of $U_N(A)$. Then there is a subset $J(N) \subset \{1, \dots, N\}$ such that

$$\frac{\#J(N)}{N} \to 1 \text{ for } N \to \infty$$

and

$$\langle Op_N(f)\psi_j, \psi_j \rangle \to \int_{\mathbb{T}^2} f$$

when $j \in J(N), N \to \infty$.

The result of Kurlberg and Rudnick [K-R] goes beyond this, showing that there is $\mathcal{N} \subset \mathbb{Z}_+$ of asymptotic density 1, such that

$$\max_j \left| \langle Op_N(f)\psi_j, \psi_j \rangle - \int_{\mathbb{T}^2} f \right| \xrightarrow{N \to \infty, N \in \mathcal{N}} 0.$$

More precisely they obtain the inequality

$$\sum_{j=1}^{N} \left| < O_{p_N}(f)\psi_j, \psi_j \rangle - \int_{\mathbb{T}^2} f \right|^4 \ll \frac{N(\log N)^{14}}{\mathrm{ord}(A, N)^2}$$

where $\mathrm{ord}(A, N)$ denotes the order of $A \pmod N$.

Next they show that

$$\mathrm{ord}(A, N) \gg N^{\frac{1}{2}} \exp\left((\log N)^\delta\right) \qquad (*)$$

for some $\delta > 0$ and N restricted to $\mathcal{N} \subset \mathbb{Z}_+$ of asymptotic density 1.

Problem ([K-R]). What if $\mathrm{ord}(A, N) < N^{\frac{1}{2}}$?

It turns out one may now deal with the case $\mathrm{ord}(A, N) > N^\varepsilon$ for any $\varepsilon > 0$. The following results are obtained in [B5].

Proposition 1. *[B6] For all $\varepsilon > 0$, there is $\delta > 0$ such that if N is prime and $\mathrm{ord}(A, N) > N^{\varepsilon}$, then*

$$\max_{\psi} |\langle T_N(n)\psi, \psi\rangle| < N^{-\delta}$$

with ψ a normalized eigenfunction of $U_N(A)$.

Theorem 2. *[B6] (N prime).*

For all $\varepsilon > 0$, there is $\delta > 0$ and a sequence $\mathcal{S}_\varepsilon$ of primes such that

$$\#\{N \in \mathcal{S}_\varepsilon | N < T\} \ll T^{\varepsilon}$$

and for given $f \in \mathcal{C}^\infty(\mathbb{T}^2)$

$$\max_{\psi} \left| \langle Op_N(f)\psi, \psi\rangle - \int_{\mathbb{T}^2} f \right| < N^{-\delta}$$

for N a sufficiently large prime outside $\mathcal{S}_\varepsilon$.

Theorem 3. *[B6] (N general).*

There is a density 1 sequence $\mathcal{N} \subset \mathbb{Z}_+$ and $\delta > 0$ such that for all observables $f \in \mathcal{C}^\infty(\mathbb{T}^2)$

$$\max_{\psi} \left| \langle Op_N(f)\psi, \psi\rangle - \int_{\mathbb{T}^2} f \right| < C_f N^{-\delta}$$

for $N \in \mathcal{N}$.

Sketch of the Proof of Proposition 1.

$N = p$ (prime).
$K =$ real quadratic field containing eigenvalues of A (units).
$O =$ maximal order of K.
$\mathcal{P} =$ prime of K lying above p.
Denote

$$K_p = O/\mathcal{P} \simeq \begin{cases} \mathbb{F}_p \text{ if } p \text{ splits} \\ \mathbb{F}_{p^2} \text{ if } p \text{ is inert.} \end{cases}$$

Diagonalize A over K_p. Thus

$$A' = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix} \qquad \varepsilon \in K_p^*.$$

Problem reduces then to an estimate on the number of solutions of the following system in K_p.

$$\begin{cases} \sum_{s=1}^{4\ell} (-1)^s \varepsilon^{j_s} = 0 \\ \sum_{s=1}^{4\ell} (-1)^s \varepsilon^{-j_s} = 0 \end{cases}$$

with $(j_1, \ldots, j_{4\ell}) \in \{1, \ldots, t\}^{4\ell}$ and $t = \mathrm{ord}_{K_p}(\varepsilon)$.

Use of exponential sum bounds with ℓ taken large enough. We distinguish 2 cases.

Split Case. $K_p = \mathbb{F}_p$.

Bounds on $\sum_{j=1}^{t} e_p(a_1\varepsilon^j + a_2\varepsilon^{-j})$. Apply theorems 3 from Sect.Ṡ8.

Inert Case. $K_p = \mathbb{F}_{p^2}$.

Bounds on $\sum_{j=1}^{t} e_p\left(Tr(a_1\varepsilon^j) + Tr(a_2\varepsilon^{-j})\right)$.

Apply Theorem 2 from Sect. 7 and Theorem 3 from Sect. 8. (2 further cases must be distinguished in view of the condition in Theorem 2, Sect. 7.)

References

[B1] J. Bourgain, *Multilinear exponential sums in prime fields under optimal entropy condition on the sources*, GAFA 18 (2009), no 5, 1477–1502.

[B2] J. Bourgain, *Mordell's exponential sum estimate revisited*, JAMS 18(2) (2005), 477–499.

[B3] J. Bourgain, The *Sum-product theorem in $\mathbb{Z}_q$, with q arbitrary*, J. Analyse 106 (2008), 1–93.

[B4] J. Bourgain, *Exponential sum estimates over subgroups of $\mathbb{Z}_q^*$, q arbitrary*, J. Analyse 97 (2005), 317–356.

[B5] J. Bourgain, *Exponential sum estimates in finite commutative rings and applications*, J. Analyse Math., 101 (2007), 325–355.

[B6] J. Bourgain, *A remark on quantum ergodicity for cat maps*, Springer, Berlin, lMN, 1910 (2007), 89–98.

[B-C1] J. Bourgain, M. Chang, *A Gauss sum estimate in arbitrary finite fields*, C.R. Math. Acad. Sci. Paris 342(9) (2006), 643–646.

[B-C2] J. Bourgain, M. Chang, *On the minimum norm of representatives of residue classes in number fields*, Duke Math. J. 138(2) (2007), 263–280.

[B-C3] J. Bourgain, M. Chang, *Exponential sum estimates over subgroups and almost subgroups of $\mathbb{Z}_Q^*$, where Q is composite with few prime factors*, GAFA 16(2) (2006), 327–366.

[B-G] J. Bourgain, M.Z. Garaev, *On a variant of sum-product estimates and explicit exponential sum bounds in prime fields*, Math. Proc. Camb. Phil. Soc. (2008).

[B-G-K] J. Bourgain, A. Glibichuk, S. Konyagin, *Estimate for the number of sums and products and for exponential sums in fields of prime order*, J. London Math. Soc. 73 (2006), 380–398.

[B-K-T] J. Bourgain, N. Katz, T. Tao, *A sum-product estimate in finite fields and applications*, GAFA 14 (2004), 27–57.

[B-L-M-V] J. Bourgain, E. Lindenstrauss, P. Michel, A. Venkatesh, *Some effective results for a, b*, ETDS, vol 29, 06 (2009), 1705–1722.

[Ga] M. Garaev, *An explicit sum-product estimate in $\mathbb{F}_p$*, IMRN (2007), no 11.

[H-B] D.R. Heath-Brown, *An estimate for Heilbronn's exponential sum*, Analytic Number Theory: The Halberstam Festschrift (B.C. Berndt, H.G. Diamond, A.J. Hildebrand, eds.), 451–463 (Progress in Mathematics 138/139, Birkhäuser-Verlag, Boston 1996).

[H-K] D.R. Heath-Brown, S.V. Konyagin, *New bounds for Gauss sum derived from k-th powers and for Heilbronn's exponential sum*, Q. J. Math. 51 (2003), 221–335.

[Ka-S] N. Katz, C.-Y. Shen, *A slight improvement to Garaev sum-product estimate*, Proc. AMS 136(7) (2008), 2499–2504.

[Kon] S.V. Konyagin, *Estimates for trigonometric sums over subgroups and for Gauss sums*, International Conference on Modern Problems of Number Theory and its Applications: Current Problems, Part III, 86–114 (Moscow State University, 2002) (in Russian).

[Kor] N. Korobov, *The distribution of digits in periodic fractions*, Math. Sb. (N.S) 89(131) (1972), 654–670.

[K-S] S.V. Konyagin, I.E. Shparlinski, *Character Sums with Exponential Functions and their Applications* (Cambridge Tracts in Mathematics 136, Cambridge University Press, Cambridge 1999).

[K-R] P. Kurlberg, Z. Rudnick, *On quantum ergodicity for linear maps of the torus*, Comm. Math. Phys. 222 (201), 201–227.

[M-V-W] H. Montgomery, R. Vaughan, T. Wooley, *Some remarks on Gauss sums associated with kth powers*, Math. Proc. Cambridge Philos. Soc. 118(1) (1995), 21–22.

[Mor] L.J. Mordell, *On a sum analogous to a Gauss sum*, Q.J. Math. 3 (1932), 161–162.

[Od] R.W.K. Odoni, *Trigonometric sums of Heilbronn's type*, Math. Proc. Cambridge Philos. Soc. 98 (1985), 389–396.

[T-V] T. Tao, V. Vu, *Additive Combinatorics*, Cambridge Studies in Advanced mathematics, vol. 105, Cambridge University Press, Cambridge (2006).

Can You Hear the Shape of a Beatty Sequence?

Ron Graham and Kevin O'Bryant

Summary Let $K(x_1, \ldots, x_d)$ be a polynomial. If you are not given the real numbers $\alpha_1, \alpha_2, \ldots, \alpha_d$, but are given the polynomial K and the sequence $a_n = K(\lfloor n\alpha_1 \rfloor, \lfloor n\alpha_2 \rfloor, \ldots, \lfloor n\alpha_d \rfloor)$, can you deduce the values of α_i? No, it turns out, in general. But with additional irrationality hypotheses and certain polynomials, it is possible. We also consider the problem of deducing α_i from the integer sequence $(\lfloor \lfloor \cdots \lfloor \lfloor n\alpha_1 \rfloor \alpha_2 \rfloor \cdots \alpha_{d-1} \rfloor \alpha_d \rfloor)_{n=1}^{\infty}$.

Keywords Beatty sequence · Generalized polynomial

Mathematics Subject Classifications (2010). 37A15, 11B75

1 Introduction

If you are given a sequence of integers $(a_n)_{n=1}^{\infty}$ and told that the sequence was generated by the formula $a_n = \lfloor n\alpha_1 \rfloor \lfloor n\alpha_2 \rfloor$ for some real numbers α_1, α_2, is it possible to determine α_1 and α_2? In other words, what are the solutions $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ to the infinite system of equations

$$\lfloor n\alpha_1 \rfloor \lfloor n\alpha_2 \rfloor = \lfloor n\beta_1 \rfloor \lfloor n\beta_2 \rfloor \qquad (n \in \mathbb{N})?$$

A *generalized polynomial* is defined to be any formula built up from the unknowns $x_1, x_2, \ldots$, the real numbers, and the operations of addition, multiplication, and the floor function. These have arisen recently in ergodic theory (e.g., [1, 3, 4]), particularly in connection with rotations on nilmanifolds.

R. Graham
University of California, San Diego, CA, USA
e-mail: graham@ucsd.edu

K. O'Bryant
The City University of New York, College of Staten Island and Graduate Center, New York, USA
e-mail: kevin@member.ams.org.

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_3,

The first problem we are concerned with is, given a sequence $(a_n)_{n=1}^{\infty}$ of integers and a generalized polynomial $G(\bar{x})$, to describe the set of $\bar{\alpha} \in \mathbb{R}^d$ such that

$$\forall n \geq 1, \qquad G(n\bar{\alpha}) = a_n.$$

A few examples will help to clarify the difficulty in dealing with generalized polynomials. First, we note that to determine real numbers from an integer sequence, we must use the tail of the sequence, i.e., limits must be involved in some form. As a first example, consider the sequence $a_n = n - 1$ and the generalized polynomial $G(\bar{x}) = \lfloor x_1 \rfloor + \lfloor x_2 \rfloor$. For *any* irrational α_1 and $\alpha_2 = 1 - \alpha_1$, we have $G(n\alpha_1, n\alpha_2) = a_n$ for all positive integers n. Another curious example is given by $G(x_1, x_2, n) = \lfloor \lfloor n\, x_1 \rfloor\, x_2 \rfloor$, which satisfies (among very many other sporadic relations)

$$\forall n \in \mathbb{Z}, \qquad G(3/7, 2/9, n) = G(1/3, 2/7, n).$$

I. Håland Knutson [personal communication] notes that

$$G(n) = \lfloor \lfloor \sqrt{2}n \rfloor 2\sqrt{2}n \rfloor - \lfloor \sqrt{2}n \rfloor^2 - 2n^2 + 1 = \begin{cases} 1, & n = 0; \\ 0, & n \in \mathbb{Z} \setminus \{0\}. \end{cases}$$

In this work, we restrict ourselves to generalized polynomials with a particular structure.

Specifically, let $K(\bar{x})$ be a (classical) polynomial, and set $a_n = K(\lfloor n\bar{\beta} \rfloor)$ (the floor function applied to each component of the vector $\bar{\beta}$) for some 'sufficiently' irrational $\bar{\beta}$. We attempt to find all nontrivial solutions to the system of equations

$$\forall n \geq 1, \qquad K(\lfloor n\alpha \rfloor) = a_n.$$

With varying success we treat linear polynomials $x_1 + \cdots + x_d$, sums of powers $x_1^r + \cdots + x_d^r$, and monomials $x_1 \cdots x_d$, and other shapes.

The second problem we address is, given d and a sequence $(a_n)_{n=1}^{\infty}$ of integers, to find all solutions to the infinite system of equations

$$\lfloor \lfloor \cdots \lfloor \lfloor n\alpha_1 \rfloor \alpha_2 \rfloor \cdots \alpha_{d-1} \rfloor \alpha_d \rfloor = a_n.$$

We were motivated by two problems[1] given in "Concrete Mathematics" [2]:

> **Comment to Bonus Problem 3.49:** Find a necessary and sufficient condition on the real numbers $0 \leq \alpha < 1$ and $0 \leq \beta < 1$ such that we can determine the unordered pair $\{\alpha, \beta\}$ from the infinite multiset of values $\{\lfloor n\alpha \rfloor + \lfloor n\beta \rfloor \mid n > 0\}$.

[1] It is plausible that their origins were in signal analysis. Consider a linear signal $(\alpha t + \gamma)_{t \in \mathbb{R}}$, that is measured at discrete times (replace $t \in \mathbb{R}$ with $n \in \mathbb{Z}_{\geq 0}$) and with finite precision (replace $\alpha n + \gamma$ with $\lfloor \alpha n + \gamma \rfloor$). Given finitely many such measurements, how accurately can you estimate α? It is not difficult to imagine a situation where several such signals are preprocessed algebraically into a single signal, and yet one still wishes to discern the original signals.

Research Problem 3.50: Find a necessary and sufficient condition on the nonnegative real numbers α and β such that we can determine α and β from the infinite multiset of values $\{\lfloor\lfloor n\alpha\rfloor\beta\rfloor\}$.

A partial solution to the first problem (with the additional assumption that $1, \alpha, \beta$ are linearly independent over $\mathbb{Q}$) has recently been published [6], and [2] itself credits a sufficient condition for the second problem to unpublished notes of William A. Veech. We provide partial answers to generalizations of both problems.

To state our theorems, it is convenient to first introduce some notation. For a vector of reals $\bar{\zeta} = \langle\zeta_1, \ldots, \zeta_d\rangle$, we define the fractional part $\{\bar{\zeta}\} = \langle\{\zeta_1\}, \ldots, \{\zeta_d\}\rangle$ (this paper contains no sets of vectors!) and floor $\lfloor\bar{\zeta}\rfloor = \langle\lfloor\zeta_1\rfloor, \ldots, \lfloor\zeta_d\rfloor\rangle$. Also, inequalities such as $\bar{\zeta} \geq 0$ are to be understood componentwise, i.e., $\zeta_1 \geq 0, \ldots, \zeta_d \geq 0$. We say that $\bar{\zeta}$ is rational if there is a nonzero vector of integers $\bar{c}$ such that the dot product $\bar{c} \cdot \bar{\zeta}$ is an integer, and otherwise say that $\bar{\zeta}$ is irrational. For a polynomial $K(x_1, \ldots, x_d)$, the expression $K(\bar{\zeta})$ is defined to be $K(\zeta_1, \ldots, \zeta_d)$. Also, $\sum\bar{\zeta} = \zeta_1 + \cdots + \zeta_d$.

Let $\bar{\zeta}, \bar{\eta} \in \mathbb{Z}^d$ both sum to 0, and let σ be a permutation of $1, 2 \ldots, d$. Let $\beta_i = \alpha_{\sigma(i)} + \zeta_i$ and $\delta_i = \gamma_{\sigma(i)} + \eta_i$. Then trivially

$$\lfloor n\alpha_1 + \gamma_1\rfloor + \cdots + \lfloor n\alpha_d + \gamma_d\rfloor = \lfloor n\beta_1 + \delta_1\rfloor + \cdots + \lfloor n\beta_d + \delta_d\rfloor$$

for all n. Our first theorem states that this is the only type of solution that is possible when $\bar{\alpha}$ is irrational. It is plausible and consistent with our experiments that the phrase "$\bar{\alpha}$ is irrational" could be weakened to "$\alpha_i + \alpha_j$ is not an integer for any i, j".

Theorem 1. *Let $K(x_1, \ldots, x_d) = x_1 + \cdots + x_d$, and $\bar{\alpha}, \bar{\gamma}, \bar{\beta}, \bar{\delta} \in \mathbb{R}^d$. If*

$$\forall n \geq 1, \qquad K(\lfloor n\bar{\alpha} + \bar{\gamma}\rfloor) = K(\lfloor n\bar{\beta} + \bar{\delta}\rfloor),$$

then either $\bar{\alpha}$ is rational, or there are lattice points $\bar{\zeta}, \bar{\eta} \in \mathbb{Z}^d$ and a permutation σ of $1, 2 \ldots, d$ with $\beta_i = \alpha_{\sigma(i)} + \zeta_i$, $\delta_i = \gamma_{\sigma(i)} + \eta_i$, and $\sum\bar{\zeta} = \sum\bar{\eta} = 0$.

Using the fact that for nonintegral α, the sequence $(|\lfloor n\alpha\rfloor|)_{n=1}^{\infty}$ contains arbitrarily large primes, we can also handle products. Note that in this case we do not need the irrationality of $\bar{\alpha}$.

Theorem 2. *Let $K(\bar{x}) = x_1x_2\cdots x_d$, and $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^d$. If*

$$\forall n \geq 1, \qquad K(\lfloor n\bar{\alpha}\rfloor) = K(\lfloor n\bar{\beta}\rfloor),$$

then either some α_i is an integer or $\{\alpha_1, \ldots, \alpha_d\} = \{\beta_1, \ldots, \beta_d\}$ (as multi-sets).

The next theorem assumes algebraic independence of the α_i, but this is used in only a very weak manner. The hypothesis could be weakened to assuming that $\bar{\alpha}$ is irrational and the α_i do not satisfy any of a specific (depending on K) small finite set of algebraic relations. In fact, we believe that the conclusion is true as long as

none of α_i are integers. Additionally, whether a particular form of S can be included in the following theorem depends on an *ad hoc* solution of a system of equations that arises. Certainly, the given list is not the extent of the method, but a general statement remains elusive.

Theorem 3. *Let $K(\bar{x}) = S(\bar{x}) + R(\bar{x})$ be a polynomial, where $S(\bar{x})$ is a symmetric polynomial of one the following types ($d \geq 2, r \geq 2$)*

$$\prod_{i=1}^{d} x_i, \quad \sum_{i=1}^{d} x_i^r, \quad \text{or} \quad \sum_{i,j=1}^{d} x_i x_j,$$

and $\deg(R) < \deg(S)$. *Assume that α_i ($1 \leq i \leq d$) are positive and do not satisfy any algebraic relations of degree less than* $\deg(S)$, *and β_i ($1 \leq i \leq d$) are positive and do not satisfy any algebraic relations of degree less than* $\deg(S)$. *If*

$$\forall n \geq 1, \qquad K(\lfloor n\bar{\alpha} \rfloor) = K(\lfloor n\bar{\beta} \rfloor),$$

then $\{\alpha_i : 1 \leq i \leq d\} = \{\beta_i : 1 \leq i \leq d\}$.

Rasmussen [6] proves the $d = 2$ and $d = 3$ cases of the following conjecture:

Conjecture 1. Suppose that $\bar{\alpha}, \bar{\beta} \in [0,1)^d$, and that both $\langle \alpha_1, \alpha_1\alpha_2, \ldots, \alpha_1\alpha_2\cdots\alpha_d \rangle$ and $\langle \beta_1, \beta_1\beta_2, \ldots, \beta_1\beta_2\cdots\beta_d \rangle$ are irrational. If

$$\lfloor \cdots \lfloor \lfloor n\alpha_1 \rfloor \alpha_2 \rfloor \cdots \alpha_d \rfloor = \lfloor \cdots \lfloor \lfloor n\beta_1 \rfloor \beta_2 \rfloor \cdots \beta_d \rfloor$$

for all $n \geq 1$, then $\bar{\alpha} = \bar{\beta}$.

We give his proofs (with corrections) in Sect. 2.4. It is certainly desirable to extend his work to $d > 3$, to weaken the irrationality condition, and to consider $\alpha_i \in \mathbb{R}$ instead of merely $\alpha_i \in [0,1)$. Using a different method, we make the following step in this direction.

Theorem 4. *Suppose that $\bar{\alpha}, \bar{\beta} \in [1,\infty) \times [2,\infty)^{d-1}$ are irrational. If*

$$\forall n \geq 1, \qquad \lfloor \cdots \lfloor \lfloor n\alpha_1 \rfloor \alpha_2 \rfloor \cdots \alpha_d \rfloor = \lfloor \cdots \lfloor \lfloor n\beta_1 \rfloor \beta_2 \rfloor \cdots \beta_d \rfloor,$$

then the sets of fractional parts are equal: $\{\{\alpha_1\}, \ldots, \{\alpha_d\}\} = \{\{\beta_1\}, \ldots, \{\beta_d\}\}$.

2 Proofs

2.1 Proof of Theorem 1

Proof. Without loss of generality, we assume that $\bar{\alpha}, \bar{\gamma}$ are in $[0,1)^d$ and that $\bar{\alpha}$ is irrational. Let $S(i) = K(\lfloor n\bar{\alpha} + \bar{\gamma} \rfloor)$, and set. Define $\Delta(i) = S(i+1) - S(i)$. Thus, $\Delta(i) \in \{0, 1, \ldots, k\}$. We say that S has an r-jump at i if $S(i+1) - S(i) = \Delta(i) = r$.

The frequency of r-jumps of S depends on the frequency that $(\{n\alpha_1 + \gamma_1\}, \ldots, \{n\alpha_k + \beta_k\})$ is in a particular subcube of $[0,1)^k$. To wit, if there are exactly r coordinates j such that

$$1 - \alpha_j \le \{i\alpha_j + \gamma_j\} < 1,$$

which is equivalent (ignoring the technical circumstance when $1 - \alpha_j - \gamma_j < 0$) to

$$1 - \alpha_j - \gamma_j \le \{i\alpha_j\} < 1 - \gamma_j,$$

then there is an r-jump at i. The volume of this region in $[0,1)^d$ is the asymptotic frequency of r-jumps of S, and is given by

$$V_r = \sum_{\substack{R \subseteq K \\ |R| = r}} \prod_{i \in R} (1 - \alpha_i) \prod_{j \in K \setminus R} \alpha_j \quad \text{where } K = \{1, 2, \ldots, k\}.$$

Consider the polynomial

$$P(z) = \prod_{i=1}^{k} \{(1 - \alpha_i)z + \alpha_i\} = \sum_{r=0}^{k} V_r z^r,$$

which is determined by S. Hence, all the roots $-\frac{\alpha_i}{1-\alpha_i}$ of P are determined by S, and therefore, so are all the values α_i.

Let $i_0, i_1, \ldots$ be the sequence of i such that $\Delta(i) = k$, which is exactly the same condition as 'for all j, $1 - \alpha_j - \gamma_j \le \{i\alpha_j\} < 1 - \gamma_j$'. By the irrationality of $\bar{\alpha}$, the closure of

$$\{(\{i_t\alpha_1\}, \ldots, \{i_t\alpha_k\}) : t = 0, 1, 2, \ldots\}$$

is the set

$$\prod_{j=1}^{k} [1 - \alpha_j - \gamma_j, 1 - \gamma_j].$$

Since we already know the α_j, we find that the γ_j are also determined. □

2.2 *Proof of Theorem 2*

Lemma 1. *If $\alpha \in \mathbb{R}$ is not an integer, then the sequence $(|\lfloor n\alpha \rfloor|)_{n=1}^{\infty}$ of nonnegative integers contains arbitrarily large prime numbers.*

Our proof works equally well to show that $(|\lfloor n\alpha + \gamma \rfloor|)_{n=1}^{\infty}$ contains large primes when α is irrational, but for rational α the conclusion would be false: the sequence $(\lfloor n\frac{15}{2} + 3 \rfloor)_{n=1}^{\infty}$ contains only one prime.

Proof. First, observe that the sequence contains *all* large positive integers, if $0 < |\alpha| \le 1$, so we assume henceforth that $|\alpha| > 1$.

First, we further assume that α is irrational and positive. We will show that $(\lfloor n\alpha + \gamma \rfloor)_{n=1}^{\infty}$ contains arbitrarily large primes. We note the oft-used and elementary criterion [5] that $k \in (\lfloor n\alpha + \gamma \rfloor)_{n=1}^{\infty}$ if and only if $k \ge \lfloor \alpha + \gamma \rfloor$ and either $\{(k-\gamma)/\alpha\} > 1 - 1/\alpha$ or $(k-\gamma)/\alpha \in \mathbb{Z}$. Thus it suffices for our purposes to show that the sequence of fractional parts $\{p/\alpha\}$ is uniformly distributed, where p goes through the prime numbers. This was shown by Vinogradov [7, Chapter XI].

If α is irrational and negative, then $|\lfloor n\alpha \rfloor| = \lfloor n|\alpha| + 1 \rfloor$, and this is the case considered in the previous paragraph.

For the remainder of the proof, we assume that $\alpha = q/p$, with $p \ge 2$ and $\gcd(p, q) = 1$. In particular,

$$\lfloor n\alpha \rfloor = \left\lfloor \frac{nq}{p} \right\rfloor.$$

It suffices for our purpose to restrict to $n \equiv r \pmod{p}$, that is, we replace n with $np + r$:

$$\left\lfloor \frac{(np+r)q}{p} \right\rfloor = nq + \left\lfloor \frac{rq}{p} \right\rfloor.$$

We have reduced the problem (by Dirichlet's theorem on the infinitude of primes in arithmetic progressions) to choosing r so that $\gcd(q, \lfloor rq/p \rfloor) = 1$. Set $r = q^{-1}$, where q^{-1} is the integer in $[2, p+1]$ with $qq^{-1} \equiv 1 \pmod{p}$; define u through $qq^{-1} = pu + 1$, and note that $\gcd(q, u) = 1$. We now have

$$\left\lfloor \frac{rq}{p} \right\rfloor = \left\lfloor \frac{q^{-1}q}{p} \right\rfloor = \left\lfloor u + \frac{1}{p} \right\rfloor = u,$$

with the last equality being our usage of $p \ge 2$, i.e., the reason we need α to be nonintegral. Since $\gcd(q, u) = 1$, we have $\gcd(q, \lfloor rq/p \rfloor) = \gcd(q, u) = 1$. □

Proof (Proof of Theorem 2). We proceed by induction on d. The claim is immediate for $d = 1$. Now assume that $d \ge 2$ and that Theorem 2 holds for $d - 1$.

Assume without loss of generality that $\alpha_1 \ge \alpha_2 \ge \cdots \alpha_d$. If $\lfloor n\alpha_1 \rfloor = q$ is prime, then it will show up in the factorization of $\prod_{i=1}^{d} \lfloor n\alpha_i \rfloor = P_n$ as a prime factor $q \ge P_n^{1/d}$ (since $\lfloor n\alpha_1 \rfloor \ge \lfloor n\alpha_i \rfloor$ for all i). Conversely, any prime factor q of P_n which is greater than or equal to $P_n^{1/d}$ must come from $\lfloor n\alpha_1 \rfloor$. Thus, we know the value of $\lfloor n\alpha_1 \rfloor$ for infinitely many values of n, and so we can determine α_1. Now, by factoring out $\lfloor n\alpha_1 \rfloor$ from each term $K(\lfloor n\bar{\alpha} \rfloor)$, we have reduced the problem to the case of $d - 1$ factors. This completes the induction step, and the theorem is proved. □

2.3 Proof of Theorem 3

A d-dimensional cube is defined as $Q_a(\bar{x}) := \{a + \sum_{j=1}^{d} \epsilon_j x_j : \epsilon_j \in \{0, 1\}\}$.

Lemma 2. *Let $d \in \mathbb{N}$, and $a, b \in \mathbb{R}, \bar{x}, \bar{y} \in \mathbb{R}^d$. If $Q = Q_a(\bar{x}) = Q_b(\bar{y})$ and $|Q| = 2^d$, then $\{|x_j| : 1 \le j \le d\} = \{|y_j| : 1 \le j \le d\}$.*

Proof. Since $|Q| = 2^d$, we know that none of x_j, y_j are 0, and that the x_j are distinct, as are the y_j. Further, note that,

$$Q = Q_a(x_1, \ldots, x_d) = Q_{\min Q}(|x_1|, \ldots, |x_d|),$$

so that we can assume without loss of generality that x_j, y_j are positive, and that $a = b = \min Q$.

The generating function of Q factors as

$$f(z) = \sum_{q \in Q} z^q = z^a \prod_{j=1}^{d} (1 + z^{x_j}) = z^a \prod_{j=1}^{d} (1 + z^{y_j}),$$

whence

$$\prod_{j=1}^{d} (1 + z^{x_j}) = \prod_{j=1}^{d} (1 + z^{y_j}) \tag{1}$$

for appropriate complex numbers z.

We will show by induction on d that such an equality implies that $\{x_j : 1 \le j \le d\} = \{y_j : 1 \le j \le d\}$. This is trivially true for $d = 1$. Now assume that it is true for $d - 1 \ge 1$.

Let $X = \max\{x_1, \ldots, x_d\}, Y = \max\{y_1, \ldots, y_d\}$. The left hand side of (1) vanishes at $z = \exp(\pi i / X)$, and so the right hand side must also vanish, i.e., $1 + \exp(\pi i y_j / X) = 0$ for some j. It follows that $y_j / X = 2k + 1$ for some integer k, and therefore, that for some j, $Y \ge y_j \ge X$. Interchanging the roles of x and y yields that some for some j, $X \ge x_j \ge Y$, and therefore $X = Y$. We can cancel out the terms on the left and right hand sides of (1) corresponding to X and Y (which are the same), and we get a product with $d - 1$ factors, completing the inductive step. □

Proof (Proof of Theorem 3). Define

$$\Delta(n) = \frac{K(\lfloor (n+1)\bar{\alpha} \rfloor) - K(\lfloor n\bar{\alpha} \rfloor)}{n^{D-1}}.$$

The set $\{\Delta(n): n \in \mathbb{N}\}$ has limit points (call the set of limit points Δ), which depend only on S and which we can describe in the following manner:

$$\Delta = \left\{ \sum_{i=1}^{d} [\alpha_i] \frac{\partial S}{\partial x_i}(\bar{\alpha}) : [\alpha_i] \in \{\lfloor \alpha_i \rfloor, \lceil \alpha_i \rceil\} \right\}.$$

We have assumed that $\bar{\alpha}$ is irrational to guarantee that all of these expressions arise as limit points, and we assumed that α_i are algebraically independent to guarantee that all of these expressions correspond to distinct real numbers. We can apply the previous lemma to learn

$$L_S := \left\{ \left| \frac{\partial S}{\partial x_i}(\bar{\alpha}) \right| \right\}.$$

From here, we apply *ad hoc* arguments that depend on the special structure of S.

If $S(\bar{x}) = \prod_{i=1}^{d} x_i$, then we have learned

$$L = \left\{ \alpha_j^{-1} \prod_{i=1}^{d} \alpha_i : 1 \le j \le d \right\}.$$

The product of all the elements of this set is just

$$\left(\prod_{i=1}^{d} \alpha_i \right)^{d-1}.$$

As $\bar{\alpha} > 0$, we can take the $(d-1)$-th root, learning the value of $\prod \alpha_i$. Dividing $\prod \alpha_i$ by each element of the set L yields the set

$$\{\alpha_j : 1 \le j \le d\}.$$

If $S(\bar{x}) = \sum_{i=1}^{d} x_i^r$, then we have learned

$$L = \{r\alpha_j^{r-1} : 1 \le j \le d\}$$

Dividing each element of L by r and then taking $(r-1)$-th roots (again using $\bar{\alpha} > 0$) yields the set

$$\{\alpha_j : 1 \le j \le d\}.$$

If $K(\bar{x}) = \sum_{i,j=1}^{d} x_i x_j$, then we have learned

$$L = \left\{ \alpha_i + \sum_{j=1}^{d} \alpha_j : 1 \le i \le d \right\}.$$

The sum of all the elements of this set is just

$$(d+1)\sum_{j=1}^{d}\alpha_j.$$

Dividing by $d+1$ yields $\sum\alpha_j$, and subtracting this from each element of L gives the set

$$\{\alpha_i : 1 \le i \le d\}.$$

□

2.4 Rasmussen's Approach to Conjecture 1

Our first proof of the $d = 2$ case is markedly different from the other proofs of this article. First, we do not assume $\langle\alpha_1, \alpha_2\rangle$ to be irrational, but $\langle\alpha_1, \alpha_1\alpha_2\rangle$. Second, the proof is by contradiction and therefore not constructive.

Suppose, by way of contradiction, that

$$s(n) = \lfloor\lfloor n\alpha_1\rfloor\alpha_2\rfloor = \lfloor\lfloor n\beta_1\rfloor\beta_2\rfloor,$$

with $\bar{\alpha} \neq \bar{\beta}$, and $\langle\alpha_1, \alpha_1\alpha_2\rangle, \langle\beta_1, \beta_1\beta_2\rangle$ are irrational. Note

$$\alpha_1\alpha_2 = \lim_{n\to\infty}\frac{s(n)}{n} = \beta_1\beta_2.$$

Suppose without loss of generality that $\beta_2 < \alpha_2$ and $\alpha_1 < \beta_1$. Since $\langle\alpha_1, \alpha_1\alpha_2\rangle$ is irrational, there exists an n such that $\{n\alpha_1\} > \frac{\alpha_2+\beta_2}{2\alpha_2}$ (note that $\frac{\alpha_2+\beta_2}{2\alpha_2} < 1$ by virtue of the assumption that $\beta_2 < \alpha_2$) and $\beta_2 < \{n\alpha_1\alpha_2\} < \frac{\alpha_2+\beta_2}{2}$. But then

$$s(n) = \lfloor\lfloor n\alpha_1\rfloor\alpha_2\rfloor = \lfloor n\alpha_1\alpha_2 - \{n\alpha_1\}\alpha_2\rfloor = \lfloor n\alpha_1\alpha_2\rfloor - 1$$

whereas, since $\{n\beta_1\beta_2\} = \{n\alpha_1\alpha_2\} > \beta_2 > \{n\beta_1\}\beta_2$,

$$s(n) = \lfloor\lfloor n\beta_1\rfloor\beta_2\rfloor = \lfloor n\beta_1\beta_2 - \{n\beta_1\}\beta_2\rfloor = \lfloor n\beta_1\beta_2\rfloor = \lfloor n\alpha_1\alpha_2\rfloor.$$

The method of Rasmussen, which works[2] for $d = 2$ and $d = 3$, might be more amenable to generalization. Define for $\bar{\alpha} \in \mathbb{R}^d$

$$T_{d,k} := \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\left(n\alpha_1\cdots\alpha_d - \lfloor\cdots\lfloor\lfloor n\alpha_1\rfloor\alpha_2\rfloor\cdots\alpha_d\rfloor\right)^k.$$

[2] In the $d = 3$ case, Rasmussen miswrote the formula for $T_{3,2}$, which erroneously led to a system of equations (using $T_{3,1}$ and $T_{3,2}$) with a unique solution. The analogous system using $T_{3,1}$ and $T_{3,3}$, however, does have a unique solution. We give this minor correction here.

Using Weyl's Criterion and straightforward integration (with which we trust *Mathematica 6.0*), we find that if $\bar{\alpha} \in [0,1)^d$ and $\langle \alpha_1, \alpha_1\alpha_2, \ldots, \alpha_1\alpha_2\cdots\alpha_d \rangle \in [0,1)^d$ is irrational, then

$$T_{2,1} = \frac{1+\alpha_2}{2},$$
$$T_{3,1} = \frac{1+\alpha_3+\alpha_2\alpha_3}{2},$$
$$T_{3,3} = \frac{1}{2}T_{3,1}\cdot\left((1+\alpha_3+\alpha_3^2)+(\alpha_3+\alpha_3^2)\alpha_2+(\alpha_3^2)\alpha_2^2\right).$$

Since both $P_d := \prod_{i=1}^d \alpha_i$ and the $T_{d,k}$ are determined by the sequence

$$(\lfloor\cdots\lfloor\lfloor n\alpha_1\rfloor\alpha_2\rfloor\cdots\alpha_d\rfloor)_{n=1}^{\infty},$$

so are the α_i: for $d = 2$

$$\alpha_2 = 2T_{2,1}-1, \qquad \alpha_1 = P_d/\alpha_2$$

and for $d = 3$

$$s = \operatorname{sgn}\left(4T_{3,1}^3 - 2T_{3,1}^2 + T_{3,1} - 2T_{3,3}\right),$$
$$\alpha_2 = \frac{-4T_{3,1}^3 - T_{3,1} + 4T_{3,3} + s\,(1-2T_{3,1})\sqrt{-12T_{3,1}^4 + 4T_{3,1}^3 - 3T_{3,1}^2 + 8T_{3,3}T_{3,1}}}{2\left(4T_{3,1}^3 - 2T_{3,1}^2 + T_{3,1} - 2T_{3,3}\right)},$$
$$\alpha_3 = \frac{2T_{3,1}-1}{1+\alpha_2},$$
$$\alpha_1 = \frac{P_3}{\alpha_2\alpha_3}.$$

We expect that this approach will work in principle for arbitrarily large d, but the practical difficulties in carrying this out are not trivial. Already, we are loathe to check the formula for $T_{3,3}$ and to solve the resulting equations by hand. *Mathematica*'s `Solve` command only gives generic solutions, while its `Reduce` command is too slow to handle $d = 4$.

The formulas given above for $T_{d,k}$ can be computed using Weyl's criterion: If $\bar{\alpha}$ is irrational, then

$$\frac{1}{N}\sum_{n=1}^{N} f(\{n\bar{\alpha}\}) = \int_{[0,1)^d} f(\bar{x})\,d\bar{x}.$$

We calculate $T_{3,1}$ as an example. By repeatedly using $\lfloor q \rfloor = q - \{q\}$ and $\{q + r\} = \{\{q\} + r\}$, we calculate

$$\begin{aligned}(n\alpha_1\alpha_2\alpha_3 &- \lfloor\lfloor\lfloor n\alpha_1\rfloor \alpha_2\rfloor\alpha_3\rfloor)\\ &= \{n\alpha_1\}\alpha_2\alpha_3 + \{\{n\alpha_1\alpha_2\} - \{n\alpha_1\}\alpha_2\}\alpha_3\\ &\quad + \{\{n\alpha_1\alpha_2\alpha_3\} - \{n\alpha_1\}\alpha_2\alpha_3 - \{\{n\alpha_1\alpha_2\} - \{n\alpha_1\}\alpha_2\}\alpha_3\}\\ &= x\alpha_2\alpha_3 + \{y - x\alpha_2\}\alpha_3 + \{z - x\alpha_2\alpha_3 - \{y - x\alpha_2\}\alpha_3\}\end{aligned}$$

where $\langle x, y, z\rangle = \langle\{n\alpha_1\}, \{n\alpha_1\alpha_2\}, \{n\alpha_1\alpha_2\alpha_3\}\rangle$. By Weyl's criterion, we get

$$\begin{aligned}T_{3,2} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} &(n\alpha_1\alpha_2\alpha_3 - \lfloor\lfloor\lfloor n\alpha_1\rfloor \alpha_2\rfloor\alpha_3\rfloor)\\ &\int_0^1\int_0^1\int_0^1 x\alpha_2\alpha_3 + \{y - x\alpha_2\}\alpha_3 + \{z - x\alpha_2\alpha_3 - \{y - x\alpha_2\}\alpha_3\}dx\,dy\,dz.\end{aligned}$$

Using $\bar{\alpha} \in [0, 1)^3$, we can eliminate the fractional parts in the above integral and get

$$T_{3,2} = \frac{1}{3} + \frac{1 + \alpha_2}{2}\alpha_3 + \frac{2 + 3\alpha_2 + 2\alpha_2^2}{6}\alpha_3^2.$$

It is clear that this method can yield a formula for $T_{d,k}$ for any d, k.

2.5 *Proof of Theorem 4*

Let $[x]_0$ be the floor of x, and $[x]_1$ be the ceiling. Let

$$T(W, \bar{\alpha}; n) := [\ldots[[n\alpha_1]_{w_1}]\alpha_2]_{w_2}\ldots\alpha_d]_{w_d},$$

where $W = w_1w_2\ldots w_k$ is a word in the alphabet $\{0, 1\}$, and $\bar{\alpha} = \langle\alpha_1, \alpha_2, \ldots, \alpha_d\rangle$. In addition to its usual meaning, let "$<$" denote the lexicographic ordering on $\{0, 1\}^d$. Let $h(W)$ be the Hamming weight of the word W, i.e., the number of `1`s in W.

Lemma 3. *If $\alpha_1, \ldots, \alpha_d$ are not integers, with $\alpha_1 > 1$ and $\alpha_i > 2$ (for $2 \le i \le d$), then*

$$W < V \Leftrightarrow T(W, \bar{\alpha}; 1) < T(V, \bar{\alpha}; 1).$$

Proof. We work by induction on d. For $d = 1$, the result obviously holds since $\alpha_1 \notin \mathbb{Z}$.

Now assume that $d \geq 2$ and that the result holds for *all* $(d-1)$-tuples . Assume that $W < V$. If $w_1 = v_1$, then we may apply the induction hypothesis by observing that

$$T(W, \langle \alpha_1, \ldots, \alpha_d \rangle; 1) = T(w_2 \cdots w_d, \langle [\alpha_1]_{w_1} \alpha_2, \alpha_3, \ldots, \alpha_d \rangle; 1)$$
$$T(V, \langle \alpha_1, \ldots, \alpha_d \rangle; 1) = T(v_2 \cdots v_d, \langle [\alpha_1]_{w_1} \alpha_2, \alpha_3, \ldots, \alpha_d \rangle; 1)$$

Thus, we may assume that $w_1 < v_1$, and so $w_1 \cdots w_{d-1} < v_1 \cdots v_{d-1}$. Since $\alpha_d > 2$, we have $[m\alpha_d]_{w_d} < [m'\alpha_d]_{v_d}$ whenever $0 < m < m'$, and by induction, we have

$$m = T(w_1 \cdots w_{d-1}, \langle \alpha_1, \ldots, \alpha_{d-1} \rangle; 1) < T(v_1 \cdots v_{d-1}, \langle \alpha_1, \ldots, \alpha_{d-1} \rangle; 1) = m'.$$

Now, we have $T(W, \bar{\alpha}; 1) = [m\alpha_d]_{w_d} < [m'\alpha_d]_{v_d} = T(V, \bar{\alpha}; 1)$. □

Lemma 4. *Suppose that* $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^d$ *are irrational, and suppose that for any pair* W, V *of words of length* d

$$T(W, \bar{\alpha}; 1) < T(V, \bar{\alpha}; 1) \Leftrightarrow T(W, \bar{\beta}; 1) < T(V, \bar{\beta}; 1),$$

and further suppose that if W *and* V *have different Hamming weight, then* $T(W, \bar{\alpha}; 1) \neq T(V, \bar{\alpha}; 1)$. *If*

$$\forall n \geq 1, \qquad T(0^d, \bar{\alpha}; n) = T(0^d, \bar{\beta}; n),$$

then $\{\{\alpha_i\} : 1 \leq i \leq d\} = \{\{\beta_i\} : 1 \leq i \leq d\}$.

Proof. Set $\Delta(n) = T(0^d, \bar{\alpha}; n+1) - T(0^d, \bar{\alpha}; n)$, and note that

$$\{\Delta(n) : n \in \mathbb{N}\} = \{T(W, \bar{\alpha}; 1) : \mathrm{len}(W) = d\}.$$

In fact, by the irrationality of $\bar{\alpha}$, the density of n such that $\Delta(n) = T(w_1 \cdots w_d, \bar{\alpha}; 1)$ is

$$V_W(\bar{\alpha}) = \prod_{\substack{i=1 \\ w_i = 0}}^{d} \{\alpha_i\} \prod_{\substack{i=1 \\ w_i = 1}}^{d} (1 - \{\alpha_i\}).$$

While for any particular W it is possible that $V_W(\bar{\alpha}) \neq V_W(\bar{\beta})$, the condition on the ordering of $T(W, \bar{\alpha}; 1), T(W, \bar{\beta}; 1)$ guarantees the set equalities for $1 \leq i \leq d$:

$$\left\{ V_W(\bar{\alpha}) : h(W) = i \right\} = \left\{ V_W(\bar{\beta}) : h(W) = i \right\}.$$

Thus, the polynomial

$$P(z) = \prod_{i=1}^{d} \Big(\{\alpha_i\} + (1 - \{\alpha_i\})z \Big) = \sum_{I \subseteq \{1,\ldots,d\}} \prod_{i \in I} \{\alpha_i\} \prod_{i \in \{1,\ldots,d\} \setminus I} (1 - \{\alpha_i\})z$$

$$= \sum_{\substack{W \\ \mathrm{len}(W)=d}} \left(V_W(\bar{\alpha}) \prod_{\substack{i=1 \\ w_i=0}}^{d} 1 \prod_{\substack{i=1 \\ w_i=1}}^{d} z \right)$$

$$= \sum_{\substack{W \\ \mathrm{len}(W)=d}} V_W(\bar{\alpha})\, z^{h(W)}$$

$$= \sum_{i=0}^{d} \left(\sum_{\substack{W \\ \mathrm{len}(W)=d,\, h(W)=i}} V_W(\bar{\alpha}) \right) z^d$$

is determined by the sequence. Therefore, the set of its roots $-\frac{\{\alpha_i\}}{1-\{\alpha_i\}}$ is also determined by the sequence. Since $x \mapsto -\frac{1-x}{x}$ is a 1-1 map, this implies that the set $\{\{\alpha_1\}, \ldots, \{\alpha_d\}\}$ is determined from the sequence, concluding the proof. □

Proof (Proof of Theorem 4). Combine Lemmas 3 and 4. □

3 Open Questions Concerning Generalized Polynomials

The meta-issue is to find an efficient algorithm that will determine whether a generalized polynomial with algebraic coefficients is identically zero on the positive integers. Humble first steps in this direction would be to completely answer the problems implied in Concrete Mathematics [2]:

Problem 1. Find a necessary and sufficient condition on the real numbers $\alpha_i, \beta_j \in [0, 1)$ such that for all positive integers n,

$$\sum_{i=1}^{d} \lfloor n\alpha_i \rfloor = \sum_{j=1}^{\ell} \lfloor n\beta_j \rfloor.$$

We suspect that this equality happens only if $d = \ell$ and for some a, b, c, d, $\alpha_a + \alpha_b = \beta_c + \beta_d = 1$, and that this, and trivial solutions, are the only way that equality can occur.

Problem 2. Find a necessary and sufficient condition on the real numbers $\alpha_i, \beta_j \in \mathbb{R}$ such that for all positive integers n,

$$\lfloor \cdots \lfloor \lfloor n\alpha_1 \rfloor \alpha_2 \rfloor \cdots \alpha_d \rfloor = \lfloor \cdots \lfloor \lfloor n\beta_1 \rfloor \beta_2 \rfloor \cdots \beta_\ell \rfloor.$$

There are very many solutions in rationals, and we do not have a guess as to their structure.

Both problems are obvious if all α, β are taken to be integers, and both are answered here if $d = \ell$ and the α, β are taken to be sufficiently irrational. The most difficult case to understand, for both questions, seems to be when the α, β are all rational, but not all integral.

Acknowledgements We thank Inger Håland Knutson for helpful comments and nice examples. This work was supported (in part) by a grant from The City University of New York PSC-CUNY Research Award Program.

References

[1] Vitaly Bergelson and Alexander Leibman, *Distribution of values of bounded generalized polynomials*, Acta Math. **198** (2007), no. 2, 155–230. **MR 2318563**

[2] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete mathematics*, 2nd ed. Addison-Wesley Publishing Company, Reading, MA, 1994. A foundation for computer science. **MR 1397498**

[3] Vitaly Bergelson, Inger J. Håland Knutson, and Randall McCutcheon, *IP-systems, generalized polynomials and recurrence*, Ergodic Theory Dynam. Systems, **26**, (2006), no. 4, 999–1019. **MR 2246589**

[4] Inger Johanne Håland, *Uniform distribution of generalized polynomials of the product type*, Acta Arith., **67** (1994), no. 1, 13–27 **MR 1292518**

[5] Kevin O'Bryant, *Fraenkel's partition and Brown's decomposition*, Integers, **3** (2003), A11, 17 pp. (electronic). **MR 2006610**

[6] Kenneth Valbjørn Rasmussen, *Ligefordelte følger i* $[0, 1]^k$ *med anvendelser*, FAMøS 18 (2004), no. 2, 35–42.

[7] I. M. Vinogradov, *The method of trigonometrical sums in the theory of numbers*, Dover Publications Inc. Mineola, NY, 2004. Translated from the Russian, revised and annotated by K. F. Roth and Anne Davenport; Reprint of the 1954 translation. **MR 2104806**

Variance of Signals and Their Finite Fourier Transforms

D.V. Chudnovsky, G.V. Chudnovsky, and T. Morgan

Summary The study of properties of the finite Fourier matrix can be traced back to I. Schur and his use of the Gauss reciprocity law to determine the spectral properties of the finite Fourier matrix. Unlike the continuous Fourier transform, there is no widely accepted eigenvector basis. We present different approaches to eigenvector construction for the finite Fourier matrix, and a new set of extremality principles for the finite Fourier transform. One of the consequences of this construction is a new discrete uncertainty principle, analogous to a classical Heisenberg–Weyl formulation.

Keywords Eigenvectors · Fourier matrix · Fractional fourier transform

Mathematics Subject Classifications (2010). Primary 11C20, 65T50, 65F15

1 Eigenvalue and Eigenvectors of the Finite Fourier Matrix

The finite Fourier matrix F of size N:

$$F = F_N = \left(e^{\frac{2\pi\sqrt{-1}}{N} ij} \right)_{i,j=0}^{N-1}$$

has eigenvalues only among four values: $\pm 1, \pm i$. For $N \geq 4$, there are always four orthogonal subspaces of eigenvectors for the finite Fourier transform. This means that there are many possible choices of eigenvector basis for the Fourier matrix.

To show why the finite Fourier F matrix has eigenvalues $\pm 1, \pm i$, following McClellan [18], notice that

$$F = F^T \quad F.F^* = F^*.F = I \tag{1}$$

D.V. Chudnovsky, G.V. Chudnovsky, and T. Morgan
Polytechnic Institute of NYU, IMAS, 6 MetroTech Center, Brooklyn, NY 11201, USA
e-mail: david@imas.poly.edu; gregory@imas.poly.edu; tmorgan@acm.org

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_4,

Table 1 Eigenvalue multiplicities-Fourier matrix of dimension N

$N =$	$\lambda = +1$	$\lambda = -1$	$\lambda = +\mathrm{i}$	$\lambda = -\mathrm{i}$
$4m$	$m+1$	m	m	$m-1$
$4m+1$	$m+1$	m	m	m
$4m+2$	$m+1$	$m+1$	m	m
$4m+3$	$m+1$	$m+1$	$m+1$	m

that is, F is symmetric and has an inverse, which is equal to the conjugate transpose of F. Since F is unitary, all of its eigenvalues are of absolute value 1 and there exists a set of N orthonormal eigenvectors for the matrix F.

Next, notice that

$$F^4 = I$$

This can be seen by considering the action of $F^2 = F.F$ on a vector, which can be seen to reverse all vector coordinates save the first. A second application of F^2 to a vector returns the original order; that is, $F^4 = I$. If λ is an eigenvalue of F, then λ^4 is an eigenvalue of F^4, but the eigenvalues of F^4 are all one, hence λ is a root of the cyclotomic polynomial $x^4-1 = 0$; that is $\pm 1, \pm i$. Moreover, the multiplicities of the eigenvalues of the Fourier matrix can be computed for arbitrary N.

The table above is due to Schur [24] and is a direct consequence of the Gauss quadratic reciprocity theorem, applied to the following trace of the Fourier matrix:

$$\sum_{j=0}^{N-1} e^{\frac{2\pi\sqrt{-1}}{N} j^2}$$

Knowledge of the value of this sum, and simple formulas for the traces of other powers F^k, together with the observation that the eigenvalues of the Fourier matrix are ± 1, $\pm i$ permits the calculation of the multiplicities of the eigenvalues as a function of the size of the Fourier matrix.

Since there are only four possible eigenvalues, there is enormous freedom possible in the selection of a basis of eigenvectors for the Fourier matrix. The dimension of each of the eigenspaces is roughly $N/4$. Any basis in such a subspace can serve as a basis of eigenvectors. This embarrassment of riches actually represents a problem, because there is no clear canonical choice to be made for the eigenvectors. This problem is particularly acute for the fractional Fourier transform, which requires a choice of an eigenvector basis before it can even be defined. On the positive side, one can choose eigenvectors of the Fourier matrix to satisfy various computational needs. For example, choosing a basis of eigenvectors which is sparse represents a possible alternative to evaluation of the discrete Fourier transform by means of the FFT techniques.

Such an alternative to FFT algorithms would be a sparse representation (or numerically sparse) for the matrix F in a suitable basis. This is not effective for the computation of the full discrete Fourier transform on a standard computer architecture, but is potentially interesting for computation with minors of the Fourier matrix, see [9].

1.1 McClellan Basis

The basis of eigenvectors described by McClellan [18] has the advantage of being relatively simple to describe; however, the basis is not orthogonal. Yarlaggada [27] provided a means of orthonormalizing these vectors.

This basis was described in the engineering signal processing literature and, as pointed out by Auslander and Tolimieri in [3], the work was done without reference to the work of Schur. For a modern description, see [17].

The basis is constructed by forming a basis of even and odd functions and then making adjustments to the vectors to convert them into eigenvectors while preserving the basis property.

For this purpose, a vector $v = (v_0, \ldots, v_{n-1})$ is said to be even if $v_i = v_{n-i}$ and odd if $v_i = -v_{n-i}$ (here the indices are taken mod n).

It is a property of the Fourier matrix F_n that ${F_n}^2$ is the vector index permutation operator $P(v) = (v_0, v_{n-1}, \ldots, v_1)$; that is $P(v)$ leaves the first vector component in place and reverses the order of the remaining components.

This property allows us to construct an eigenvector for any even vector v as follows

$$F_n \cdot v \pm v$$

for eigenvalues ± 1, respectively.

Similarly, any odd vector, u, can be used to construct an eigenvector of the Fourier matrix by taking

$$\sqrt{-1} F_n \cdot u \pm u$$

for eigenvalues $\mp i$ respectively.

The McClellan basis is constructed as follows. Let

$$\nu = \left\lceil \frac{n}{2} \right\rceil$$

and construct the vectors of dimension n.

$$u_1 = (1, 0, \ldots, 0), u_2 = (1, 0, \ldots, 1)$$

and

$$u_j = \left(\ldots, \underset{\substack{\uparrow \\ j^{\text{th}}\text{position}}}{1}, \ldots, \underset{\substack{\uparrow \\ n-2+j^{\text{th}}}}{1}, \ldots 0 \right) \text{ for } j = 3, \ldots \nu$$

and

$$v_1 = (0, 1, 0, \ldots, -1), v_2 = (0, 0, 1, 0, \ldots, -1)$$

and

$$v_j = \left(\ldots, \underset{\substack{\uparrow \\ j+1^{\text{th}}\text{position}}}{1}, \ldots, \underset{\substack{\uparrow \\ n+1-j^{\text{th}}}}{-1}, \ldots 0 \right) \text{ for } j = 3, \ldots n-v$$

Here is an example of the vectors u_j, v_j for $n = 7$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}$$

Notice that the vectors u_j are even, and the vectors v_j are odd. The McClellan basis is then formed by evaluating $F_n \cdot u_j \pm u_j$ for the vectors u_j and $\sqrt{-1} F_n \cdot v_j \pm v_j$ where the choice of addition/subtraction is made in accordance with the known multiplicities of the eigenvalues $\pm 1, \pm i$.

1.2 Carlitz/Morton Basis

These eigenvectors are formed by setting vector coordinates from the character basis mod n. These eigenvectors are difficult to construct, due to the dependency on the factoring of n, on the solution to the discrete logarithm problem, and on the computation of Gauss sums. While the basis was first presented by Morton [20], it is derived from a relation between a set of vectors first described by Carlitz [7]. See also a more recent paper [12] for important applications. The vectors are formed using coordinate entries, which are the values of the mod n characters evaluated at the coordinate number.

These vectors have the property that the action of the Fourier matrix on the vectors yields the conjugate of the vectors times a constant expressed by a Gauss sum. A suitable linear combination of a vector with its conjugate thus yields an eigenvector for the Fourier matrix.

For ease of exposition, we describe Morton's construction of the vectors for a basis of n vectors, whenever n is a prime.

Let α be a primitive root mod n and for any $0 < k \leq n-1$, let $g(k)$ satisfy

$$g(k) = r \text{ when } \alpha^r = k,$$

and define the characters

$$\chi_r(k) = e^{\frac{2\pi\sqrt{-1}rg(k)}{n-1}}$$

then form the vectors

$$u_r = (\chi_r(0)\dots\chi_r(n-1)),\ v_r = \left(\chi_{(n-1-r \bmod n-1)}(0)\dots\chi_{(n-1-r \bmod n-1)}(n-1)\right)$$

the vectors u_r and v_r have the property that the action of the discrete Fourier transform on the vectors u_r and v_r produces vectors which are functions of the original vector and its conjugate. An eigenvector of the discrete Fourier transform can then be constructed by applying suitable adjustment factors as follows.

Let us define

$$\lambda_r = \sqrt{n\chi_r(-1)}$$

and the Gauss sum

$$\tau(\chi_r) = \sum_{k=0}^{n-1} \chi_{(p-1-r \bmod n-1)}(k) e^{\frac{2\pi\sqrt{-1}k}{n}}.$$

It can then be shown that the vectors:

$$u_r + \frac{\lambda_r}{\tau(\chi_r)}v_r,\ v_r - \frac{\lambda_r}{\tau(\chi_r)}v_r$$

are each eigenvectors of the discrete Fourier transform. Some intricate book keeping allows the determination of the appropriate choice of either u_r or v_r for successive eigenvalues. A natural modification is needed to make these eigenvectors real.

The construction above is a consequence not of special properties of the characters, but of a more basic commutativity relationships between combinations of some permutation matrices and the matrix F_n. For example, for the permutation matrix P induced by the modular scaling $x \leftarrow x \times k$ for $(k, n) = 1$, the matrix $P + P^{-1}$ commutes with F_n, providing essentially the same eigenvectors as above. More complex objects built from a larger class of permutations provide sparse matrices commuting with F_n that are tridiagonal, but with permuted rows and columns.

1.3 Dickinson–Steiglitz or Hofstatder Basis

Two separate groups arrived at a similar set of eigenvectors for the Fourier matrix, Harper–Hofstadter [13, 15] (physics) and Dickinson–Steiglitz [10] (fractional fourier transform). These eigenvectors are analogous to the basis functions for the continuous Fourier transform. They are based on discrete analogs of the Hermite functions for the Fourier transform. This basis is found using a periodic tridiagonal matrix commuting with the Fourier matrix. The eigenvectors of this commuting matrix are no longer (completely) degenerate, and are also eigenvectors of the

Fourier matrix. The use of tridiagonal and other sparse matrices is one of the tools used in the study of other related matrices and transforms, particularly the q-Fourier transforms, see [9].

The importance of an eigenvector basis analogous to the Hermite functions is best illustrated by its application to the discrete fractional Fourier transform. In order to compute the discrete fractional transform, one has to compute the fractional power F^{α} of the Fourier matrix F, where α is an arbitrary real parameter.

The standard definition of a fractional power of a matrix A, is

$$U \cdot \Lambda^{\alpha} U^{-1}$$

where U is the eigenvector matrix of A and Λ is a diagonal matrix formed from the eigenvalues of A. If the eigenvalues of A are distinct, then this representation is unique. If A has eigenvalues with multiplicity greater than one, then a 'natural' choice of eigenvectors must be determined.

In the case of the matrix F, there are only four distinct eigenvalues, $\{1,-1, i,- i\}$, yielding a large multiplicity for matrices of any size.

Traditional choices for U in this case are based on finding a matrix H, often called Harper's matrix, which commutes with F and which has (mostly) distinct eigenvalues. Harper originally defined this matrix in [13]. A good reference for the modern application of Harper's matrix in the context of the discrete fractional Fourier transform can be found in Ozatakas [22]:

$$H = \frac{\pi}{(i\,2\pi)^2} \begin{pmatrix} -2 & 1 & 0 & 0 & 1 \\ 1 & 2\cos(\frac{2\pi}{N}) - 4 & 1 & \cdots & 0 \\ 0 & 1 & 2\cos(\frac{2\pi 2}{N}) - 4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 2\cos(\frac{2\pi(N-1)}{N}) - 4 \end{pmatrix}$$

In this form, the matrix has analogies to the equivalent eigenfunctions of the continuous Fourier transform, written in the Hermite basis. This matrix commutes with the Fourier matrix, thus its eigenvectors are also eigenvectors of the Fourier matrix. This matrix exists in the literature in many equivalent forms (in the sense that the alternative forms commute with the Fourier matrix and have distinct eigenvalues). Another early appearance of this matrix is in Hofstadter [15]. Hofstadter used a matrix of this form in similar applications and showed that the matrix commutes with the Fourier matrix. Dickinson and Steiglitz [10] rediscovered a matrix of this form for application to the fractional Fourier transform. Before the paper of Dickinson and Steiglitz, the bases used for the study of the fractional Fourier transform were unsatisfactory, often creating numerical instabilities in the evaluation of the fractional Fourier transform in applications. Ozaktas and Kutay [22] provide a useful table of correspondences between the eigenvectors of H and the eigenfunctions of the continuous Fourier transform. Notice that the matrix H can have repeated eigenvectors when N is a multiple of 4 and some means is required to uniquely establish an eigenvector set in this case.

2 Discrete Analogs

The Dickinson–Steiglitz basis is motivated by a desire for a discrete analog for the Hermite function basis of the Fourier transform.

Properties desired for the discrete analog of the Gaussian are: they should tend to the Gaussian in the limit as N goes to infinity; should have zero crossing properties matching the Hermite functions, at least for lower order eigenstates; should have some analog of the creation/annihilation operators; should posses some analog of the differential equation.

Notice that it is not possible to have all of the above, for otherwise, the discrete Fourier transform would be a completely integrable system, which it is not.

Even though it is impossible to maintain all properties for a discrete analog of the Hermite functions, some can be met. The Dickinson–Steiglitz eigenvectors are orthogonal, are the solutions to a q-difference equation analog of the harmonic oscillator equation and have eigenvalues which are either $\pm\ 1$ or $\pm\ i$. The Dickinson–Steiglitz eigenvectors arrive at discrete analogs for the Hermite functions indirectly, by the solution of the eigen problem for the tridiagonal matrix H (defined above) commuting with the Fourier matrix.

The eigenvectors found in this way form an orthogonal basis of eigenforms for the discrete Fourier transform; are solutions of a difference analog of the differential equations for the harmonic operator and the eigenvalues of these discrete analogs of the Hermite functions are either $\pm\ 1$ or $\pm\ i$.

One would also like to use eigenvectors with an analytic definition, as compared to numerical or complex algebraic methods required to produce the eigenvectors for the Hofstadter matrix – cf. with the theta functions below.

We also investigated the possibility of an alternative definition of the discrete fractional Fourier transform by means of an expansion in power series of the q-matrix and proceeding to the limit $q \to$ root of unity, see [9].

3 Theta Function Expressions for the Fourier Eigenvectors

We present the general expression for the eigenvector of the Fourier matrix of order n with the eigenvalue i^k in terms of the θ-functions with an arbitrary period τ. A set of eigenvectors of the Fourier matrix built from derivatives of the θ-functions was presented by M.L. Mehta [19]. We use expression, arising from the general Weil formalism for multidimensional Θ-function, see Mumford lectures on theta-functions [21]; see also [17].

Because there are different notations used for the theta-functions, we chose notations accepted for *Mathematica* expressions, similar to those adopted by Whittaker and Watson [26]. In these notations the most general (one-dimensional) theta function is:

$$\theta(x, \tau) = \sum_{m \in \mathbb{Z}} \mathrm{e}^{\pi \mathrm{i} \tau m^2 + 2\pi \mathrm{i} m x},$$

for $\Im(\tau) > 0$. This expression coincides with the Jacobi's theta function $\vartheta_3(x,q)$, if defined as follows:

$$\vartheta_3(x,q) = \theta(x,\tau) \quad \text{for} \quad q = \mathrm{e}^{\pi \mathrm{i} \tau}.$$

In *Mathematica* $\vartheta_3(x,q)$ is called "EllipticTheta[3,πx,q]".

For further use below, we also provide the expressions of three other Jacobi's theta function $\vartheta_k(x,q)$, for $k = 1,2,4$ in terms of $\theta(x,\tau)$ as follows:

$$\begin{aligned}
\vartheta_1(x,\tau) &= -\mathrm{i} \cdot q^{\frac{1}{4}} \cdot \mathrm{e}^{\pi \mathrm{i} x} \cdot \theta\left(x + \frac{1}{2} + \frac{\tau}{2}, \tau\right), \\
\vartheta_2(x,\tau) &= q^{\frac{1}{4}} \cdot \mathrm{e}^{\pi \mathrm{i} x} \cdot \theta\left(x + \frac{\tau}{2}, \tau\right), \\
\vartheta_4(x,\tau) &= \theta\left(x + \frac{1}{2}, \tau\right).
\end{aligned}$$

Following these notations (for a fixed n), the vector $\Phi(k,x,\tau) = \left(\Phi_j(k,x,\tau)\right)_{j=0}^{n-1}$ is an eigenvector of the Fourier matrix F_n with the eigenvalue i^k if its components are defined as follows:

$$\begin{aligned}
\Phi_j(k,x,\tau) = {} & \theta\left(x + \frac{j}{n}\tau, \tau\right) \cdot \mathrm{e}^{\pi \mathrm{i}\left(\frac{j}{n}\right)^2 \tau + 2\pi \mathrm{i} \frac{j}{n} x} \\
& + (-1)^k \cdot \theta\left(x - \frac{j}{n}\tau, \tau\right) \cdot \mathrm{e}^{\pi \mathrm{i}\left(\frac{j}{n}\right)^2 \tau - 2\pi \mathrm{i} \frac{j}{n} x} \\
& + \frac{1}{\sqrt{n}} \cdot \left((-\mathrm{i})^k \cdot \theta\left(\frac{x+j}{n}, \frac{\tau}{n^2}\right) + (-\mathrm{i})^{3k} \cdot \theta\left(\frac{x-j}{n}, \frac{\tau}{n^2}\right) \right)
\end{aligned}$$

for $j = 0, \ldots, n-1$. Using Jacobi's imaginary transformation identity:

$$\theta\left(\frac{x}{\tau}, -\frac{1}{\tau}\right) = (-\mathrm{i}\tau)^{1/2} \mathrm{e}^{\mathrm{i} \frac{\pi x^2}{\tau}} \cdot \theta(x,\tau),$$

we can re-write the formulas in a more "symmetric form" (with respect to τ):

$$\begin{aligned}
\Phi_j(k,x,\tau) = {} & \theta\left(x + \frac{j}{n}\tau, \tau\right) \cdot \mathrm{e}^{\pi \mathrm{i}\left(\frac{j}{n}\right)^2 \tau + 2\pi \mathrm{i} \frac{j}{n} x} \\
& + (-1)^k \cdot \theta\left(x - \frac{j}{n}\tau, \tau\right) \cdot \mathrm{e}^{\pi \mathrm{i}\left(\frac{j}{n}\right)^2 \tau - 2\pi \mathrm{i} \frac{j}{n} x} \\
& + \frac{\sqrt{n}}{\sqrt{-\mathrm{i}\tau}} \cdot \left((-\mathrm{i})^k \cdot \theta\left(\frac{(x+j)\cdot n}{\tau}, \frac{-n^2}{\tau}\right) \cdot \mathrm{e}^{-\mathrm{i}\pi \frac{(x+j)^2}{\tau}} \right. \\
& \left. + (-\mathrm{i})^{3k} \theta\left(\frac{(x-j)\cdot n}{\tau}, \frac{-n^2}{\tau}\right) \cdot \mathrm{e}^{-\mathrm{i}\pi \frac{(x-j)^2}{\tau}} \right)
\end{aligned}$$

for $j = 0, \ldots, n-1$.

Clearly, to have the θ-functions with the same τ (or q), we need to set

$$\tau = \mathrm{i} \cdot n.$$

Under this assumption, we have more natural expressions for the eigenvector components:

$$\begin{aligned}\Phi_j(k,x) = {} & \theta(x+\mathrm{i}\cdot j, \mathrm{i}\cdot n)\cdot \mathrm{e}^{\frac{-\pi j^2}{n}+2\pi \mathrm{i}\frac{j}{n}x} + (-1)^k \cdot \theta(x-\mathrm{i}\cdot j, \mathrm{i}\cdot n)\cdot \mathrm{e}^{\frac{-\pi j^2}{n}-2\pi \mathrm{i}\frac{j}{n}x} \\ & + \Bigg((-\mathrm{i})^k \cdot \theta(\mathrm{i}\cdot (x+j), \mathrm{i}\cdot n)\cdot \mathrm{e}^{-\pi\frac{(x+j)^2}{n}} \\ & \quad + (-\mathrm{i})^{3k}(-1)^k \cdot \theta(\mathrm{i}\cdot (x-j), \mathrm{i}\cdot n)\cdot \mathrm{e}^{-\pi\frac{(x-j)^2}{n}} \Bigg)\end{aligned}$$

for $j = 0, \ldots, n-1$. Here, one can take x among the set of integers $x = 0, \ldots, n-1$ of the cardinality not larger than $m(i^k)$ – the multiplicity of the eigenvalue i^k of F_n (see Table 1 for the explicit expression). In cases $k = 1, 2, 3$, one has to exclude the zero case of $x = 0$. Then vectors

$$(\Phi_j(k,x))_{j=0}^{n-1}$$

are eigenvectors of F_n for these various sets of x.

The simplest case – for $k = 0$ (an eigenvalue of 1 for F_n) – arises at $x = 0$ when we get a very simple eigenvector:

$$\Xi = (\xi_j)_{j=0}^{n-1},$$

$$\xi_j = \theta(\mathrm{i}\cdot j, \mathrm{i}\cdot n)\cdot \mathrm{e}^{\frac{-\pi j^2}{n}}, \quad j = 0, \ldots, n-1.$$

This eigenvector can be considered as the "elliptic analogue" of the Gaussian bound state for the F_n. What is interesting is that using simple linear transformation of the vector Ξ, one can derive all eigenvectors $(\Phi_j(k,x))_{j=0}^{n-1}$. Specifically, if we want to present only the real-value eigenvectors, we get for integers $x = 0, \ldots, n-1$ the following expressions of the corresponding eigenvectors for all 4 eigenvalues i^k, $k = 0, 1, 2, 3$:

$$\Phi(x) = (\Phi_j(x))_{j=0}^{n-1}:$$

$$\Phi_j(x) = \xi_{j+x} + \xi_{j-x} + 2\cdot\cos\left(\frac{2\pi jx}{n}\right)\cdot\xi_j; \quad k = 0;\ x \geq 0$$

$$\Phi_j(x) = \xi_{j+x} - \xi_{j-x} - 2\cdot\sin\left(\frac{2\pi jx}{n}\right)\cdot\xi_j; \quad k = 1;\ x > 0$$

$$\Phi_j(x) = \xi_{j+x} + \xi_{j-x} - 2\cdot\cos\left(\frac{2\pi jx}{n}\right)\cdot\xi_j; \quad k = 2;\ x > 0$$

$$\Phi_j(x) = \xi_{j+x} - \xi_{j-x} + 2\cdot\sin\left(\frac{2\pi jx}{n}\right)\cdot\xi_j; \quad k = 3;\ x > 0$$

These expressions that show how to generate all eigenvectors from the bound state Ξ are very similar to the "creation operators" that generate higher Hermite functions – the eigenfunctions of the continuous Fourier transform – from the Gaussian (the bound state of the continuous Fourier transform). Unfortunately, the basis $\Phi(x)$ thus generated is not an orthogonal basis, and has to be orthogonalized using the Gram–Schmidt recursive orthogonalization procedure.

It does not mean, though, that such a construction is completely "nonexplicit". In the orthogonalization procedure, one needs to evaluate the Hankel determinants made of the self-correlations of the elements of the vector Ξ – like this:

$$< \Xi, \Xi >_m = \sum_{i=0}^{n-1} \xi_j \cdot \xi_{j+m},$$

where ξ_k is defined for $k \bmod n$ due to the quasi-periodicity properties of $\theta(x, \tau)$. Similarly, one needs the "shifted" self-correlations $\sum_{i=0}^{n-1} \xi_j \cdot \xi_{j+m} \cdot \mathrm{e}^{\frac{2\pi i j x}{n}}$.

Interestingly enough, these self-correlation can be expressed in a slightly simpler form using various forms of addition laws for θ-functions. To describe the relations between self-correlation coefficients, we will represent the vector Ξ in a slightly different form, as a vector Υ:

$$\Upsilon = \left(\theta\left(\frac{j}{n}, \frac{i}{n}\right)\right)_{j=0}^{n-1},$$

simply related to Ξ:

$$\Upsilon = \sqrt{n} \cdot \Xi.$$

The addition formulas that we need arise from a particular representation of the elliptic curve using torsion subgroup of order n^2 of all division points of order n. We refer to [8] for the description. Here, we will present only one such formula, using the conventional set of Jacobi ϑ-functions (corresponding to division points of order 2).

For example, we have the following expression for the Υ self-correlations, whenever n is even:

$$\langle \Upsilon, \Upsilon \rangle_{2m} = \langle \Upsilon, \Upsilon \rangle \cdot \frac{\left(\vartheta_2\left(\frac{m}{n}, \frac{i}{n}\right)^2 + \vartheta_1\left(\frac{m}{n}, \frac{i}{n}\right)^2\right)}{\vartheta_2\left(0, \frac{i}{n}\right)^2}.$$

In fact, the correlations $\langle \Upsilon, \Upsilon \rangle_m$ and $\langle \Xi, \Xi \rangle_m$ can be expressed explicitly, but differently for odd and even n. For even n, we get:

$$\langle \Xi, \Xi \rangle_0 = \|\Xi\|^2 = \sqrt{\frac{n}{2}} \cdot \theta\left(0, \frac{in}{2}\right)^2,$$

while for an odd n:

$$\langle \Xi, \Xi \rangle_0 = \|\Xi\|^2 = \sqrt{\frac{n}{2}} \cdot \left(\theta \left(0, \frac{in}{2} \right)^2 - 4 \cdot \theta(0, 4in) \cdot \vartheta_2(0, 4in) \right).$$

Alternatively, one can differentiate expressions $\Phi_j(k, x)$ in x, and then set $x = 0$ to get the basis of eigenvectors for F_n build from derivatives of the theta-function values at torsion points, see [19]. Nevertheless, the theta-function expressions for higher eigenvectors cannot qualify to be called explicit expressions. Moreover, the orthogonalization of this basis is very complex compared to the above.

All various expressions of the eigenvectors in terms of the theta-functions presented here (particularly those in terms of the Ξ or Υ vectors or similar derivatives) are not transcendental expressions, despite their appearance. The normalized eigenvectors – for example, normalized by the norm $\|\Xi\|^2$, or $\langle \Xi, \Xi \rangle_0$ -have their components as algebraic numbers. These algebraic numbers are elements of the algebraic number field, generated by addition of torsion points of order n (for even n) for an elliptic curve corresponding to the periods 1 and i, i.e., isogenous to the lemniscate elliptic curve $y^2 = x^4 - 1$.

4 Variational Principles for the Determination of Eigenfunctions of the Discrete Fourier Transform

It would be desirable to have a natural definition of an eigenfunction basis for the discrete Fourier transform, using a variational principle based on a quadratic positive definite form. In this formulation, the eigenfunctions sought will be provided by successive extremal values of the Rayliegh–Ritz ratio:

$$R_M(v) \quad = \quad \frac{v \cdot M \cdot v^*}{v \cdot v^*}$$

with subsequent identification of the corresponding extremal vectors as the eigenvectors of the Fourier matrix. One can restrict the discussion below to the case of v with real valued components, although everything below is completely extended to arbitrary complex vectors v. We establish this limitation on v in anticipation of establishing various uncertainty principle inequalities analogous to the uncertainty principles for integrable functions defined on the real axis. The discussion following can also be generalized to multivectors, which would establish analogies for uncertainty principles for real valued functions in $\mathbb{R}^m$.

If the vectors which are extrema of the Rayliegh–Ritz ratio corresponding to the matrix M are to be eigenvectors of the discrete Fourier transform, then M must commute with F_N. We also want the expression for the ratio $R_M(v)$ to be symmetric

in the vector v and its discrete Fourier transform, $\hat{v}$. Introducing a new matrix A, these conditions can be stated as follows:

$$\frac{v \cdot A \cdot v^* + \hat{v} \cdot A \cdot \hat{v}^*}{v \cdot v^*} = \frac{v \cdot M \cdot v^*}{v \cdot v^*},$$
$$F_n \cdot M = M \cdot F_n$$

Here we mean by $\hat{v}$ the discrete Fourier transform of v, $\hat{v} = F_n \cdot v$. This defines the matrix M in terms of the matrix A, as follows:

$$M = A + F_n \cdot A \cdot {F_n}^*.$$

If M is to commute with F_N, then the following condition must hold for the matrix A:

$${F_n}^2 \cdot A = A \cdot {F_n}^2$$

This symmetry condition for the matrix $A = (a_{i,j})_{i,j=0}^{n-1}$, is equivalent to the following condition on the components of A:

$$a_{n-i,j} = a_{i,n-j}$$

where we understand the indices mod n for $i, j = 0, \ldots, n-1$. Diagonal matrices satisfying this symmetry condition are of most interest since A, a general matrix satisfying this symmetry condition, can be reduced by a unitary transformation to a diagonal matrix. We will consider in this section only diagonal matrices A satisfying this symmetry condition. If A is diagonal and symmetric in the sense above, we may characterize the class of all matrices M derived from A in terms of circulant matrices.

If we take a general circulant matrix $C = (c_{i-j \bmod n})_{i,j=0}^{n-1}$, then it is well-known that C can be represented in the form

$$C = F_N \cdot W \cdot {F_N}^*$$

where $W = \mathrm{diag}(w)$ is a diagonal matrix with the vector w being a discrete Fourier transform of a vector c. Using this notation, the general form of the matrix M can now be written as

$$W + C$$

where W has to satisfy the symmetry condition on the vector w, $w_{n-i} = w_i$. In order to make the correspondence with the case of the continuous Fourier transform more direct, we need to normalize the Rayleigh–Ritz ratio $R_M(v)$. We also want to write the expression in a more symmetric form, extending the vector v of length n by periodicity mod n to arbitrary indices i by $v_i = v_{i \bmod n}$. Care is also needed in establishing the upper and lower limits of summation for n odd and for n even. To do so, define the "nonnegative" indices i in the following range:

$$i = 0, \ldots, n_+, \quad \text{for} \quad n_+ = \left\lceil \frac{n-1}{2} \right\rceil$$

and the "negative" indices i are defined in the range

$$i = -n_-, \ldots, -1, \quad \text{for} \quad n_- = \left\lfloor \frac{n-1}{2} \right\rfloor$$

Notice that for n odd, $n_- = n_+ = \frac{n-1}{2}$. We introduce the following weighted norm, which can be thought of as a weighted variance for a vector v, with weights w:

$$V_w(v) = \frac{1}{n} \sum_{k=-n_-}^{n_+} w(k)|v_k|^2 \quad \text{for} \quad v = (v_i)_{i=0}^{n-1}.$$

Here because of the symmetry condition on w imposed above, $w(-j) = w(j)$. With these definitions, and substituting $M = W + C$, the Rayleigh–Ritz ratio can now be written, up to the factor $\frac{1}{n}$, as:

$$\frac{V_w(v) + V_w(\hat{v})}{\|v\|^2}$$

The matrix M_w corresponding to this weighted variance has the form

$$M_w = \frac{1}{n} \cdot \left(m_{i,j}\right)_{i,j=-n_-}^{n_+} \quad \text{for } m_{i,j} = \frac{1}{n} \left(\sum_{k=-n_-}^{n_+} w(k) \cdot \cos \frac{2\pi k(i-j)}{n} + \delta_{i,j} \cdot w(i) \right).$$

Notice that by setting weights $w_2(k) = k^2$, the "weighted variance" $V_{w_2}(v)$ thus corresponds to the standard definition for the variance of a discrete function v_k:

$$V(v) = \frac{1}{n} \sum_{k=-n_-}^{n_+} k^2 |v_k|^2$$

The most noteworthy appearance of variance in harmonic analysis occurs in the Heisenberg Uncertainty Principle. We formulate the "sum version" of the Heisenberg Uncertainty Principle as follows:

$$V(f) + V(\hat{f}) \geq \frac{\|f\|^2}{2\pi}, V(f) = \int_{-\infty}^{\infty} x^2 \cdot |f(x)|^2 \mathrm{dx}.$$

This "sum version" of the Uncertainty Principle is in fact equivalent to the more familiar "product version" of the Uncertainty Principle:

$$V(f) \cdot V(\hat{f}) \geq \frac{\|f\|^4}{16\pi^2}$$

The standard reduction of the "sum version" to the "product version" is done by applying the "sum version" to dilated functions, $f_c(x) = f(c \cdot x)$. This way one gets an inequality involving a quadratic polynomial in c^2 resulting from the dilation. For our purposes, the crucial consequences of the Uncertainty Principle are:

- The fact that it provides a canonical characterization of the Gaussian function, as that function which achieves equality in either form of the Uncertainty Principle. This characterization establishes the Gaussian (up to a normalization constant) to be $e^{-\pi x^2}$.
- Equality in the sum version of the Uncertainty Principle and the successive extremal vectors of the corresponding Rayleigh–Ritz ratio characterize the Hermite functions.

We will use these characterizations to define discrete analogues of the Gaussian and Hermite functions.

5 Discrete Uncertainty Principle

The search for discrete analogs of the Hermite functions leads naturally to a desire to find a reasonable discrete analog of the Heisenberg Uncertainty Principle. Notice that this is a variation on the more widely discussed discrete uncertainty principles involving the size of the support of a vector and its discrete Fourier transform. Our new analog involves the introduction of weights/distances over and above the basic structure of functions over modular integers. With these preliminaries, we are in a position to state the discrete analog of the sum form of the Heisenberg Uncertainty Principle.

This is the Discrete Uncertainty Principle:

For an arbitrary vector v of real numbers of length n and its discrete Fourier transform, $\hat{v}$, we have:

$$V(v) + V(\hat{v}) \geq \lambda_1 \cdot \|v\|^2,$$

where numerically

$$\lambda_1 = \frac{1}{2\pi} - e^{-\frac{n\pi}{2}} \cdot \sqrt{n} \cdot (1 + o(1)).$$

Here λ_1 is the lowest eigenvalue of the matrix M_{w_2}, defined above, for the case of the quadratic weight. Moreover, as in the Heisenberg Uncertainty Principle, the inequality is attained by an extremal function which is a reasonable discrete analog of the Gaussian function.

Similarly, analogs of higher order Hermite functions arise from the higher order eigenvectors of the matrix M_{w_2} for quadratic weights. The most interesting analog is the second eigenvector $v2$ of M_{w_2}, which is an odd vector; $v2_{-i} = -v2_i$. $v2$ is

an extremal function for the discrete analog of the Heisenberg Uncertainty Principle for odd functions, with

$$V(v) + V(\hat{v}) \geq \lambda_2 \cdot \|v\|^2,$$

where again numerically

$$\lambda_2 = \frac{3}{2\pi} - e^{-\frac{n\pi}{2}} \cdot O\left(\sqrt{n}\right).$$

For odd vectors v only, satisfying $v(-i) = -v(i)$ for $i \bmod n$, the inequality is attained only for v as a multiple of $v2$.

There is no obvious discrete analogue of the product version of the Heisenberg Uncertainty Principle in this formulation, because δ- functions are in our space of eligible vectors. For these δ- function or for vectors with all components equal, the variances $V(v)$ or $V(\hat{v})$ are zeros. This excludes any nontrivial lower bound on the product $V(v) \cdot V(\hat{v})$. A product-like analogue of the Heisenberg Uncertainty Principle can still be formulated for odd vectors v (which cannot have zero variance). For these vectors, using the λ_2 expression above, the discrete analog of the product form Heisenberg Uncertainty Principle is:

$$V(v) \cdot V(\hat{v}) \geq \left(\frac{9}{16\pi^2} - e^{-\frac{n\pi}{2}} \cdot O\left(\sqrt{n}\right)\right) \|v\|^4$$

This product version also can be reduced to the sum version, as in the continuous case. The main difference is that, since one cannot "dilate" the discrete function by an arbitrary constant $c > 0$, there is a need to consider the modified variational principle with the variance of v and $\hat{v}$ represented with factors c and $1/c$, creating the matrix $c \cdot W + \frac{1}{c} \cdot C$ for the quadratic weights $w_2(i) = i^2$. The second eigenvalue of this matrix is bound by the same λ_2, giving the product version of the discrete uncertainty principle for the odd vectors v.

The product version of the standard Uncertainty Principle can be still formulated in the discrete case, if one adds an $\varepsilon > 0$ to the quadratic weight $w(i) = i^2$. With ε of the order of $e^{-O(n)}$, one achieves a familiar "product form" of the Discrete Uncertainty Principle.

6 Explicit Expressions for the Matrix M_{w_2} in the Discrete Version of the Uncertainty Principle

In general, the "extremality matrices" M_w for weights $w(i)$ might look rather complex, but for most interesting weight classes these expression can be greatly simplified. The most interesting among them is the matrix M_{w_2} in the Discrete Version of the Uncertainty Principle, corresponding to the quadratic weight $w_2(i) = i^2$ and the classical definition of the variance $V(v)$ of the vector v. The expressions of elements of M_{w_2} are slightly different for even and odd n.

This is how the matrix M_{w_2} looks in the case of the even n:

$$\frac{1}{n}\begin{pmatrix} \frac{1}{12}\left(n^2+2\right) & -\frac{1}{2}\csc^2\left(\frac{\pi}{n}\right) & \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{2}\csc^2\left(\frac{\pi}{n}\right) & \frac{1}{12}\left(n^2+2\right)+1 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \cdots \\ \frac{1}{2}(-1)^i\csc^2\left(\frac{i\pi}{n}\right) & \cdots & i^2+\frac{1}{12}\left(n^2+2\right) & \cdots & \frac{1}{2}(-1)^{i-j}\csc^2\left(\frac{(i-j)\pi}{n}\right) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

This is the formal expression of the elements of M_{w_2} in the case of *even* n:

$$\left(M_{w_2}\right)_{i,j} = \frac{1}{n}\begin{cases} \min(i,n-i)^2+\frac{1}{12}\left(n^2+2\right) & i=j \\ \frac{1}{2}(-1)^{i-j}\csc^2\left(\frac{(i-j)\pi}{n}\right) & \text{otherwise} \end{cases}$$

The next example of the matrix $2\cdot n\cdot M_{w_2}$ for the case of $n=14$ uses the following shorthand $s_i = \frac{1}{\sin^2(\frac{i\pi}{14})}$:

$$\begin{pmatrix} 33 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 \\ -s_1 & 35 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 \\ s_2 & -s_1 & 41 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 \\ -s_3 & s_2 & -s_1 & 51 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 \\ s_3 & -s_3 & s_2 & -s_1 & 65 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 \\ -s_2 & s_3 & -s_3 & s_2 & -s_1 & 83 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 \\ s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 105 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 \\ -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 131 & -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 \\ s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 105 & -s_1 & s_2 & -s_3 & s_3 & -s_2 \\ -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 83 & -s_1 & s_2 & -s_3 & s_3 \\ s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 65 & -s_1 & s_2 & -s_3 \\ -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 51 & -s_1 & s_2 \\ s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 41 & -s_1 \\ -s_1 & s_2 & -s_3 & s_3 & -s_2 & s_1 & -1 & s_1 & -s_2 & s_3 & -s_3 & s_2 & -s_1 & 35 \end{pmatrix}$$

This is the matrix M_{w_2} for *odd* n:

$$\frac{1}{n}\begin{pmatrix} \frac{1}{12}\left(-1+n^2\right) & \frac{-\cos\left(\frac{\pi}{n}\right)}{2\sin^2\left(\frac{\pi}{n}\right)} & \cdots & \cdots & \cdots & \cdots \\ \frac{-\cos\left(\frac{\pi}{n}\right)}{2\sin^2\left(\frac{\pi}{n}\right)} & 1+\frac{1}{12}\left(-1+n^2\right) & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \cdots \\ \frac{(-1)^i\cos\left(\frac{i\pi}{n}\right)}{2\sin^2\left(\frac{i\pi}{n}\right)} & \cdots & i^2+\frac{1}{12}\left(-1+n^2\right) & \cdots & \frac{(-1)^{i-j}\cos\left(\frac{(i-j)\pi}{n}\right)}{2\sin^2\left(\frac{(i-j)\pi}{n}\right)} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

This is the formal expression of the elements of M_{w_2} for odd n:

$$\left(M_{w_2}\right)_{i,j} = \frac{1}{n}\begin{cases} \min(i,n-i)^2+\frac{1}{12}\left(n^2-1\right) & i=j \\ \dfrac{(-1)^{i-j}\cos\left(\frac{(i-j)\pi}{n}\right)}{2\sin^2\left(\frac{(i-j)\pi}{n}\right)} & \text{otherwise} \end{cases}$$

Even though the matrix M_{w_2} is clearly dense, for very large n most elements in the matrix become very small compared to the main diagonal elements. Only the near-diagonal (and corner) elements of the matrix stay large (of the order of n).

7 Theta Function Bounds for Minimal Eigenvalues in the Discrete Uncertainty Principle

In any formulation of the eigenvalue problem using a variational principle (like Rayleigh–Ritz), one can get good bounds on the minimal eigenvalue using a "trial" function. Applying the Rayleigh–Ritz quotient to the trial function one gets an upper bound on the lowest eigenvalue (and also the lower bound on the highest eigenvalue) of the matrix problem, arising from the extremality condition.

In all cases of the eigenvalue λ_1 of the matrix M_{w_2} in the discrete version of the uncertainty principle, we seek the bound state to be as close as possible to the Gaussian $e^{-\pi x^2}$. We can take the discretized Gaussian with $x = \frac{j}{\sqrt{n}}$ naturally as follows:

$$Gaus[n] = \left(e^{-\pi\cdot\frac{j^2}{n}}\right)_{j=-\left\lfloor\frac{n-1}{2}\right\rfloor}^{\left\lceil\frac{n-1}{2}\right\rceil},$$

and use it as a trial function. Of course, for this to work well, we need a vector that is an eigenvector of the discrete Fourier matrix, that is "very close" to this trial function. We actually have this vector, defined above in terms of the θ-function, as Ξ. It turns out that the trial vector Ξ is better than the vector $Gaus[n]$ since it provides a slightly smaller value of the Rayleigh–Ritz quotient.

We present the analysis for the case of the variational problem defined by the discrete Heisenberg uncertainty principle with the Rayleigh–Ritz quotient of $\frac{V(v)+V(\hat{v})}{\|v\|^2}$. In the analysis below, we take an error term as a parameter q for $\tau = \frac{ni}{2}$:

$$q = e^{-\frac{\pi n}{2}}, \quad \tau = \frac{ni}{2}.$$

To start with, the norm of the trial vector Ξ has a closed form expression:

$$\|\Xi\|^2 = \sqrt{\frac{n}{2}} \cdot \theta\left(0, \frac{in}{2}\right)^2 \text{ for } n \text{ even}$$

$$\|\Xi\|^2 = \sqrt{\frac{n}{2}} \cdot \left(\theta\left(0, \frac{in}{2}\right)^2 - 4 \cdot \theta(0, 4in) \cdot \vartheta_2(0, 4in)\right) \text{ for } n \text{ odd}$$

Thus,

$$\|\Xi\|^2 = \sqrt{\frac{n}{2}} \cdot \left(1 + 4q + O(q^2)\right).$$

Now, one needs to bound the variance $V(\Xi)$ of Ξ from above. For this, one needs to approximate the elements ξ_j of Ξ. Here:

$$\xi_j = e^{-\frac{j^2\pi}{n}} \cdot \theta(ij, in) = e^{-\frac{j^2\pi}{n}} \cdot \left(1 + \sum_{k=-\infty, k\neq 0}^{\infty} e^{-\pi k^2 n - 2\pi k j}\right).$$

In the sum above, the term with $k = -sign(j)$ provides the largest contribution. This gives the following tight estimate:

$$\xi_j = e^{-\frac{j^2\pi}{n}} \cdot \left(1 + e^{2\pi|j| - \pi n} + O(q^2)\right)$$

for $|j| \leq \frac{n}{2}$. This implies:

$$\xi_j^2 = e^{-\frac{2j^2\pi}{n}} \cdot \left(1 + 2e^{2\pi|j| - \pi n} + e^{4\pi|j| - 2\pi n} + O(q^2)\right)$$

for $|j| \leq \frac{n}{2}$.

Set as before

$$m = \frac{n}{2}.$$

Then, using the expressions for ξ_j^2 above, we estimate $V(\Xi)$ by:

$$\sum_{|j|\le\frac{n}{2}} j^2\cdot e^{-\frac{2j^2\pi}{n}}\cdot\left(1+e^{2\pi|j|-\pi n}+O(q^2)\right)^2.$$

The sum that we are going to approximate $V(\Xi)$ with is:

$$U_m=\sum_{0\le j\le m} j^2\cdot e^{-\frac{j^2\pi}{m}}\cdot\left(1+e^{2\pi|j|-\pi 2m}\right)^2.$$

In these notations,

$$n\cdot V(\Xi)=2\cdot U_m+O\left(n^{2+\frac{1}{2}}\cdot q^2\right).$$

To estimate U_m, apply the Poisson summation formula to the following "constrained shifted Gaussian" $GU_m(x)$:

$$GU_m(x)=e^{-\frac{\pi x^2}{m}}\left(1+e^{2\pi x-2m\pi}\right)^2x^2$$

for $x\in[0,m]$. We set $GU_m(x)=0$ for $x\notin[0,m]$.

The application of the Poisson summation formula follows the standard method of the proof of the Jacobi's imaginary transformation identity, using *Mathematica's* facility for manipulation with the Fourier transforms, series expansions, and symbolic manipulation with various combinations of the error-function.

As a result, we get the following bound on the U_m:

$$U_m=\frac{n^{\frac{3}{2}}}{8\sqrt{2}\pi}+\frac{e^{-\frac{n\pi}{2}}n^2}{2\pi}+O(n^{\frac{5}{2}}q^2)+O\left(n^{\frac{3}{2}}q\right).$$

Using the expression above of $\cdot V(\Xi)$ in terms of $\cdot U_m$, we get the bound on $V(\hat{\Xi})+V(\Xi)=2\cdot V(\Xi)$:

$$V(\hat{\Xi})+V(\Xi)=\frac{1}{2\pi}\cdot\sqrt{\frac{n}{2}}+q\cdot\frac{2n}{\pi}+O\left(n^{\frac{3}{2}}q^2\right)+O\left(n^{\frac{1}{2}}q\right).$$

Combining this with the estimate of the norm $\|\Xi\|^2$ presented above, we get the value of the Rayleigh–Ritz quotient at $v=\Xi$:

$$\frac{V(\hat{\Xi})+V(\Xi)}{\|\Xi\|^2}\ge\frac{\frac{1}{2\pi}\cdot\sqrt{\frac{n}{2}}+q\cdot\frac{2n}{\pi}+O\left(n^{\frac{3}{2}}q^2\right)+O\left(n^{\frac{1}{2}}q\right)}{\sqrt{\frac{n}{2}}\cdot(1+4q+O(q^2))}.$$

This finally gives the upper bound on the smallest eigenvalue λ_1 of the matrix M_{w_2} in the case of the quadratic weight, that is exponentially (in n) close to its value:

$$\lambda_1 \leq \frac{1}{2\pi} - e^{-\frac{\pi n}{2}} \cdot O(\sqrt{n}).$$

Here, the constant at the $O(\sqrt{n})$ term computed above is

$$\frac{2\sqrt{2}}{\pi},$$

rather close to an asymptotic value of 1 at $n \to \infty$.

8 Numerical Evaluation of the Minimal Eigenvalues in the Discrete Uncertainty Principle

Numerically, as we see below, the actual asymptotics of the smallest eigenvalue λ_1 of the matrix M_{w_2} in the case of the quadratic weight is

$$\lambda_1 \sim \frac{1}{2\pi} - e^{-\frac{\pi n}{2}} \sqrt{n}$$

as $n \to \infty$.

In fact, one gets the following lower bound for the discrete uncertainty principle:

$$\frac{V(\hat{Gaus}) + V(Gaus)}{\|Gaus\|^2} \geq \lambda_1 \geq \frac{1}{2\pi} - 0.00007..$$

for all $n \geq 8$.

The list of evaluated $\frac{1}{2\pi} - \lambda_1$ as a function of n is presented as the Table 2.

9 Extensions of the Heisenberg–Weyl Inequality in the Continuous and Discrete Cases

The uncertainty principle for functions $f(x)$ in $L^2(\mathbf{R})$ and the continuous Fourier transform is often called the Heisenberg–Weyl inequality because the first proof of it belongs to H. Weyl [25].

Since in recent decades, interest in the discrete Fourier transforms and their applications rose steadily, demands for discrete analogues of the uncertainty principle rose as well. There appeared various attempts to bring to the periodic or discrete word extremality inequalities similar to Heisenber–Weyl. We refer to the recent review of Prestin et al. [23] on the better analogues of the uncertainty principle for the periodic functions, defined on the unit circle, represented by infinite Fourier series. Similar results were obtained in papers Ishii–Furukawa [16],

Table 2 Minimal eigenvalues as a function of N

n	$\frac{1}{2\pi} - \lambda_1$	$e^{-\frac{\pi n}{2}}$	$e^{-\frac{\pi n}{2}} \cdot \sqrt{n}$
3	1.82717×10^{-2}	8.98329×10^{-3}	1.55595×10^{-2}
4	2.22527×10^{-3}	1.86744×10^{-3}	3.73489×10^{-3}
5	9.6173×10^{-4}	3.88203×10^{-4}	8.68049×10^{-4}
6	1.42612×10^{-4}	8.06995×10^{-5}	1.97673×10^{-4}
7	4.57888×10^{-5}	1.67758×10^{-5}	4.43845×10^{-5}
8	7.63603×10^{-6}	3.48734×10^{-6}	9.86369×10^{-6}
9	2.16777×10^{-6}	7.24947×10^{-7}	2.17484×10^{-6}
10	3.81625×10^{-7}	1.50702×10^{-7}	4.76561×10^{-7}
11	1.01386×10^{-7}	3.13278×10^{-8}	1.03903×10^{-7}
12	1.84782×10^{-8}	6.51241×10^{-9}	2.25597×10^{-8}
13	4.69947×10^{-9}	1.3538×10^{-9}	4.88119×10^{-9}
14	8.74983×10^{-10}	2.81427×10^{-10}	1.053×10^{-9}
15	2.16012×10^{-10}	5.85029×10^{-11}	2.26581×10^{-10}
16	4.08776×10^{-11}	1.21616×10^{-11}	4.86462×10^{-11}
17	9.86688×10^{-12}	2.52814×10^{-12}	1.04238×10^{-11}
18	1.88865×10^{-12}	5.25549×10^{-13}	2.22971×10^{-12}
19	4.48225×10^{-13}	1.09251×10^{-13}	4.76213×10^{-13}
20	8.66078×10^{-14}	2.2711×10^{-14}	1.01567×10^{-13}
30	1.62883×10^{-20}	3.42259×10^{-21}	1.87463×10^{-20}
40	2.8606×10^{-27}	5.1579×10^{-28}	3.26214×10^{-27}
50	4.84578×10^{-34}	7.77304×10^{-35}	5.49637×10^{-34}
60	8.02836×10^{-41}	1.17141×10^{-41}	9.07371×10^{-41}
70	1.31012×10^{-47}	1.76534×10^{-48}	1.47699×10^{-47}
80	2.1147×10^{-54}	2.66039×10^{-55}	2.37953×10^{-54}
90	3.38513×10^{-61}	4.00926×10^{-62}	3.80352×10^{-61}
100	5.38369×10^{-68}	6.04202×10^{-69}	6.04202×10^{-68}
200	4.62423×10^{-136}	3.6506×10^{-137}	5.16273×10^{-136}
300	3.4278×10^{-204}	2.2057×10^{-205}	3.82039×10^{-204}

A. Grunbaum [1], Calvez–Vilbe [6]. They all require additional assumptions on the periodic functions $f(\theta)$ periodic on $[0, 2\pi]$, like $f(\pi) = f(-\pi) = 0$, and, moreover, the extremality inequalities are always sharp; the equality is never achieved. Consequently, the extremal function – the analogue of the Gaussian – is never defined.

This is one of the more recent attempts on the discrete version of the uncertainty principle for the discrete case, due to A. Grunbaum [2]:

$$
\begin{aligned}
&4 \sum \sin^2\left(\frac{2\pi j}{N}]\right) |a_j|^2 \sum \sin^2\left(\frac{2\pi k}{N}\right) |a_k|^2 \\
&\geq \frac{1}{4}\left(\sum_{j=0}^{N-1}\left(\sin\left(\frac{2\pi j}{N}\right) - \sin\left(\frac{2\pi(j+1)}{N}\right)\right)(a_j \bar{a}_{j+1} + a_{j+1}\bar{a}_j)\right)^2 \\
&-\frac{1}{4}\left(\sum_{j=0}^{N-1}\left(\sin\left(\frac{2\pi j}{N}\right) + \sin\left(\frac{2\pi(j+1)}{N}\right)\right)(a_j \bar{a}_{j+1} - a_{j+1}\bar{a}_j)\right)^2
\end{aligned}
$$

The discrete analogue of the uncertainty principle presented in this paper very well matches the classical version, but, more importantly, it gives an ability to conduct large scale numerical experiments in the search for various similar extremality inequalities that are hard to conduct in the continuous case. A very tempting target here is the search for general inequalities of the Heisenberg–Weyl style for general classes of weights, and particularly for higher order moments.

Attempts to generalize the uncertainty principle (or Heisenberg–Weyl inequality) to moments higher than two are the subject of active study in the last 50 years, since the fundamental paper of J.I. Hirschman (1957) [14]. For more recent results see Folland, Sitaram [11]. The general form of such generalizations is:

$$\int_{-\infty}^{\infty} |x|^{\alpha} |f(x)|^2 \mathrm{d}x + \int_{-\infty}^{\infty} |x|^{\alpha} |\hat{f}(x)|^2 \mathrm{d}x \geq K_{\alpha} \cdot \|f\|^2$$

for a constant $K_{\alpha} > 0$. There is also a similar nontrivial inequality with $\log(x)$ replacing $|x|^{\alpha}$ belonging to Beckner [4, 5]. We will call this logarithmic case the case of $\alpha = 0$. In that case, Beckner proved that $K_0 = \psi(1/4) - \ln \pi$, but he also proved that the inequality is sharp – the equality is never achieved in the $L^2(\mathbf{R})$ space of functions, and thus there is no natural extremal function in this case. The precise value of the constant K_{α} is unknown for any other $\alpha \neq 0, 2$. All known bounds on K_{α} in these cases are known to be imprecise.

In the discrete case, one can take analogously the weight w to be $w_{\alpha}(i) = |i|^{\alpha}$. It turned out that in these cases there is a natural scaling of the lowest eigenvalue $\lambda_{1,\alpha}$ of the matrix $M_{w_{\alpha}}$ in n:

$$\lambda_{1,\alpha} \sim n^{\frac{\alpha-1}{2}} \cdot k_{\alpha}.$$

Here $\lambda_{1,\alpha}$ represents naturally the low bound in the discrete analogue of uncertainty principle with the weight $|x|^{\alpha}$, and thus the constant k_{α} is an analogue of yet unknown continuous constant K_{α}. It turns out that for *even* integers α, the convergence of $\lambda_{1,\alpha}$ to $n^{\frac{\alpha-1}{2}} \cdot k_{\alpha}$ is geometric in n, giving a good indication that the extremal function does exists. The arithmetic nature of the constant k_{α} is a mystery even in the simplest but most interesting nontrivial case of $\alpha = 4$.

For the quartic weight $w_4(i) = i^4$, the constant k_4 is numerically

$$k_4 = 0.017689769692136367634471484772200089760776281782338073296581121608...$$

Looking at limited degrees and heights, the standard technique of finding linear relations did not show any expression for k_4 as a multiplicative combination of rational powers of classical numbers such as $2, 3, 5, \pi, \Gamma(1/3), \Gamma(1/4), \gamma, e^{\gamma}, \zeta(3)$. To get a firm handle on the important "continuous case" constant K_4, one needs to understand its discrete incarnation k_4 and the corresponding extremal function arising from the bound state of M_{w_4}.

9.1 The Dickenson–Steiglitz Basis as Derived from Variational Principles

The general form of a variational principle presented above, using a matrix M, derived from symmetric weights w, is sufficient to subsume the Dickenson–Steiglitz basis. The weights in this case arise from the simplest positive periodic sequence. If one defines

$$w_i = 4 \cdot \sin^2 \frac{\pi i}{n},$$

then the matrix M_w resulting from this choice of weights is proportional to the standard Harper (Dickenson–Steiglitz) matrix.

It is satisfying that the general formulation of the variational principle is capable of naturally describing an often favored eigenvector basis for the Fourier matrix. The asymptotics of the lowest eigenvalue – the one which corresponds to the discrete analog of the Gaussian/bound state of the discrete Fourier transform – gives a good insight into the scaling of the eigenstates.

While in the case of polynomial weights, the scaling of the lowest eigenvalue is geometric, in the case of the Harper matrix, the scaling is sublinear. Using the notation for the weight w above, the smallest eigenvalue λ_w for the matrix M_w corresponding to the Harper (Dickenson–Steiglitz) matrix, scales linearly with n and asymptotically very slowly approaches 2π. For example, for $n = 1,000$, numerically one obtains only

$$\frac{1}{2} n \cdot \lambda_w = 3.13912\ldots$$

This stands in the striking contrast with the scaling of λ_w for weights $w(i)$ that are powers of i.

We would like to point out that a different form of the uncertainty principle involving discrete Fourier eigenvectors, concentrated at subsets S of $\{0, \ldots, n-1\}$, also can be reduced to the weighted framework. In this case, the weights $w(i)$ arise from characteristic functions of subsets S.

References

1. Alberto Grünbaum, F.: Trying to beat Heisenberg. Lect. Notes Pure Appl. Math. **39**, 657–665 (1990)
2. Alberto Grünbaum, F.: The Heisenbergy inequality for the discrete Fourier transform. Appl. Comput. Harmon. Anal. **15**(2), 163–167 (2003)
3. Auslander, L., Tolimieri, R.: Is computing with the finite Fourier transform pure or applied mathematics? Bull. Amer. Math. Soc. (N.S.) **1**(6), 847–897 (1979)
4. Beckner, W.: Pitt's inequality and the uncertainty principle. Proc. Amer. Math. Soc. **123**, 1897–1905 (1995)
5. Beckner, W.: Pitt's inequality with sharp error estimates. arXiv:math/0701939v2 (2007)

6. Calvez, L., Vilbe, P.: On the uncertainty principle in discrete signals. IEEE Trans. Circuits Syst. II: Analog Digital Signal Process. **39**(6), 394–395 (1992)
7. Carlitz, L.: Some cyclotomic matrices. Acta Arithmetica **5**, 293–308 (1959)
8. Chudnovsky, D.V., Chudnovsky, G.V.: Some remarks on theta functions and S-matrices. In: Classical and quantum models and arithmetic problems, Lecture notes in Pure and Applied Mathematics, 117–214 Marcel Dekker, New York (1984)
9. Chudnovsky, D.V., Chudnovsky, G.V., Morgan, T.: Sparse Sets in Time and Frequency related to Diophantine Problems and Integrable Systems. In: this volume, Springer, New York (2009)
10. Dickinson, B.W., Steiglitz, K.: Eigenvalues and eigenvectors of the discrete Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **30**(1), 25–31 (1982)
11. Folland, G., Sitaram, A.: The uncertainty principle: A mathematical survey. J. Fourier Anal. Appl. **3**(3), 207–238 (1997)
12. Gutzwiller, M.: Physics and arithmetic chaos in the Fourier transform. In: The Mathematical Beauty of Physics, A Memorial Volume for Claude Itzykson, World Scientific, Singapore, 258–280 (1997)
13. Harper, P.G.: Single band motion of conduction electrons in a uniform magnetic field. Proc. Phys. Soc. A **68**, 874–878 (1955)
14. Hirschman, J.I.: A note on entropy. Am. J. Math. **79**, 152–156 (1957)
15. Hofstadter Douglas, R.: Energy levels and wave functions of Bloch electrons in rational and irrational magnetic fields. Phys. Rev. B **14**(6), 2239–2249 (1976)
16. Ishii, R., Furukawa, K.: The uncertainty principle in discrete signals. IEEE Trans. Circuits Syst. **33**(10), 1032–1034 (1986)
17. Matveev, V.B.: Intertwining relations between the Fourier transform and discrete Fourier transform, the related functional identities and beyond. Inverse Probl. **17**(4), 633–657 (2001)
18. McClellan, J., Parks, T.: Eigenvalue and eigenvector decomposition of the discrete Fourier transform. IEEE Trans. Audio Electroacoustics **20**(1), 66–74 (1972)
19. Mehta, M.L.: Eigenvalues and eigenvectors of the finite Fourier transform. J. Math. Phys. **28**(4), 781–785 (1987)
20. Morton, P.: On the eigenvectors of Schur's matrix. J. Number Theor. **12**, 122–127 (1980)
21. Mumford, D.: Tata lectures on theta I, vol. 28. Birkhäuser Boston Inc., Boston, MA (1983)
22. Ozaktas, H.M., Kutay, M.A., Zalevsky, Z.: The fractional Fourier transform with applications in optics and signal processing. Wiley, Chichester, New York (2001)
23. Prestin, J., Quak, E., Rauhut, H., Selig, K.: The connection of uncertainty principles for functions on the circle and on the real line. J. Fourier Anal. Appl. **9**(4), 387–409 (2003)
24. Schur, I.: Über die Gaussschen summen. J. Gött. Nachr. pp. 147–153 (1921)
25. Weyl, H.: The Theory of Groups and Quantum Mechanics. Dover, New York (1950)
26. Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. Cambridge University Press, Cambridge (1927)
27. Yarlagadda, R.: A note on the eigenvectors of DFT matrices. IEEE Trans. Acoust. Speech Signal Process. **25**(6), 586–589 (1977)

Sparse Sets in Time and Frequency Related to Diophantine Problems and Integrable Systems

D.V. Chudnovsky, G.V. Chudnovsky, and T. Morgan

Summary Reconstruction of signals and their Fourier transforms lead to the theory of prolate functions, developed by Slepian and Pollack. We look at prior contributions by Szegö to these problems. We present a unified framework for solutions of Szegö like problems for signals supported by an arbitrary union of intervals, using the techniques of Garnier isomonodromy equations. New classes of completely integrable equations and Darboux–Backlund transformations that arise from this framework are similar to the problems encountered in transcendental number theory. A particular example for the Hilbert matrix is studied in detail.

Keywords Hilbert matrix · Hankel matrices · Padé approximations

Mathematics Subject Classifications (2010). Primary 11C20, 15B05, 65F15

The Hilbert Matrix and Related Operators

In 1936, Szegö considered the Hilbert matrix as a particular example of a general extremality problem, the simultaneous concentration of polynomials on intervals and arcs in the complex plane [20].

In the modern era, the extremality study of simultaneously space and frequency band limited signals was initiated by Slepian and Pollack in 1960's [18,19] who discovered a differential operator, which commuted with integral operators describing such simultaneously space and frequency band limited signals.

The connection of their studies with the Szegö problem was mentioned in their paper [7].

D.V. Chudnovsky, G.V. Chudnovsky, and T. Morgan
Polytechnic Institute of NYU, IMAS, 6 MetroTech Center, Brooklyn, NY 11201, USA
e-mail: david@imas.poly.edu; gregory@imas.poly.edu; tmorgan@acm.org

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_5,

It is important to note that the first occurrence of commuting integral and differential operators in the context of extremality studies seems to belong to M. Rosenblum [14, 15]. Rosenblum represented the infinite Hilbert matrix as an equivalent integral operator. For this, Rosenblum used an orthogonal basis consisting of functions $e^{-x/2} \cdot L_n(x)$ (where $L_n(x)$ is a Laguerre polynomial). Rosenblum determined the spectrum of this operator using a commuting differential operator, which was explicitly constructed in terms of the generalized hypergeometric equation. It seems that Rosenblum's contribution is a seminal one, which precedes all other similar "miracles" as Slepian put it in 1961 paper [19]. It is especially interesting in practical applications because a problem which is well known to be ill conditioned in its original formulation is converted to a much better conditioned form, which is also numerically more efficient to solve. We discovered the papers of Rosenblum after much study on the finite sections of the Hilbert matrix, and its commuting differential operator which was completed in 2005. It should be noted that the operator which was found by the study of the finite sections of the Hilbert matrix, when extended to the full Hilbert matrix, differs entirely from the Rosenblum operator due to the difference in the choice of base. While Rosenblum uses a basis suited to the Whittaker function (hypergeometric function ${}_0F_2$), the basis used in this study of the Hilbert matrix and its finite sections is the polynomial basis. The choice of the polynomial basis results in the ${}_2F_1$ hypergeometric function. The two bases and the corresponding integral operators are related by an integral transform with an exponential kernel. An alternative representation of matrices commuting with the Hilbert matrix was later given by Sawyer and Grunbaum [1, 16].

The Hilbert matrix analysis leads to a more general case of minor of finite Fourier transform matrix, arising from arbitrary set C of intervals. These are the objects that naturally occur in the study of signal reconstruction. Investigation of spectral properties of these finite matrices led to a class of number-theoretic subjects: Padé approximations to classical transcendental functions (linear combinations of logarithms and q-logarithms) and isomonodromy transformations. It provides a better look at those ill-conditioned matrices and their deep analytical properties.

General Prolate Functions

In the most general (n-dimensional) case, the story of functions band-limited in space and/or frequency looks as follows. For a bounded subset Ω in n-dimensional space, we are looking at the subspace R_Ω of $L_2(R^n)$ of functions $f(x)$ that can be represented in the form:

$$f(x) = \left(\frac{1}{2\pi}\right)^{n/2} \cdot \int \cdots \int_{\Omega} e^{-i(x.u)} \cdot s(u)\, du \tag{1}$$

for a standard scalar product $(x.u)$ in R^n. Thus, R_Ω is the space of functions $\{f(x)\}$, whose frequency is band-limited to the set Ω. Let P_Ω be the projection

from $L_2(R^n)$ to R_Ω – i.e. P_Ω is the operator of band-limiting to Ω. We also introduce "space-limited" functions and the corresponding projections. If M is a subset of R^n, we have a very simple operator of restriction of functions to M. This operator, D_M maps functions $f(x)$ from $L_2(R^n)$ to the following one:

$$D_M f(x) = f(x) \cdot \chi_M(x), \tag{2}$$

where $\chi_M(x)$ is the characteristic function of the set M. The combined projection operator $P_\Omega D_M$ is an isomorphism $R_\Omega \to R_\Omega$. There is a natural band-limited basis of R_Ω associated with this operator. This system of functions, $\{\psi_k\}$ satisfies the following properties (commonly referred to as double-orthogonality properties).

1. System $\{\psi_k\}$ is complete in R_Ω.
2. System $\{\psi_k\}$ is orthogonal in $L_2(R^n)$.
3. System $\{\psi_k\}$ is orthogonal on $L_2(M)$ in the following way:

$$\int \cdots \int_M \psi_k(x) \cdot \psi_l(x) \mathrm{d}x = 0; \; k \neq l. \tag{3}$$

4. System $\{\psi_k\}$ is the complete system of eigenfunctions of the "double projection" operator $P_\Omega D_M$:

$$P_\Omega D_M \psi_k = \lambda_k \cdot \psi_k. \tag{4}$$

The "double projection" operator $P_\Omega D_M$ has actually a very natural integral representation:

$$P_\Omega D_M(f(x)) = \int \cdots \int_M K_\Omega(x-y) \cdot f(y) \, \mathrm{d}y, \tag{5}$$

where the kernel $K_\Omega(x-y)$ is defined in terms of the spectral support Ω:

$$K_\Omega(x-y) = \left(\frac{1}{2\pi}\right)^{n/2} \cdot \int \cdots \int_\Omega \mathrm{e}^{-\mathrm{i}((x-y).u)} \, \mathrm{d}u. \tag{6}$$

Generally speaking, this integral equation is incredibly complex, as it ties together space and frequency. Of course, this is just a "continuous" version of the symmetric square

$$C \cdot C^H \tag{7}$$

of the general n by m minor of the discrete Fourier matrix

$$C = \left(\mathrm{e}^{2\pi\sqrt{-1}x_i y_j/N}\right)_{i=1,j=1}^{i=n,j=m} \tag{8}$$

However, in the incredibly "lucky accident", as its discoverers put it, in the cases of the Ω and M being balls (or intervals in the case of $n = 1$) the problem turned out to be reducible to a well studied classical one.

In the series of remarkable papers started in 1961, D. Slepian, H. Landau, and H. Pollak found and developed an extraordinary detailed theory of eigenfunction of the projection operator $P_\Omega D_M$ for intervals

$$\Omega = [-W, W];\ \ M = [-T, T];\ n = 1. \tag{9}$$

They called this theory a theory of prolate functions because it turned out that the eigenfunctions ψ_k are actually solutions (eigenfunctions) of a classical prolate spheroidal wave equation. Specifically, in the 1-dimensional case when $\Omega = [-W, W]$ and $M = [-T, T]$, if we put $x = Tz$, and $c = W \cdot T$, we get the following representation of the $P_\Omega D_M(f(x))$ integral operator:

$$P_\Omega D_M(\psi(z)) = \int_{-1}^{1} \frac{\sin c(z-u)}{\pi(z-u)} \cdot \psi(u)\,\mathrm{d}u. \tag{10}$$

This operator commutes with the prolate spheroidal linear differential operator:

$$P_z = \frac{\mathrm{d}}{\mathrm{d}z}\left(1-z^2\right)\frac{\mathrm{d}}{\mathrm{d}z} - c^2 z^2. \tag{11}$$

This means that the eigenfunctions $\psi_k(z)$ are actually eigenfunctions of the differential operator P_z, or, in other words, are solutions of the prolate spheroidal wave equation:

$$(1-z^2)\cdot\psi_k''(z) - 2z\cdot\psi_k'(z) + \left(\chi_k - c^2 z^2\right)\cdot\psi_k(z) = 0. \tag{12}$$

This representation is quite significant for many reasons. One of them, rather important in practical applications, is that of inherit ill-conditioning of the integral operator $P_\Omega D_M$ that have almost all eigenvalues λ_k very close to degenerate; almost all of them cluster at $\lambda = 0$ and $\lambda = 1$ (exponentially close to these points). Eigenvalues χ_k are, on the other hand, very well separated. In the discrete case, the well conditioning of the prolate operator commuting with the standard projection operator allows for the numeric computation of the prolate eigenforms (sequences).

General Prolate Functions and Commuting Differential Operators

Clearly having the second, better conditioned, linear problem that defines the same eigenfunctions is very important both for theoretical and applied development of the theory of general space and frequency limited functions. For this, one needs the second (say, differential) operator commuting with the main integral one. For example, the best quantative versions of the uncertainty principle in the case of an interval in space and the interval in frequency were developed by Slepian, Landau, Pollak

using prolate spheroidal wave operators. The theory was generalized by them virtually unchanged to the case of n-dimensional balls, where the original sinc kernel $K_\Omega(x-y)$:

$$\frac{\sin(c\cdot(x-y))}{x-y} \tag{13}$$

is replaced by the Bessel-like kernel

$$J_N(cxy)\sqrt{xy}. \tag{14}$$

There were various attempts in 1960s–1980s to extend this commuting "miracle" to other cases. The new kernel found was the Airy kernel $Ai(x+y)$.

Work of J. Morrison in the 1960s [12] and Grunbaum showed that the cases of commuting differential operators (or sparse matrices in the discrete cases) are basically reduced to the known ones. Of course, these days we call the cases of nontrivial commuting differential or integral operators completely integrable. These studies started with Burnchall-Chaundy in 1920s, and are known and studied under various names (Lax pairs, etc.). It is generally understood that such cases are very rare and always somehow fall into one or another class of well-studied classes with very rich internal structures.

Modern approach to operators with kernel similar to those mentioned above, including Airy-kernel use methods of isospectral and isomonodromy deformation equation applied to the Fredholm operators – see among others papers [2, 9, 21, 22].

Szegö Problem and Concentrated Polynomials

In fact, all this business started with Szegö. One can trace the field of dually concentrated functions to a remarkable paper Szegö [20]. There Szegö asks the following question. Let C_1, C_2 be Jordan curves in the complex plane, bounding regions E_1, E_2. What is the maximum value of the following ratio:

$$M_n(C_1,C_2)=\frac{\int\limits_{C_1}|P(z)|^2\,dz}{\int\limits_{C_2}|P(z)|^2\,dz} \tag{15}$$

among all polynomials $P(z)$ of degree n?

Then $M_n(C_1,C_2)$ can be interpreted as the "energy ratio", and $P(z)$ as a polynomial having its energy most concentrated in C_1 at the expense of its energy in C_2. Using normalized orthogonal polynomials $q_\nu(z)$ on C_2, the ratio becomes the maximum of the quadratic form:

$$M_n=\max\sum_{\mu,\nu=0}^{n}x_\mu\bar{x}_\nu(l)^{-1}\int_{C_1}q_\mu(z)\overline{q_\nu(z)}\,|\mathrm{d}z| \tag{16}$$

provided that the variables x_ν satisfy the condition $\sum_{\nu=0}^{n} |x_\nu|^2 \leq 1$) (or $= 1$). In other words, M_n is the greatest characteristic value of the Hermitian form on the right side. Here l denotes the length of C_1.

Szegö made a very general prediction of the asymptotics of $M_n(C_1, C_2)$ as $n \to \infty$ in terms of the maximum of the Green's function of C_2 on the curve C_1.

It has an easier interpretation in the case of simply-connected C_2, where we map the exterior of C_2 to the exterior of the unit circle and are looking at the greatest distance, ρ, of the image of C_1 under that mapping, to the unit circle. Then asymptotically $\sqrt{M_n^{1/n}}$ is ρ.

A particular case that was treated in great detail is that of the interval C_1 and a circle C_2. Specifically, let C_1 is an interval $(0, 1)$, and C_2 is the unit circle. One gets then the quadratic form:

$$\int_0^1 (x_0 + x_1 t + x_2 t^2 + \cdots + x_n t^n)^2 \mathrm{d}t = \sum_{\mu,\nu=0}^{n} \frac{x_\mu x_\nu}{\mu + \nu + 1} \tag{17}$$

and Szegö obtains for the smallest characteristic value

$$\lambda_n \cong 2^{15/4} \pi^{3/2} n^{1/2} \cdot (2^{1/2} - 1)^{4n+4}. \tag{18}$$

There is another similar asymptotics in the same paper [20]. Let λ_n be the smallest characteristic value of the quadratic form

$$\frac{1}{2} \int_{-1}^{1} (x_0 + x_1 t + x_2 t^2 + \cdots + x_n t^n)^2 \mathrm{d}t = \sum_{\mu,\nu=0}^{n\,*} \frac{x_\mu x_\nu}{\mu + \nu + 1}, \tag{19}$$

where $\sum^*$ indicates that the summation is extended only over even values of $\mu + \nu$. Then

$$\lambda_n \cong 2^{9/4} \pi^{3/2} n^{1/2} (2^{1/2} - 1)^{2n+3}. \tag{20}$$

The matrix occurring in the first problem is known as Hilbert matrix. Since its largest eigenvalue is very close to π, this matrix is notoriously ill-conditioned.

For example, for $n = 100$, the smallest eigenvalue of Hilbert matrix is

$$1.71 \cdot 10^{-152} \tag{21}$$

(as accurately predicted by Szegö formula).

The Hilbert matrix is a favorite subject of study, as it provides an excellent prototype of numerical instability. While the inverse of Hilbert matrix is well-known analytically, not much was known of analytic properties of its eigenfunction and eigenvalues.

Hilbert Matrix and a Commuting Differential Operator

The eigenvectors of the Hilbert matrix

$$H[n] = \left(\frac{1}{i+j+1}\right)_{i,j,=0}^{n} \tag{22}$$

can be described using a "very natural" 4-th order Fuchsian linear differential operator:

$$\begin{aligned} L_n^{\{4\}} = {} & x^3 \cdot (x-1)^2 \cdot \frac{\mathrm{d}^4}{\mathrm{d}x^4} + 2 \cdot x^2 \cdot (5x-3) \cdot (x-1) \cdot \frac{\mathrm{d}^3}{\mathrm{d}x^3} \\ & + x \cdot (6 - (n^2+2n)(x-1)^2 + 4x(6x-7)) \cdot \frac{\mathrm{d}^2}{\mathrm{d}x^2} \\ & + (-n(n+2) + 4(-2+n(n+2))x - 3(-4+n(n+2))x^2) \cdot \frac{\mathrm{d}^1}{\mathrm{d}x^1} \\ & + (C - n(n+2)x) \cdot \frac{\mathrm{d}^0}{\mathrm{d}x^0} \end{aligned}$$

The relationship of this operator $L_n^{\{4\}}$ with eigenvectors of $H[n]$ is rather straightforward. The linear differential equation

$$L_n^{\{4\}} Q = 0 \tag{23}$$

has a polynomial (in x) solution $Q(x)$ of degree n when and only when the vector of coefficients of $Q(x)$ is the eigenvector of $H[n]$:

$$H[n] \cdot v = \lambda \cdot v, \tag{24}$$

where $v = (v_i)_{i=0}^n$ and $Q(x) = \sum_{i=0}^n v_i \cdot x^i$. The relationship between the commuting eigenvalues – C (the accessory parameter of the Fuchsian l.d.e.) and λ (the matrix eigenvalue) reveals the monodromy of $L_n^{\{4\}}$ and the role of Padé approximations. Namely, if $Q_n(x)$ is a polynomial solution of $L_n^{\{4\}} Q = 0$, then there are two solutions of this equation of the form:

$$f_1 = Q_n \cdot \log\left(\frac{x-1}{x}\right) + P_{n-1} \tag{25}$$

$$f_2 = \frac{Q_n^1}{x^{n+1}} \cdot \log(x-1) + \frac{P_{n-1}^1}{x^n} \tag{26}$$

with f_1 providing a rational (Padé-type) approximation to the logarithm $\log(1-1/x)$ at $x \to \infty$, and f_2 providing an approximation to the logarithm $\log(1-x)$ at

$x \to 0$. Then the eigenfunction condition, determining the original λ eigenvalue is the connectivity formula:

$$Q_n^1(x) = \lambda \cdot Q_n\left(\frac{1}{x}\right) x^n. \tag{27}$$

The Hilbert eigenvalue problem thus becomes an over-convergence Padé-like approximation problem: Find a polynomial $Q_n(x)$ of degree n such that the following linear form has an order $n+2$ of zero

$$\log\left(1-\frac{1}{x}\right)\cdot Q_n(x) - P_{n-1}(x) - \lambda \cdot Q_n\left(\frac{1}{x}\right) x^{-1} = O\left(x^{-n-2}\right) \tag{28}$$

at $x \to \infty$. (Here $P_{n-1}(x)$ is a simple function of $Q_n(x)$). If in the right hand side the order of approximation would have been a standard, $O(x^{-n-1})$, it would have been an ordinary Padé approximation problem with nothing more complicated than Legendre polynomials. The differential equation formulation achieves the following goal: replaces an ill-conditioned problem with the equivalent well-conditioned, and a dense matrix with a commuting sparse one, arising from Padé-like approximation problem.

Szegö Problem and Arbitrary Unions of Intervals

The most interesting and full of application case of the Szegö problem is that of sets C_1, C_2 which are the unions of intervals. In 1978 Slepian and Gilbert [7] tried to find differential operators commuting with the concentration problem. They found that there were just 2 such cases: of C_1, C_2 single intervals with C_2 centrally positioned inside C_1 (i.e., $[-a, a]$ inside $[-1, 1]$ for $a < 1$), and of C_1, C_2 adjacent (i.e., $[-1, 1]$ and $[1, a]$ for $a > 1$).

While it is true that there is no commuting differential operators in the multiple interval cases both in Szegö problem, and in the general space/frequency bandlimiting problem, it does not mean that the problems are not completely integrable in some sense. In fact, these problems are reduced to a classical isomonodromy deformation problem, and to problems that were actively studied by the Chudnovskys in 1980 [3, 4].

The simplest way to enter this area is to start with a well-known definition of Padé approximation to a function $f(x)$, analytic at $x = 0$. The rational function $P_n(x)/Q_n(x)$ is called a (diagonal) Padé approximation to $f(x)$ (of order n) if $P_n(x)/Q_n(x)$ expanded at $x = 0$ matches first $2n+1$ terms of the series expansion of $f(x)$ at $x = 0$, or

$$Q_n(x) \cdot f(x) - P_n(x) = O(x^{2n+1}). \tag{29}$$

If the order is exactly x^{2n+1}, the Padé approximation is called a normal (or perfect, see K. Mahler's definition [11] in general multifunction case).

Szegö problem can be naturally formulated in terms of Padé approximation to the most general logarithmic function

$$f(x) = \sum_{i=1}^{m} w_i \log(1 - a_i x). \tag{30}$$

Only when $m = 2$ this is an explicitly solvable problem (Legendre polynomials) when Padé approximations are always normal.

It turns out the Szegö problem is equivalent to finding cases of non-normality for a fixed set $\{a_i\}_{i=1}^m$ of singularities, made from the ends of intervals comprising C_1 and C_2. Namely, when

$$C_1 = \bigcup_{i_1=1}^{d_1} (b_{i_1}, c_{i_1});\ C_2 = \bigcup_{i_2=1}^{d_2} (d_{i_2}, e_{i_2}), \tag{31}$$

we will associate weight $w(.)$ with the ends of these intervals:

$$w(b_{i_1}) = 1, w(c_{i_1}) = -1, w(d_{i_2}) = \lambda, w(e_{i_2}) = -\lambda. \tag{32}$$

If $\{a_i\}_{i=1}^m$ is set of these ends, and

$$f(x) = \sum_{i=1}^{m} w(a_i) \log(1 - a_i x), \tag{33}$$

then λ is an eigenvalue in $M_n(C_1, C_2)$ problem if and only if the Padé approximation $P_n(x)/Q_n(x)$ of order n to $f(x)$ is non-normal (and the polynomial $Q_n(x)$ is then the most concentrated polynomial).

This characterization of Szegö problem immediately associates with the $M_n(C_1, C_2)$ problem a Fuchsian linear differential equation of the second order. Its fundamental system of solutions is simply

$$Q_n(x) \text{ and } Q_n(x) \cdot f(x) - P_n(x). \tag{34}$$

This Fuchsian equation exists naturally over the moduli space of intervals (essentially two copies of the hyperelliptic moduli space).

Garnier Isomonodromy Deformation Equations

The Fuchsian equations for the Szegö problem start with the set $\{a_i\}_{i=1}^m$ of ends of intervals. We can always assume, without loss of generality, that

$$a_1 = 0,\ a_2 = 1. \tag{35}$$

Thus, we have m regular (logarithmic) singularities. We also have $m-2$ apparent singularities $\{b_k\}_{k=1}^{m-2}$ – these are the singularities of the equation, but not of any solutions.

The apparent singularities correspond to spurious zeros of $Q_n(x)$ outside of C_1, C_2. The Fuchsian equation has the following form:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}Y + \left(\sum_{i=1}^{m}\frac{1}{x-a_i} - \sum_{k=1}^{m-2}\frac{1}{x-b_k}\right)\frac{\mathrm{d}}{\mathrm{d}x}Y + \frac{p_{2m-4}(x)}{\prod(x-a_i)\prod(x-b_k)}Y = 0 \tag{36}$$

There are $m-2$ free parameters $\{c_j\}$ in $p_{2m-4}(x)$, known as accessory parameters. They are usually defined as residues

$$c_j = Res_{x=b_j}\frac{p_{2m-4}(x)}{\prod(x-a_i)\prod(x-b_k)}. \tag{37}$$

The main point of this equation is that its monodromy group depends only on the number m of logarithmic singularities (of $f(x)$), and thus the dependence in this equation on actual position of singularities $\{a_i\}$ defines isomonodromy deformation equation. The particular system of isomonodromy deformation system for 2nd order Fuchsian equations is called Garnier system, after Garnier papers of 1912–1919 [6, 13].

The Garnier system can be written in the Hamiltonian form describing the dependencies of all parameters (apparent singularities and accessory parameters) on actual singularities a_j, [13]:

$$\frac{\partial b_k}{\partial a_j} = \frac{\partial K_j}{\partial c_k}; \tag{38}$$

$$\frac{\partial c_k}{\partial a_j} = -\frac{\partial K_j}{\partial b_k}. \tag{39}$$

The Hamiltonians K_j have the following explicit form (in our particular case of logarithmic singularities):

$$T(x) = \prod(x-a_i);\ L(x) = \prod(x-b_i) \tag{40}$$

$$K_j = -\frac{L(a_j)}{T'(a_j)}\left(\sum_{l=1}^{m-2}\frac{T(b_l)}{L'(b_l)(b_l-a_j)}\left(c_l^2 + c_l\sum_{i=1}^{m-2}\frac{\delta_{ij}}{b_l-a_i}\right) + n(n+1)\right). \tag{41}$$

Here K_i are natural parameters of the Fuchsian equation:

$$K_i = Res_{x=a_i}\frac{p_{2m-4}(x)}{\prod(x-a_i)\prod(x-b_k)}, \tag{42}$$

and $j = 2, \ldots, m$.

In the case of $m = 3$, the Garnier system is the celebrated Painlevé VI equation. Isomonodromy deformation systems like Garnier posses many special symmetry properties in addition to having $m-2$ commuting Hamiltonians (K_j). In particular, they posses birational transformations, known as Darboux–Backlund or Schelsinger transformations.

In our particular case, these are explicit nonlinear algebraic transformations that relate the parameters a_i, b_j, c_k for a given n to the similar parameters for $n+1$ (or $n-1$). These are essentially nonlinear recurrences on such parameters (as a function of the degree n of the polynomial $Q_n(x)$). These recurrence relations allow for a fine analysis of the eigenvalue and eigenfunction behavior and additional asymptotic terms. It also allows us to have a very interesting view on the global behavior of the eigenvalues as a function of singularity positions $\{a_i\}$.

There is another important consequence of the Fuchsian equation associated with the solution $Q_n(z)$ of the $M_n(C_1, C_2)$ problem. It shows that the zeros of $Q_n(z)$ on the complex plane are governed by Heune-Stiltjes theory of zeros of polynomial solutions of Fuchsian equations as electrostatic particles arising from the minimization of the energy functional

$$\prod_{i \neq j} |z_i - z_j| \cdot \prod_{i,k} |z_i - a_k|^2 \cdot \prod_{i,l} |z_i - b_l|^{-2}. \tag{43}$$

Here $\{z_j\}_{j=1}^n$ are roots of $Q_n(z) = 0$. Their location shows the true concentration patterns for different geometries of C_1, C_2.

Arbitrary union of intervals can be handled this way; including the intervals with complex ends. There the sets C are defined as minimal capacity set in the complex plane containing the end points. One can actually visualize these sets by looking at the accumulation of zeros of $Q_n(z)$ in the complex plane. Circles can be added to sets C (like in the Hilbert matrix case).

Figure 1 shows the location of the zeros of the generating polynomial for eigenvectors 1,20,40,60,80 and 100 of the Hilbert matrix of dimension 100, when the eigenvectors are arranged in decreasing order of their corresponding eigenvalue. Notice how the zeros of the generating polynomial are located near the unit circle in the complex plane and along the real unit interval. The lower number eigenvectors have more zeros near the unit circle, the higher number eigenvectors have more zeros in the unit interval. This pattern of appearance of the zeros of the generating polynomial provides a visualization of the underlying structure of differential equations which govern the eigenvectors of the Hilbert matrix.

Explicit Expressions for the Non-Linear Darboux Transform

The Darboux-Schlesinger-Garnier-Backlund [13, 17] (or simply Darboux) general transformations show the birational algebraic transformations (that are also canonical with respect to the Hamiltonian structure) between solutions of Fuchsian differential equations with the same monodromy group and same real singularities

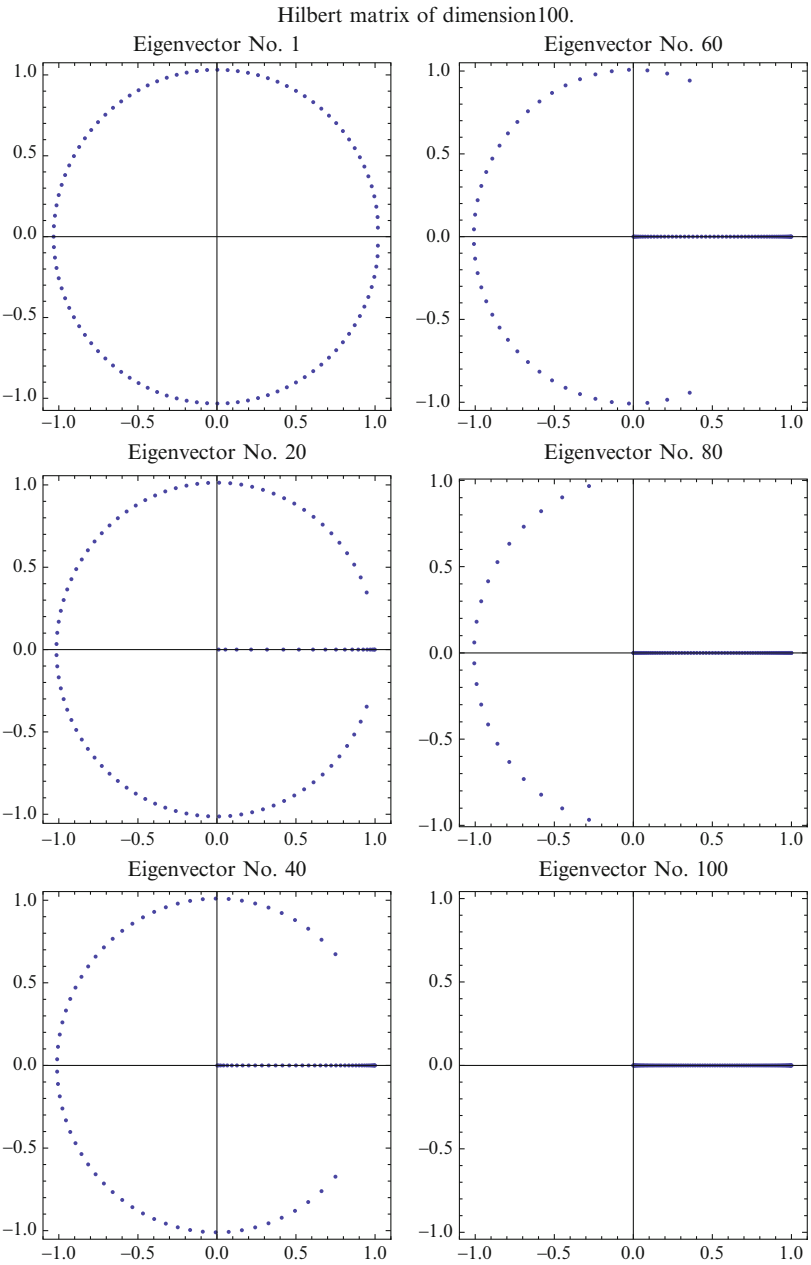


Figure 1 Location of zeros of generating polynomials for selected eigenvectors

when local exponents of the equation are simply transformed by exchanges or addition/subtraction of integers. These relations were called by Riemann contiguous relations (after Euler and Gauss formulae in hypergeometric case).

Derivation of these formulas is tedious but elementary and is reduced to solving linear equations. It requires only matching linear forms in solutions of Fuchsian equations with rational function coefficients to correct orders of singularities. The canonical Hamiltonian representation of Garnier system requires a canonical transformation

$$s_i = \frac{a_i}{a_i - 1}, q_i = \frac{a_i L(a_i)}{T'(a_i)},\ p_i = \sum_{k=1}^{m-2} \frac{(1-a_i)T(b_k)c_k}{b_k(b_k-1)(b_k-a_i)L'(b_k)}. \tag{44}$$

We will present Darboux transformation in the most general form, where the local exponents at all singularities, real and apparent, have the most general matrix form

$$\begin{pmatrix} x=0 & x=1 & x=\infty & x=a_i & x=b_k \\ 0 & 0 & \alpha & 0 & 0 \\ \kappa_0 & \kappa_1 & \alpha+\kappa_\infty & \theta_i & 2 \end{pmatrix} \tag{45}$$

Here $\alpha = -\frac{1}{2}(\kappa_0 + \kappa_1 + \kappa_\infty + \sum \theta_i - 1)$. Of course in our special case $\kappa_0 = \kappa_1 = \theta_i = 0$ and $\kappa_\infty = 2n + 1$. We describe these Darboux transforms in canonical variables $\mathbf{p}, \mathbf{q}, \mathbf{s}$ showing also the effect of transformations on exponents of the Fuchsian equations:

$$\kappa_1 \rightleftharpoons \kappa_0 :$$

$$Q_i = \frac{q_i}{s_i};\ p_i = s_i p_i;\ S_i = \frac{1}{s_i}$$

$$\kappa_1 \rightleftharpoons \kappa_\infty :$$

$$Q_i = \frac{q_i}{g_1 - 1};\ p_i = (g_1 - 1)(p_i - \alpha - \sum p_j q_j); S_i = \frac{s_i}{s_i - 1}$$

$$\kappa_\infty \to -\kappa_\infty :$$

$$Q_i = q_i;\ P_i = p_i;$$

$$\kappa_1 \to -\kappa_1 :$$

$$Q_i = q_i;\ P_i = p_i - \frac{\kappa_1}{g_1 - 1};$$

$$\kappa_0 \to -\kappa_0 :$$

$$Q_i = q_i;\ P_i = p_i - \frac{\kappa_0}{(g_s - 1)s_i};$$

$$\theta_j \to -\theta_j :$$

$$Q_i = q_i;\ P_j = p_j - \frac{\theta_j}{q_j};\ P_i = p_i (j \neq i)$$

$$(\kappa_0, \kappa_1, \kappa_\infty, \theta_1, \ldots, \theta_m) \to (-\kappa_0 + 1, -\kappa_1 + 1, -\kappa_\infty, -\theta_1, \ldots, -\theta_m) :$$

$$Q_i = \frac{s_i p_i (q_i p_i - \theta_i)}{(\alpha + \sum_j q_j p_j)(\alpha + \kappa_\infty + \sum_j q_j p_j)};\ Q_i P_i = -q_i p_i \tag{46}$$

In the formulas above,

$$g_1 = \sum_i q_i, \ g_s = \sum_i \frac{q_i}{s_i}. \tag{47}$$

In our particular case, we need a combination of transformations above giving us the "addition of orders to Padé approximation"

$$(\kappa_0, \kappa_1, \kappa_\infty, \theta_1, \ldots, \theta_m) \to (\kappa_0, \kappa_1, \kappa_\infty + 2, \theta_1, \ldots, \theta_m). \tag{48}$$

Note the heavy nonlinearity of these Darboux transformations. Basically, only in the "trivial" cases, reducible to hypergeometric functions, the recurrence relations connecting the n and $n+1$ cases are explicit. Otherwise, these complicated equations make the corresponding Padé approximations to have the heights of the these approximants to grow with n as C^{n^2}. In particular, this makes the Padé approximation to a linear combination of logarithms useless for diophantine approximation applications (unless it reduces to a single logarithm).

Darboux Transformation for $m = 3$ Case

These are the simplest formulas for $m = 3$ (Painleve VI) with one singularity $a_3 = t$; one apparent singularity λ and an accessory parameter μ. It shows the complexity of the transformation from n to $n-1$.

$$\mu' = \frac{(\lambda^3\mu^3-\lambda^2(t+1)\mu^3+n^2(t\mu+\mu+2n)+\lambda(\mu^3t-3\mu n^2))(\mu^4\lambda^6-2\mu^4(t+1)\lambda^5+}{(\mu^2\lambda^3-\mu(t\mu+\mu+2n)\lambda^2+(t\mu^2+2n\mu+n^2)\lambda+}$$
$$\frac{(\mu^4(t^2+4t+1)-6\mu^2n^2)\lambda^4-2\mu(t(t+1)\mu^3-4n^2(t+1)\mu-4n^3)\lambda^3+}{n^2(t-1))(\mu^2\lambda^3-\mu(t\mu+\mu+2n)\lambda^2+(t\mu^2+2nt\mu+n^2)\lambda-}$$
$$\frac{(t^2\mu^4-2n^2(t^2+5t+1)\mu^2-8n^3(t+1)\mu-3n^4)\lambda^2+2n^2(t(t+1)\mu^2+4nt\mu+n^2(t+1))\lambda+n^4(t-1)^2)}{n^2(t-1))(\mu^2\lambda^3-\mu(t\mu+\mu+2n)\lambda^2+(t\mu^2+2n(t+1)\mu+n^2)\lambda-n(tn+n+2\mu t))} \tag{49}$$

$$\lambda' = \frac{\lambda((\lambda-1)(n-\lambda\mu)^2-((\lambda-1)\lambda\mu^2-2(\lambda-1)n\mu+n^2)t)^2}{(\lambda-1)((\lambda-1)\lambda^2\mu^2+2(\lambda-1)\lambda n\mu-(3\lambda+1)n^2)(n-\lambda\mu)^2+}$$
$$\frac{}{(n^2-(\lambda-1)\lambda\mu^2)^2t^2-2(\lambda-1)((\lambda-1)\lambda^3\mu^4-\lambda(4\lambda-1)n^2\mu^2+4\lambda n^3\mu-n^4)t} \tag{50}$$

The poles of λ and μ give the location of eigenvalues for 2-interval sets. See [5] for the discussion of zeros and poles of accessory parameters and apparent singularities as eigenvalue problems.

Generalized Prolate Functions and Another Isomonodromy Problem

Szegö problem in the case of arbitrary unions of intervals was reduced to the Padé approximations and isomonodromy deformation (Garnier) equations. The same is true for the generalized prolate functions whenever sets Ω and M are unions of intervals. The eigenvalue problem there also leads to the non-normal Padé-like approximation problem; the Fuchsian differential equations, and Garnier isomonodromy deformation system of nonlinear differential equations. The main difference here is that the auxiliary function we are approximating is

$$f(x) = \sum w_i \cdot \log \frac{x - a_i}{x - \frac{1}{a_i}} \tag{51}$$

where a_i are points on the unit circle. The function $f(x)$ is approximated simultaneously at $x = 0$ and $x = \infty$:

$$Q(x) \cdot f(x) - P(x) = O(x^{-1});\ x \to \infty \tag{52}$$

$$Q(x) \cdot f(x) - P(x) + \lambda \cdot Q(x) = O(x^n);\ x \to 0. \tag{53}$$

Generalized Prolate Matrices

The most general solution of Szegö problem reduces to the analysis of the following matrix (corresponding to arbitrary weights)

$$M_N = \left(\sum_{\alpha=1}^{m} \frac{w_\alpha a_\alpha^{i+j+1}}{i + j + 1} \right)_{i,j=0}^{N} \tag{54}$$

and solving for

$$M_N \bar{v} = 0. \tag{55}$$

The case of two interval unions C_1 and C_2 can also be reduced to the (generalized) eigenvalue problem

$$A_{C_1} \bar{v} = \lambda A_{C2} \bar{v}, \tag{56}$$

where the A_C matrix has the form

$$A_C = \left(\sum_k \frac{c_k^{i+j+1} - b_k^{i+j+1}}{i + j + 1} \right)_{i,j=0}^{N} \tag{57}$$

when $C = \bigcup_k [b_k, c_k]$. The prolate kernel

$$\frac{\sin(ax)}{x} \tag{58}$$

leads to the following class of matrices for the arbitrary union of intervals

$$P_{C,N} = \left(\sum_{\alpha=1}^{m} \frac{a_\alpha^{i-j} - a_\alpha^{j-i}}{i-j} \right)_{i,j=0}^{N}. \tag{59}$$

Here we get the completely integrable case as well, involving only second order Fuchsian linear differential equations with real singularities at $x = a_\alpha$ and $x = a_\alpha^{-1}$. The alternative formulation of the problem involves 2 – point Padé approximations to

$$\sum_\alpha \log \frac{x - a_\alpha}{x - a_\alpha^{-1}} \tag{60}$$

at both $x = 0$ and $x = \infty$.

Eigenvalue Problems for Hankel Matrices and Fourth Order Differential Equations

The function $f(x)$ is called a symbol of the Hankel matrix $H = (h_{i+j+1})_{i,j=0}^{n-1}$ if the expansion of $f(x)$ at $x = \infty$ has the form:

$$f(x) = \sum_{m=0}^{\infty} h_m \cdot x^{-m-1}.$$

As we saw in examples of the preceding sections, one can construct second order differential equations that describes the generalized eigenvalue problem for Hankel matrices, whose symbols satisfy (inhomogeneous) first order linear differential equations over $\mathbf{C}(x)$. This does not work, however, for the standard Hankel eigenvalue problem. The reason for this is that the multiplication

$$Q(x) \cdot f(x)$$

correctly describes the action of the Hankel matrix with the symbol f, on the vector whose generating function is $Q(x)$, but the right hand side in the standard eigenvalue problem corresponds to the vector with the generating function of the form:

$$\lambda \frac{Q\left(\frac{1}{x}\right)}{x}$$

That is why one needs fourth order differential equation to describe eigenvectors of the Hilbert matrix. In this section, we will show how this construction can be generalized to a more general class of Hankel matrices.

Let $f = f(x)$ be a solution of the first order inhomogeneous linear differential equation, and $k = k(x)$ be the solution of the corresponding homogeneous equation:

$$f' = a \cdot f + b,$$
$$k' = a \cdot k$$

To express the eigenvalue problem for the Hankel matrix whose symbol corresponds to the expansion of $f(x)$ we need to consider both $f(x)$ and $f(\frac{1}{x})$, effectively expanding each at $x = \infty$ and $x = 0$, respectively. For this we introduce the auxiliary objects:

$$f^-(x) = f\left(\frac{1}{x}\right); k^-(x) = k\left(\frac{1}{x}\right); Q^-(x) = Q\left(\frac{1}{x}\right); \quad P^-(x) = P\left(\frac{1}{x}\right).$$

We need the direct sum of two second order linear differential equations, describing the approximations to $f(x)$ and $f(\frac{1}{x})$. The first equation has a familiar basis:

$$R(x) = Q(x) \cdot f(x) - P(x), \quad Q(x) \cdot k(x)$$

The second equation has as a basis a transformed set of functions, scaled by a new function $u = u(x)$:

$$u(x) \cdot (Q^-(x) \cdot f^-(x) - P^-(x)), \quad u(x) \cdot (Q^-(x) \cdot k^-(x)).$$

We use the same rational approximation $\frac{P}{Q}$ to f, but expanded in x and $\frac{1}{x}$, meaning identical expansions at $x = \infty$ and $x = 0$. We look at the fourth order linear differential equation, which is the direct sum of these two equations. Its basis of solutions is simply the union of these two sets of functions. We use two linear forms in these solutions that would have a zero of a high order at $x = \infty$ and at $x = 0$. To get a linear form that would give us a Hankel eigenvalue problem:

$$f(x) \cdot Q(x) - P(x) - \lambda \cdot \frac{Q^-(x)}{x},$$

and its counterpart, with x replaced by $\frac{1}{x}$, we need to build it from the following two linear forms that mix solutions from two linear differential equations:

$$f \cdot Q - P - \lambda \cdot (k^- Q^-) \cdot u$$
$$u \cdot (f^- \cdot Q^- - P^-) - \lambda^- \cdot (kQ)$$

For these linear forms to be equivalent under the transformation: $x \leftrightarrow \frac{1}{x}$, one needs the following conditions to be satisfied:

$$u(x) = \frac{k(x)}{x},$$
$$k(x) \cdot k^{-}(x) = k(x) \cdot k(\tfrac{1}{x}) = \text{const}.$$

It is the last condition that put the restriction on the class of functions $f(x)$ among all possible solutions of first order inhomogeneous linear differential equations that lead to the explicit second commuting, auxiliary, Hankel eigenvector problem. These conditions mean that k has the following form:

$$k(x) = x^{\omega} \cdot \prod_{\alpha \in A} \left(\frac{x-\alpha}{\alpha \cdot x - 1}\right)^{\nu_\alpha} \cdot \mathrm{e}^{r(x)},$$

where $A \subset \mathbf{C}$ such that $0 \notin A$ and for $\alpha \in A$, $\frac{1}{\alpha} \notin A$, and $r(\frac{1}{x}) = -r(x)$ for $r(x) \in \mathbf{C}(x)$.

The fourth order linear differential equations that arises from this choice of $u(x)$ and $k(x)$ can be represented as a Wronskian of five functions

$$\{y(x), Q(x) \cdot f(x) - P(x), Q(x) \cdot k(x), u(x) \cdot (Q^{-}(x) \cdot f^{-}(x) - P^{-}(x)), u(x) \cdot (Q^{-}(x) \cdot k^{-}(x))\}.$$

Here $y(x)$ is a function satisfying the fourth order differential equation, and $\frac{P(x)}{Q(x)}$ is a rational function approximating $f(x)$ at $x = \infty$ in the following sense:

$$f(x) \cdot Q(x) - P(x) - \lambda \cdot \frac{Q^{-}(x)}{x} = O(x^{-N-2}).$$

Here $Q(x)$ is a polynomial of degree N. The order of approximation, $N + 2$ at $x = \infty$ is one greater than the value expected for a generic rational approximation. Attaining the extra order of (non-normal) approximation in this approximation problem is the meaning of the eigenvalue problem in this context.

The resulting fourth order differential equation in $y(x)$:

$$\sum_{i=0}^{4} C_i(x) \cdot y^{(i)}(x) = 0$$

has rational function coefficients $C_i(x)$.

The fourth order differential equations above can be represented, after multiplication by x^{2N}, as a differential equation with rational coefficients of bounded degree. The number of apparent singularities introduced in this equation depends on the degrees of rational functions $a(x)$ and $b(x)$.

The simplest case is just

$$k(x) = x^{\omega}$$

corresponding to a classical hypergeometric function:

$$h_{\omega}(x) = \sum_{m=0}^{\infty} \frac{x^m}{m+\omega}.$$

This gives us (using arbitrary inhomogeneous terms $b(x)$) arbitrary linear combinations

$$f\ (x) = \sum_{\alpha=1}^{m} w_{\alpha} h_{\omega}(a_{\alpha} \cdot x).$$

For the corresponding Hankel matrices with the elements

$$\sum_{\alpha=1}^{m} w_{\alpha} \frac{a_{\alpha}^{i+j+1} - 1}{i+j+\omega}$$

the eigenvector problem on the vector v posses a "commuting auxiliary" eigenvalue formulation, represented by the fourth order linear differential equation on the generating function $Q(x)$ of v. When $m = 1$, there are no apparent singularities, just as in the case of the standard Hilbert matrix above. For $m > 1$, there are $m-1$ apparent singularities and m accessory parameters. These accessory parameters themselves satisfy nonlinear differential equations in apparent singularities. Such equations represent a class of fourth order "Painleve-like" transcendents. Unlike the Garnier equations which are associated with the isomonodromy of second order linear differential equations, these transcendents arise from the isomonodromy deformations of special classes of fourth order linear differential equations.

Notice that the original eigenvalue λ had disappeared from the coefficients of the (fourth order) differential equation. Its meaning lies only in the connection formulas, that connect solutions defined by their expansions at $x = 0$ and $x = \infty$.

Variational Principles and q: Difference Equations

We considered a generalization of discrete prolate functions, which are simultaneously concentrated on arbitrary unions of intervals in space and frequency. We saw that the generalized discrete prolate functions arise from the following sets of matrices

$$P_{C,N} = \left(\sum_{\alpha=1}^{m} \frac{a_{\alpha}^{i-j} - a_{\alpha}^{j-i}}{i-j} \right)_{i,j=0}^{N}.$$

These matrices also arise naturally from the variational principle of concentrated Fourier polynomials as follows:

Analogous to the polynomials in the Szegö case, we consider finite Fourier polynomials of the form

$$F(z) = \sum_{j=0}^{N} x_j e^{2\pi i j z}$$

The variational principle that is applicable to functions of this form is

$$\frac{\int_C |F(z)|^2 \mathrm{d}z}{\int_0^1 |F(z)|^2 \mathrm{d}z}$$

where C is the union of arcs in the complex plane. The classical discrete prolate case arises when C is an arc on the unit circle defined by the bandwidth limitation. The integral in the denominator in the equation above can be replaced by an integral of the same form but over another union of intervals, yielding the most general form of simultaneous concentration problem with solutions by generalized discrete prolate functions. The corresponding extremality problem is solved by means of the generalized eigenvalue problem for linear combinations of matrices the form $P_{C,N}$.

The most general form of discrete prolate-like functions arise from the minors of the q- generalizations of the Fourier matrix,

$$\left(q^{i \cdot j}\right)$$

when q is a root of unity. These generalizations lead naturally to matrices of the form of the symmetric square of the q-Fourier matrix.

A variational principle in this case exists, but requires the use of the q-calculus, the so-called Jackson integral.

The indefinite Jackson integral [10] of the function $f(x)$ is defined as

$$\int f(x) \mathrm{d}_q x = (1-q)\, x \sum_{k=0}^{\infty} q^k f\left(q^k x\right)$$

Using this integral, we can write the following new variational principle for polynomials

$$P(z) = \sum_{i=0}^{n} x_i \cdot z^i$$

in the form:

$$\frac{\int_C |P(z)|^2 \mathrm{d}_q z}{\|P(z)\|^2}.$$

Here C can be an arbitrary union of arcs (intervals) in the complex plane. The corresponding matrices are linear combinations of the following ones:

$$\left(\frac{Q_\alpha^{i-j}-1}{q^{i-j}-1}\right)_{i,j=0}^{n}$$

where $Q_\alpha = q^{N_\alpha}$, for Q_α being the ends of the intervals from the set C.

These objects also correspond to integrable systems, but this time described by (second order) q-difference equations. The isomonodromy equations here are relatively new systems known as "q-Garnier systems" [8].

More complex are q-generalizations of the Hilbert matrix, also arising from the minors of the q-Fourier matrix, whose elements have the form

$$\frac{Q^{i+j+1}-1}{q^{i+j+1}-1} \quad \text{or} \quad \sum_\alpha w_\alpha \frac{Q_\alpha^{i+j+1}-1}{q^{i+j+1}-1}.$$

These also give rise to q-difference isomonodromy equations.

References

1. Alberto Grünbaum, F.: A remark on Hilbert's matrix. Linear Algebra Appl. **43**, 119–124 (1982)
2. Adler, M., Shiota, T., van Moerbeke, P.: Random matrices, Virasoro algebras, and noncommutative KP. Duke Math. J. **94**(2), 379–431 (1998)
3. Chudnovsky, D.: Riemann monodromy problem, isomonodromy deformation equations and completely integrable system. In: Bifurcation phenomena in mathematical physics and related topics, pp. 385–448. D. Reidel, Dordrecht (1980)
4. Chudnovsky, G.: Padé approximation and Riemann monodromy problem. In: Bifurcation phenomena in mathematical physics and related topics, pp. 449–410. D. Reidel, Dordrecht (1980)
5. Chudnovsky, D., Chudnovsky, G.: Explicit continued fractions and quantum gravity. Acta Appl. Math. **36**(1–2), 167–185 (1994)
6. Garnier, R.: Sur une classe de systèmes différentiels Abéliens déduits de la théorie des équations linéaires. Circolo Math. Palermo **43**, 155–191 (1919)
7. Gilbert, E., Slepian, D.: Doubly orthogonal concentrated polynomials. SIAM J. Math Anal. **8**, 290–319 (1977)
8. Hidetaka, S.: A q-analog of the Garnier system. Funkcialaj Ekvacioj **48**, 273–297 (2005)
9. Its, A.R., Tracy, C.A., Widom, H.: Random words, Toeplitz determinants and integrable systems II. Physica D **152–153**, 199–224 (2001)
10. Jackson, F.H.: On q-definite integrals. Quart. J. Pure Appl. Math. **41**, 193–203 (1910)
11. Mahler, K.: Perfect systems. Compos. Math. pp. 95–166 (1968)
12. Morrison, J.A.: On the eigenfunctions corresponding to the bandpass kernel. Quart. Appl. Math. **21**, 13–19 (1963)
13. Okamoto, K.: Isomonodromic deformation and Painleve equation, and the Garnier system. J. Fac. Sci. Univ. Tokyo **Sec. IA**(33), 575–618 (1986)
14. Rosenblum, M.: On the Hilbert matrix I. Proc. Amer. Math. Soc. **9**(1), 137–140 (1958)
15. Rosenblum, M.: On the Hilbert matrix II. Proc. Amer. Math. Soc. **9**(4), 581–585 (1958)

16. Sawyer, W.W.: On the matrix on elements 1/(r+s-1). Can. Math. Bull. **17**(2), 297–298 (1974)
17. Schlesinger, L.: Uber eine klasse von differentialsystemen beliebliger ordnung mit festen kritischer punkten. J. Math. **141**, 96–145 (1912)
18. Slepian, D.: Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V the discrete case. Bell Syst. Tech. J. **57**, 1371–1429 (1978)
19. Slepian, D., Pollak, H.: Prolate spheroidal wave functions, Fourier analysis and uncertainty: Part I. Bell Sys. Tech. J. **40**, 43–64 (1961)
20. Szegö, G.: On some Hermitian forms associated with two given curves of the complex plan. Trans. Am. Math. Soc. **40**(3), 450–461 (1936)
21. Tracy, C.A., Widom, H.: Level spacing distributions and the Bessel kernel. Comm. Math. Phys. **161**, 289–309 (1994)
22. Tracy, C.A., Widom, H.: Fredholm determinants, differential equations and matrix models. Comm. Math. Phys. **163**, 33–72 (1994)

Addition Theorems in Acyclic Semigroups

Javier Cilleruelo, Yahya O. Hamidoune, and Oriol Serra

Summary We give a necessary and sufficient condition on a given family $\mathscr{A}$ of finite subsets of integers for the Cauchy–Davenport inequality

$$|\mathscr{A} + \mathscr{B}| \geq |\mathscr{A}| + |\mathscr{B}| - 1,$$

to hold for any family $\mathscr{B}$ of finite subsets of integers. We also describe the extremal families for this inequality. We prove this result in the general context of acyclic semigroups, which also contain the semigroup of sequences of elements in an ordered group.

Keywords Addition theorems · Semigroups

Mathematics Subject Classifications (2010). 11P70

1 Introduction

Recently, some additive results have been considered in the setting of the semigroup of subsets of integers, see e.g. [1] where Sidon sets are generalized to this context. Following this direction introduced by two of the authors, in this paper we consider

J. Cilleruelo
Instituto de Ciencias Matemáticas CSIC-UAM-UC3M-UAM, and Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Madrid, Spain
e-mail: franciscojavier.cilleruelo@uam.es

Y.O. Hamidoune
UER Combinatoire, Univ. Paris VI, Paris, France
e-mail: hamidoune@math.jussieu.fr

O. Serra
Dept. Matemàtica Aplicada 4, Univ. Politècnica de Catalunya, Spain
e-mail: oserra@ma4.upc.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_6,

some extensions of Cauchy–Davenport and Vosper theorems, see e.g. [5], to the semigroup of subsets of integers, but we shall do the proof in the more general context of acyclic semigroups.

We denote a semigroup $(M, \cdot)$, where '$\cdot$' is a binary associative operation on the set M with a neutral element, simply by M. The semigroup M is *acyclic* if

M1: $y \cdot x$ implies $y = 1$, for every $x \in M$.
M2: $x \cdot y$ implies $x = y = 1$, for every $x, y \in M$.

Our basic examples are the following ones. Let G be an ordered group and let $P = G_+ = \{x : x \geq 1\}$. The set M of finite subsets of P with the product

$$a \cdot b = \{\alpha\beta : \alpha \in a, \beta \in b\}$$

is an acyclic semigroup with neutral element$\{1\}$, where 1 is the neutral element of G. We call M the *sumset* semigroup of G.

For our second example, let P^I denote the set of functions from a set I to P with the induced product

$$(f \cdot g)(y) = f(y) \cdot g(y)$$

is an acyclic semigroup with neutral element the constant function 1. In particular, if $|I| = 1$, then P^I is isomorphic to P (as semigroups). For $|I| \geq 2$, P^I is the semigroup of sequences of elements in P indexed by I. In particular, if $P = \mathbb{N}$, then $\mathbb{N}^I$ is the free abelian monoid generated by I, an important object in factorization theory, see for instance [2].

Acyclic semigroups have the following important property.

Lemma 1. *For any finite nonempty subset S of an acyclic semigroup M and for every $x \neq 1$, we have $xS \neq S$.*

Proof. Suppose that $xS = S$. Take $a \in S$. Then $x^j a \in S$ for all j by induction. Since S is finite, we have $x^j a = x^{k+j} a$ for some j and $k > 0$. By axiom M2, we have $x^k = 1$ and then $x = 1$. □

2 Cayley Graphs on Semigroups

Let M be a semigroup. Let S be a finite subset of M. The Cayley graph $\mathrm{Cay}(M, S)$ of S in M has the elements of M as vertices, and there is an arc (x, y) colored $s \in S$ whenever $y = xs$. Note that the resulting graph is oriented and edge–colored, and it may have parallel arcs. If $1 \in S$, then it has a loop at every vertex.

If M is an acyclic semigroup and $1 \in S$, then the only finite directed cycles in the Cayley graph $\mathrm{Cay}(M, S)$ are the loops. This fact motivates the terminology. In what follows, we assume that M is an acyclic semigroup.

We shall write $\delta(S) = \min\{|xS| : x \in M\}$, the minimum out-degree of a vertex in $\mathrm{Cay}(M, S)$. A subset S will be called *regular* if $\delta(S) = |S|$. We say that S is *biregular* if in addition $|Sx| = |S|$ for every $x \in M$.

We are interested in obtaining lower bounds for the cardinality of the product of two sets in M. To this end, we use the isoperimetric method, see e.g. [3,4].

For a positive integer k and a finite set $S \subset M$ with $1 \in S$, denote by

$$\kappa_k(S) = \min\{|XS| - |X| : |X| \geq k\}$$

the k-th isoperimetric connectivity of S.

It follows from the definition that, for every pair X and S of finite sets in M with $|X| \geq k$, we have

$$|XS| \geq |X| + \kappa_k(S).$$

Note also that $\kappa_i(S) \leq \kappa_{i+1}(S)$ for each $i \geq 1$. A subset $F \subset M$ with $|F| \geq k$ is said to be a *k–fragment* of S if

$$|FS| - |S| = \kappa_k(S).$$

A k–fragment of S with minimal cardinality will be called a *k–atom* of S.

Lemma 2. *Let F and S be finite nonempty subsets of an acyclic semigroup M with $1 \in S$. There is an element $a \in F$ such that $|(F \setminus a)S| \leq |FS| - 1$.*

In particular, every k–fragment of S in M contains a k–atom of S with cardinality k.

Proof. Consider the subgraph of $\mathrm{Cay}(M, S)$ induced on F. Since the graph has no directed cycles, (except for the loops) there is an element $a \in F$ with indegree $\delta^-(a) = 1$ (just the loop). It follows that $a \in (FS) \setminus (F \setminus a)S$.

Now suppose that F is a k–fragment, so that $|FS| = |F| + \kappa_k(S)$. Let A be the smallest k–fragment contained in F. Suppose that $|A| > k$. By the first part of the Lemma, there is $a \in A$ such that $|(A \setminus a)S| \leq |AS| - 1 = |A| - 1 + \kappa_k(S)$, contradicting the minimality of $|A|$. Hence $|A| = k$ and A is a k–atom. □

Theorem 1 (Cauchy–Davenport for acyclic semigroups). *Let S be a finite subset of acyclic semigroup M with $1 \in S$. For every nonempty finite subset X of M, we have*

$$|XS| \geq |X| + \delta(S) - 1,$$

and the inequality is best possible. In particular, we have

$$|XS| \geq |X| + |S| - 1$$

for each finite subset $X \subset M$ if and only if S is regular.

Proof. By Lemma 2, there is a 1–fragment with cardinality one, say $X = \{a\}$. Then, by the definition, $|XS| - |X| \geq \kappa_1(S) = |aS| - 1 \geq \delta(S) - 1$.

By taking $x \in M$ such that $|xS| = \delta(S)$ and $X = \{x\}$, we see that the equality holds. □

Note that, without the assumption $1 \in S$ the best one can say in general is just $|XS| \geq 1$ in contrast with the trivial bound $|XS| \geq \max\{|X|, |S|\}$ in a group. The following example illustrates this remark.

Example 1. Consider the sumset semigroup of the integers. For a subset $A = \{a_1 < a_2 < \cdots < a_n\}$, we denote by $d(A) = \max_{1 \leq i < n}(a_{i+1} - a_i)$ the length of the largest gap in A.

Let $\mathscr{A} = \{A_1, \cdots, A_k\}$ be a collection of subsets of integers with gaps of length at most k, namely $\max_i d(A_i) \leq k$, and with $\min(A_i) = m$, $\max(A_i) = M$ for each i. Let $P = \{0, 1, \cdots, k\}\}$. We have $\mathscr{A} + P = \{m, m+1, M+k\}$, and hence

$$|\mathscr{A} + P| = 1.$$

However $|\mathscr{A}|$ can be arbitrarily large.

Note that if $S = \{\{0\}, P\}$, then $\delta(S) = 2$ and $|\mathscr{A} + S| = |\mathscr{A}| + 1$.

Let M be the sumset semigroup of the integers. One can characterize the sets for which the classical Cauchy–Davenport inequality holds in M. Define a partial order in M by

$$x \preceq y \Leftrightarrow \begin{cases} x = y, & \text{or} \\ \min(x) < \min(y), & \text{or} \\ \min(x) = \min(y) \text{ and } \max(x) < \max(y). \end{cases}$$

Proposition 1. *Let M be the sumset semigroup of the integers. Then S is regular if and only if S is a chain.*

Proof. For any $z \in M$, we observe that $(\min(zx), \max(zx)) = (\min(z) + \min(x), \max(z) + \max(x))$. If all the pairs $(\min(x), \max(x))$, $x \in S$, are all distinct, then the pairs $(\min(zx), \max(zx))$, $x \in S$, are all distinct. Thus all the elements zx, $x \in S$ are also distinct and $|zS| = |S|$.

On the other hand, if $(\min(x), \max(x)) = (\min(y), \max(y))$ for distinct elements $x, y \in S$, we have that $xz = yz = [\min(x), \max(x) + k]$ when $z = [0, k]$ and $k \geq \max(x) - \min(x)$. Thus $\delta(S) < |S|$. □

As a consequence of the above Proposition, the classical Cauchy–Davenport inequality holds for chains in the sumset semigroup of the integers.

Corollary 1. *Let S be a chain in the sumset semigroup of the integers with $1 \in S$. Then, for each finite nonempty subset $X \subset M$,*

$$|XS| \geq |X| + |S| - 1.$$

The following example shows that there are antichains in the sumset semigroup of the integers with $|\mathscr{A} + \mathscr{A}| = |\mathscr{A}| + 1$, the minimum possible value given by Theorem 1.

Example 2. Let $\mathscr{A}_0$ be an arbitrary family of sets of integers in the interval $[m, M]$. Let $\mathscr{A}_1 = \{[2m - M, m] \cup A \cup [M, 2M - m],\ A \in \mathscr{A}_0\}$. Note that, for every pair $A, A' \in \mathscr{A}_1$, we have $A + A' = [4m - 2M, 4M - 2m]$. By setting $\mathscr{A} = \mathscr{A}_1 \cup \{0\}$, we have a family with $|\mathscr{A} + \mathscr{A}| = |\mathscr{A}| + 1$.

We conjecture that Theorem 1 holds in the semigroup of finite sequences of elements from a torsion–free group:

Conjecture 1. Let G be a torsion–free group and I a finite set. Then for every nonempty finite subsets $S, T \subset G^I$ with $1 \in S$, we have

$$|ST| \geq |T| + \delta(S) - 1.$$

3 Vosper's Theorem

We next analyze the case of equality in the Cauchy–Davenport theorem for acyclic abelian semigroups.

A set $P \subset M$ of the form $P = a\{1, r, r^2, \ldots r^{k-1}\}$ is called an $r-$ progression.

Lemma 3. *Let S be a biregular finite nonempty subset of an acyclic semigroup M with $1 \in S$ and let $u \in M \setminus \{1\}$. If*

$$|\{1, u\}S| = |S| + 1,$$

then uS is an u–progression.

Proof. Since $\delta(S) = |S|$, we have $|uS| = |S|$, which implies

$$|S \cap uS| = 2|S| - |\{1, u\}S| = |S| - 1.$$

It follows that the subgraph $\Gamma(uS)$ of $\Gamma = Cay(M, \{1, u\})$ induced by uS contains $|S| - 1$ arcs. Since S is biregular, we cannot have $su = s'u$ for a pair of distinct elements $s, s' \in S$, so that the indegree of every element in $\Gamma(uS)$ is at most one. Since $\Gamma(uS)$ is acyclic, it is a path of length $|S|$. This implies that uS is an u–progression. □

Theorem 2 (Vosper Theorem for acyclic semigroups). *Let M be an abelian acyclic semigroup. Let S be a regular nonempty finite subset of M with $1 \in S$ and $|S| \geq 2$. Let X be a finite subset of M with $|X| \geq 2$. If*

$$|XS| = |X| + |S| - 1,$$

then one of the following conditions holds:

(i) There are $u, v \in X$ such that $uS^* = vS^*$,
(ii) There is $u \in M$ such that uS is an r–progression for some $r \in M$. Moreover, if X is also regular, then there is $u' \in M$ such that $u'X$ is an r–progression as well.

Proof. By the definition, we have $\kappa_2(S) \leq |S|-1$. By Theorem 1, since S is regular, we have $\kappa_2(S) = |S|-1$. By Lemma 2, there is a 2–atom of S with cardinality two contained in X. Thus there are $u, v \in X$ with $|\{u, v\}S| = |S| + 1$. We consider two cases.

Case 1. $v \notin (uS)$ and $u \notin (vS)$. In this case (i) holds.

Case 2. $v \in uS$ or $u \in vS$. We may assume that $v = us$ for some $s \in S$. Then $|\{u, us\}S| = |\{1, s\}(uS)| = |S|+1$. By Lemma 3, uS is an r–progression for some r, say $uS = a\{1, r, \cdots, r^{k-1}\}$.

Now if X is also regular, then Xu is a regular set and $|(Xu)S| = |X| + |S| - 1$. We can write $(Xu)S = Xa\{1, r, \ldots, r^{k-1}\} = Xa\{1, r\}\cdots\{1, r\}$. Since $|(Xu)S| = |X|+|S|-1$, we have $|Xa\{1, r\}| = |aX\{1, r\}| = |X|+1$ and we likewise conclude that aX is an r–progression.

□

In the sumset semigroup of the integers, both conclusions in the above Theorem may hold as illustrated by the following example.

Example 3. Let

$$\mathscr{A} = \{\{0\}, \{0, 3, 6, 9\}, \{0, 2, 3, 6, 9, 10\}, \{0, 1, 4, 7, 9, 11\}\}.$$

Since $\mathscr{A}$ is a chain, it is biregular. Now let

$$\mathscr{B} = \{\{0, 1, 2, 3, 4, 5, 6\}, \{0, 1, 3, 4, 5, 6\}\}.$$

We have $|\mathscr{A} + \mathscr{B}| = |\mathscr{A}| + |\mathscr{B}| - 1$ and $\mathscr{A}$ is not a progression.

References

1. J. Cilleruelo and O. Serra, Sidon families of k-sets, preprint (2008).
2. A. Geroldinger and F. Halter-Koch, *Non-Unique Factorizations. Algebraic, Combinatorial and Analytic Theory*, Pure and Applied Mathematics, vol. 278, Chapman & Hall/CRC, Boca Raton, 2006.
3. Yahya Ould Hamidoune, Some additive applications of the isopermetric approach, *Annales de l'institut Fourier*, 58 no. 6 (2008), p. 2007–2036 arXiv:0706.0635.
4. O. Serra, An isoperimetric method for the small sumset problem. *Surveys in combinatorics* 2005, 119–152, *London Math. Soc. Lecture Note Ser.*, 327, Cambridge University Press, Cambridge, 2005.
5. T. Tao and V.H. Vu, *Additive Combinatorics*, Cambridge Studies in Advanced Mathematics 105, Cambridge Press University, Cambridge, 2006.

Small Sumsets in Free Products of $\mathbb{Z}/2\mathbb{Z}$

Shalom Eliahou and Cédric Lecouvey

Summary Let G be a group. For positive integers $r, s \leq |G|$, let $\mu_G(r, s)$ denote the smallest possible size of a sumset (or product set) $AB = \{ab \mid a \in A, b \in B\}$ for any subsets $A, B \subset G$ subject to $|A| = r$, $|B| = s$. The behavior of $\mu_G(r, s)$ is unknown for the free product G of groups G_i, except if the factors G_i are all isomorphic to $\mathbb{Z}$, in which case $\mu_G(r, s) = r + s - 1$ by a theorem of Kemperman for torsion-free groups (1956). In this paper, we settle the case of a free product G whose factors G_i are all isomorphic to $\mathbb{Z}/2\mathbb{Z}$, and prove that $\mu_G(r, s) = r + s - 2$ or $r + s - 1$, depending on whether r and s are both even or not.

Keywords Additive combinatorics · Cauchy subsets · Reduced words

Mathematics Subject Classifications (2010). 11B13, 11P70, 20E06

1 Introduction

Let G be a group written multiplicatively. Given subsets $A, B \subset G$, we denote by

$$AB = \{ab \mid a \in A, b \in B\}$$

the *product set* of A, B. A classical question in additive number theory consists in determining the smallest possible cardinality $\mu_G(r, s)$ of a product AB of subsets $A, B \subset G$ subject to $|A| = r, |B| = s$ for any given integers $r, s \geq 1$.

S. Eliahou and C. Lecouvey
Univ Lille Nord de France, F-5900 Lille, France; ULCO, LMPA Joseph Liouville,
F-62228 Calais, France; FR CNRS 2956, France
e-mail: eliahou@lmpa.univ-littoral.fr; lecouvey@lmpa.univ-littoral.fr

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_7,

The oldest result addressing this question is the Cauchy–Davenport theorem for cyclic groups of prime order [2, 3]. More recently, the function $\mu_G(r,s)$ has been determined for all abelian groups [4]. Not much is known on $\mu_G(r,s)$ for nonabelian groups, besides for dihedral groups [7] and torsion-free groups [9] (See Sect. 2). The latter of course include free groups, i.e., free products of copies of $\mathbb{Z}$.

In this paper, we determine $\mu_G(r,s)$ for the free product G of copies of $\mathbb{Z}/2\mathbb{Z}$. This is achieved by an explicit construction of small product sets in Sect. 4, together with a proof of optimality in Sect. 6, using theorems of Olson and of Hamidoune recalled in Sect. 5. In the process, we also use the Kurosh subgroup theorem for free products of groups, recalled in Sect. 3. Our main result reads as follows.

Theorem 1. *Let G be the free product of any collection of two or more copies of $\mathbb{Z}/2\mathbb{Z}$. Then, for all positive integers $r,s \geq 1$, we have*

$$\mu_G(r,s) = \begin{cases} r+s-2 & \text{if } r \equiv s \equiv 0 \bmod 2, \\ r+s-1 & \text{otherwise.} \end{cases}$$

To the best of our knowledge, the case of free groups and the above result are the only instances of free products of groups for which the function $\mu_G(r,s)$ is exactly known so far.

We refer to [12] for background on additive number theory and to [6] for a survey on the function $\mu_G(r,s)$. We close this section with a few generalities on $\mu_G(r,s)$, valid for any group G and positive integers $r,s \leq |G|$, namely: $\mu_G(1,s) = s$, $\mu_G(r,s) = \mu_G(s,r)$, $\mu_G(r,s) \geq \max(r,s)$, and finally $\mu_G(r,r) = r$ if and only if G contains a subgroup of order r.

2 The Function $\kappa_G(r,s)$

We start by recalling the behavior of $\mu_G(r,s)$ in case G is abelian, or a dihedral group, or a torsion-free group. In these cases, $\mu_G(r,s)$ can be exactly modeled by a numerical function $\kappa_G(r,s)$ involving the set of orders of finite subgroups of G. As we shall see, this function is also hidden in Theorem 1.

Notation. Let G be a group. We denote by

- $\mathscr{H}(G)$ the set of orders of finite subgroups of G,
- $\kappa_G(r,s) = \min_{h \in \mathscr{H}(G)} (\lceil r/h \rceil + \lceil s/h \rceil - 1)\, h$.

For example, if G is torsion-free, then $\mathscr{H}(G) = \{1\}$ and $\kappa_G(r,s) = r+s-1$. If G is finite abelian of order n, then $\mathscr{H}(G) = \{d \in \mathbb{N} \mid d \text{ divides } n\}$. In particular, if $n = p$ is prime, then $\kappa_G(r,s) = \min\{r+s-1, p\}$; this is precisely the formula determining $\mu_{\mathbb{Z}/p\mathbb{Z}}(r,s)$ in the Cauchy–Davenport theorem.

Theorem 2. *Let G be an arbitrary abelian group, or a dihedral group, or a torsion-free group. Then, for all positive integers $r,s \leq |G|$, we have*

$$\mu_G(r,s) = \kappa_G(r,s)\ .$$

This equality is proved in [4] for abelian groups, in [7] for dihedral groups, and in [9] for torsion-free groups.

In this paper, we shall show that $\mu_G(r,s)$ can also be exactly modeled by $\kappa_G(r,s)$ when G is the free product of any collection of copies of $\mathbb{Z}/2\mathbb{Z}$. This is the true content of Theorem 1. More precisely, we shall prove that for such a group G, we have

$$\mu_G(r,s) = \kappa_G(r,s) = \begin{cases} r+s-2 & \text{if } r \equiv s \equiv 0 \bmod 2\,, \\ r+s-1 & \text{otherwise.} \end{cases}$$

The second equality follows from the Kurosh subgroup theorem for free products and is settled in the next section. The proof of the first equality, in fact of Theorem 1, is longer and achieved in Sects. 4 and 6.

It should be noted at this point that $\mu_G(r,s)$ cannot always be exactly modeled by $\kappa_G(r,s)$. The smallest counterexample occurs with the unique nonabelian group of order 21 [5, p. 246].

3 Free products of groups

Let $G = *G_i$ be the free product of a collection of groups G_i. Recall that as a set, the free product G consists of all *reduced words* $w = g_1 \cdots g_k$ of length $k \geq 0$, where each letter g_j belongs to some factor G_i and is distinct from 1, and where consecutive letters g_j, g_{j+1} belong to distinct factors. Note that a free product with at least two factors $G_i \neq \{1\}$ contains elements of infinite order, e.g., any reduced word $w = g_1 g_2$ of length 2. We refer to the book of Lyndon–Schupp for extensive information on free products [11].

In order to understand $\kappa_G(r,s)$ for the free product G, we need the classical Kurosh subgroup theorem [10].

Theorem 3 (Kurosh). *Let $G = *G_i$ be the free product of groups G_i. Let $H \leq G$ be a subgroup of G. Then H is a free product $H = F*(*H_j)$, where F is a free group and each H_j is the intersection of H with a conjugate of some factor G_i of G.*

This theorem allows us to easily express $\kappa_G(r,s)$ in terms of the $\kappa_{G_i}(r,s)$.

Proposition 1. *Let $G = *_{i\in I} G_i$ be the free product of a collection of groups $(G_i)_{i\in I}$. Then, for all integers $r, s \geq 1$, we have*

$$\kappa_G(r,s) = \min_{i\in I} \kappa_{G_i}(r,s)\,.$$

Proof. By the Kurosh subgroup theorem, every finite subgroup of G must be a conjugate of a finite subgroup of one of the factors G_i. Indeed, a free product

$H = F*(*H_j)$ as in the theorem of Kurosh, with more than one nontrivial factor, would be infinite. This implies the formula

$$\mathscr{H}(G) = \bigcup_{i \in I} \mathscr{H}(G_i) \,, \tag{1}$$

and the result follows from the definition of $\kappa_G(r,s)$. □

As a direct application, assume that each G_i is isomorphic to the cyclic group $\mathbb{Z}/n\mathbb{Z}$, for some fixed integer n. It follows that

$$\kappa_G(r,s) = \min_{d|n} (\lceil r/d \rceil + \lceil s/d \rceil - 1)d = \kappa_{\mathbb{Z}/n\mathbb{Z}}(r,s) \tag{2}$$

for all integers $r, s \geq 1$. Specializing to the case $n = 2$, we get the following expression.

Corollary 1. *Let G be the free product of a nonempty collection of copies of $\mathbb{Z}/2\mathbb{Z}$. Then for all $r, s \geq 1$, we have*

$$\kappa_G(r,s) = \begin{cases} r+s-2 \text{ if } r \equiv s \equiv 0 \bmod 2 \,, \\ r+s-1 \text{ otherwise.} \end{cases}$$

Proof. It follows from Proposition 1 and formula (2) that $\kappa_G(r,s) = \kappa_{\mathbb{Z}/2\mathbb{Z}}(r,s) = \min\{r+s-1, (\lceil r/2 \rceil + \lceil s/2 \rceil - 1)2\}$. The only occurrence of a strict inequality

$$(\lceil r/2 \rceil + \lceil s/2 \rceil - 1)2 < r+s-1$$

is when r, s are both even, in which case we get $\kappa_G(r,s) = r+s-2$. □

4 Proof of $\mu_G(r,s) \leq \kappa_G(r,s)$

In this section, we focus on the group $G = *_i \, \mathbb{Z}/2\mathbb{Z}$, the free product of a collection of at least two copies of $\mathbb{Z}/2\mathbb{Z}$. We shall prove the inequality

$$\mu_G(r,s) \leq \kappa_G(r,s) \tag{3}$$

of Theorem 1 (combined with Corollary 1), for all $r, s \geq 1$. This will be achieved by an explicit construction of small product sets in G.

We start with a general upper bound on $\mu_F(r,s)$ for any group F containing a copy of $\mathbb{Z}$.

Lemma 1. *Let F be a group containing an element x of infinite order. Then for all $r, s \geq 1$, we have*

$$\mu_F(r,s) \leq r+s-1 \,.$$

Proof. Let $A = \{1, x, \ldots, x^{r-1}\}$ and $B = \{1, x, \ldots, x^{s-1}\}$. Then the sets A, B satisfy $|A| = r$, $|B| = s$ and $|AB| = |\{1, x, \ldots, x^{r+s-2}\}| = r + s - 1$. This implies the stated inequality. □

We are now ready to prove inequality (3).

- Assume first that $r \not\equiv 0$ or $s \not\equiv 0 \bmod 2$. In that case, Corollary 1 yields $\kappa_G(r, s) = r + s - 1$. Since G contains elements of infinite order, we may apply Lemma 1. This yields $\mu_G(r, s) \leq r + s - 1 = \kappa_G(r, s)$, as desired.
- Assume now $r \equiv s \equiv 0 \bmod 2$. We must then prove that $\mu_G(r, s) \leq r + s - 2$, and this will require somewhat subtler sets. Their construction will take place in a subgroup G_0 of G consisting of the free product of just two copies of $\mathbb{Z}/2\mathbb{Z}$. We shall use the following presentation of G_0 by generators and relations:

$$G_0 = \langle a, b \mid a^2 = b^2 = 1 \rangle \ .$$

Since G_0 is a subgroup of G, it follows that

$$\mu_G(r, s) \leq \mu_{G_0}(r, s) \tag{4}$$

for all $r, s \geq 1$.

Given that $\kappa_{G_0}(r, s) = \kappa_G(r, s)$ by Corollary 1, in order to prove inequality (3), it suffices to prove that $\mu_{G_0}(r, s) \leq \kappa_{G_0}(r, s)$, for all $r, s \geq 1$.

As a set, the group G_0 consists of all *reduced words* in a, b, i.e., in this case, words alternating the letters a and b, such as $ababa$ for instance. There is a length function

$$l : G_0 \to \mathbb{N}$$

defined by $l(w) = t$ if $w = x_1 \cdots x_t$ is a reduced word with t letters $x_i \in \{a, b\}$. This function induces an obvious metric on G_0, namely $d(w_1, w_2) = l(w_1 w_2^{-1})$ for all $w_1, w_2 \in G_0$.

For $k \geq 1$, there are exactly two reduced words of length k in G_0, namely:

1. $(ab)^{k/2}$ and $(ba)^{k/2}$ if k is even,
2. $a(ba)^{(k-1)/2}$ and $b(ab)^{(k-1)/2}$ if k is odd.

For every $m \geq 0$, we shall denote by $\mathscr{B}(m)$ the set of words in G_0 of length at most m. Note that $\mathscr{B}(m)$ is the ball of radius m centered at 1 for the above metric on G_0.

The product set of two balls $\mathscr{B}(k_1)$, $\mathscr{B}(k_2)$ will be needed below. It is given by the following obvious formula, valid for all $k_1, k_2 \geq 0$:

$$\mathscr{B}(k_1)\mathscr{B}(k_2) \ = \ \mathscr{B}(k_1 + k_2) \ . \tag{5}$$

Actually, the analogous formula holds in any free product of groups. Indeed, a reduced word of length at most $k_1 + k_2$ is the product of two reduced words of lengths at most k_1 and k_2, respectively.

Here are two more useful results concerning these balls. We denote by $\langle a \rangle$ the subgroup of G_0 generated by a, i.e., $\langle a \rangle = \{1, a\}$.

Lemma 2. *For all integers $k \geq 0$, we have*

$$|\mathscr{B}(k)| = 2k + 1 \,,$$
$$|\langle a \rangle \mathscr{B}(k)| = 2k + 2 \,.$$

Moreover, $\mathscr{B}(k)$ is contained in $\langle a \rangle \mathscr{B}(k)$, and the set difference $\langle a \rangle \mathscr{B}(k) \setminus \mathscr{B}(k)$ consists of the unique reduced word of length $k + 1$ starting with the letter a.

Proof. The empty word 1 is the only word of length $t = 0$. For $1 \leq t \leq k$, there are exactly two words of length t in G_0, namely the one starting with a and that starting with b. This proves that $|\mathscr{B}(k)| = 2k + 1$.

For the second assertion, we have $\langle a \rangle \mathscr{B}(k) = \mathscr{B}(k) \cup a\mathscr{B}(k)$. Now, all words in $a\mathscr{B}(k)$ are already contained in $\mathscr{B}(k)$, except the unique reduced word of length $k + 1$ starting with a. This shows that $|\langle a \rangle \mathscr{B}(k)| = |\mathscr{B}(k)| + 1$, and we are done. □

There are obvious analogous statements for $\mathscr{B}(k)\langle a \rangle$ and $\mathscr{B}(k)\langle b \rangle$, that we shall implicitly use in the proof below.

Lemma 3. *For all integers $k \geq 0$, we have the switching formulas*

$$\langle a \rangle \mathscr{B}(2k) = \mathscr{B}(2k)\langle a \rangle \,,$$
$$\langle a \rangle \mathscr{B}(2k + 1) = \mathscr{B}(2k + 1)\langle b \rangle \,.$$

Proof. By Lemma 2, we know that for $t \geq 0$, the product set $\langle a \rangle \mathscr{B}(t)$ consists of $\mathscr{B}(t)$ plus one more element, namely the unique reduced word of length $t + 1$ starting with a.

Assume first that $t = 2k$ is even. Then the unique reduced word of length $2k + 1$ starting with a necessarily ends with a as well. This establishes the equality $\langle a \rangle \mathscr{B}(2k) = \mathscr{B}(2k)\langle a \rangle$. If now $t = 2k + 1$ is odd, then the unique reduced word of length $2k+2$ starting with a ends with b. Therefore $\langle a \rangle \mathscr{B}(2k+1) = \mathscr{B}(2k+1)\langle b \rangle$, as claimed. □

We are now ready to prove that if r, s are both even, then $\mu_{G_0}(r, s) \leq r+s-2 = \kappa_{G_0}(r, s)$. For this, it is convenient to look separately at the cases $r \equiv s \bmod 4$ and $r \not\equiv s \bmod 4$.

Case $r \equiv s \bmod 4$. Set $A = \langle a \rangle \mathscr{B}(r/2 - 1)$ and $B = \mathscr{B}(s/2 - 1)\langle a \rangle$. Then by Lemma 2, we have $|A| = r$, $|B| = s$ as required. Now, computing the product set of A, B, we get

$$\begin{aligned} AB &= \langle a \rangle \mathscr{B}(r/2 + s/2 - 2)\langle a \rangle \quad \text{by Formula (5)} \\ &= \langle a \rangle^2 \mathscr{B}(r/2 + s/2 - 2) \quad \text{since } r/2 + s/2 \text{ is even and by Lemma 3} \\ &= \langle a \rangle \mathscr{B}(r/2 + s/2 - 2) \quad \text{since } \langle a \rangle \text{ is a subgroup.} \end{aligned}$$

Case $r \not\equiv s \bmod 4$. Here, set $A = \langle a \rangle \mathscr{B}(r/2-1)$ and $B = \mathscr{B}(s/2-1)\langle b \rangle$. Then again $|A| = r$, $|B| = s$, and computing AB we get

$$\begin{aligned} AB &= \langle a \rangle \mathscr{B}(r/2+s/2-2)\langle b \rangle \quad \text{by Formula (5)} \\ &= \langle a \rangle^2 \mathscr{B}(r/2+s/2-2) \quad \text{since } r/2+s/2 \text{ is odd and by Lemma 3} \\ &= \langle a \rangle \mathscr{B}(r/2+s/2-2). \end{aligned}$$

In both cases, we get $AB = \langle a \rangle \mathscr{B}(r/2+s/2-2)$, and it follows from Lemma 2 that $|AB| = r+s-2$. We conclude that $\mu_G(r,s) \leq r+s-2 = \kappa_G(r,s)$ in the present case, as desired.

5 Optimality

In order to prove that the upper bound obtained in the preceding section is optimal, we need some general tools. We shall use theorems of Olson [13] and of Hamidoune [8], recalled below. We also briefly indicate how Kemperman's theorem on torsion-free groups easily follows from each of them.

Theorem 4 (Olson). *Let A, B be two finite subsets of a group G. There exists a nonempty subset S of AB and a finite subgroup H of G such that*

$$|AB| \geq |S| \geq |A| + |B| - |H|$$

and either $HS = S$ or $SH = S$.

This easily implies Kemperman's theorem in [9].

Theorem 5 (Kemperman). *Let G be a torsion-free group. Then for all $r, s \geq 1$, we have*

$$\mu_G(r,s) = r+s-1 \ .$$

Proof. Lemma 1 gives $\mu_G(r,s) \leq r+s-1$. In order to prove that this bound is optimal, let $A, B \subset G$ have cardinality r, s respectively. By the above theorem of Olson, we have

$$|AB| \geq |A| + |B| - |H|$$

for some finite subgroup $H \leq G$. But then $H = \{1\}$ since G a torsion-free. Thus $|AB| \geq r+s-1$, as desired. □

We now turn to the promised theorem of Hamidoune.

Definition 1. A finite subset A in a group G is said to be a Cauchy subset if, for all finite subset $B \subset G$, one has

$$|AB| \geq \min\{|A| + |B| - 1, |G|\} \ .$$

In order to test the property for $A \subset G$ to be a Cauchy subset, it suffices to consider product sets AH and HA where H is a finite subgroup of G. The first instance of such a statement goes back to Mann for finite abelian groups [1], which easily implies both the Cauchy–Davenport theorem and an extension due to Chowla. Here we need the following version [8].

Theorem 6 (Hamidoune). *Let G be an infinite group, and let A be a finite subset of G containing 1. Then A is a Cauchy subset of G if and only if, for all finite subgroups $H \leq G$, one has*

$$|AH| \geq |A| + |H| - 1 \, .$$

There is a corresponding version for finite groups in [8], but we won't need it here. It is obvious that the above theorem also implies Kemperman's theorem.

6 Proof of $\mu_G(r,s) \geq \kappa_G(r,s)$

Here again, let $G = *_i \ \mathbb{Z}/2\mathbb{Z}$ be the free product of copies of $\mathbb{Z}/2\mathbb{Z}$. We now set out to prove the inequality $\mu_G(r,s) \geq \kappa_G(r,s)$ of Theorem 1. Recall from Corollary 1 that $\kappa_G(r,s) = r + s - 2$ or $r + s - 1$, according as r, s are both even or not.

In order to prove this inequality, we shall use the theorems of Olson and of Hamidoune recalled in the preceding section.

Our first claim is that $\mu_G(r,s) \geq r + s - 2$ for all $r, s \geq 1$. Indeed, let $A, B \subset G$ with $|A| = r$, $|B| = s$. Then, by Olson's theorem, there is a subset $\emptyset \neq S \subset AB$, and a finite subgroup $H \leq G$ such that

$$|AB| \geq |S| \geq |A| + |B| - |H| \, , \tag{6}$$

and $SH = S$ or $HS = S$. Since, by the theorem of Kurosh and (1), the finite subgroups of G have order 1 or 2 only, inequality (6) gives $|AB| \geq |A| + |B| - 2$. This already shows that $\mu_G(r,s) \geq \kappa_G(r,s)$ for all r, s even.

It remains to show that $\mu_G(r,s) \geq r + s - 1$ if r, s are not both even. To start with, the finer statement $SH = S$ or $HS = S$ of Olson's theorem allows us to treat the case where $r + s$ is odd. Indeed, assume for instance that $|A|$ is odd and $|B|$ is even. If $|H| = 1$ in Olson's theorem, then $|AB| \geq |A| + |B| - 1$ as required. If now $|H| = 2$, then $|S|$ is even since S is stabilized by H. But then, the inequality

$$|S| \geq |A| + |B| - 2$$

may be refined into

$$|S| \geq |A| + |B| - 1 \, ,$$

since $|S|$ is even and $|A| + |B| - 2$ is odd here. This yields $\mu_G(r,s) \geq r + s - 1 = \kappa_G(r,s)$ if $r + s$ is odd.

In order to settle the remaining case where r, s are both odd, we appeal to Hamidoune's theorem. Let $A \subset G$ satisfy $|A| = r$ odd. Then we claim that A is a Cauchy subset of G, i.e., that $|AB| \geq |A| + |B| - 1$ for all nonempty finite subsets $B \subset G$. Note that this claim will imply $\mu_G(r, s) \geq \kappa_G(r, s) = r + s - 1$ for all $s \geq 1$, whether odd or not (but still with r odd). In particular, the case $r + s$ odd will follow from both Olson's and Hamidoune's theorems.

To prove the claim that A is a Cauchy subset if $|A|$ is odd, Hamidoune's test in Section 5 says that it suffices to check that

$$|AH| \geq |A| + |H| - 1$$

for all finite subgroups $H \leq G$. If $|H| = 1$, we are done. If $|H| = 2$, then $|AH|$ must be even, and since $|AH| \geq |A|$ and $|A|$ is odd, it follows that $|AH| \geq |A| + 1 = |A| + |H| - 1$, as desired.

The proof of Theorem 1 is now complete.

References

1. Mann, H. B.: An addition theorem of Abelian groups for sets of elements. Proc. Amer. Math. Soc. **4**, 423 (1953)
2. Cauchy, A. L.: Recherches sur les nombres. J. École Polytechnique **9**, 99–123 (1813)
3. Davenport, H.: On the addition of residue classes. J. Lond. Math. Soc. **10**, 30–32 (1935)
4. Eliahou, S., Kervaire, M.: Minimal sumsets in infinite Abelian groups. J. Algebra **287**, 449–457 (2005)
5. Eliahou, S., Kervaire, M.: Some results on minimal sumset sizes in finite non-Abelian groups. J. Number Theory **124**, 234-247 (2007)
6. Eliahou, S., Kervaire, M.: Some extensions of the Cauchy-Davenport Theorem. Electron. Notes Discrete Math. **28**, 557-564 (2007)
7. Eliahou, S., Kervaire, M.: Minimal sumsets in finite solvable groups. Discrete Math. **310**, 471–479 (2010)
8. Hamidoune, Y. O.: On small subset product in a group. In: Structure theory of set addition. Astérisque **258**, pp. 281–308 (1999)
9. Kemperman, J. H. B.: On complexes in a semigroup. Indag. Math. **18**, 247–254 (1956)
10. Kurosh, A. G.: Die Untergruppen der freien Produkte von beliebigen Gruppen. Math. Ann. **109**, 647–660 (1934)
11. Lyndon, R. C., Schupp, P. E.: Combinatorial Group Theory. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 89. Springer, Berlin (1977)
12. Nathanson, M. B.: Additive Number Theory: Inverse Problems and the Geometry of Sumsets, Graduate Text in Mathematics 165, Springer, New York (1996)
13. Olson, J. E.: On the sum of two sets in a group. J. Number Theory **18**, 110–120 (1984)

A Combinatorial Approach to Sums of Two Squares and Related Problems

Christian Elsholtz

Dedicated to Melvyn Nathanson. With many thanks for his beautiful expositions in additive number theory, emphasizing elementary methods.

Summary In this paper, we study elementary approaches to classical theorems on representations of primes of the form $ax^2 + by^2$, in particular the two squares theorem. While most approaches make use of quadratic residues, we study a route initiated by Liouville and simplified by Heath–Brown and Zagier.

Keywords Binary quadratic forms · Fermat's two squares theorem

Mathematics Subject Classifications (2010). 11A41, 11E25

1 Introduction

In this paper we study elementary approaches to classical theorems on representations of primes of the form $ax^2 + by^2$, in particular the two squares theorem.

1.1 The Sums of Two Squares Theorem

Theorem 1. *A positive integer n can be written as a sum of two integer squares, if and only if the canonical prime factorization $n = p_1^{\gamma_1} \cdots p_r^{\gamma_r}$ (where the p_i are distinct primes) satisfies the condition: if $p_i \equiv 3 \bmod 4$, then γ_i is even.*

In order to prove this theorem, one proves the following theorem and several minor lemmata.

C. Elsholtz
Institut für Mathematik A, Technische Universität Graz,
Steyrergasse 30, A-8010 Graz, Austria
e-mail: elsholtz@math.tugraz.at

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_8,

Theorem 2. *A prime* $p \equiv 1 \bmod 4$ *can be written as* $p = x^2 + y^2$.

Wells [38] includes Theorem 2 in a list of the ten most beautiful results in mathematics.

In his "Apology", Hardy [18] writes: "Another famous and beautiful theorem is Fermat's 'two square' theorem... All the primes of the first class" [i.e., 1 mod 4]... "can be expressed as the sum of two integral squares... This is Fermat's theorem, which is ranked, very justly, as one of the finest of arithmetic. Unfortunately, there is no proof within the comprehension of anybody but a fairly expert mathematician."

In this paper, we discuss quite elementary proofs, and it would be interesting to know if Hardy would also have written this about the types of proof (and its simplifications), discussed in Sects. 1.2, 1.3, and 1.6.2.

The history of the theorems above is described in detail in Dickson [8] (volume 2, chapter VI) and also in Edwards [10]. Already Diophant discussed representations of integers as a sum of two squares, and, by slightly altering the text, Jacobi interpreted Diophant's writing in such a way that Diophant possibly essentially knew and was able to prove: if a square-free number n is a sum of two squares, then neither n nor any factor of n is of the form $4k - 1$, (see [8], page 236).

The first correct statement of the necessary and sufficient conditions for writing an integer as a sum of two integer squares, without a proof, might have been by Albert Girard. The theorem is also often attributed to Fermat, who wrote he had a proof. His proof is not known to us, even though in this case it is believed he had the right methods to prove the theorem indeed. Euler eventually gave the first proof that has survived.

Since $p = 2 = 1^2 + 1^2$, and since all squares are of the from 0 or 1 mod 4 so that no number $n \equiv 3 \bmod 4$ can be a sum of two squares, Theorem 2 implies

Corollary 1. *A prime* p *can be written as* $p = x^2 + y^2$ *if and only if* $p = 2$ *or* $p \equiv 1 \bmod 4$.

Lemma 1. *If* $m = x_1^2 + y_1^2$ *and* $n = x_2^2 + y_2^2$ *can be written as sums of two integer squares, then their product* mn *can also be written in this form.*

Proof of Lemma 1: This follows immediately from the identity $mn = (x_1x_2 - y_1y_2)^2 + (x_1y_2 + x_2y_1)^2$, an identity which can be motivated by means of complex numbers:

$$\begin{aligned} mn &= ((x_1 + y_1i)(x_1 - y_1i))((x_2 + y_2i)(x_2 - y_2i)) \\ &= ((x_1 + y_1i)(x_2 + y_2i))((x_1 - y_1i)(x_2 - y_2i)) \\ &= (x_1x_2 - y_1y_2 + i(x_1y_2 + x_2y_1))((x_1x_2 - y_1y_2 - i(x_1y_2 + x_2y_1)) \\ &= (x_1x_2 - y_1y_2)^2 + (x_1y_2 + x_2y_1)^2. \end{aligned}$$

□

Lemma 2. *If* n *is divisible by a prime* $p \equiv 3 \bmod 4$, *and* $n = x^2 + y^2$, *then* $x \equiv y \equiv 0 \bmod p$.

Proof of Lemma 2: By Fermat's little theorem: Let p be prime and x an integer, then

$$x^{p-1} \bmod p = \begin{cases} 0 & \text{if } x \equiv 0 \bmod p \\ 1 & \text{if } x \not\equiv 0 \bmod p. \end{cases}$$

If $p \equiv 3 \bmod 4$, then

$$\left(x^2 + y^2\right)\left(x^{p-3} - x^{p-5}y^2 + x^{p-7}y^4 \mp \cdots + y^{p-3}\right) = x^{p-1} + y^{p-1}.$$

Since $x^2 + y^2 \equiv 0 \bmod p$, one also has $x^{p-1} + y^{p-1} \equiv 0 \bmod p$. As $p > 2$, we must have, by Fermat's observation above, that $x \equiv y \equiv 0 \bmod p$. □

The above lemmata reduce the proof of Theorem 1 to a proof of Theorem 2.

There is a multitude of proofs of Theorem 2. Most of these use quite essentially the fact that for a prime $p \equiv 1 \bmod 4$ there is a solution of $x^2 \equiv -1 \bmod p$. This follows for example from $x = \frac{p-1}{2}!$ or $x = g^{\frac{p-1}{4}}$, where g is a generating element of the group $(\mathbb{Z}/p\mathbb{Z})^\times$ or g is a nonresidue modulo p. However, checking the details in this calculation from first principles is already half of the proof.

The methods involved in these various proofs include, e.g., congruence computations, Minkowski's theorem, the pigeon hole principle, properties of Gaussian integers, continued fractions, and so on. The book by Hardy and Wright [19] gives several different proofs. For other proofs, see also [6, 33, 40].

A very different second type of proof goes back to Liouville. In a series of eighteen papers, Liouville describes a quite general method, a special case of which gives Theorem 2. Liouville's work is described in the books by Bachmann [3], Dickson [8], Uspensky and Heaslet [34], Venkov [36], and Nathanson [29].

This special case was considerably simplified by Heath–Brown [20]. Zagier [41] reformulated Heath–Brown's proof to write it in one sentence, however, leaving elementary calculations to the reader.

This proof has generated a considerable literature explaining the proof for teaching purposes [5, 12, 31, 35, 39] or extending it to related results: [4, 13, 14, 16, 21–23, 32]. The collection of beautiful proofs "Proofs from the BOOK" by Aigner and Ziegler [1] explains in its first edition Zagier's version of the proof, but changed to Heath–Brown's version for the 2nd edition.

A key ingredient is an ingenious choice of a set which allows a partition into orbits of length 1 or 2. In this way, a simple parity check guarantees the decomposition into two squares. The reader who is familiar with Liouville's method will appreciate the simplifications made by Heath–Brown and Zagier. Still, the proof is quite mysterious. We make an attempt to demystify the proof, i.e., explain how the details can be motivated.

In addition to the study of this second type of proof, we apply the idea of orbits of length 1 or 2 to a proof based on lattice points, which is more in the spirit of the first type of proof. After reviewing the history of these, i.e., discuss contributions by Lucas, Grace, and others we present in Sect. 1.6 a quite short version of the proof, which admittedly also requires some routine checking, as is the case with the proofs by Zagier and Heath–Brown.

1.2 Zagier's Proof

Here is the famous one-sentence-proof for primes $p = 4k + 1$, quoting from Zagier [41].

"The involution on the finite set $S = \{(x, y, z) \in \mathbb{N}^3 : x^2 + 4yz = p\}$ defined by

$$(x, y, z) \mapsto \begin{cases} (x + 2z, z, y - x - z) & \text{if } x < y - z \\ (2y - x, y, x - y + z) & \text{if } y - z < x < 2y \\ (x - 2y, x - y + z, y) & \text{if } x > 2y \end{cases}$$

has exactly one fixed point, so $|S|$ is odd and the involution defined by $(x, y, z) \to (x, z, y)$ also has a fixed point." □

Quite a few routine checks are necessary to verify all these implicit claims. For the reader's ease, we would like to add that the first map, α (say), defines a partition $S = S_1 \cup S_2 \cup S_3$ with $S_1 = \{(x, y, z) \in S : x < y - z\}, S_2 = \{(x, y, z) \in S : y - z < x < 2y\}, S_3 = \{(x, y, z) \in S : x > 2y\}$. There are no solutions with $y - z = x$ or $x = 2y$, since otherwise $x^2 + 4yz$ is not a prime. Solutions with $x < y - z$ are mapped to solutions with $x > 2y$, and vice versa. Solutions with $y - z < x < 2y$ are mapped to solutions with the very same property. That is $\alpha(S_1) = S_3, \alpha(S_3) = S_1, \alpha(S_2) = S_2$. Thus, fixed points of α must lie in S_2 and therefore satisfy $(x, y, z) = (2y - x, y, x - y + z)$, i.e., $x = y$. Since p is prime, the only fixed point is $(1, 1, (p - 1)/4)$.

Writing out all details, which we do not do here, makes the proof actually quite a bit longer.

1.3 Heath–Brown's Proof

Heath-Brown reformulated Liouville's work in 1971. His version [20] appeared in 1984 in a student magazine, issued by the undergraduate mathematical society at Oxford University. Meanwhile a retyped version is available, see the bibliography. Since Heath–Brown's proof was slightly different, we describe his proof briefly.

Let us define

$$X_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 2 & -1 \end{pmatrix}.$$

Define the sets

$$S = \{(x, y, z) \in \mathbb{Z}^3 : p = 4xy + z^2, \quad x, y > 0\},$$

$$T = \{(x, y, z) \in S : z > 0\}, \quad U = \{(x, y, z) \in S : x + z > y\}.$$

One can check that $X_1^2 = X_2^2 = X_3^2 = I$. Moreover, X_1 maps S to itself, X_2 maps T to itself, and X_3 maps U to itself. One also verifies that $|T| = |X_1 T|$ and $|U| = |X_1 U|$. Since S is the disjoint union of T and $X_1 T$, it follows that $|S| = |T| + |X_1 T| = 2|T|$ and similarly $|S| = 2|U|$. This implies $|T| = |U|$. Since the map X_3 acting on U has exactly one orbit of length 1 (for $y = z = 1$), and since all other orbits have length two, we find that $|U|$ must be odd. So, $|T|$ is also odd, and the action of X_2 on T must have an orbit of length 1, i.e., there is a fixed point with $x = y$, giving $p = 4x^2 + z^2$.

This is an impressive example that the right choice of a set, group action and orbit counting can simplify existing proofs. Another example of this principle is McKay's proof [28] of a Theorem of Cauchy in group theory.

1.4 Grace' Lattice Point Proof

In this section, we describe a proof based on lattice points, due to Grace [17]. It is one of the proofs in Hardy and Wright's book [19].

The proof starts with the fact that $a^2 \equiv -1 \bmod p$ has a solution. Take those lattice points in $\mathbb{Z} \times \mathbb{Z}$ with $ax \equiv y \bmod p$. Note that if (x, y) and (x', y') belongs to the set, then also $(x \pm x', y \pm y')$ belong to it, so that the set of these points define a discrete lattice. Let $P_1 = (x, y)$ be one of the points with minimal distance to the origin $P_0 = (0, 0)$. Since $-ay \equiv x \bmod p$, the point $P_2 = (-y, x)$ also belongs to the lattice. These points together with $P_3 = (x - y, x + y)$ define the fundamental domain. Observe that there are no further lattice points in this fundamental domain, since otherwise the distance from $(0, 0)$ to (x, y) was not minimal. Also observe that in this situation the fundamental domain is not only a parallelogram, but even a square.

In a very large circle about the origin, the proportion of points belonging to the lattice is $\frac{1}{p}$ so that the area of the fundamental domain is p. Hence the side lengths of the square satisfy by Pythagoras' theorem: $x^2 + y^2 = p$.

The lattices can also be understood as coming from the problem of regular solutions of placing p non-taking queens on a $p \times p$ chessboard, with reduction modulo p, i.e., a chessboard on a torus. This approach has been studied by Polya [30], Kraitchik [24] and Larson [25]. These proofs also make use of counting the lengths of orbits and are similar in spirit to those discussed below.

1.5 Lucas' Work on Regular Satins

In 1867, Édouard Lucas [26] had similar ideas on regular "Satin" squares which were thought of in connection with patterns of fabrics. As Decaillot [7] writes, in France at that time there was a group of mathematicians writing as accessible as possible for a wide audience.

Without assuming that there is a solution of $a^2 \equiv -1 \bmod p$, he considered those integer lattices with slopes $2, 3, \ldots, \frac{p-1}{2}$. He paired off those lattices with slopes s_i and s_j where $s_i s_j \equiv \pm 1$. For a given s_i there is a unique s_j in this set. He interpreted this in terms of the geometric pattern. Starting with an odd number of lattices, one lattice remains. This remaining lattice is associated to itself, and has a square unit.

In this paper, Lucas did not actually conclude the two squares theorem, namely that a prime $p \equiv 1 \bmod 4$ is a sum of two squares, but rather the opposite.

The reason for this apparently comes from the historical background. The question, for which moduli regular lattices exist, was asked by Édouard Gand, also in 1867, in connection with fabric patterns, and Gand's question was answered by Lucas.

However, there is some indirect evidence that Lucas later actually proved Theorem 2 using this method. Dickson [8] (Volume 2, page 245) gives [27] (which does not contain that proof) and Aubry [2] as references. Decaillot [7] mentions a comment by Aubry in Fermat's collected works [15] (note 27 of the 4th volume). Here, Aubry writes that the two squares theorem is "perhaps the most beautiful of all of Fermat's theorems", and Aubry refers to a graphical proof by Lucas.

Decaillot [7] constructed a proof that possibly was the one given by Lucas. It is very similar to the proof by Grace discussed above.

1.6 A Short Proof

In this section we aim to modify the two approaches above to assemble a proof which can be formulated in one sentence. However, as is the case with Zagier's proof, several additional words of explanations are appropriate, and several routine calculations required. The author believes that memorizing this proof may be easier than memorizing Zagier's proof.

1.6.1 The Long Version

- Let $p \equiv 1 \bmod 4$ be a prime and let $S = \left\{2, 3, \ldots, \frac{p-1}{2}\right\}$. For $z \in S$, let us define the lattices

$$L_z = \{(x \bmod p,\ zx \bmod p) : 0 \le x < p\}$$

as subsets of $\mathbb{Z}_p \times \mathbb{Z}_p$ (which can be thought of as a torus). To see that these are lattices take any two points $(x \bmod p,\ xz \bmod p)$ and $(y \bmod p,\ yz \bmod p)$. The sum $(x + y \bmod p, (x + y)z \bmod p)$ is again in S_z and the same follows for integer multiples $(\lambda x \bmod p, \lambda xz \bmod p)$.
For $p \equiv 1 \bmod 4$, the number $|S| = \frac{p-1}{2} - 1$ of lattices is odd. For a better understanding, we draw these for $p = 17$ (Figs. 1–7).

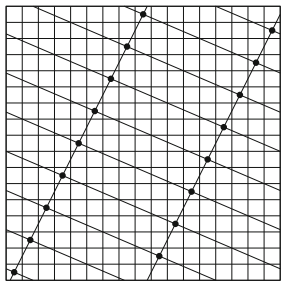


Figure 1 Lattice L_2

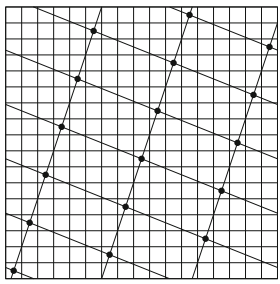


Figure 2 Lattice L_3

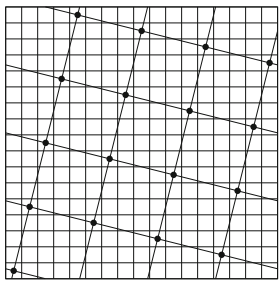


Figure 3 Lattice L_4

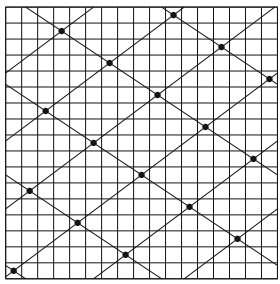


Figure 4 Lattice L_5

Figure 5 Lattice L_6

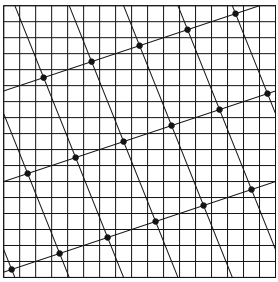


Figure 6 Lattice L_7

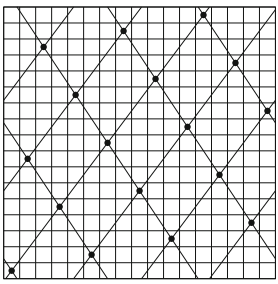
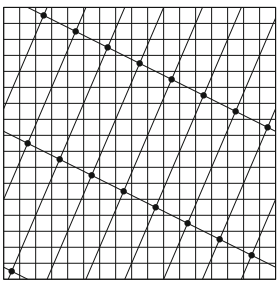


Figure 7 Lattice L_8



In the pictures, we include the parallelograms which define fundamental domains of the lattices. A fundamental domain is a parallelogram, spanned by a point and two of its 4 closest neighbours in two linear independent directions. In this sense, each point uniquely corresponds to a fundamental domain, so that there are p fundamental domain, and for a given lattice all of these parallelograms are congruent, understood modulo p.

- But the fundamental domains for different lattices are in general not congruent to each other. In the above example with $p = 17$, the shape of the fundamental domain is the same for L_2 and L_8, for L_3 and L_6, for L_5 and L_7. The lattice L_4 (which turns out to deliver the solution $x^2 \equiv -1 \bmod 17$ and finally the decomposition of $17 = 1^2+4^2$) does not have a corresponding partner. Generally this can be described by means of the following map: Let $S = \{2 \le a \le \frac{p-1}{2}\}$. Let $f : S \to S$ with

$$a \mapsto \begin{cases} a^{-1} \bmod p & \text{if } 2 \le (a^{-1} \bmod p) \le \frac{p-1}{2}, \\ -a^{-1} \bmod p & \text{otherwise.} \end{cases}$$

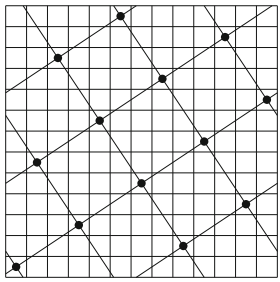


Figure 8 L_z with $z^2 \equiv -1 \bmod p$ being a fixed point, here $p = 13, z = 5$

Observe that for $p = 17$ one has that $f(2) = 8, f(8) = 2, f(3) = 6, f(6) = 3, f(5) = 7, f(7) = 5, f(4) = 4$. Here the representatives of the residue classes modulo p are assumed to be in the interval $0 \leq b < p$. It can be easily checked that f is an involution. We have to show that for all $a \in S$: $f(f(a)) = a$. If the first alternative holds for the inner argument, then also at the second time so that $f(f(a)) = f(a^{-1}) = (a^{-1})^{-1} = a$ and similarly $f(f(a)) = f(-a^{-1}) = -(-a^{-1})^{-1} = a$. Since $|S|$ is odd, there must be an odd number (i.e., at least one) of elements with $a = f(a)$. Since $-1, 1 \notin S$, it follows that $(a+1)(a-1) \equiv 0 \bmod p$ has no solution in S which implies that $a \equiv a^{-1} \bmod p$ has no solution. But then there must be an element with $a \equiv -a^{-1} \bmod p$. It is this element which satisfies $a^2 \equiv -1 \bmod p$, but we better leave it as $a \equiv -a^{-1} \bmod p$. In this form we see that the slopes a and $-a^{-1}$ of the sides of the parallelogram are orthogonal. The lattice is invariant under the map f which means it is invariant under a rotation by 90°. This proves why for prime $p \equiv 1 \bmod 4$ there must be a lattice amongst the lattices L_z, of which the fundamental domain is a square (Fig. 8).

- Since the fundamental domains are defined by a point and its closest neighbours, the fundamental domains do not contain any lattice point in their interior. Thus the fundamental domains cover the $p \times p$ board without overlap. Since for each of the p points there is exactly one fundamental domain, its area is $\frac{p^2}{p} = p$, so that the length of a side is $\sqrt{p}$. An alternative argument here could be the one by Grace [17].
- Finally, an application of Pythagoras' theorem to the grid decomposition of the base side of the square shows that $p = (\sqrt{p})^2 = a^2 + b^2$ holds.

It seems particularly pleasant that we did not explicitly need the solution of $a^2 \equiv -1 \bmod p$, but could rather directly conclude from $a \equiv -a^{-1} \bmod p$ that the parallelogram is a square.

1.6.2 A Short Version of the Proof

Having said all this, the reader can see that the following one sentence version of the proof, written in the spirit of Zagier's proof [41], contains essentially all the necessary information and is perhaps easier to work with, or memorize, than other

proofs of this theorem. The amount of hidden routine checking may be comparable with that in Heath-Brown's or Zagier's version.

The involution on the finite set $S = \left\{2 \le a \le \frac{p-1}{2}\right\}$ defined by

$$a \mapsto \begin{cases} a^{-1} \bmod p & \text{if } 2 \le (a^{-1} \bmod p) \le \frac{p-1}{2}, \\ -a^{-1} \bmod p & \text{otherwise,} \end{cases}$$

has at least one fixed point z, so the fundamental domain of the lattice defined by

$$L_z = \{(x,\ zx \bmod p), 0 \le x < p\}$$

is a square with area p, so that the two squares theorem follows by an application of Pythagoras' theorem. □

2 How Zagier's Involution can be Motivated

We will give two explanations, how Zagier's map can be motivated. One was found by the present author, and was described in [12–14]. We will show that this approach gives a method to search systematically for proofs of related theorems on quadratic forms.

An alternative motivation can be found in lecture notes by E.W. Dijkstra.

2.1 First Motivaton

It is possible to *construct* the "complicated" involution by means of some fairly easy assumptions, (see also [12]). These assumptions ensure that the final mapping would be as simple as possible.

If we look for a mapping that

I) can be described by a matrix $B = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$, with integer entries which are independent of k, (linearity),

II) maps the solutions (in positive integers) of $p = 4k + 1 = x^2 + 4yz$ onto such solutions, (invariance),

III) has the easiest solution, namely $(1, 1, k)$ as its only fixed point, (simplicity),

then we are uniquely led to $B = \begin{pmatrix} -1 & 2 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$.

This can be seen as follows: Property (III) gives

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ k \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 1 \\ 1 \\ k \end{pmatrix}.$$

In particular, $a + b + ck = 1$. But since the coefficients are supposed to be independent of k, we have that

$$\begin{aligned} a + b + ck = 1 &\Rightarrow c = 0, \;\; a + b = 1 \\ d + e + fk = 1 &\Rightarrow f = 0, \, d + e = 1 \\ g + h + ik = k &\Rightarrow i = 1, \;\; g + h = 0. \end{aligned}$$

Property (II) gives

$$(x')^2 + 4y'z' = (ax + by + cz)^2 + 4(dx + ey + fz)(gx + hy + iz) \stackrel{!}{=} x^2 + 4yz.$$

Hence, a comparison of the coefficients shows that

$$\begin{array}{ll} x^2 & a^2 + 4dg = 1 \\ xy & 2ab + 4(dh + eg) = 0 \\ \vdots & \vdots \\ yz \text{ with } c = f = 0, i = 1 : & ei = 1 \Rightarrow e = 1, \Rightarrow d = 0. \end{array}$$

Now $a^2 + 4dg = 1$ is simplified to $a^2 = 1$. Suppose that $a = 1$. Then $b = 0$ and from $2ab + 4(dh + eg) = 0$ we get $g = 0$ and finally $h = 0$. Then, the matrix would be the identity matrix I. This is not what we want, since the map shall have only one fixed point.

Thus, $a = -1$, and so $b = 2, g = 1$, and finally $h = -1$. So, we have found the matrix B.

Surprisingly, we did not even need that our map shall be an involution but we can readily check that $B^2 = I$.

This only works for $-x + 2y > 0$ and $x - y + z > 0$. For the other cases one apparently needs a different matrix. Let us see how we can manipulate B to yield a corresponding row condition $x - 2y > 0$. We look for a matrix X which turns the row conditions of B into $(1, -2, 0)$ and $(1, -1, 1)$. Let $X = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. The matrices $A = BX$ and $C = XB$ cover all cases. Let $A = BX = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ -1 & 1 & -1 \end{pmatrix}$ and $C = XB = \begin{pmatrix} 1 & -2 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. We see that the row conditions perfectly fit to each other and induce a partition of all solutions.

Alternatively, one can find these matrices A and C by choosing small primes ($p = 13, 17, 29$) and observing that here the sets of solutions with $-x + 2y < 0$ or $x - y + z < 0$ only have one or two elements. For $p = 13$, we find that $(1, 3, 1)$ must be mapped to $(3, 1, 1)$ and vice versa. For $p = 17$, we find that $(1, 4, 1)$ must be mapped to $(3, 1, 2)$ and vice versa. For $p = 29$, there are two possibilities. One excludes by the partially known mapping that $(1, 7, 1)$ is mapped to $(5, 1, 1)$ and finds that $(1, 7, 1)$ is mapped to $(3, 1, 5)$, from which A and C uniquely follow.

Even though we did not know about the partition of S into three sets, we have found the map $\alpha : S \to S$ with

$$\alpha = \begin{cases} \alpha_1 & \text{decribed by matrix A, if } -x + y - z > 0 \\ \alpha_2 & \text{described by matrix B, if } -x + 2y > 0 \text{ and } x - y + z > 0 \\ \alpha_3 & \text{decribed by matrix C, if } x - 2y > 0 \text{ and } x - y + z > 0. \end{cases}$$

This is precisely the mapping given by Zagier. Of course, α as a whole is not a linear map, so that property (I) is not strictly satisfied. We obtain in this way the easiest involution, α, with the required property, namely that we know the set of fixed points.

Let us remark that the intersection into three subsets was caused since we work with positive x, y, z. In Heath–Brown's version negative values are allowed, and so he did not need this division into three cases.

Zagier's second mapping, β (say), with $\beta : S \to S$ and $(x, y, z) \mapsto (x, z, y)$ corresponds to the matrix

$$Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

2.2 Making the Proof Constructive

In his paper, Zagier mentioned the proof only shows the existence of the solution. Combining the two involutions α and β, we can give a constructive proof. Starting with the only fixed point of α, and iterating $\beta, \alpha \ldots$ we must arrive at a period.

$$(1, 1, k) \xrightarrow{\beta} (1, k, 1) \xrightarrow{\alpha} (3, 1, k-2) \xrightarrow{\beta} \cdots \xrightarrow{\beta} (3, 1, k-2) \xrightarrow{\alpha} (1, k, 1) \xrightarrow{\beta} (1, 1, k).$$

Since the maps are bijective, there is no preperiod. So, we eventually come back to $(1, 1, k)$ with β. The number of elements in the period is even. By symmetry, there must be another fixed point in the middle of the cycle. Since there is only one fixed point of α, this iteration constructs a fixed point of β, that is a solution of $p = x^2 + 4y^2$.

Applying this algorithm to a composite nonsquare integer $n = 4k + 1$ the very same argument shows that any cycle containing $(1, 1, k)$ must also contain another fixed point. Since n is no longer prime we may well come to another fixed point of α which corresponds to a factorization of n. To see that this can happen, let us concentrate on products of two distinct primes $n = p_1 p_2$ with $p_1 \equiv p_2 \equiv 3 \bmod 4$. Here β does not have a fixed point, since n cannot be written as a sum of two squares. Hence, in this case the iteration $\beta, \alpha, \beta \ldots$ must eventually come to another fixed point of α which corresponds to $x = y$, i.e., a factorization of n.

This algorithm for finding the decomposition into 2 squares is very slow. For some details, see Bagchi [4]. Shiu [32] describes how one can accelerate this algorithm. It turns out to have an interpretation in the theory of continued fractions. A fast algorithm is described by Wagon [37].

2.3 A Motivation Due to Dijkstra

A different, and very elegant derivation of Zagier's map was also given by Dijkstra [9]. His notes are written in the language of a computer scientist and are extraordinary detailed. I will try keeping the flavor of his exposition, but will have to shorten his account. After some general remarks on involutions Dijkstra concludes that to write p as a sum of two integer squares it is enough to look at

$$(x, y) : x^2 + 4y^2 = p. \qquad (*)$$

In order to establish the desired correspondence between solutions of this equation and the fixed points of an involution, "we do something with which every computer scientist is very familiar: replacing in a target relation" $(*)$ "something by a fresh variable". Dijkstra refers to "Leibniz' principle" (informally: substituting equals for equals) to rewrite $(*)$ as

$$(x, y, z) : x^2 + 4yz = p \text{ and } y = z.$$

Let $S = \{(x, y, z) : x, y, z \in \mathbb{N} : x^2 + 4yz = p\}$. Exploiting the symmetry in y and z, Dijkstra chooses a first involution inv_0 by $S \to S : (x, y, z) \mapsto (x, z, y)$. The fixed points of inv_0 satisfy $y = z$. Hence it is enough to show that inv_0 has at least one fixed point. In order to do this one intends to construct a second involution inv_1 on S, which has exactly one fixed point.

Next, Dijkstra gathers some elementary facts:
$x > 0, y > 0, z > 0, x \neq \pm(y - z)$, since p is odd and not a square.

Next, "can we think of operators on (x, y, z) for which $x^2 + 4yz = p$ is an invariant", i.e., an operator which maps solutions of S onto such solutions?

Dijkstra then studies operators of the type

$$(x, y, z) \mapsto (x + \Delta x, y + \Delta y, z + \Delta z).$$

Here Dijkstra implicitly assumes that Δ is an operator, for which

$$\Delta f(x) = f(x + \Delta x) - f(x)$$

so that for example $\Delta(x^2) = (x + \Delta x)^2 - x^2 = 2x\Delta x + (\Delta x)^2$.

Since $\Delta x = 0$ would too easily lead back to inv_0, he assumes $\Delta x \neq 0$. Since for all elements of S, x is odd, Δx is even, so that $\Delta x = 2b$, say.

The invariance assumption $\Delta : S \to S$, i.e., $(x')^2 + 4y'z' = p$ means that

$$\Delta(x^2 + 4yz) = 0.$$

So,

$$\begin{aligned}
\Delta(x^2 + 4yz) &= 0 \\
\Delta(x^2) &= -4\Delta(yz) \\
2x(\Delta x) + (\Delta x)^2 &= -4\left((y + \Delta y)(z + \Delta z) - yz\right) \\
b(x + b) &= -y\Delta z - z\Delta y - \Delta y\Delta z.
\end{aligned}$$

In order to simplify this expression Dijkstra chooses $\Delta y = 0$ and arrives at

$$b(x + b) = -y\Delta z.$$

He remarks that this choice does not restrict the generality, since one could arrive at any "move" with $\Delta y \neq 0, \Delta z \neq 0$ by means of two single moves.

Now, the last equation suggests the following four possibilities:

1. $b = -y, x + b = \Delta z$, giving $(x, y, z) \mapsto (x - 2y, y, z + x - y)$
2. $b = y, x + b = -\Delta z$, giving $(x, y, z) \mapsto (x + 2y, y, z - x - y)$
3. $b = \Delta z, x + b = -y$, giving $(x, y, z) \mapsto (-x - 2y, y, z - x - y)$
4. $b = -\Delta z, x + b = y$, giving $(x, y, z) \mapsto (2y - x, y, z + x - y)$

In order to satisfy the invariance of $x > 0, y > 0, z > 0$, one sees that the third case above with $x' = -x - 2y$ can be discarded from consideration.

So far, we have not yet used the fact inv_1 is supposed to have exactly one fixed point. Now, for a fixed point $(x, y, z) = (x', y', z')$. Here $x = x'$ and $y > 0$ mean that the only remaining case is the 4th case above. Here $x = 2y - x$ shows that a fixed point can only occur if $x = y$ so that $p = x^2 + 4yz = x(x + 4z)$ implies that $z = \frac{p-1}{4}$, giving the unique fixed point $\left(1, 1, \frac{p-1}{4}\right)$.

Dijkstra then completes the construction of the involution inv_1 for those solutions for which $y > z + x$ or $x > 2y$, respectively.

2.4 Comparison

Comparing both constructions in Sects. 2.1 and 2.3, it can be observed that the principle to keep the construction as simple as possible, but also as general as necessary is quite successful. While in my motivation in Sect. 2.1, the choice of the fixed point

$(1, 1, k)$ quickly led to the entries $c = f = 0, i = 1$ of the matrix B, and then the invariance of the quadratic form delivered the additional entries. Dijkstra's choice of $\Delta y = 0$, in the language of Sect. 2.1 quickly led to $d = f = 0, e = 1$, and then the invariance of the form and consideration of the fixed point completed the entries of B.

Let us finally ask: is there any application (other than the two squares theorem itself) of the fact discovered by this combinatorial proof that the number of solutions (x, y, z) of a given type (say for $p = x^2 + 4yz$ with $x < y - z$) equals the number of solutions of another type (say here $x > 2y$)? If so, that could be of interest also for the generalizations considered below.

3 Generalization of the Method

One can ask for similar involutions α for related question on $p = sx^2 + tyz$, where s and t are fixed constants. For example, it is well known that for a prime p the following holds

$$p \equiv 1, 3 \bmod 8 \Leftrightarrow p = x^2 + 2y^2 \text{ in positive integers.}$$

It would be nice to have an easy proof of this theorem by the idea of the Heath-Brown—Zagier proof.

Such generalizations were found by the current author in 1996, see [13], and also by Jackson [21–23] and Generalov [16].

Here we shall derive the following results:

Theorem 3. *Let p denote a prime.*

(a) For $p = 8k + 3$ there is a solution of $p = x^2 + 2y^2$ in positive integers.
(b) For $p = 8k + 7$ there is a solution of $p = x^2 - 2y^2$ in positive integers.
(c) For $p = 8k + 5$ there is a representation as $p = x^2 + y^2$. (A new proof!)

Theorem 4. *Let p denote a prime.*

(a) For $p = 12k + 7$ there is a solution of $p = 3x^2 + 4y^2$ in positive integers.
(b) For $p = 12k + 11$ there is a solution of $p = 3x^2 - 4y^2$ in positive integers.

Generalizing the approach of Sect. 2.1 one can prove that the matrix $B = \begin{pmatrix} -1 & 2\frac{m}{n} & 0 \\ 0 & 1 & 0 \\ 4\frac{sm}{tn} & -4\frac{sm^2}{tn^2} & 1 \end{pmatrix}$ maps solutions of $p = sx^2 + tyz$ to such solutions and has the fixed point (m, n, k'). Here m, n, s, and t are fixed nonnegative integers. So $k' = \dfrac{p - sm^2}{tn}$. We note that again $B^2 = I$. Unfortunately, in the general case the boundaries induced by the rows, namely $-x + 2\frac{m}{n} > 0$ and $4\frac{sm}{tn}x - 4\frac{sm^2}{tn^2}y + z > 0$, do not induce such a balanced three-partition of the set of solutions.

However, it is possible to construct mappings for $p = x^2 + 2y^2$ and $p = 3x^2 + 4y^2$ consisting of even more matrices. As before, these matrices are generated by B and X.

Note, even though the occurring matrices will be more complicated, the idea of the proof is still the same. The justification of the properties of the map α can -in principle- be left to an automatic system since it requires elementary calculations only.

As before, we try, if $A = BX$ can be useful. As above we use $X = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

At this point, we do not worry about the boundaries or a partition of the set of all solutions.

Geometrically, we can expect that $|\det A| = 1$, since we should not map bijectively a large region to a small one and vice versa.

Consider the eigenvalues of

$$A = BX = \begin{pmatrix} -1 & 2\frac{m}{n} & 0 \\ 0 & 1 & 0 \\ 4\frac{sm}{tn} & -4\frac{sm^2}{tn^2} & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 2\frac{m}{n} \\ 0 & 0 & 1 \\ -4\frac{sm}{tn} & 1 & -4\frac{sm^2}{tn^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & a \\ 0 & 0 & 1 \\ -c & 1 & -d \end{pmatrix}, \text{ say.}$$

Noting that $ac = 2d$ we find

$$\begin{aligned} 0 &= (1-\lambda)(0-\lambda)(-d-\lambda) - (1-\lambda) - (-c)(0-\lambda)a \\ &= (\lambda+1)(\lambda^2 + (d-2)\lambda + 1). \end{aligned}$$

We find that $\lambda_1 = -1$ and $\lambda_{2,3} = -\frac{d-2}{2} \pm \sqrt{\left(\frac{d-2}{2}\right)^2 - 1}$.

For integers $d \geq 5$ or $d \leq -1$, the values of $\lambda_{2,3}$ are real but irrational numbers. So the order of A is infinite, and there is little hope of finding a suitable map consisting of finitely many parts. So $d = 0, 1, 2, 3, 4$ and in these cases $|\lambda_1| = |\lambda_2| = |\lambda_3| = 1$. This justifies our expectation that $\det A = 1$.

Recall that $d = \dfrac{4sm^2}{tn^2}$. Since we want to represent primes with $p = sx^2 + tyz = sm^2 + tnz$ we may assume that $\gcd(sm, tn) = 1$. We shall systematically consider all cases.

3.1 $d = 0$

For $d = 0$ we have $sm = 0$, so that $p = tnz$. This case is of no interest.

3.2 $d = 1$

Here $d = \dfrac{4sm^2}{tn^2} = 1$, and $(s,t) = (s,n) = (t,m) = (m,n) = 1$. Hence there are two possibilities:

- $s = m = n = 1$, $t = 4$. This is precisely the case of Heath–Brown's and Zagier's proof.
- $s = m = t = 1$, $n = 2$.

  In $p = x^2 + yz$ the solution with $p = x^2 + y^2 = y^2 + x^2$ is counted twice. In order to make the original argument work we need to break the symmetry. This can be done by assuming y and z to be even.

  The involution α is generated by

$$B = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix},$$

  $A = BX$, and $C = A^{-1}$.

  This gives the following variant of the proof of the two squares theorem:

  The involution on the finite set $S = \{(x,y,z) \in \mathbb{N} \times 2\mathbb{N} \times 2\mathbb{N} : x^2 + yz = p\}$ defined by

$$(x,y,z) \mapsto \begin{cases} (x+z, z, -2x+y-z) & \text{if } 2x+z < y \\ (-x+y, y, 2x-y+z) & \text{if } x < y < 2x+z \\ (x-y, 2x-y+z, y) & \text{if } y < x \end{cases}$$

  has exactly one fixed point, so $|S|$ is odd and the involution defined by $(x,y,z) \to (x,z,y)$ also has a fixed point.

3.3 $d = 2$

3.3.1 The Case $p = x^2 + 2yz$

Here, we consider the case $d = \dfrac{4sm^2}{tn^2} = 2$. By the coprime condition $(sm^2, tn^2) = 1$, we necessarily have that $s = m = n = 1$, $t = 2$.

Empirically one observes that the number of fixed points varies with the residue classes modulo 8:

(a) primes $p \equiv 3 \bmod 8$ induce 1 fixed point.
(b) primes $p \equiv 7 \bmod 8$ induce 2 fixed points.
(c) primes $p \equiv 5 \bmod 8$ induce 2 fixed points.
(d) primes $p \equiv 1 \bmod 8$ induce 3 fixed points.

Case (a) was also proved by Jackson [21] and Generalov [16]. They also observed (d), but did not prove it by elementary methods. We shall prove (a), (b), and (c) which corresponds to our Theorem 3 (a,b,c). Unfortunately, we do not see either a convenient way to prove (d) without appealing to the theory of quadratic forms.

Let $S = \{(x, y, z) \in \mathbb{N}^3 : x^2 + 2yz = p\}$. The one sentence proof is as before with the following map $\alpha : S \to S$.

$$\alpha = \begin{cases} A = BX & = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ -2 & 1 & -2 \end{pmatrix} \text{ if } -2x + y - 2z > 0 \\ E = -XA^2 & = \begin{pmatrix} -3 & 2 & -2 \\ -2 & 2 & -1 \\ 2 & -1 & 2 \end{pmatrix} \begin{cases} \text{if } -3x + 2y - 2z > 0 \\ \text{and } 2x - y + 2z > 0 \\ (\text{then } -2x + 2y - z > 0 \text{ is implied.}) \end{cases} \\ D = -A^2 & = \begin{pmatrix} 3 & -2 & 2 \\ 2 & -1 & 2 \\ -2 & 2 & -1 \end{pmatrix} \begin{cases} \text{if } 3x - 2y + 2z > 0 \\ \text{and } -2x + 2y - z > 0 \\ (\text{then } 2x - y + 2z > 0 \text{ is implied.}) \end{cases} \\ B = XA^3 & = \begin{pmatrix} -1 & 2 & 0 \\ 0 & 1 & 0 \\ 2 & -2 & 1 \end{pmatrix} \text{ if } -x + 2y > 0 \text{ and } 2x - 2y + z > 0 \\ C = A^{-1} = A^3 = XB & = \begin{pmatrix} 1 & -2 & 0 \\ 2 & -2 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{array}{l} \text{if } x - 2y > 0, \\ (2x - 2y + z > 0 \text{ follows trivially.}) \end{array} \end{cases}$$

Note that this map makes use of all matrices of the form $(-1)^{j+1}A^j$, $(j=1, 2, 3)$, and $(-1)^{j+1}XA^j$, $(j = 2, j = 3)$. Note that also $A^4 = I$. The matrix XA is of no use, since this contains an impossible row condition $-x - 2z > 0$.

The map above is equivalent to that given by Jackson and Generalov, here in Jackson's notation [21]:

$$(x,y,z)\mapsto\begin{cases}(x-2y,z+2x-2y,y) & \text{if } y<\frac{x}{2}\\ (2y-x,y,2x-2y+z) & \text{if } \frac{x}{2}<y<x+\frac{z}{2}\\ (3x-2y+2z,2x-y+2z,-2x+2y-z) & \text{if } x+\frac{z}{2}<y<\frac{3}{2}x+z\\ (-3x+2y-2z,-2x+2y-z,2x-y+2z) & \text{if } \frac{3}{2}x+z<y<2x+2z\\ (x+2z,z,-2x+y-2z) & \text{if } 2x+2z<y.\end{cases}$$

Let us call the subsets of S that correspond to the matrices A, B, C, D, E by $\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D}, \mathscr{E}$. For a complete proof, we have to show that

1. $\alpha : S \to S$, i.e., α maps (x, y, z) with $p = x^2 + 2yz$ to (x', y', z') with $p = x'^2 + 2y'z'$,
2. $\alpha^2 = id$,
3. the boundaries ($x - 2y = 0, 2x - 2y + z = 0$ etc.) are never attained,
4. the sets $\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D}, \mathscr{E}$ induce a partition of the set of all solutions,
5. there is only one fixed point.

3.3.2 Proof of Theorem 3

1. Since all parts of the mapping α are generated by $-I, X$ and B it suffices to prove the first property for $-I, X$ and B. It is obvious for $-I$ and X. For B we have:

$$(x')^2 + 2y'z' = (-x + 2y)^2 + 2(y)(2x - 2y + z) = x^2 + 2yz = p.$$

2. Note that A maps the region $\mathscr{A}$ to the region $\mathscr{C}$. Because of $C = A^{-1}$ the region $\mathscr{C}$ is mapped to the region $\mathscr{A}$. The first assertion follows from $x' - 2y' = (x + 2z) - 2z > 0$ and $2x' - 2y' + z' = 2(x + 2z) - 2z + (-2x + y - 2z) = y > 0$. For the second assertion, we need that $-2x' + y' - 2z' > 0$ with $x' = x - 2y, y = 2x - 2y + z, z' = y$ and so $-2x' + y' - 2z' = z > 0$. Note that $B^2 = D^2 = E^2 = I$. So the matrix B maps the set $\mathscr{B}$ onto $\mathscr{B}$. The same holds for $D : \mathscr{D} \to \mathscr{D}$ and $E : \mathscr{E} \to \mathscr{E}$.
3. Suppose the boundaries are attained. This will lead to a contradiction.
   a. For the boundaries in the first row, namely $x - 2y = 0, -x + 2y = 0, 3x - 2y + 2z = 0, -3x + 2y - 2z = 0$, it would follow that x is even. This contradicts $p = x^2 + 2yz$, since p is odd.
   b. $-2x + y - 2z = 0$: $p = x^2 + 2yz = x^2 + 2(2x + 2z)z = (x + 2z)^2$. This contradicts the primality of p.

c. $2x - 2y + z = 0$: $p = x^2 + 2yz = x^2 + 2y(2y - 2x) = (x - 2y)^2$, contradicting the primality of p.
d. $2x - y + 2z = 0$: See (b).
e. $-2x + 2y - z = 0$: See (c).

4. It follows easily from the boundaries in Jackson's notation (given above) that α induces a partition of S.
5. We now look for the fixed points of α. Here we distinguish between the various cases depending on the residue class modulo 8. We see that A and C cannot have fixed points, since the set $\mathscr{A}$ is mapped onto $\mathscr{C}$ and the other way around. Suppose that (x, y, z) is a fixed point of B, then

$$B\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -x + 2y \\ y \\ 2x - 2y + z \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Hence $x = y$. Because of $p = x^2 + 2yz$ this is only possible for $x = y = 1$. Hence B has precisely one fixed point.

For the matrix D, we find that

$$D\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3x - 2y + 2z \\ 2x - y + 2z \\ -2x + 2y - z \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Hence $y = x+z$, and therefore, $p = x^2+2yz = x^2+2(x+z)z = (x+z)^2+z^2$.

Similarly,

$$E\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -3x + 2y - 2z \\ -2x + 2y - z \\ 2x - y + 2z \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Hence $y = 2x + z$, which implies that $p = x^2 + 2yz = x^2 + 2(2x + z)z = (x + 2z)^2 - 2z^2$.

If $p \equiv 3 \bmod 8$, there is no fixed point coming from D and E. To see this, recall that squares modulo 8 only take the values $0, 1, 4$. So, the only fixed point is in $\mathscr{B}$, and so $|S|$ is odd. As before, the involution β must have an odd number of fixed points. Hence there is at least one fixed point with $y = z$, leading to the solution of $p = 8k + 3 = x^2 + 2y^2$.

The same consideration of the values of squares modulo 8 shows:

If $p \equiv 7 \bmod 8$, we have again the trivial fixed point of B. There cannot be a fixed point from D. Since there cannot be a representation $p = x^2+2y^2$, we see that there must be a fixed point coming from E. So, $p \equiv 7 \bmod 8$ can be written as $p = x^2 - 2y^2$. This proves theorem 3b).

If $p \equiv 5 \bmod 8$, there cannot be a fixed point of E. Since $p = x^2 + 2y^2$ is impossible, there must be a fixed point of D, hence p has a representation of the form $x^2 + y^2$. This gives a new proof for one half of the two squares theorem, here Theorem 3(c).

If $p \equiv 1 \bmod 8$, we have a fixed point of B and (by the two squares theorem) of D. In order to prove the existence of the representation $p = x^2 + 2y^2$, it is enough to prove that there is (precisely) one fixed point of E. We do not see how to prove this with the methods of this paper. For this reason we did not state a theorem for the case $p \equiv 1 \bmod 8$.

3.4 $d = 3$

3.4.1 The Case $p = 3x^2 + 4y^2$

Here we deal with the case $d = \dfrac{4sm^2}{tn^2} = 3$. We have again two sub-cases.

- $s = 3$, $m = n = 1$, $t = 4$.
- $s = 3$, $m = t = 1$, $n = 2$, with even y and z.

As above in the case $d = 1$, both of these sub-cases are equivalent. We will thus concentrate on the first case.

The form $p = 3x^2 + 4yz$ represents only primes $p \equiv 3 \bmod 4$, hence we consider $p = 12k + 7$ and $p = 12k + 11$. We will proceed as in the case $d = 2$.

The general form of our matrix B is now

$$B = \begin{pmatrix} -1 & 2 & 0 \\ 0 & 1 & 0 \\ 3 & -3 & 1 \end{pmatrix}, A = BX = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ -3 & 1 & -3 \end{pmatrix}.$$

In view of $A^6 = I$, we consider the 9 matrices

$$(-1)^{j+1}A^j, (j = 1, \ldots, 5) \text{ and } (-1)^{j+1}XA^j, (j = 2, \ldots, 5).$$

(As before, the matrix XA is of no use, in view of the row condition $-x - 2z > 0$.)

$$-A^2 = \begin{pmatrix} 5 & -2 & 4 \\ 3 & -1 & 3 \\ -6 & 3 & -4 \end{pmatrix}, A^3 = \begin{pmatrix} 7 & -4 & 4 \\ 6 & -3 & 4 \\ -6 & 4 & -3 \end{pmatrix}, -A^4 = \begin{pmatrix} 5 & -4 & 2 \\ 6 & -4 & 3 \\ -3 & 3 & -1 \end{pmatrix},$$

$$A^5 = A^{-1} = XB = \begin{pmatrix} 1 & -2 & 0 \\ 3 & -3 & 1 \\ 0 & 1 & 0 \end{pmatrix}, D = -XA^2 = \begin{pmatrix} -5 & 2 & -4 \\ -6 & 3 & -4 \\ 3 & -1 & 3 \end{pmatrix},$$

$$E = XA^3 = \begin{pmatrix} -7 & 4 & -4 \\ -6 & 4 & -3 \\ 6 & -3 & 4 \end{pmatrix}, F = -XA^4 = \begin{pmatrix} -5 & 4 & -2 \\ -3 & 3 & -1 \\ 6 & -4 & 3 \end{pmatrix}, B = XA^5.$$

The corresponding boundaries are induced by the matrices themselves: for example, the matrix $\begin{pmatrix} -5 & 4 & -2 \\ -3 & 3 & -1 \\ 6 & -4 & 3 \end{pmatrix}$ corresponds to $-5x+4y-2z>0$, $-3x+3y-z>0$, $6x-4y+3z>0$.

Hence the map α is:

$$(x,y,z)\to \begin{cases} (x-2y,3x-3y+z,y) & \text{if } y<\frac{x}{2} \\ (-x+2y,y,3x-3y+z) & \text{if } \frac{x}{2}<y<x+\frac{z}{3} \\ (5x-4y+2z,6x-4y+3z,-3x+3y-z) & \text{if } x+\frac{z}{3}<y<\frac{5}{4}x+\frac{z}{2} \\ (-5x+4y-2z,-3x+3y-z,6x-4y+3z) & \text{if } \frac{5}{4}x+\frac{z}{2}<y<\frac{3}{2}x+\frac{3}{4}z \\ (7x-4y+4z,6x-3y+4z,-6x+4y-3z) & \text{if } \frac{3}{2}x+\frac{3}{4}z<y<\frac{7}{4}x+z \\ (-7x+4y-4z,-6x+4y-3z,6x-3y+4z) & \text{if } \frac{7}{4}x+z<y<2x+\frac{4}{3}z \\ (5x-2y+4z,3x-y+3z,-6x+3y-4z) & \text{if } 2x+\frac{4}{3}z<y<\frac{5}{2}x+2z \\ (-5x+2y-4z,-6x+3y-4z,3x-y+3z) & \text{if } \frac{5}{2}x+2z<y<3x+3z \\ (x+2z,z,-3x+y-3z) & \text{if } 3x+3z<y. \end{cases}$$

In order to prove theorem 4, we shall show: For primes $p\equiv 7 \bmod 12$, there is one fixed point of α. For primes $p\equiv 11 \bmod 12$, there are two fixed points of α.

3.4.2 Proof of Theorem 4

Suppose that the boundaries are attained. This will lead to a contradiction. Note that for odd primes $p=3x^2+4yz$ the value of x is odd. This excludes the boundaries $-x+2y=0$, $5x-2y+4z=0$, $5x-4y+2z=0$, and $7x-4y+4z=0$. Since $p=3x^2+4yz$ is prime ($p>3$), we can deduce that y and z are not divisible by 3. This excludes the boundaries $3x-3y+z=0$, $3x-y+3z=0$, $6x-4y+3z=0$, and $6x-3y+4z=0$.

Now let us look at the fixed points: The mappings $A, -A^2, -A^4, A^5$ cannot have any fixed points, (since A maps the region A onto the region A^5 etc.). The matrices B, A^3, D, E, F are involutions. So we have to check their fixed points.

- As before, B has precisely one fixed point: $(1,1,k'=\frac{p-3}{4})$.
- For A^3 the fixed point condition $A^3\begin{pmatrix} x \\ y \\ z \end{pmatrix}\stackrel{!}{=}\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ simplifies to: $3x-2y+2z=0$.

  This is a contradiction since x is odd.

- Similarly, for D, we need to look at $3x - y + 2z = 0$. Consider the equation $p = 3x^2 + 4yz = 3x^2 + 12xz + 8z^2 = 3(x + 2z)^2 - 4z^2$ modulo 3. With $z^2 \equiv 0, 1 \bmod 3$ and for $p \equiv 7 \bmod 12$, we find that $1 = 2z^2 \bmod 3$, a contradiction.
- For E, the fixed point condition is $2x - y + z = 0$. We look at $p = 3x^2 + 4yz = 3x^2 + 8xz + 4z^2 = 4(x + z)^2 - x^2 = (3x + 2z)(x + 2z)$, contradicting the primality of p.
- Finally, for F we have to look at $3x - 2y + z = 0$, and plug this into our ternary form $p = 3x^2 + 4yz = 3x^2 - 12xy + 8y^2 = 3(x - 2y)^2 - 4y^2$. Again, we consider this modulo 3: For $p = 12k + 7$ and with $2y^2 = 1 \bmod 3$ we see that there cannot be a fixed point.

We find that for $p = 12k + 7$ there is only the trivial fixed point of B, namely $(1, 1, (p - 3)/4)$. By the standard argument p can be written as $p = 3x^2 + 4y^2$.

Since $p = 12k + 11$ cannot be written as $p = 3x^2 + 4y^2$, there must be a fixed point of D or F. Any such fixed point induces a representation of the type $p = 3x^2 - 4y^2$, (see the analysis of these cases above).

This proves theorem 4.

3.5 $d = 4$

$d = \dfrac{4sm^2}{tn^2} = 4$. Here necessarily $s = t = m = n = 1$, and therefore

$$B = \begin{pmatrix} -1 & 2 & 0 \\ 0 & 1 & 0 \\ 4 & -4 & 1 \end{pmatrix}.$$

This matrix generates an infinite partition. Since in the case $s = t = 1$, we do not expect anything new, we do not pursue this case further.

4 On Infinite but Incomplete Mappings

One can also consider corresponding mappings induced by B and X for other values of d. We cannot expect that the number of required matrices is finite.

Consider $p = 3x^2 + 2yz = 24k + 5$. Generate the matrices with

$$B = \begin{pmatrix} -1 & 2 & 0 \\ 0 & 1 & 0 \\ 6 & -6 & 1 \end{pmatrix}.$$

Take

$$A = BX \text{ and } C = A^{-1} = XB.$$

$$A = BX, -A^2, A^3, -A^4, A^5 \text{ etc.}$$

$$C = XB, -C^2, C^3, -C^4, \text{ etc.}$$

$$B, -BC, BC^2, -BC^3, BC^4 \text{ etc.}$$

$$-XA^2, XA^3, -XA^4 \text{ etc.}$$

(Note: $-X$ and XA are again omitted.) The matrix A does not have a finite order. This can easily be seen by looking at the eigenvalues of A, namely $-1, -2 - \sqrt{3}, -2 + \sqrt{3}$.

Taking infinitely many of these matrices, we see: The "region" of each matrix becomes smaller and smaller.

For the powers of $-A$ the row conditions come arbitrarily close to:

$$\left(3 + \sqrt{3}\right) x - y + \left(2 + \sqrt{3}\right) z > 0$$

and

$$-\left(3 + \sqrt{3}\right) x + y - \left(2 + \sqrt{3}\right) z > 0.$$

There are similar row conditions for the other series of matrices. The series

$$C = XB, -C^2, C^3, -C^4, \text{ etc.}$$

corresponds to

$$B, -BC, BC^2, -BC^3, BC^4 \text{ etc.}$$

in that respect that the row conditions of the first and third row are the same and the condition of the first row is reversed. Similarly, the two series

$$A = BX, -A^2, A^3, -A^4, A^5 \text{ etc.}$$

and

$$-XA^2, XA^3, -XA^4 \text{ etc.}$$

have associated boundaries. This latter series tends to a row condition of

$$\left(3 + \sqrt{3}\right) x - \left(2 + \sqrt{3}\right) y + z > 0$$

and

$$-\left(3 + \sqrt{3}\right) x + \left(2 + \sqrt{3}\right) y - z > 0.$$

Unfortunately, the two boundaries

$$\left(3 + \sqrt{3}\right) x - y + \left(2 + \sqrt{3}\right) z > 0$$

and

$$\left(3+\sqrt{3}\right)x-\left(2+\sqrt{3}\right)y+z>0$$

do not correspond. Hence there is a gap in between the regions of these series.

One would need further matrices to close this gap in order to proceed.

Looking at the conditions $ax-by+cz>0$ and $ax-cy+bz>0$, we see how incidental the above described finite mappings are.

In the case studied by Zagier, we have $a=b=c$. So there are no problems at all. In the case $p=x^2+2yz$, we had $2x-y+2z$. Here $a=c$ so we still do not clearly see, what the condition in the general case is.

In the case $p=3x^2+4yz$, we had $6x-3y+4z$ and $6x-4y+3z$. Here, we see the importance of the matrices $A^3=BXBXBX$ and $XA^3=XBXBXBX$ with both rows, $6,-3,4$ and $-6,4,-3$. These matrices are the "turning point", reversing the y and z coordinate. We have a complete cycle: $(-3,1,-3)\Rightarrow(3,-1,3)\Rightarrow(-6,3,-4)\Rightarrow(6,-3,4)\Rightarrow(-6,4,-3)\Rightarrow(6,-4,3)\Rightarrow(-3,3,-1)\Rightarrow(3,-3,1)$. These matrices can be discovered by a sub-matrix (omit the first row and column) of the form $\begin{pmatrix} a & -b \\ -b & a \end{pmatrix}$.

In the incomplete mapping above, there are no such "turning points".

Acknowledgements The author is grateful to B. Artmann and D. Spalt for introducing him to Zagier's proof and for the challenge to understand how the proof could have been found. Further thanks goes to A.M. Decaillot for clarifying a question on Lucas' work. Sections 2.1 and 2.2 were found in 1990, Sect. 3 in 1996, and Sect. 1.6 in 2001, see also [11–14].

References

1. Aigner, M., Ziegler, G.M.: Proofs from THE BOOK, 2nd edition, Springer, Berlin, 2001.
2. Aubry, A.: Les principes de la géométrie des quinconces, L'Enseignement Mathématique 13 (1911), 187–203.
3. Bachmann, P.: Niedere Zahlentheorie, reprint by Chelsea Publishing Co., New York, 1968, originally published 1902/1910.
4. Bagchi, B.: Fermat's two squares theorem revisited. Resonance 4 (7) (1999), 59–67.
5. Barbeau, E. J.: Polynomials. Problem Books in Mathematics. Springer, New York, 1995.
6. Clarke, F.W., Everitt, W.N., Littlejohn, L.L., Vorster, S.J.R.: H. J. S. Smith and the Fermat Two Squares Theorem, Am. Math. Mon. 106(7) (1999), 652–665.
7. Décaillot A.-M.: Géométrie des tissus. Mosaïques. Échiquiers. Mathématiques curieuses et utiles. Revue d'histoire des mathématiques 8. Vol. 2. (2002), 145–206.
8. Dickson, L. E.: History of the theory of numbers, reprint by Chelsea Publishing Co., New York, 1966, originally published 1919–1923.
9. Dijkstra, E.W.: A derivation of a proof by D. Zagier, Manuscript EWD 1154, 1993, available at http://www.cs.utexas.edu/users/EWD/welcome.html
10. Edwards, H. M.: A genetic introduction to algebraic number theory. Springer, New York, 1996.
11. Elsholtz, C.: Primzahlen der Form $4k+1$ sind Summe zweier Quadrate, contribution to "Bundeswettbewerb Jugend forscht" (German National Contest for Young Scientists), 1990/91, published in [12].
12. Elsholtz, C.: Primzahlen der Form $4k+1$ sind Summe zweier Quadrate, Mathematiklehren, no. 62, February 1994, pp. 58–61.

13. Elsholtz, C.: The Liouville—Heath-Brown—Zagier proof of the two squares theorem (Preprint 2001/10, Institut für Mathematik, TU Clausthal, Germany).
14. Elsholtz, C.: Kombinatorische Beweise des Zweiquadratesatzes, Mathematische Semesterberichte 50 (2003), 77–93.
15. Œuvres de Fermat, ed: Paul Tannery and Charles Henry, Paris: Gauthier-Villars et fils 1891–1912.
16. Generalov, A.I.: A combinatorial proof of Euler-Fermat's theorem on the representation of the primes $p = 8k + 3$ by the quadratic form $x^2 + 2y^2$, J. Math. Sci. 140 (2007), 690–691.
17. Grace, J.H.: The four square theorem, J. London Math. Soc. 2 (1927), 3–8.
18. Hardy, G.H.: A Mathematician's Apology, Cambridge University Press, 1940.
19. Hardy, G.H.; Wright, E.M.: An introduction to the theory of numbers. 5th edition. Oxford University Press, New York, 1979.
20. Heath-Brown, D.R.: Fermat's two squares theorem, Invariant, 1984, 3–5. Available at http://eprints.maths.ox.ac.uk/677/
21. Jackson, T.: A Short Proof That Every Prime $p = 3(\mathrm{mod}8)$ is of the Form $x^2 + 2y^2$, Amer. Math. Mon. 107 (2000), 447.
22. Jackson, T.: Automorphs and involutions. Tatra Mt. Math. Publ. 20 (2000), 59–63.
23. Jackson, T.: Direct proofs of some of Euler's results. Number theory (Turku, 1999), 163–166, de Gruyter, Berlin, 2001.
24. Kraitchik, M.: Mathematical Recreations. Allen & Unwin, London, 1943.
25. Larson, L.C.: A theorem about primes proved on a chessboard. Math. Mag. 50 (2) (1977), 69–74,
26. Lucas, É.: Application de l'arithmétique à la construction de l'armure des satins réguliers. Paris, 1867. Available online at http://edouardlucas.free.fr/gb/index.html
27. Lucas, É.: Les principes fondamentaux de la géométrie des tissus, Congrès de l'Association française pour l'avancement des sciences 40 (1911), 72–87, (based on an article in L'Ingegnere Civile, Torino, 1880). Available at http://www.biodiversitylibrary.org/item/27373#5
28. McKay, James H.: Another proof of Cauchy's group theorem. Amer. Math. Mon. 66 (1959), 119.
29. Nathanson, M.B.: Elementary methods in number theory. Graduate Texts in Mathematics, 195. Springer, New York, 2000.
30. Pólya, G.: Über die "doppelt-periodischen" Lösungen des n-Damen Problems, in: Ahrens, W. Mathematische Unterhaltungen und Spiele, Teubner, Leipzig, Volume II, 2nd edition 1918, 364–374.
31. Shirali, S.: On Fermat's Two-Square Theorem. Resonance 2(3) (1997), 69–73.
32. Shiu, P.: Involutions associated with sums of two squares. Publ. Inst. Math. (Beograd) (N.S.) 59(73) (1996), 18–30.
33. Tikhomirov, V.: Quantum, May/June 1994, pp. 5–7.
34. Uspensky, J. V.; Heaslet, M. A.: Elementary Number Theory. McGraw-Hill Book Company, New York and London, 1939.
35. Varouchas, I.: Une démonstration élémentaire du théorème des deux carrés, I.R.E.M. Bull. 6 (1984), 31–39.
36. Venkov, B.A.: Elementary Number Theory, Wolters-Noordhoff, Groningen, 1970, (originally published in Russian, 1937).
37. Wagon, S.: The Euclidean algorithm strikes again. Amer. Math. Mon. 97 (1990), 125–129.
38. Wells, D.: Are these the most beautiful? Math. Intelligencer 12(3) (1990), 37–41.
39. Williams, K.S.: Heath-Brown's elementary proof of the Girard-Fermat theorem, Carleton Coordinates, (1985), 4–5.
40. Winter, H.: Der Zwei-Quadrate-Satz von Fermat - eine Studie zur Heuristik des Beweisens. Math. Semesterber. 50 (2003), 191-235.
41. Zagier, D.: A one-sentence proof that every prime $p \equiv 1 \pmod 4$ is a sum of two squares, Amer. Math. Mon. 97(2) (1990), 144.

A Note on Elkin's Improvement of Behrend's Construction

Ben Green and Julia Wolf

To Mel Nathanson

Summary We provide a short proof of a recent result of Elkin in which large subsets of $\{1, \dots, N\}$ free of three-term progressions are constructed.

Keywords Arithmetic progressions · Roth's theorem

Mathematics Subject Classifications (2010). Primary 11B25, Secondary 11B75

1 Introduction

Write $r_3(N)$ for the cardinality of the largest subset of $\{1, \dots, N\}$ not containing three distinct elements in arithmetic progression. A famous construction of Behrend [1] shows, when analysed carefully, that

$$r_3(N) \gg \frac{1}{\log^{1/4} N} \cdot \frac{N}{2^{2\sqrt{2}\sqrt{\log_2 N}}}.$$

In a recent preprint [2], Elkin was able to improve this 62-year old bound to

$$r_3(N) \gg \log^{1/4} N \cdot \frac{N}{2^{2\sqrt{2}\sqrt{\log_2 N}}}.$$

B. Green
Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, England
e-mail: b.j.green@dpmms.cam.ac.uk

J. Wolf
Mathematical Sciences Research Institute, 17 Gauss Way, Berkeley, CA 94720, USA
e-mail: julia.wolf@cantab.net

The first author holds a Leverhulme Prize and is grateful to the Leverhulme Trust for their support. This paper was written while the authors were attending the special semester in ergodic theory and additive combinatorics at MSRI.

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_9,

Our aim in this note is to provide a short proof of Elkin's result. It should be noted that the only advantage of our approach is brevity: it is based on ideas morally close to those of Elkin, and moreover, his argument is more constructive than ours.

Throughout the paper, $0 < c < 1 < C$ denote absolute constants which may vary from line to line. We write $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ for the d-dimensional torus.

2 The Proof

Let d be an integer to be determined later, and let $\delta \in (0, 1/10)$ be a small parameter (we will have $d \sim C\sqrt{\log N}$ and $\delta \sim \exp(-C\sqrt{\log N})$). Given $\theta, \alpha \in \mathbb{T}^d$, write $\Psi_{\theta,\alpha} : \{1, \dots, N\} \to \mathbb{T}^d$ for the map $n \mapsto \theta n + \alpha (\mathrm{mod}\, 1)$.

Lemma 2.1. *Suppose that n is an integer. Then $\Psi_{\theta,\alpha}(n)$ is uniformly distributed on $\mathbb{T}^d$ as θ, α vary uniformly and independently over $\mathbb{T}^d$. Moreover, if n and n' are distinct positive integers, then the pair $(\Psi_{\theta,\alpha}(n), \Psi_{\theta,\alpha}(n'))$ is uniformly distributed on $\mathbb{T}^d \times \mathbb{T}^d$ as θ, α vary uniformly and independently over $\mathbb{T}^d$.*

Proof. Only the second statement requires an argument to be given. Perhaps the easiest proof is via Fourier analysis, noting that

$$\int \mathrm{e}^{2\pi \mathrm{i}(k\cdot(\theta n+\alpha)+k'\cdot(\theta n'+\alpha))}\, \mathrm{d}\theta\, \mathrm{d}\alpha = 0$$

unless $k + k' = kn + k'n' = 0$. Provided that k and k' are not both zero, this cannot happen for distinct positive integers n, n'. Since the exponentials $\mathrm{e}^{2\pi\mathrm{i}(kx+k'x')}$ are dense in $L^2(\mathbb{T}^d \times \mathbb{T}^d)$, the result follows. □

Let us identify $\mathbb{T}^d$ with $[0, 1)^d$ in the obvious way. For each $r \leqslant \frac{1}{2}\sqrt{d}$, write $S(r)$ for the region

$$\{x \in [0, 1/2]^d : r - \delta \leqslant \|x\|_2 \leqslant r\}.$$

Lemma 2.2. *There is some choice of r for which* $\mathrm{vol}(S(r)) \geqslant c\delta 2^{-d}$.

Proof. First note that if $(x_1, \dots, x_d)$ is chosen at random from $[0, 1/2]^d$ then, with probability at least c, we have $|\|x\|_2 - \sqrt{d/12}| \leqslant C$. This is a consequence of standard tail estimates for sums of independent identically distributed random variables, of which $\|x\|_2^2 = \sum_{i=1}^d x_i^2$ is an example. The statement of the lemma then immediately follows from the pigeonhole principle. □

Write $S := S(r)$ for the choice of r whose existence is guaranteed by the preceding lemma; thus $\mathrm{vol}(S) \geqslant c\delta 2^{-d}$. Write $\tilde{S}$ for the same set S but considered now as a subset of $[0, 1/2]^d \subseteq \mathbb{R}^d$. Since there is no "wraparound", the three-term progressions in S and $\tilde{S}$ coincide and henceforth we abuse notation, regarding S as a subset of $\mathbb{R}^d$ and dropping the tildes. (To use the additive combinatorics jargon, S and $\tilde{S}$ are *Freiman isomorphic*.) Suppose that (x, y) is a pair for which $x - y, x$ and $x + y$ lie in S. By the parallelogram law

$$2\|x\|_2^2 + 2\|y\|_2^2 = \|x + y\|_2^2 + \|x - y\|_2^2$$

and straightforward algebra we have

$$\|y\|_2 \leqslant \sqrt{r^2 - (r - \delta)^2} \leqslant \sqrt{2\delta r}.$$

It follows from the formula for the volume of a sphere in $\mathbb{R}^d$ that the volume of the set $B \subseteq \mathbb{T}^d \times \mathbb{T}^d$ in which each such pair (x, y) must lie is at most $\mathrm{vol}(S)C^d(\delta/\sqrt{d})^{d/2}$.

The next lemma is an easy observation based on Lemma 2.1.

Lemma 2.3. *Suppose that N is even. Define $A_{\theta,\alpha} := \{n \in [N] : \Psi_{\theta,\alpha}(n) \in S\}$. Then*

$$\mathbb{E}_{\theta,\alpha}|A_{\theta,\alpha}| = N\mathrm{vol}(S)$$

whilst the expected number of nontrivial three-term arithmetic progressions in $A_{\theta,\alpha}$ is

$$\mathbb{E}_{\theta,\alpha}T(A_{\theta,\alpha}) = \frac{1}{4}N(N-5)\mathrm{vol}(B).$$

Proof. The first statement is an immediate consequence of the first part of Lemma 2.1. Now each nontrivial three-term progression is of the form $(n - d, n, n + d)$ with $d \neq 0$. Since N is even there are $N(N-5)/4$ choices for n and d, and each of the consequent progressions lies inside $A_{\theta,\alpha}$ with probability $\mathrm{vol}(B)$ by the second part of Lemma 2.1. □

To finish the argument, we just have to choose parameters so that

$$\frac{1}{3}\mathrm{vol}(S) \geqslant \frac{1}{4}(N-5)\mathrm{vol}(B). \tag{2.1}$$

Then, we shall have

$$\mathbb{E}\left(\frac{2}{3}|A_{\theta,\alpha}| - T(A_{\theta,\alpha})\right) \geqslant \frac{1}{3}N\mathrm{vol}(S).$$

In particular, there is a specific choice of $A := A_{\theta,\alpha}$ for which both $T(A) \leqslant 2|A|/3$ and $|A| \geqslant \frac{1}{2}N\mathrm{vol}(S)$. Deleting up to two thirds of the elements of A, we are left with a set of size at least $\frac{1}{6}N\mathrm{vol}(S)$ that is free of three-term arithmetic progressions.

To do this, it suffices to have $C^d(\delta/\sqrt{d})^{d/2} \leqslant c/N$, which can certainly be achieved by taking $\delta := c\sqrt{d}N^{-2/d}$. For this choice of parameters we have, by the earlier lower bound on $\mathrm{vol}(S)$, that

$$|A| \geqslant \frac{1}{6}N\,\mathrm{vol}(S) \geqslant c\sqrt{d}2^{-d}N^{1-2/d}.$$

Choosing $d := \lceil \sqrt{2\log_2 N} \rceil$ we recover Elkin's bound. □

3 A Question of Graham

The authors did not set out to try and find a simpler proof of Elkin's result. Rather, our concern was with a question of Ron Graham (personal communication to the first-named author, see also [3, 4]). Defining $W(2;3,k)$ to be the smallest N such that any red-V-blue colouring of $[N]$ contains either a three-term red progression or a k-term blue progression, Graham asked whether $W(2;3;k) < k^A$ for some absolute constant A or, even more ambitiously, whether $W(2;3,k) \leqslant Ck^2$. Our initial feeling was that the answer was surely no, and that a counterexample might be found by modifying the Behrend example in such a way that its complement does not contain long progressions. Reinterpreting the Behrend construction in the way that we have done here, it seems reasonably clear that it is not possible to provide a negative answer to Graham's question in this way.

Acknowledgement The authors are grateful to Tom Sanders for helpful conversations.

References

[1] F. Behrend. *On sets of integers which contain no three terms in arithmetic progression,* Proc. Nat. Acad. Sci., **32**:331–332, 1946
[2] M. Elkin, *An improved construction of progression-free sets,* available at http://arxiv.org/abs/0801.4310
[3] R. Graham, *On the growth of a van der Waerden-like function,* Integers, **6**:#A29, 2006
[4] B. Landman, A. Robertson and C. Culver, *Some new exact van der Waerden numbers,* Integers, **5**(2):#A10, 2005

Distinct Matroid Base Weights and Additive Theory

Y.O. Hamidoune and I.P. da Silva

Summary Let M be a matroid on a set E and let $w : E \longrightarrow G$ be a weight function, where G is a cyclic group. Assuming that $w(E)$ satisfies the Pollard's Condition (i.e., Every non-zero element of $w(E) - w(E)$ generates G), we obtain a formula for the number of distinct base weights. If $|G|$ is a prime, our result coincides with a result of Schrijver and Seymour.

We also describe Equality cases in this formula. In the prime case, our result generalizes Vosper's Theorem.

Keywords Additive inequalities · Vosper's theorem · Weighted matroid

Mathematics Subject Classifications (2010). 11P70, 05B35

1 Introduction

Let G be a finite cyclic group and let A, B be nonempty subsets of G. The starting point of Minkowski set sum estimation is the inequality $|A + B| \geq \min(|G|, |A| + |B| - 1)$, where $|G|$ is a prime, proved by Cauchy [2] and rediscovered by Davenport [4]. The first generalization of this result, due to Chowla [3], states that in a cyclic group G, $|A + B| \geq \min(|G|, |A| + |B| - 1)$, if there is a $b \in B$ such that every non-zero element of $B - b$ generates G. In order to generalize his extension of the Cauchy–Davenport Theorem [11] to composite moduli, Pollard introduced in [12] the following more sophisticated Chowla type condition: Every non-zero element of $B - B$ generates G.

Y.O. Hamidoune
Université Pierre et Marie Curie, E. Combinatoire, Case 189, 4 Place Jussieu, 75005 Paris, France
e-mail: hamidoune@math.jussieu.fr

I.P. da Silva
CELC/Universidade de Lisboa, Faculdade de Ciências, Campo Grande,
edifício C6 - Piso 2, 1749-016 Lisboa, Portugal
e-mail: isilva@cii.fc.ul.pt

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_10,

Equality cases of the Cauchy–Davenport were determined by Vosper in [16, 17]. Vosper's Theorem was generalized by Kemperman [9]. We need only a light form of Kemperman's result stated in the beginning of Kemperman's paper.

We need the following combination of Chowla and Kemperman results:

Theorem A (Chowla [3], Kemperman [9]). *Let A, B be non-empty subsets of a cyclic group G such that for some $b \in B$, every non-zero element of $B - b$ generates G. Then*

(i) $|A + B| \geq \min\{|G|, |A| + |B| - 1\}$.
(ii) If $|A + B| = |A| + |B| - 1$ *and* $\min(|A|, |B)| \geq 2$, *then* $A + B$ *is an arithmetic progression.*

(i) Is Chowla's Theorem [3]. Kemperman proved (ii) in [9]. A shortly proved generalization of this result to non-abelian groups is obtained in [8].

Zero-sum problems form another developing area in Additive Combinatorics having several applications. The Erdős–Ginzburg–Ziv Theorem [6] was the starting point of this area. This result states that a sequence of elements of an abelian group G with length $\geq 2|G| - 1$ contains a zero-sum subsequence of length $= |G|$.

The reader may find some details on these two areas of Additive Combinatorics in the text books: Nathanson [10], Geroldinger–Halter–Koch [7] and Tao–Vu [15]. More specific questions may be found in Caro's survey paper [1].

The notion of a matroid was introduced by Whitney in 1935 as a generalization of a matrix. Two pioneer works connecting matroids and Additive Combinatorics are due to Schrijver–Seymour [13] and Dias da Silva–Nathanson [5]. Recently, in [14], orientability of matroids is naturally related with an open problem on Bernoulli matrices.

Stating the first result requires some vocabulary:

Let E be a finite set. The set of the subsets of E will be denoted by 2^E.

A *matroid* over E is an ordered pair $(E, \mathcal{B})$ where $\mathcal{B} \subseteq 2^E$ satisfies the following axioms:

(B1) $\mathcal{B} \neq \emptyset$.
(B2) For all $B, B' \in \mathcal{B}$, if $B \subseteq B'$ then $B = B'$.
(B3) For all $B, B' \in \mathcal{B}$ and $x \in B \setminus B'$, there is a $y \in B' \setminus B$ such that $(B \setminus \{x\}) \cup \{y\} \in \mathcal{B}$.

A set belonging to $\mathcal{B}$ is called a *base* of the matroid M.

The *rank* of a subset $A \subseteq E$ is by definition

$$r_M(A) := \max\{|B \cap A| : B \text{ is a base of } M\}.$$

We write $r(M) = r(E)$. The reference to M could be omitted. A *hyperplane* of the matroid M is a maximal subset of E with rank $= r(M) - 1$.

The *uniform* matroid of rank r on a set E is by definition $\mathcal{U}_r(E) = (E, \binom{E}{r})$, where $\binom{E}{r}$ is the set of all r-subsets of E. Let M be a matroid on E and let N be a matroid on F. We define the direct sum:

$$M \oplus N = (E \times \{0\} \cup F \times \{1\}, \{B \times \{0\} \cup C \times \{1\}: \\ B \text{ is a base of } M \text{ and } C \text{ is a base of } N\}).$$

Let $w : E \longrightarrow G$ be a weight function, where G is an abelian group. The weight of a subset X is by definition

$$X^w := \sum_{x \in X} w(x).$$

The set of distinct base weights of a matroid M is

$$M^w := \{B^w : B \text{ is a basis of } M\}.$$

Suppose now $|G| = p$ is a prime number. Schrijver and Seymour proved that $|M^w| \geq \min(p, \sum_{g \in G} r(w^{-1}(g)) - r(M) + 1)$.

Let A and B be subsets of G. Define $w : A \times \{0\} \cup B \times \{1\}$, by the relation $w(x, y) = x$. Then,

$$(\mathcal{U}_1(A) \oplus \mathcal{U}_1(B))^w = A + B.$$

Applying their result to this matroid, Schrijver and Seymour obtained the Cauchy–Davenport Theorem.

Let $x_1, \ldots, x_{2p-1} \in G$. Consider the uniform matroid $M = \mathcal{U}_p(E)$, of rank p over the set $E = \{1, \ldots, 2p-1\}$, with weight function $w(i) = x_i$. In order to prove the Erdős–Ginzburg–Ziv Theorem [6], one may clearly assume that no element is repeated p times. In particular for every $g \in G$, $r(w^{-1}(g)) = |w^{-1}(g)|$. Applying their result to this matroid, Schrijver and Seymour obtained

$$|M^w| \geq \min\left(|G|, \sum_{g \in G} r(w^{-1}(g)) - r(M) + 1\right) \\ = \min\left(p, \sum_{g \in G} |w^{-1}(g)| - p + 1\right) = p.$$

In particular, $0 \in M^w$ as stated by the Erdős-Ginzburg-Ziv Theorem [6].

In the present work, we prove the following result:

Theorem 1. *Let G be a cyclic group, M be a matroid on a finite set E with $r(M) \geq 1$ and let $w : E \longrightarrow G$ be a weight function. Assume moreover that every nonzero element of $w(E) - w(E)$ generates G. Then*

$$|M^w| \geq \min\left(|G|, \sum_{g \in G} r(w^{-1}(g)) - r(M) + 1\right), \tag{1}$$

where M^w denotes the set of distinct base weights. Moreover, if Equality holds in (1), then one of the following conditions holds:

(i) $r(M) = 1$ or M^w is an arithmetic progression.
(ii) There is a hyperplane H of M such that $M^w = g + (M/H)^w$, for some $g \in G$.

If G has a prime order, then the condition on $w(E) - w(E)$ holds trivially. In this case, (1) reduces to the result of Schrijver–Seymour.

2 Terminology and Preliminaries

Let M be a matroid on a finite set E. One may see easily from the definitions that all bases of a matroid have the same cardinality. A *circuit* of M is a minimal set not contained in a base. A loop is an element x such that $\{x\}$ is a circuit. By the definition bases contain no loop. The closure of a subset $A \subseteq E$ is by definition

$$cl(A) = \{x \in A : r(A \cup x) = r(A)\}.$$

Note that an element $x \in cl(A)$ if and only if $x \in A$, or there is circuit C such $x \in C$ and $C \backslash \{x\} \subseteq A$.

Given a matroid M on a set E and a subset $A \subseteq E$. Put

$$\mathcal{B}/A := \{B \backslash A : B \text{ is a base of } M \text{ with } |B \cap A| = r(A)\}.$$

One may see easily that $M/A = (E \backslash A, \mathcal{B}/A)$ is a matroid on $E \backslash A$. We say that this matroid is obtained from M contracting A. Notice that

$$r_{M/A}(X) = r_M(X \cup A) - r_M(A).$$

Recall the following easy lemma:

Lemma B. *Let M be a matroid on a finite set E and let U, V be disjoint subsets of E. Then*

- *M/U and $M/cl(U)$ have the same bases. In particular, $(M/U)^w = (M/cl(U))^w$.*
- *$(M/U)/V = M/(U \cup V)$.*

For more details on matroids, the reader may refer to one of the text books: Welsh [18] or White [19].

For $u \in E$, we put

$$G_u := \{g \in G : u \in cl(w^{-1}(g))\}.$$

We use also the following lemma:

Lemma C (Schrijver and Seymour in [13]). *Let M be a matroid on a finite set E and let $w : E \longrightarrow G$ be a weight function. Then for every nonloop element $u \in E$,*

$$(M/u)^w + G_u \subseteq M^w.$$

Proof. Take a base B of M/u and an element $g \in G_u$. If $g = w(u)$, then by definition of contraction, $B \cup \{u\}$ is a base of M and $B^w + w(u) \in M^w$. If $g \neq w(u)$, there is a circuit C containing u such that $\emptyset \neq C\backslash\{u\} \subseteq w^{-1}(g)$. For some $v \in C\backslash\{u\}$, the subset $B \cup \{v\}$ must be a base of M otherwise $C\backslash\{v\} \subseteq cl(B)$, implying that $u \in cl(B)$, in contradiction with the assumption that B is a base of M/u. Therefore, $(B \cup \{v\})^w = B^w + g \in M^w$. □

3 Proof of the Main Result

We shall now prove our result:

Proof of Theorem 1: First, we prove (1) by induction on the rank of M. The result holds trivially if $r(M) = 1$. Since $r(M) \geq 1$, M contains a non-loop element. Take an arbitrary non-loop element y.

$$\begin{aligned} |M^w| &\geq |(M/y)^w + G_y| \\ &\geq |(M/y)^w| + |G_y| - 1 \\ &\geq \sum_{g \in G} r(w^{-1}(g)) - r(M) + 1. \end{aligned} \tag{2}$$

The first inequality follows from Lemma C, the second follows by Theorem A and the third is a direct consequence of the definitions of M/y and G_y. This proves the first part of the theorem.

Suppose now that Equality holds in (1) and that Condition (i) is not satisfied. In particular, $r(M) \geq 2$. Also $|M^w| \geq 2$, otherwise M^w is a progression, a contradiction.

We claim that there exits a non-loop element $u \in E$ such that $|(M/u)^w| \geq 2$. Assume on the contrary that for every non-loop element $u \in E$, we have $|(M/u)^w| = 1$. Then every pair of bases B_1, B_2 of M with $B_1^w \neq B_2^w$ satisfies $B_1 \cap B_2 = \emptyset$ otherwise for every $z \in B_1 \cap B_2$, $|(M/z)^w| \geq 2$. Now, for every $z \in B_1$, there is $z' \in B_2$ such that $C = (B_1\backslash\{z\}) \cup \{z'\}$ is a base of M. For such a base C, $B_1 \cap C \neq \emptyset$, $B_2 \cap C \neq \emptyset$, and we must have $B_1^w = C^w = B_2^w$, a contradiction.

Applying the chain of inequalities proving (2) with $y = u$. We have

$$|M^w| = |(M/u)^w + G_u| = |(M/u)^w| + |G_u| - 1. \tag{3}$$

Note that $w(E\backslash\{u\}) \subset w(E)$. Thus $w(E\backslash\{u\})$ verifies the Pollard condition. If $|G_u| \geq 2$ Theorem A implies that M^w is a progression and thus M satisfies Condition (i) of the theorem, contradicting our assumption on M. We must have $|G_u| = 1$.

Thus, $G_u = \{w(u)\}$ and $M^w = w(u) + (M/u)^w$.

Since the translate of a progression is a progression, $(M/u)^w$ is not a progression. By Lemma B, (M/u) and $M/cl(u)$ have the same bases and thus the result holds if $r(M) = 2$. If $r(M) > 2$, then by the Induction hypothesis there is a hyperplane H of M/u such that $(M/u)^w = (M/u/H)^w = (M/(cl(\{u\} \cup H))^w$, and (ii) holds. □

Corollary 2 (Vosper's Theorem [16,17]). *Let p be a prime and let A, B be subsets of $\mathbb{Z}_p$ such that $|A|, |B| \geq 2$.*

If $|A + B| = |A| + |B| - 1 < p$ then one of the following holds:

(i) where $c = Zp\backslash(A + B)$
(ii) A and B are arithmetic progressions with the same difference.

Proof. Consider the matroid $N = (\mathcal{U}_1(A) \oplus \mathcal{U}_1(B))$ and its weight function w defined in the Introduction. $H = A \times \{0\}$ and $H' = B \times \{1\}$ are the hyperplanes of N and we have $N^w = A + B$.

If $|N^w| = |A|+|B|-1$, then Theorem 1 says that N must satisfy one of its conditions (i) or (ii). Since by hypothesis $|A|, |B| \geq 2$ we have $|N^w| > \max(|A|, |B|) \geq |(N/H)^w|, |(N/H')^w|$ and we conclude that N^w must be an arithmetic progression with difference d. Without loss of generality, we may take $d = 1$.

Case 1. $|A + B| = p - 1$. Put $\{c\} = \mathbb{Z}_p\backslash(A + B)$. We have $c - A \subset (\mathbb{Z}_p\backslash B)$. Since these sets have the same cardinality, we have $c - A = (\mathbb{Z}_p\backslash B)$.

Case 2. $|A + B| < p - 1$.

We have $|A + B + \{0, 1\}| = |A + B| + 1 = |A| + |B| < p$.

We must have $|A + \{0, 1\}| = |A| + 1$, since otherwise by the Cauchy-Davenport Theorem,

$$\begin{aligned} |A + B| + 1 &= |A + B + \{0, 1\}| \\ &= |A + \{0, 1\} + B| \\ &\geq (|A| + 2) + |B| - 1 = |A| + |B| + 1, \end{aligned}$$

a contradiction. It follows that A is an arithmetic progression with difference 1. Similarly B is an arithmetic progression with difference 1. □

References

1. Caro, Yair Zero-sum problems, a survey. *Discrete Math.* 152 (1996), no. 1–3, 93–113.
2. A. L. Cauchy, Recherches sur les nombres, *J. Ecole Polytechnique* 9 (1813), 99–116.
3. I. Chowla, A theorem on the addition of residue classes: Applications to the number $\Gamma(k)$ in Waring's problem, *Proc. Indian Acad. Sc.*, Section A, no. 1 (1935) 242–243.

4. H. Davenport, On the addition of residue classes, *J. Lond. Math. Soc.* 10 (1935), 30–32.
5. J.A. Dias da Silva and M.B. Nathanson, Maximal Sidon sets and matroids, *Discrete Math.* 309 (2009), no. 13, 4489–4494.
6. P. Erdős, A. Ginzburg and A. Ziv, A theorem in additive number theory, *Bull. Res. Council Israel 10F* (1961), 41–43.
7. A. Geroldinger and F. Halter-Koch, *Non-unique factorizations. Algebraic, combinatorial and analytic theory. Pure and Applied Mathematics (Boca Raton)*, 278. Chapman & Hall/CRC, Boca Raton, FL, 2006. xxii+700 pp.
8. Y.O. Hamidoune, An isoperimetric method in additive theory. *J. Algebra* 179 (1996), no. 2, 622–630.
9. J. H. B. Kemperman, On small sumsets in abeliangroups, *Acta Math.* 103 (1960), 66–88.
10. M. B. Nathanson, *Additive Number Theory. Inverse problems and the geometry of sumsets*, Grad. Texts in Math. 165, Springer, Berlin, Heidelberg, 1996.
11. J. M. Pollard, A generalisation of the theorem of Cauchy and Davenport, *J. Lond. Math. Soc.* (2) 8 (1974), 460–462.
12. J. M. Pollard, Addition properties of residue classes, *J. Lond. Math. Soc.* (2) 11 (1975), no. 2, 147–152.
13. A. Schrijver, P. D. Seymour, Spanning trees of different weights. *Polyhedral combinatorics* (Morristown, NJ, 1989), 281–288, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 1, Amer. Math. Soc., Providence, RI, 1990.
14. I. P. F. da Silva, Orientability of Cubes, *Discrete Math.* 308 (2008), 3574–3585.
15. T. Tao and V.H. Vu, *Additive Combinatorics*, 105. Cambridge Studies in Advanced Mathematics, Cambridge Press University, Cambridge, 2006.
16. G. Vosper, The critical pairs of subsets of a group of prime order, *J. Lond. Math. Soc.* 31 (1956), 200–205.
17. G. Vosper, Addendum to The critical pairs of subsets of a group of prime order, *J. Lond. Math. Soc.* 31 (1956), 280–282.
18. D. J. A. Welsh, *Matroid theory*, Academic Press, London, 1976.
19. White, N. (ed.), *Theory of matroids*, Cambridge University Press, Cambridge, 1986.

The Postage Stamp Problem and Essential Subsets in Integer Bases

Peter Hegarty

Dedicated to Melvyn B. Nathanson on the occasion of his 60th birthday

Summary Plagne recently determined the asymptotic behavior of the function $E(h)$, which counts the maximum possible number of essential elements in an additive basis for $\mathbb{N}$ of order h. Here, we extend his investigations by studying asymptotic behavior of the function $E(h,k)$, which counts the maximum possible number of essential subsets of size k, in a basis of order h. For a fixed k and with h going to infinity, we show that $E(h,k) = \Theta_k\left([h^k/\log h]^{1/(k+1)}\right)$. The determination of a more precise asymptotic formula is shown to depend on the solution of the well-known 'postage stamp problem' in finite cyclic groups. On the other hand, with h fixed and k going to infinity, we show that $E(h,k) \sim (h-1)\frac{\log k}{\log\log k}$.

Keywords Additive basis · Essential subset

Mathematics Subject Classifications (2000). 11B13 (primary), 11B34 (secondary)

1 Essential Subsets of Integer Bases

Let S be a countable abelian semigroup, written additively, h be a positive integer and $A \subseteq S$. The h-fold sumset hA consists of all $s \in S$, which can be expressed as a sum of exactly h not necessarily distinct elements of A. If S is infinite, we write $hA \sim S$ if all but finitely many elements of S lie in hA. In that case, A is said to be a *basis of order* h[1] if $hA \sim S$ but $(h-1)A \not\sim S$. If S is finite, then a basis A of order h must satisfy $hA = S$ and $(h-1)A \neq S$. The two semigroups of interest in this paper (and in most of the additive number theory literature) are $S = \mathbb{N}$, the set of positive integers, and $S = \mathbb{Z}_n$, the set of residue classes modulo a positive integer n.

[1] In the literature, the term *asymptotic basis* is common.

P. Hegarty
Mathematical Sciences, Chalmers University of Technology and Göteborg University, Göteborg, Sweden
e-mail: hegarty@chalmers.se

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_11,

Now, suppose A is a basis of some order for $\mathbb{N}$, a so-called *integer basis*. A finite subset E of A is said to be an *essential subset* of A if $A\backslash E$ is no longer a basis of any order, and the set E is minimal with this property. In the case when E is a singleton set, $E = \{a\}$ say, we say that a is an *essential element* of A.

A fundamental result of Erdős and Graham [EG] states that every integer basis possesses only finitely many essential elements. Grekos [G] refined this observation by showing that the number of essential elements in a basis of order h is bounded by a function of h only. Let $E(h)$ denote the maximum possible number of essential elements in a basis of order h. Two recent papers have left us with a very good understanding of this function. In 2007, Deschamps and Farhi [DF] proved that

$$E(h) \leq c\sqrt{\frac{h}{\log h}}, \tag{1.1}$$

with $c = 30\sqrt{\frac{\log 1564}{1564}} \approx 2.05$, and gave an example to show that this is the best possible universal constant. That left the question of asymptotic behavior and, in 2008, Plagne [P] completed the picture by showing that

$$E(h) \sim 2\sqrt{\frac{h}{\log h}}. \tag{1.2}$$

Most of his paper was in fact devoted to verifying that the asymptotic behavior of $E(h)$ is regular.

Deschamps and Farhi appear to be the first people to study essential subsets in an integer basis of arbitrary size. They generalized the Erdős–Graham result by showing that any basis possesses only finitely many essential subsets. However, the number of these cannot be bounded purely in terms of the order of the basis, as the following example from their paper shows. Let $s \geq 1$ and $p_1, \ldots, p_s$ denote the first s prime numbers. Put $\mathscr{P} := \prod_{i=1}^{s} p_i$ and take

$$A = \mathscr{P} \cdot \mathbb{N} \cup \{1, 2, \ldots, \mathscr{P} - 1\} \tag{1.3}$$

Clearly, A is a basis of order 2, but it possesses s different essential subsets, namely the sets

$$E_i = \{x \in \{1, \ldots, \mathscr{P} - 1\} : (x, p_i) = 1\}, \quad i = 1, \ldots, s. \tag{1.4}$$

Note, however, that as s increases in this example, so do the sizes of the essential subsets E_i (and drastically so!). Deschamps and Farhi suggested that the right generalization of (1.1) would be an upper bound for the number of essential subsets of a given size in a basis of a given order. In other words, the function $E(h, k)$, which denotes the maximum possible number of essential subsets of size k in an integer basis of order h, should be well-defined. In [He], the present author proved that this

is the case, but made no attempt to obtain precise estimates. Motivated by Plagne's subsequent work, we will in this paper prove the following two results:

Theorem 1.1. *For each fixed $h > 0$, as $k \to \infty$ we have*

$$E(h,k) \sim (h-1)\frac{\log k}{\log\log k}. \tag{1.5}$$

Theorem 1.2. *Let the function $f(h,k)$ be given by*

$$f(h,k) := \frac{k+1}{k^2} \cdot \sqrt[k+1]{k} \cdot \left(\frac{h^k}{\log h}\right)^{\frac{1}{k+1}}. \tag{1.6}$$

Then, for each fixed k, as $h \to \infty$ we have

(i)

$$E(h,k) \gtrsim f(h,k). \tag{1.7}$$

(ii) There is a number $\underline{R}(k) \in (1/e, 1)$, to be defined below, such that

$$E(h,k) \lesssim \left(\frac{1}{\underline{R}(k)}\right)^{\frac{k}{k+1}} f(h,k). \tag{1.8}$$

The problem of estimating the function $E(h,k)$ is intimately connected with the well-known Postage Stamp Problem (PSP), this being the popular name for the general problem of finding bases which are, in some sense, the most economical possible. In Sect. 2 we present an overview of this problem and, in particular, define the numbers $\underline{R}(k)$ appearing in (1.8) above. Note that the exact values of these numbers are not known for any $k > 1$. Theorems 1.1 and 1.2 are proven in Sects. 3 and 4 respectively. All our proofs build on the ideas in previous papers on this subject and are supplemented by ingredients of a mostly technical nature. That of Theorem 1.2 is modeled closely on Plagne's [P]. The main technical problem he faced was to show that the function $E(h,1)$ behaved regularly, and in his case this was basically due to the unsatisfactory state of current knowledge concerning the distribution of primes in short intervals. When $k > 1$ that state of affairs continues to create difficulties, but they will turn out to be less serious than those arising from the gaps in our current understanding of the PSP. These gaps mean that, not only can we not compute exactly the numbers $\underline{R}(k)$, but we will be unable to prove rigorously what we strongly believe to be true, namely:

Conjecture 1.3 *With notation as in Theorem 1.2, we have in fact that*

$$E(h,k) \sim \left(\frac{1}{\underline{R}(k)}\right)^{\frac{k}{k+1}} f(h,k). \tag{1.9}$$

We will summarise these outstanding issues in Sect. 5.

2 The Postage Stamp Problem

For an up-to-date and much more thorough exposition of the material in this section, including an explanation of the name 'PSP', see [HJ2]. A more concise, but older, exposition can be found in [AB].

Let positive integers h, k be given. The *postage stamp number* $n(h,k)$ is the largest integer n such that there exists a k-element set A of positive integers satisfying $hA_0 \supseteq \{0,1,\ldots,n\}$, where $A_0 = A \cup \{0\}$. The problem of determining the numbers $n(h,k)$ is usually traced back to a 1937 paper of Rohrbach [R]. Historically, two special cases have attracted most attention: either h is fixed and $k \to \infty$ or vice versa. The two cases seem to be about equally difficult and the current state of knowledge is about the same in both. For our applications to essential subsets of bases, it turns out however that we can make do with much less information in the case when h is fixed. The following estimate, already proven by Rohrbach and valid for any h and k, will suffice:

$$\left(\frac{k}{h}\right)^h \le n(h,k) \le \binom{h+k}{h}. \tag{2.1}$$

The upper bound in (2.1) is obtained by a simple counting argument, and the lower bound is developed constructively. Regarding the former, observe that for h fixed and k going to infinity,

$$\binom{h+k}{h} = \frac{k^h}{h!} + O(k^{h-1}). \tag{2.2}$$

Now let us turn instead to the case when k is fixed and $h \to \infty$. Stöhr [S] proved the following analogue of Rohrbach's estimates:

$$\left(\left\lfloor \frac{h}{k} \right\rfloor + 1\right)^k \le n(h,k) \le \binom{h+k}{k}. \tag{2.3}$$

Let[2]

$$s(h,k) := \frac{1}{k}\left(\frac{n(h,k)}{h^k}\right)^{-1/k}. \tag{2.4}$$

For each fixed k, the limit

$$s(k) := \lim_{h\to\infty} s(h,k) \tag{2.5}$$

[2] We have not seen the numbers defined in equations (2.4), (2.5), (2.10), (2.11), (2.17) and (2.18) introduced explicitly in the existing literature on the PSP.

is known to exist [K] and it follows easily from (2.3) that, for each k,

$$\frac{1}{e} < s(k) \leq 1. \tag{2.6}$$

Only three values are known:

$$s(1) = 1, \quad s(2) = 1, \quad s(3) = \sqrt[3]{3/4}. \tag{2.7}$$

The first of these is trivial, the second due to Stöhr and the third to Hofmeister [Ho]. For general k the best-known lower bound on $s(k)$ tends to $1/e$ as $k \to \infty$, but for upper bounds it follows from work of Mrose [M] that

$$\limsup_{k\to\infty} s(k) \leq \frac{1}{\sqrt[4]{2}}. \tag{2.8}$$

In more recent times, the PSP has received more attention in the setting of finite cyclic groups, partly because it can then be formulated in terms of diameters of so-called Cayley graphs, which has applications in the theory of communication networks. We let $N(h,k)$ denote the largest integer N such that there exists a k-element subset A of $\mathbb{Z}_N \backslash \{0\}$ satisfying $hA_0 = \mathbb{Z}_N$, where $A_0 = A \cup \{0\}$. It is trivial that

$$N(h,k) \geq n(h,k) - 1. \tag{2.9}$$

Bounds similar to (2.1) and (2.3) can be easily obtained, so that if we define

$$S(h,k) := \frac{1}{k}\left(\frac{N(h,k)}{h^k}\right)^{-1/k}, \tag{2.10}$$

$$\underline{S}(k) := \liminf_{h\to\infty} S(h,k), \quad \overline{S}(k) := \limsup_{h\to\infty} S(h,k), \tag{2.11}$$

then it can be shown that

$$1/e < \underline{S}(k) \leq \overline{S}(k) \leq 1. \tag{2.12}$$

When the limit exists in (2.11), we denote it $S(k)$. Existence of the limit does not seem to be known in general. Intuitively, the reason why the numbers $N(h,k)$ are more awkward to handle than the $n(h,k)$ is as follows: If A is a set of integers such that $hA \supseteq \{0,1,\ldots,n\}$, then naturally $hA \supseteq \{0,1,\ldots,m\}$ for any $m < n$ also. But the corresponding statement for $\mathbb{Z}_n$ and $\mathbb{Z}_m$ need not be true.

For $k = 2$, it is known that the limit exists and that

$$S(2) = \sqrt{2/3}. \tag{2.13}$$

The first rigorous proof of this result seems to be in [HJ1]. No other values of $\underline{S}(k), \overline{S}(k)$ are known. Once again, no general lower bound is known which doesn't tend to $1/e$ as $k \to \infty$. The current record for general upper bounds seems to be due to Su [Su] :

$$\limsup_{k\to\infty} \overline{S}(k) \leq \sqrt[5]{\frac{17^5}{5^5 \cdot 7^4}}. \tag{2.14}$$

There is a natural 'dual' to the numbers $N(h,k)$. This time, let N, k be given positive integers, with $N > k$. We define $h(N,k)$[3] to be the smallest positive integer h such that there exists a basis for $\mathbb{Z}_N$ of order h containing $k+1$ elements. For applications to Cayley graphs and also, as we shall see, to essential subsets of bases, the numbers $h(N,k)$ are a more natural choice to work with than the $N(h,k)$. The duality between the two is expressed by the easy relations

$$t \leq N(h(t,k),k), \quad \text{and} \quad h \geq h(N(h,k),k), \quad \text{for any } t,h,k \in \mathbb{N}. \tag{2.15}$$

A dual to (2.3) proven by Wang and Coppersmith [WC] is the double inequality

$$\sqrt[k]{k!\, N} - \frac{k+1}{2} \leq h(N,k) \leq k \cdot (\sqrt[k]{N} - 1). \tag{2.16}$$

The natural counterparts to the numbers $S(h,k), \underline{S}(k), \overline{S}(k)$ are thus

$$R(h,k) := \frac{h(N,k)}{k \cdot \sqrt[k]{N}}, \tag{2.17}$$

$$\underline{R}(k) := \liminf_{h\to\infty} R(h,k), \quad \overline{R}(k) := \limsup_{h\to\infty} R(h,k). \tag{2.18}$$

The numbers $\underline{R}(k)$ are those appearing in Theorem 1.2. From (2.16), we have

$$1/e < \underline{R}(k) \leq \overline{R}(k) \leq 1. \tag{2.19}$$

Again it is natural to conjecture that the limits always exist and then that $R(k) = S(k)$. All we can immediately deduce from (2.15), however, is that

$$\overline{R}(k) \geq \underline{S}(k) \quad \text{and} \quad \underline{R}(k) \leq \overline{S}(k). \tag{2.20}$$

Apart from what can then be deduced from (2.13), (2.14), and (2.20), very little seems to be known, though it was shown in [WC] that $\underline{R}(2) = S(2) = \sqrt{2/3}$. In particular, existence of the limits $R(k)$ does not seem to be known for a single value of $k > 1$. The subtle difficulty in handling the numbers $N(h,k)$ referred to above is thus fully reflected in the $h(N,k)$. Tables of values computed in [HJ1] show that $h(N,k)$ is not even a nondecreasing function of N.

[3] The notation $d(N,k)$ is common in the literature, since these numbers can be interpreted as *diameters* of Cayley graphs.

3 Proof of Theorem 1.1

Let A be a basis for $\mathbb{N}$ of order h with s essential subsets of size k, say $E_1, \ldots, E_s$. We think of h as being fixed and k, s large. Let $E := \cup_i E_i$, $E_0 := E \cup \{0\}$ and, for each i,

$$d_i := \text{GCD } \{a - a' : a, a' \in A \backslash E_i\}. \tag{3.1}$$

Then each $d_i > 1$ and these numbers are relatively prime ([DF], Lemma 12). So if the d_i are in increasing order, then $d_i \geq p_i$, the i:th prime. Let $d := \prod_i d_i$. Thus,

$$d \geq \prod_{i=1}^{s} p_i \gtrsim \left(\frac{s \log s}{\alpha}\right)^s, \tag{3.2}$$

for some absolute constant $\alpha > 0$. This latter estimate for the product of the first s primes is well-known : see, for example, [Rob].

Next, let $\alpha_1, \ldots, \alpha_s$ be numbers such that $a \equiv \alpha_i \pmod{d_i}$ for all $a \in A \backslash E_i$. Without loss of generality, each $\alpha_i = 0$ (otherwise, choose a negative integer α such that $\alpha \equiv \alpha_i \pmod{d_i}$ for each i, and replace A by the shifted set $A - \alpha$). Now since A is a basis for $\mathbb{N}$ of order h, the numbers in E_0 must, when considered modulo d, form a basis for $\mathbb{Z}_d$ of order at most $h - 1$. Thus,

$$d \leq N(h - 1, ks). \tag{3.3}$$

From (3.2), (3.3), (2.9), (2.1), and (2.2), it is easily verified that

$$s \lesssim (h - 1)\frac{\log k}{\log \log k}, \tag{3.4}$$

which proves that the right-hand side of (1.5) is asymptotically an upper bound for $E(h, k)$.

For the lower bound, we turn the above argument on its head. Let h be given and k a very large integer. We wish to construct a subset A of $\mathbb{N}$, which is a basis of order h and has about $(h - 1)\frac{\log k}{\log \log k}$ essential subsets of size k. Our example is modeled on that in [DF], and presented in Sect. 1. To begin with, let s be the largest integer such that

$$(hs) \cdot \left(\prod_{i=1}^{s} p_i\right)^{\frac{1}{h-1}} \leq k. \tag{3.5}$$

From (3.2) we have

$$s \sim (h - 1) \cdot \left(\frac{\log k}{\log \log k}\right). \tag{3.6}$$

Put $\mathscr{P} := \prod_{i=1}^{s} p_i$. By the left-hand inequality in (2.1), there exists a set $F \subseteq \{1, \ldots, \mathscr{P} - 1\}$ with

$$|F| \leq (h - 1) \cdot \mathscr{P}^{\frac{1}{h-1}} \leq k \tag{3.7}$$

and such that, considered modulo $\mathscr{P}$, F_0 is a basis for $\mathbb{Z}_{\mathscr{P}}$ of order $h-1$. For each $i = 1, \ldots, s$, let $F_i := \{x \in F : (x, p_i) = 1\}$. Thus, $|F_i| \leq k$ for each i also. We wish to augment the set F to a set E, still contained inside $\{1, \ldots, \mathscr{P} - 1\}$, such that two conditions are satisfied:

(i) E_0 is still a basis of order $h-1$ for $\mathbb{Z}_{\mathscr{P}}$, i.e., it is not a basis of strictly smaller order,
(ii) $|E_i| = k$, for $i = 1, \ldots, s$.

Note that, for sufficiently large k, (i) will follow from (ii) by the choice of s. Let $\mathscr{G} := \{1, \ldots, \mathscr{P} - 1\} \backslash F$ and, for each i,

$$\mathscr{G}_i := \{x \in \mathscr{G} : p_i | x \text{ and } (x, p_j) = 1 \text{ for all } j \neq i\}. \tag{3.8}$$

Note that the sets $\mathscr{G}_i$ are pairwise disjoint and that, from (3.7) and Mertens theorem,

$$|\mathscr{G}_i| = \Theta\left(\frac{\mathscr{P}}{\log s}\right), \quad \text{for } i = 1, \ldots, s. \tag{3.9}$$

Put $f_i := |F_i|$. First of all, add in at most $s-2$ multiples of $p_{s-1}p_s$ from $\mathscr{G}$ to F so that at this point

$$\textstyle\sum_{i=1}^{s} f_i \text{ is a multiple of } s-1. \tag{3.10}$$

Now we want to throw in g_i elements of $\mathscr{G}_i$ so that, for each i,

$$f_i + \sum_{j \neq i} g_j = k. \tag{3.11}$$

The unique solution to the linear system (3.11) is

$$g_i = \frac{k + (s-1)f_i - \sum_{i=1}^{s} f_i}{s-1} \tag{3.12}$$

and, by (3.5), (3.7), (3.9), and (3.10), the right-hand side of (3.12) is a positive integer less than $|\mathscr{G}_i|$ for each i, as desired. The set E now consists of F together with all the numbers we have thrown in during the above process and, by construction, it satisfies (ii). Finally, then, let $A \subseteq \mathbb{N}$ be given by

$$A = (\mathscr{P} \cdot \mathbb{N}) \cup E. \tag{3.13}$$

Since E is a basis of order $h-1$ for $\mathbb{Z}_{\mathscr{P}}$, it follows that A is an integer basis of order h. By construction, it has s essential subsets of size k, namely the sets $E_1, \ldots, E_s$. From (3.6), we thus have what we want, and so the proof of Theorem 1.1 is complete.

4 Proof of Theorem 1.2

First we consider the upper bound (1.8). As in the previous section, let A be a basis for $\mathbb{N}$ of order h with s essential subsets of size k, say $E_1, \ldots, E_s$. This time we think of k as being fixed and h very large. Let

$$E_i = \{a_{i,j} : j = 1, \ldots, k\}, \quad i = 1, \ldots, s. \tag{4.1}$$

Let the numbers d_i be as in (3.1), $d := \prod_i d_i$ and $A^* := A \setminus (\cup_i E_i)$. As before, we can argue that, without loss of generality, $a \equiv 0 \pmod{d}$ for all $a \in A^*$. Now, with the numbers $h(\cdot, \cdot)$ defined as in Sect. 2, we claim that

$$h \geq \sum_{i=1}^{s} h(d_i, k). \tag{4.2}$$

To see this, first note that, by definition of the numbers $h(d_i, k)$, there exist integers x_i such that, for each i, no representation

$$x_i \equiv \sum_{j=1}^{k} \gamma_{i,j} a_{i,j} \pmod{d_i} \tag{4.3}$$

exists satisfying

$$\gamma_{i,j} \in \mathbb{N}_0, \quad \sum_j \gamma_{i,j} < h(d_i, k). \tag{4.4}$$

Now let x be any positive integer satisfying $x \in hA$ and $x \equiv x_i \pmod{d_i}$ for $i = 1, \ldots, s$. Since $x \in hA$ there exists a representation

$$x = \sum_{i=1}^{s} \sum_{j=1}^{k} \gamma_{i,j} a_{i,j} + \sum_{a \in A^0} a, \tag{4.5}$$

where A^0 is some multisubset of A^*, each $\gamma_{i,j} \geq 0$ and

$$h = |A^0| + \sum_{i,j} \gamma_{i,j}. \tag{4.6}$$

But reducing (4.5) modulo d_i gives a congruence of the form (4.3) for each i. Thus (4.2) follows from (4.4) and (4.6).

Now let $h \to \infty$. Then

$$h \geq \sum_{i=1}^{s} h(d_i, k) \gtrsim k \cdot \underline{R}(k) \cdot \sum_{i=1}^{s} \sqrt[k]{d_i} \tag{4.7}$$

and

$$\sum_{i=1}^{s} \sqrt[k]{d_i} \geq \sum_{i=1}^{s} \sqrt[k]{p_i} \gtrsim \sum_{i=1}^{s} \sqrt[k]{i \log i}$$
$$\gtrsim \int_1^s (x \log x)^{1/k} \, \mathrm{d}x \gtrsim \frac{k}{k+1} \left(s^{k+1} \log s \right)^{1/k}, \tag{4.8}$$

where the integral has been easily estimated using partial integration. Summarizing, we have shown that

$$(s^{k+1} \log s)^{1/k} \lesssim \left(\frac{k+1}{k^2} \frac{1}{\underline{R}(k)} \right) h. \tag{4.9}$$

Choosing the set A so that $s = E(h,k)$, this is easily checked to yield (1.8).

So to the lower bound (1.7). Once again, we wish to turn the above argument on its head. In [P], the author considered the case $k = 1$. To show that the function $E(h,1)$ behaved regularly, he needed to know that every sufficiently large positive integer could be expressed as $\sum(p-1)$, the sum being over a particular type of set of prime numbers. In the present context, one should think of $p-1$ as being the number $h(p,1)$. To generalize the argument directly and prove Conjecture 1.3, it would suffice that, for each $k > 1$, every sufficiently large integer could be expressed as $\sum h(p,k)$, the sum being over a similar set of primes with the additional property that the numbers $R(p,k)$ approach $\underline{R}(k)$ as $p \to \infty$. Of course, if we also knew that the limits $R(k)$ existed, then we wouldn't need to worry about the latter bit. We do not see how to carry out this procedure, given the current state of knowledge about extremal bases in finite cyclic groups, though we strongly believe it can be done, perhaps with some small modifications. Instead, we prove the weaker inequality (1.7) by constructing, for all large primes, a large number of bases for $\mathbb{Z}_p$ all of which are fairly close to extremal (Theorem 4.4). These bases are sufficiently plentiful to allow us to deal easily with further technical issues concerning the distribution of primes in short intervals (Theorem 4.3). Now to the details. We begin with a pair of lemmas.

The first is a result of Alon and Frieman also used in [P]. Recall the following notations : If X is a finite subset of $\mathbb{N}$ then $\Sigma(X)$ denotes the collection of all subset sums from X. If $q \in \mathbb{N}$ then we denote $X(q) := \{x \in X : q|x\}$. We also set

$$S_X := \sum_{x \in X} x \tag{4.10}$$

and

$$B_X := \sqrt{\sum_{x \in X} x^2}. \tag{4.11}$$

Then there is the following result :

Lemma 4.1 *[AF] For each $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that if $n \geq n_0$ and $X \subseteq \{1, \ldots, n\}$ satisfies $|X| > n^{2/3+\epsilon}$ and $|X \backslash X(q)| \geq n^{2/3}$ for each $q \geq 2$, then*

$$\left\{ \left\lceil \frac{1}{2} S_X - \frac{1}{2} B_X \right\rceil, \ldots, \left\lfloor \frac{1}{2} S_X + \frac{1}{2} B_X \right\rfloor \right\} \subseteq \Sigma(X). \tag{4.12}$$

Our second lemma will be a rather general result about the representability of sufficiently large integers as a certain type of subset sum in a sufficiently dense multisubset of $\mathbb{N}$. Here we need to make precise some terminology.

By a *multisubset* A of $\mathbb{N}$, we mean a collection of positive integers where repetitions are allowed. We assume that each integer occurs only finitely many times in a multisubset. If $a_1 \leq a_2 \leq$ are the elements of A written in some nondecreasing order, then we denote this by $A = (a_i)$. We shall say that A is *weakly increasing* if the following holds : for each $\epsilon > 0$ there exists $\delta > 0$ such that, for all $n \gg_\epsilon 0$,

$$\frac{a_{\lfloor (1+\epsilon)n \rfloor}}{a_n} > 1 + \delta. \tag{4.13}$$

If A is a multisubset of $\mathbb{N}$ we denote by $A^{\#}$ the subset of $\mathbb{N}$ consisting of all those numbers which appear at least once in A. Now recall that if $X \subseteq \mathbb{N}$, the *lower asymptotic density* of X, denoted $\underline{d}(X)$, is defined as

$$\underline{d}(X) = \liminf_{n \to \infty} \frac{|X \cap [1, n]|}{n}. \tag{4.14}$$

Our lemma is the following :

Lemma 4.2 *Let $A = (a_i)$ be a weakly increasing multisubset of $\mathbb{N}$ such that $\underline{d}(A^{\#}) = 1$. Let $\epsilon > 0$. Then for all $h \gg_\epsilon 0$, there exists some representation of h as a sum*

$$h = \sum_{i=1}^{n} a_i + \sum_{j \in \mathscr{J}} a_j, \tag{4.15}$$

where $\mathscr{J} \subseteq A^{\#} \cap [a_n, (1+\epsilon)a_n]$. Here n depends on h, but $n \to \infty$ as $h \to \infty$.

Proof. Fix $\epsilon > 0$. For each $n > 0$ set

$$A_n^{\#} := A^{\#} \cap [a_n, a_{\lfloor (1+\epsilon)n \rfloor}]. \tag{4.16}$$

Now define the sequence $(u_n)_{n=1}^{\infty}$ by

$$u_n := \sum_{i=1}^{n} a_i + \frac{1}{2} \sum_{j \in A_n^{\#}} a_j. \tag{4.17}$$

The sequence u_n is evidently increasing and, if $n' := \lfloor (1+\epsilon)n \rfloor$, then

$$u_{n+1} - u_n \leq a_{n+1} + \frac{1}{2}\left(a_{n'+1} + a_{n'+2}\right). \tag{4.18}$$

Since $\underline{d}(A^{\#}) = 1$, it follows that

$$u_{n+1} - u_n \leq (1 + O(\epsilon))a_n. \tag{4.19}$$

Now let h be a very large integer (how large h needs to be will become clear in what follows). Let n be the largest integer such that $u_n < h$. Put $h' = h - u_n$. By (4.19), we have that, in the notation of (4.10),

$$\left| h' - \frac{1}{2} S_{A_n^{\#}} \right| = O(a_n). \tag{4.20}$$

Since A is weakly increasing, when h and thus n are sufficiently large, there exists $\delta > 0$ such that

$$\frac{a_{\lfloor (1+\epsilon)n \rfloor}}{a_n} > 1 + \delta. \tag{4.21}$$

Furthermore, since $\underline{d}(A^{\#}) = 1$ then for any $\delta' > 0$ and $h >> 0$, the set $A_n^{\#}$ will contain at least the fraction $1 - \delta'$ of all numbers in the interval $[a_n, a_{\lfloor (1+\epsilon)n \rfloor}]$. What all of this means is that $A_n^{\#}$ will satisfy the hypotheses of Lemma 4.1 and moreover that, in the notation of (4.11), $B_{A_n^{\#}} = \Omega(a_n^{3/2})$. Hence, by Lemma 4.1 and (4.20) it follows that, provided h is sufficiently large, there is a subset $\mathscr{J} \subseteq A_n^{\#}$ such that $h' = \sum_{a_j \in \mathscr{J}} a_j$. From the definition of h', this implies (4.15) and so the proof of the lemma is complete. □

Let $\mathbb{P} = (p_i)$ denote the sequence of primes, as usual. We now have:

Theorem 4.3. *Let k be a positive integer and $\epsilon > 0$. Then for all integers $h \gg_{k,\epsilon} 0$, there exists a representation*

$$h = \sum_{i=1}^{n} \lfloor \sqrt[k]{p} \rfloor + \sum_{j \in \mathscr{J}} \lfloor \sqrt[k]{p_j} \rfloor, \tag{4.22}$$

where $\mathscr{J} \subseteq \{n+1, \ldots, \lfloor (1+\epsilon)n \rfloor\}$.

Proof. Fix k and ϵ. Let A denote the multisubset of $\mathbb{N}$ consisting of the integer parts of the k:th roots of all the primes. To prove the theorem, we just need to verify that A satisfies the hypotheses of Lemma 4.2. Clearly, A is weakly increasing. It is also the case that $\underline{d}(A^{\#}) = 1$, in other words, that almost every positive integer is the integer part of the k:th root of some prime. While it is generally believed that, in fact, $A^{\#} = \mathbb{N}$, for any $k \geq 2$, what is known for certain is that $\mathbb{N} \setminus A^{\#}$ is finite for

any $k \geq 3$, and that $\underline{d}(A^{\#}) = 1$ for $k = 2$. These facts are easy consequences of the following two well-known theorems respectively (in each case, the exponents given are the smallest that have been arrived at to date, to the best of our knowledge):

RESULT 1 [H-B]: As $n \to \infty$ one has

$$\pi(n+t) - \pi(n) \sim \frac{t}{\log n}, \tag{4.23}$$

whenever $n^{7/12} \leq t \leq n$.

RESULT 2 [J]: For each $\epsilon > 0$, there is a prime in the interval $(n, n + n^{1/20+\epsilon})$ for almost all positive integers n.

Thus, our set A does indeed satisfy the hypotheses of Lemma 4.2, and thus the proof of Theorem 4.3 is complete. □

The above will take care of the technicalities arising from the distribution of the primes. We now turn to the construction of reasonably efficient bases in finite cyclic groups.

Theorem 4.4. *Let $k \geq 2$ be an integer. There exists an absolute constant $c > 0$, independent of k, such that, for all primes $p \gg_k 0$, and all s such that $0 \leq s < c\lfloor \sqrt[k]{p} \rfloor$, there exists a set A of k nonzero elements of $\mathbb{Z}_p$ such that $A \cup \{0\}$ is a basis for $\mathbb{Z}_p$ of order $k \cdot \lfloor \sqrt[k]{p} \rfloor + s$.*

Remark 4.5. This is overkill for our purposes. It would suffice for us to know that there exist $(k+1)$-element bases for $\mathbb{Z}_p$ of order $k \cdot \lfloor \sqrt[k]{p} \rfloor + s$ for each $s \in \{0, 1\}$. But we think the result as stated may be of independent interest – see Sect. 5.

Proof. Fix $k \geq 2$ and let p be a prime. Let $x := \lfloor \sqrt[k]{p} \rfloor$ and $\epsilon := \sqrt[k]{p} - x$. Thus $\epsilon \in (0, 1)$. Our goal is to construct, for some constant $c > 0$ and all each $s \in \{0, 1, \ldots, \lfloor cx \rfloor\}$, a subset $A \subseteq \mathbb{Z}_p^{\times}$ of size k such that $A_0 := A \cup \{0\}$ is a basis for $\mathbb{Z}_p$ of order $kx + s$. By the binomial theorem,

$$x^k = p - \sum_{j=1}^{k} \binom{k}{j} \epsilon^j x^{k-j}. \tag{4.24}$$

In particular, it is clear that, for $p \gg 0$ we will have

$$p - (k+1)x^{k-1} < x < p. \tag{4.25}$$

First consider $A := \{1, x, x^2, \ldots, x^{k-1}\}$. Then A_0 is a basis of order $(kx - k) + u$, where u is the smallest integer such that

$$(x-2)\cdot 1 + (x-1)\cdot x + (x-1)\cdot x^2 + \cdots + (x-1)\cdot x^{k-2} + (x+u)\cdot x^{k-1} \geq p-1. \tag{4.26}$$

The left-hand side of (4.26) is just $x^k + (u+1)x^{k-1} - 2$. Hence, if $p \gg 0$, (4.25) implies that $0 \le u \le k$. Thus, A_0 is a basis of order $kx - j$ for some $j \in \{0, 1, \dots, k\}$. Now let t be any integer and consider

$$A_t := \{1, x, x^2, \dots, x^{k-2}, x^{k-2}(x-t)\}. \tag{4.27}$$

If t is small compared to x then $A_{t,0}$ will be a basis of order $(kx-k)+(u_t-t)$, where u_t is the smallest integer such that

$$\begin{aligned}(x-2)\cdot 1 + (x-1)\cdot x + (x-1)\cdot x^2 + \cdots + (x-1)\cdot x^{k-3} \\ + (x-t-1)\cdot x^{k-2} + (x+u_t)\cdot x^{k-2}(x-t) \ge p-1.\end{aligned} \tag{4.28}$$

Let $v_t := u_t - t$. We have already seen above that $0 \le v_0 \le k$. The theorem will be proved if we can show that there are values of t for which v_t takes on each of the values $k, k+1, \dots, k+\lfloor cx \rfloor$, for some absolute constant $c > 0$. After some tedious computation where we make use of (4.24), the inequality (4.28) reduces to

$$x^{k-2}\left[(u_t+1)(x-t) - tx\right] \ge 1 + \sum_{j=1}^{k} \binom{k}{j} \epsilon^j x^{k-j}. \tag{4.29}$$

Note that the right-hand side is independent of t. Denote it simply by Σ and note from (4.24) that $x^k + \Sigma = p + 1$. Then from (4.29) we easily deduce that $v_t = \lceil f(t) \rceil$, where the real-valued function f of one variable is given by

$$f(\xi) = \frac{\Sigma + \xi x^{k-1}}{x^{k-2}(x-\xi)} - (\xi + 1). \tag{4.30}$$

One easily computes that

$$f'(\xi) = \frac{p+1}{x^{k-2}(x-\xi)^2} - 1, \tag{4.31}$$

hence that

$$f'(\xi) = \frac{1 + o_p(1)}{(1-\xi/x)^2} - 1. \tag{4.32}$$

Thus f is increasing in the range $0 \le \xi < x$, $f'(\xi) = \Theta(1)$ when $\xi = \Theta(x)$ and $f'(\xi) \le 1 + o_p(1)$ when $\xi/x \le 1 - 1/\sqrt{2}$. It follows easily that, as t increases, the integer-valued quantity v_t takes on a sequence of $\Theta(x)$ consecutive values, starting at v_0. This suffices to prove Theorem 4.4. □

Now we are ready to prove inequality (1.7). Let $k \ge 2$ be a fixed integer. Let h be a positive integer and write $h = kh_1 + s$ where $0 \le s < k$. Let $\epsilon > 0$. If $h \gg_{\epsilon,k} 0$ then, by Theorem 4.3 there exists a representation

$$h_1 = \sum_{i=1}^{n} \lfloor \sqrt[k]{p_i} \rfloor + \sum_{j \in \mathscr{J}} \lfloor \sqrt[k]{p_j} \rfloor, \tag{4.33}$$

where $\mathscr{J} \subseteq \{n+1, \ldots, \lfloor (1+\epsilon)n \rfloor\}$. For each prime $p_i > p_k$ in this sum we wish to choose a k-element subset A_i of $\{1, 2, \ldots, p_i - 1\}$ such that, if we identify A_i with a subset of $\mathbb{Z}_p$ and let r_i denote the order of $A_i \cup \{0\}$ as a basis for $\mathbb{Z}_p$, then

$$r_i = k \cdot \sqrt[k]{p_i} + O(1), \tag{4.34}$$

and

$$\sum r_i = h. \tag{4.35}$$

From (4.33) and Theorem 4.4 (see Remark 4.5 in fact), it is clear that such a choice is possible, for sufficiently large h. Set $\mathscr{I} := \{1, \ldots, n\} \cup \mathscr{J}$, $\mathscr{P} := \prod_{i \in \mathscr{I}} p_i$ and, for each i, $\mathscr{P}_i := \mathscr{P}/p_i$. For each $i \in \mathscr{I} \backslash \{1, \ldots, k\}$ set

$$A_i := \{a_{i,j} : j = 1, \ldots, k\}, \tag{4.36}$$

and

$$E_i := \{a_{i,j} \mathscr{P}_i : j = 1, \ldots, k\}. \tag{4.37}$$

Now consider the subset $A \subseteq \mathbb{N}$ given by

$$A = \mathscr{P} \cdot \mathbb{N} \cup \left(\bigcup_{i \in \mathscr{I} \backslash \{1,,k\}} E_i \right). \tag{4.38}$$

By construction, the set A is a basis for $\mathbb{N}$ of order h and contains $|\mathscr{I}| - k$ essential subsets of size k, namely each of the sets E_i. The proofs of these assertions are similar to those of the corresponding assertions in [P] (see page 9 of that paper), so we do not include them. For the purpose of obtaining the right-hand side of (1.7) as a lower bound for the asymptotic behavior of $E(h, k)$, it now suffices to show that

$$|\mathscr{I}| - k \geq (1 - O(\epsilon)) \left(\frac{k+1}{k^2} \cdot \sqrt[k+1]{k} \right) \left(\frac{h^k}{\log h} \right)^{\frac{1}{k+1}}. \tag{4.39}$$

First, it is obvious that

$$|\mathscr{I}| - k = (1 + O(\epsilon))n. \tag{4.40}$$

Second, it follows from (4.33) and (4.34) that

$$h \leq (1 + O(\epsilon)) \cdot k \cdot \sum_{i=1}^{\lfloor (1+\epsilon)n \rfloor} \sqrt[k]{p_i}. \tag{4.41}$$

Hence if we can show that

$$\sum_{i=1}^{n} \sqrt[k]{p_i} \sim \frac{k}{k+1} \left(n^{k+1} \log n \right)^{1/k}, \tag{4.42}$$

then this and (4.40)–(4.41) are easily seen to imply (4.39). But (4.42) has already been established in (4.8), and so our proof of Theorem 1.2 is complete.

5 Discussion

We have seen that an entirely satisfactory estimate for the function $E(h,k)$ cannot be obtained without significant progress on the PSP in the case when k is fixed and $h \to \infty$. Specifically, one needs to know the numbers $\underline{R}(k)$ given by (2.18). Even then, a subtle technicality arises in attempting to rigorously prove Conjecture 1.3, as was alluded to in Sect. 4. It is possible, though highly unlikely, that not all sufficiently large integers can be expressed as sums $\sum h(p,k)$ over certain sets of primes, as in Theorem 4.3. For example, it could happen that $h(p,k)$ was a multiple of k for every p. Note that the upper bound in (2.16) has this property, and it was just this fact that necessitated the long detour via Theorem 4.4 when trying to prove (1.7). Theorem 4.4 may be independently interesting in the sense that one can ask a very general question as to what are the possible orders of an arbitrary $(k+1)$-element basis for $\mathbb{Z}_n$. A special case would be to ask for the best-possible c in the statement of that theorem. Does $c \to \infty$ as p does? For the proof of Conjecture 1.3, one would instead like to know what is the largest possible $C = C(p,k)$ such that there exists a $(k+1)$-element basis for $\mathbb{Z}_p$ of order $h(p,k)+s$, for every $0 \le s \le C(p,k)$. Can we take $C(p,k) = \Omega\left(\sqrt[k]{p}\right)$?

Acknowledgment I thank Alain Plagne for very helpful discussions and Melvyn Nathanson for some literature tips on the PSP. This work was completed while I was visiting City University of New York, and I thank them for their hospitality. My research is partly supported by a grant from the Swedish Research Council (Vetenskapsrådet).

References

[AB] R. Alter and J. A. Barrett, *A postage stamp problem*, Amer. Math. Monthly **87** (1980), 206–210.

[AF] N. Alon and G. A. Freiman, *On sums of subsets of a set of integers*, Combinatorica **8** (1988), 297–306.

[DF] B. Deschamps and B. Farhi, *Essentialité dans les bases additives* (French), J. Number Theory **123** (2007), 170–192.

[EG] P. Erdős and R. L. Graham, *On bases with an exact order*, Acta Arith. **37** (1980), 201–207.

[G] G. Grekos, *Sur l'ordre d'une base additive* (French), Séminaire de théorie des nombres de Bordeaux, Année 1987/88, exposé 31.

[H-B] D. R. Heath-Brown, *The number of primes in a short interval*, J. Reine Angew. Math. **389** (1988), 22–63.

[He] P. V. Hegarty, *Essentialities in additive bases*. Preprint available at http://arxiv.org/abs/0802.2928

[Ho] G. Hofmeister, *Asymptotische Abschatzungen für dreielementige Extremalbasen in natürlichen Zahlen* (German), J. Reine Angew. Math. **232** (1968), 77–101.

[HJ1] D. F. Hsu and X. D. Jia, *Extremal problems in the construction of distributed loop networks*, SIAM J. Disc. Math. **7** (1994), 57–71.

[HJ2] D. F. Hsu and X. D. Jia, *Additive bases and extremal problems in groups, graphs and networks*, Util. Math. **66** (2004), 61–91.

[J] C. Jia, *Almost all short intervals containing primes*, Acta Arith. **76** (1996), No. 1, 21–84.

[K] C. Kirfel, *On extremal bases for the h-range problem, II*. Report 55, Department of Mathematics, University of Bergen, Norway (1990). A copy of the paper canbe obtained from the author upon request.

[M] A. Mrose, *Untere Schranken für die Reichweiten von Extremalbasen fester Ordnung* (German), Abh. Math. Sem. Univ. Hamburg **48** (1979), 118–124.

[P] A. Plagne, *Sur le nombre d'éléments exceptionnels d'une base additive* (French), J. Reine Angew. Math. **618** (2008), xx–yy.

[R] H. Rohrbach, *Ein Beitrag zur additiven Zahlentheorie* (German), Math. Z. **42** (1937), 1–30.

[Rob] G. Robin, *Estimation de la fonction de Tchebychev*, Acta Arith. **42** (1983), 367–389.

[S] A. Stöhr, *Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe I* (German), J. Reine Angew. Math. **194** (1955), 40–65.

[Su] W. Su, *A combinatorial problem in the construction of distributed loop networks*. Master's Thesis, Southwest Texas State University (1993).

[WC] C. K. Wong and D. Coppersmith, *A combinatorial problem related to multimode memory organisation*, J. Assoc. Comp. Mach. **21** (1974), No. 3, 392–402.

A Universal Stein-Tomas Restriction Estimate for Measures in Three Dimensions

Alex Iosevich and Svetlana Roudenko

Summary We study restriction estimates in $\mathbb{R}^3$ for surfaces given as graphs of low regularity functions. We obtain a "universal" mixed-norm estimate for the extension operator $f \to \widehat{f\mu}$ in $\mathbb{R}^3$. We also prove that this estimate holds for any Frostman measure supported on a compact set of Hausdorff dimension greater than two. The approach is geometric and is influenced by a connection with the Falconer distance problem.

Keywords Restriction estimates · Measures

Mathematics Subject Classifications (2010). 42B

1 Introduction

The classical Stein-Tomas restriction theorem says that if μ is the Lebesgue measure on S^{d-1}, the unit sphere in $\mathbb{R}^d$, or, more generally, on a smooth convex surface with everywhere nonvanishing curvature in $\mathbb{R}^d$, then

$$\|\widehat{f\mu}\|_{L^{\frac{2(d+1)}{d-1}}(\mathbb{R}^d)} \lesssim \|f\|_{L^2(S^{d-1})}, \tag{1}$$

where here, and throughout, $X \lesssim Y$ means that there exists $C > 0$ such that $X \leq CY$.

It is shown in [6] (see also [5]) that if the Gaussian curvature is allowed to vanish, (1) does not hold. Nevertheless, there is hope of obtaining (1) for all reasonable surfaces by modifying the surface carried measure in some universal way. For example, if μ_0 is the Lebesgue measure on a convex compact smooth surface Γ, of

A. Iosevich
University of Missouri, Columbia, MO 65211, USA
e-mail: iosevich@math.missouri.edu

S. Roudenko
Arizona State University, Tempe, AZ 85287, USA
e-mail: svetlana@math.asu.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_12,

finite type, in the sense that the order of contact with every tangent line is finite, and $\mathrm{d}\mu(x) = |S|^{\frac{1}{d+1}}(x)\mathrm{d}\mu_0(x)$, then one can check using standard techniques that the estimate (1) holds. The situation becomes much more complicated in the general context. See, for example, [2] and [8] for some very nice results in this direction where smooth radial hyper-surfaces are considered. See also [7] for some related work in the two-dimensional context.

In [1], the authors took a different point of view. Instead of imposing a fixed measure on the family of surfaces, they considered mixed norm restriction theorems corresponding to convex curves under rotations. The approach was heavily tied to the average decay estimates of the Fourier transform of the Lebesgue measure on convex curves, due to Podkorytov [9], which made the convexity assumption difficult to by-pass. In this paper, we take a geometric point of view which allows us to consider a much more general collection of surfaces. Our main result is the following.

Theorem 1. *Let μ be a Frostman measure on a compact $E \subset \mathbb{R}^3$ of Hausdorff dimension greater or equal to two. Suppose that for any $\delta > 0$ and ε sufficiently small with respect to δ,*

$$\mu \times \mu\{(u,v) : \delta \le |u-v| \le \delta(1+\varepsilon)\} \le C(\delta)\varepsilon. \tag{2}$$

Given $\theta \in \mathrm{SO}(3)$, the special orthogonal group in 3d over $\mathbb{R}$, define the measure $\mathrm{d}\mu_\theta$ via its action on a function $g : \mathbb{R}^3 \to \mathbb{R}$ by the formula

$$\int g(x)\,\mathrm{d}\mu_\theta(x) = \int g(\theta x)\,\mathrm{d}\mu(x).$$

Then,

$$\left(\int \left|\int_{\mathrm{SO}(3)} |\widehat{f\mu_\theta(x)}|^2 \mathrm{d}H(\theta)\right|^2 \mathrm{d}x\right)^{\frac{1}{4}} \lesssim \left[\int_{\mathrm{SO}(3)} \left(\int |f(x)|^2 \mathrm{d}\mu_\theta(x)\right)^2 \mathrm{d}H(\theta)\right]^{\frac{1}{4}}, \tag{3}$$

where $\mathrm{d}H(\theta)$ is the Haar measure on $\mathrm{SO}(3)$.

Remark 1. If μ is the Lebesgue measure on the sphere centered at the origin, rotations become superfluous, and we recover the classical Stein-Tomas restriction theorem:

$$||\widehat{f\mu}||_{L^4(\mathbb{R}^3)} \lesssim ||f||_{L^2(S^2)}.$$

The main point of this paper is in the following two statements:

Corollary 1. *Let μ is the Frostman measure on any compact subset of $\mathbb{R}^3$ of Hausdorff dimension greater than two. Then, the conclusion of Theorem 1 holds.*

Corollary 2. *Let μ be the Lebesgue measure on a compact surface in $\mathbb{R}^3$ given as a graph of a Lipschitz function. Then, the conclusion of Theorem 1 holds.*

While the conclusion of Theorem 1 holds for Lipschitz functions, as in Corollary 2, the condition (2), and consequently, the conclusion of Theorem 1, also holds for many measures supported on sets that are far from Lipschitz or even rectifiable in any sense. For example, consider a sequence of positive integers $\{q_i\}$ such that $q_1 = 2$ and $q_{i+1} > q_i^i$. Let E_q denote the $q^{-d/s}$, $0 < s < d$, neighborhood of $q^{-1}(\mathbb{Z}^d \cap [0, q]^d)$. Let $E_s = \cap_{i=1}^{\infty} E_{q_i}$. By standard geometric measure theory (see e.g. [3], Chap. 8), the Hausdorff dimension of E_s is s. Let $s = d - 1$. One can check by a direct calculation that (2) holds. See for example [4].

The example in the previous paragraph arises in the Falconer distance problem which asks whether the Lebesgue measure of the set of distances $\Delta(E) = \{|x-y| : x, y \in E\}$ is positive provided that the Hausdorff dimension of E is greater than $d/2$. Falconer [4] proved that the conclusion holds if the Hausdorff dimension of E is greater than $(d+1)/2$ and they key estimate he used was

$$\mu \times \mu\{(x, y) : 1 \leq |x - y| \leq 1 + \varepsilon\} \leq C\varepsilon$$

provided that the Hausdorff dimension of E is greater than $(d+1)/2$, where μ is a Frostman measure on E. This estimate and its refinements under additional regularity hypotheses will play a key role in the proof of Theorem 1, Corollary 1, and Corollary 2 below.

2 Reduction to the Key Geometric Estimate

We write this section in $\mathbb{R}^d$, for the sake of completeness, though we only use it in the three dimensional case in the sequel. Let

$$T(f, g)(x) = \int_{\mathrm{SO}(d)} f\mu_\theta * \widetilde{g\mu}_\theta(x)\, \mathrm{d}H_d(\theta),$$

where H_d is the Haar (probability) measure on $\mathrm{SO}(d)$ and $\tilde{f}(x) = \bar{f}(-x)$.

On one hand,

$$\begin{aligned} ||T(f,g)||_{L^1(\mathbb{R}^d)} &\leq \int_{\mathrm{SO}(d)} ||f||_{L^1(\mu_\theta)} \cdot ||g||_{L^1(\mu_\theta)} \mathrm{d}H_d(\theta) \\ &\leq \left(\int_{\mathrm{SO}(d)} ||f||^2_{L^1(\mu_\theta)} \mathrm{d}H_d(\theta) \right)^{\frac{1}{2}} \cdot \left(\int_{\mathrm{SO}(d)} ||g||^2_{L^1(\mu_\theta)} \mathrm{d}H_d(\theta) \right)^{\frac{1}{2}} \\ &\leq ||f||_{L^2(\mathrm{SO}(d))(L^1(\mu_\theta))} \cdot ||g||_{L^2(\mathrm{SO}(d)(L^1(\mu_\theta))}, \end{aligned}$$

since convolution of two L^1 functions is in L^1 by Fubini.

On the other hand,

$$||T(f,g)||_{L^\infty(\mathbb{R}^d)} \leq ||f||_{L^\infty(\mathrm{SO}(d))(L^\infty(\mu_\theta))} \cdot ||g||_{L^\infty(\mathrm{SO}(d))(L^\infty(\mu_\theta))} \cdot \sup_{x\in\mathbb{R}^d} \left| \int_{\mathrm{SO}(d)} \mu_\theta * \tilde{\mu}_\theta(x)\, \mathrm{d}H_d(\theta) \right|.$$

It follows by interpolation and setting $f = g$ that if

$$\sup_{x\in\mathbb{R}^d} \left| \int_{\mathrm{SO}(d)} \mu_\theta * \tilde{\mu}_\theta(x)\, \mathrm{d}H_d(\theta) \right| \lesssim 1, \tag{4}$$

then

$$\left(\int_{\mathbb{R}^d} \left(\int_{\mathrm{SO}(d)} |\widehat{f\mu_\theta}(x)|^2 \mathrm{d}H_d(\theta) \right)^2 \mathrm{d}x \right)^{\frac{1}{4}} \lesssim ||f||_{L^4(\mathrm{SO}(d))(L^2(\mu_\theta))}. \tag{5}$$

This reduces matters to the study of (4) and this is what the remainder of the paper is about.

3 Proof of Theorem 1 and Corollary 1

From now on, we assume that $d = 3$. A useful fact is that SO(3) is characterized as the set of all "direct" rotations in $\mathbb{R}^3$. For simplicity, we write $\mathrm{d}H(\theta)$ instead of $\mathrm{d}H_3(\theta)$. Observe that $\mu_\theta * \tilde{\mu}_\theta(x) = \mu * \tilde{\mu}(\theta x)$. Thus, by the Fourier inversion formula

$$\int_{\mathrm{SO}(3)} \mu_\theta * \tilde{\mu}_\theta(x)\, \mathrm{d}H(\theta) = \int_{\mathbb{R}^3} \int_{\mathrm{SO}(3)} \mathrm{e}^{2\pi \mathrm{i}\theta x\cdot\xi} |\widehat{\mu}(\xi)|^2\, \mathrm{d}H(\theta)\, \mathrm{d}\xi \tag{6}$$

$$= c \int \widehat{\sigma}(|x|\xi)\, |\widehat{\mu}(\xi)|^2\, \mathrm{d}\xi \tag{7}$$

$$= \lim_{\varepsilon\to 0} \varepsilon^{-1} \mu\times\mu\{(u,v) : |x|\leq|u-v|\leq|x|(1+\varepsilon)\}, \tag{8}$$

where σ is the Lebesgue surface measure on S^2. In the second to last line above, we have used the following observation.

Lemma 1. *With the notation above,*

$$\int \mathrm{e}^{-2\pi \mathrm{i}\theta x\cdot\xi} \mathrm{d}H(\theta) = c\, \widehat{\sigma}(|x|\xi),$$

where σ is the Lebesgue measure on the unit sphere.

The proof is by rotation invariance. Any two rotationally invariant measures on

$$S^2 = \mathrm{SO}(3)/\mathrm{SO}(2) \tag{9}$$

are constant multiples of each other.[1] It follows that for any $\omega \in S^2$,

$$\int_{\mathrm{SO}(3)} f(\theta\omega)\mathrm{d}H(\theta) = c\int_{S^2} f(e)d\sigma(e),$$

where σ, once again, is the Lebesgue measure on S^2 and c does not depend on f. By rotation invariance, we may assume that $x = |x|(0, 0, 1)$ and the lemma follows by choosing f to be the exponential function.

With Lemma 1 in tow, the problem reduces to showing that

$$\mu \times \mu\{(u, v) : |x| \le |u - v| \le |x|(1 + \varepsilon)\} \lesssim \varepsilon, \tag{10}$$

which is exactly the assumption (2), and this completes the proof of Theorem 1.

To prove Corollary 1, recall that we may assume that $|x| \gtrsim 1$. By the method of stationary phase (see e.g., [10]), we get

$$|\widehat{\sigma}(\xi)| \lesssim |\xi|^{-1},$$

and hence, the expression (7) is

$$\lesssim \int |\xi|^{-1}|\widehat{\mu}(\xi)|^2\,\mathrm{d}\xi = c\int\int |x - y|^{-2}\mathrm{d}\mu(x)\mathrm{d}\mu(y),$$

which certainly converges if μ is a Frostman measure on a set of Hausdorff dimension greater than two. This approach fails to work for two dimensional sets and this is where the geometric assumption (measures supported on graphs of Lipschitz functions) will play a key role. We now turn to the proof of Corollary 2.

4 Geometric Estimates: Proof of Corollary 2

In this section, we establish (10), or the assumption (1), for measures supported on graphs of Lipschitz functions. We may assume that μ is the Lebesgue measure on a graph of a Lipschitz function G. We may make an a priori assumption that $G \in C^1$ and that $|\nabla G(x_1, x_2)|$ is uniformly bounded from above, due to the Lipschitz hypothesis.

[1] Observe that (9) holds for any dimension $d \ge 2$, i.e., $S^{d-1} = \mathrm{SO}(d)/\mathrm{SO}(d-1)$, and hence, Lemma 1 can be generalized to $d \ge 2$. For the purpose of this note, this generalization is not necessary.

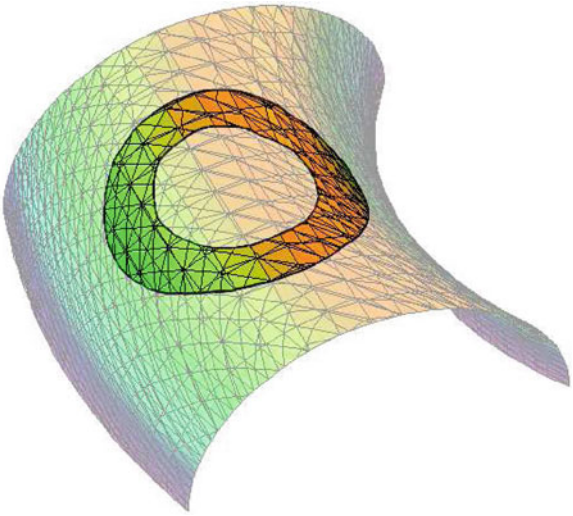


Figure 1 The set S_x in the equation (11)

Let S denote the Lipschitz surface under consideration and define (Fig. 1)

$$S_x = \{y \in S : \delta \leq |x - y| \leq \delta(1 + \varepsilon)\} \subset \mathbb{R}^3. \tag{11}$$

We shall carry out the remainder of the calculations with $\delta = 1$ for the sake of clarity.

We need to prove that

$$\int \int_{\{(x,y)\in S\times S:1\leq|x-y|\leq 1+\varepsilon\}} \mathrm{d}\mu(x)\mathrm{d}\mu(y) = \int_S \mu(S_x)\mathrm{d}\mu(x) \leq C\varepsilon,$$

which amounts to showing that

$$\int \int_{\mathscr{M}_\varepsilon} \sqrt{1 + |\nabla G(x_1, x_2)|^2}\sqrt{1 + |\nabla G(y_1, y_2)|^2}\mathrm{d}x'\mathrm{d}y' \leq C\varepsilon,$$

where B is the unit ball, $x' = (x_1, x_2)$, $y' = (y_1, y_2)$ and

$$\mathscr{M}_\varepsilon = \{(x, y) \in B\times B : 1 \leq |x_1-y_1|+|x_2-y_2|+|G(x_1, x_2)-G(y_1, y_2)| \leq 1+\varepsilon\}.$$

By the Lipschitz hypothesis, this amounts to showing that

$$|\{(x', y') \in B \times B : 1 \leq |x_1 - y_1| + |x_2 - y_2| + |G(x_1, x_2) - G(y_1, y_2)| \leq 1 + \varepsilon\}| \leq C\varepsilon, \tag{12}$$

where $|\cdot|$ denotes the two dimensional Lebesgue measure. Since G is Lipschitz,

$$\begin{aligned}&|x_1 - y_1| + |x_2 - y_2| + |G(x_1, x_2) - G(y_1, y_2)| \\ &\quad \le C(|x_1 - y_1| + |x_2 - y_2| + |x' - y'|) \\ &\quad \le C'(|x_1 - y_1| + |x_2 - y_2|).\end{aligned}$$

It follows that for each fixed $y \in B$, the set

$$\{x \in B : 1 \le |x_1 - y_1| + |x_2 - y_2| + |G(x_1, x_2) - G(y_1, y_2)| \le 1 + \varepsilon\} \quad (13)$$

can be covered by a finite number of finitely overlapping rectangles with dimensions C_1 by $C_2\varepsilon$, where C_1 and C_2 are uniform constants. More precisely, for a fixed $y \in B$, let $\phi(x_1, x_2) = |x_1-y_1|+|x_2-y_2|+|G(x_1, x_2)-G(y_1, y_2)|$, a bi-Lipschitz function on B. Note that $|x_1-y_1|+|x_2-y_2| \le \phi(x_1, x_2) \le C(|x_1-y_1|+|x_2-y_2|)$ for some $C > 1$. Without loss of generality, by choosing local coordinates we may assume that $\phi(0, 1) = 1$ and also that in some neighborhood U of $(0, 1)$, say $U = \{x \in B : |x-(0, 1)| < 1/2\}$, we have $\dfrac{\partial \phi}{\partial x_2} \ge c > 0$. Hence, estimating the measure of the set in (13) amounts to estimating

$$\int\int_{1 \le \phi(x_1,x_2) \le 1+\varepsilon} \mathrm{d}x_1 \, \mathrm{d}x_2. \quad (14)$$

Changing variables in (14) to $y_1 = \phi(x_1, x_2)$ and $y_2 = x_1/x_2$ (observe that for $x \in U$ the variable y_2 is well-defined), we obtain

$$\int_{|y_2-1| \le \frac{1}{2}} \int_1^{1+\varepsilon} \frac{x_2^2}{x \cdot \nabla \phi(x)} \, \mathrm{d}y_1 \, \mathrm{d}y_2, \quad (15)$$

where $x \in U$, and the integrand is the Jacobian of this substitution. Therefore, if we show that

$$|x \cdot \nabla \phi(x)| \ge c > 0 \quad \text{for} \quad x \in U, \quad (16)$$

then the expression in (15) will be bounded by $c\varepsilon$. To do that we parameterize level curves of ϕ (for given r, $\phi(x_1, x_2) = r$, $x \in B$) by $(t, \gamma_r(t))$. Observe that $\gamma_r(0) \sim 1$ and $\gamma_1(0) = 1$.

Differentiating $\phi(t, \gamma_r(t)) = r$, we get

$$\gamma_r'(t) = -\frac{\partial_{x_1} \phi}{\partial_{x_2} \phi}, \quad (17)$$

and thus, using the assumption on $\partial_{x_2} \phi$, we also have the bound

$$|\gamma_r'(t)| \le \left| \frac{\partial_{x_1} \phi}{\partial_{x_2} \phi} \right| \le c. \quad (18)$$

Now, for $x \in U$ using the parametrization, we obtain

$$\begin{aligned} x \cdot \nabla\phi(x) &= t\,\partial_{x_1}\phi(t,\gamma_r(t)) + \gamma_r(t)\partial_{x_2}\phi(t,\gamma_r(t)) \\ &= \partial_{x_2}\phi(t,\gamma_r(t))[-t\gamma_r'(t) + \gamma_r(t)] \geq c > 0, \end{aligned}$$

since by assumption $|\partial_{x_2}\phi(x)_{|U}| \geq c > 0$ and the expression in the square brackets for t close to 0 is bounded below by $(1 - ct) > 0$ by virtue of the bound (18) and $\gamma_r(0) \sim 1$, thus, finishing the argument.

Acknowledgements A. I. thanks Michael Loss of Georgia Institute of Technology for a helpful suggestion regarding the regularity assumptions in the main result. He also thanks Andreas Seeger for a very helpful conversation about Haar measures. A.I. was partially supported by the NSF grant DMS-0456306. S. R. was partially supported by the NSF grant DMS-0531337. Part of this work was done while both authors participated in the MTBI Program at the Arizona State University and are grateful for their support.

References

1. Brandolini, L., Iosevich, A., Travaglini, G.: Spherical means and the restriction phenomenon. J. Fourier Anal. Appl., **7**, no. 4, 359–372 (2001)
2. Carbery, A., Kenig, C., Ziesler, S.: Restriction for flat surfaces of revolution on $\mathbf{R}^3$. Proc. Am. Math. Soc., **135**, no. 6, 1905–1914 (2007)
3. Falconer, K. J.: The geometry of fractal sets. Cambridge Tracts in Mathematics, 85. Cambridge University Press, Cambridge (1986)
4. Falconer, K. J.: On the Hausdorff dimensions of distance sets. Mathematika, **32**, no.2, 206–212 (1985)
5. Iosevich, A.: Fourier transform, L^2 restriction theorem, and scaling. Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat., (8) **2**, no. 2, 383–387 (1999)
6. Iosevich, A., Lu, G.: Sharpness results and Knapp's homogeneity argument. Can. Math. Bull., **43**, no. 1, 63–68 (2000)
7. Oberlin, D.: Fourier restriction for affine arclength measures in the plane. Proc. Amer. Math. Soc., **129**, no. 11, 3303–3305 (2001)
8. Oberlin, D.: A uniform Fourier restriction theorem for surfaces in $\mathbb{R}^3$. Proc. Am. Math. Soc., **132**, no. 4, 1195–1199 (2004)
9. Podkorytov, A. N.: On the asymptotics of the Fourier transform on a convex curve. Vestnik Leningrad. Univ. Mat. Mekh. Astronom. 1991, vyp. 2, 50–57, 125 (in Russian); translation in Vestnik Leningrad Univ. Math. 24, no. 2, 57–65 (1991)
10. Stein, E. M.: Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals. Princeton Mathematical Series, 43. Monographs in Harmonic Analysis, III. Princeton University Press, Princeton, NJ (1993)

On the Exact Order of Asymptotic Bases and Bases for Finite Cyclic Groups

Xingde Jia

Dedicated to Professor Melvyn B. Nathanson

Summary Let h be a positive integer, and A a set of nonnegative integers. A is called an *exact asymptotic basis* of order h if every sufficiently large positive integer can be written as a sum of h not necessarily distinct elements from A. The smallest such h is called the *exact order* of A, denoted by $g(A)$. A subset $A - F$ of an asymptotic basis of order h may not be an asymptotic basis of any order. When $A-F$ is again an asymptotic basis, the exact order $g(A-F)$ may increase. Nathanson [48] studied how much larger the exact order $g(A - F)$ when finitely many elements are removed from an asymptotic basis of order h. Nathanson defines, for any given positive integers h and k,

$$G_k(h) = \max_{\substack{A \\ g(A)\le h}} \max_{F\in I_k(A)} g(A - F),$$

where $I_k(A) = \{|F| \, |F| = k \text{ and } g(A-F) < \infty\}$. Many results have been proved since Nathanson's question was first asked in 1984. This function $G_k(h)$ is also closely related to interconnection network designs in network theory. This paper is a brief survey on this and few other related problems. G. Grekos [11] has a recent survey on a related problem.

Keywords Additive bases · Asymptotic bases · Exact asymptotic bases · Extremal bases · Finite cyclic groups · Postage stamp problem

Mathematics Subject Classifications (2010). 05D99, 11B13, 11P05, 11P70.

X. Jia
Texas State University, San Marcos, Texas 78666, USA
e-mail: jia@txstate.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_13,

1 Exact Asymptotic Bases

Let $\mathbb{N}$ be the set of all nonnegative integers. Let h be a positive integer, and $A \subseteq \mathbb{N}$ a set of nonnegative integers. Let hA denote the set of all sums of h not necessarily distinct elements from A. For notations and concepts without definitions here, the reader is referred to the books by [50–52]. Another wonderful reference is the book sequences by Halberstam and Roth [14]. G. Grekos [10] has a recent survey on a related problem.

Definition 1. A set $A \subseteq \mathbb{N}$ of nonnegative integers is called an *exact asymptotic basis* of order h if $\mathbb{N} - hA$ is a finite set. In order words, A is called an exact asymptotic basis of order h if every sufficiently large positive integer can be written as a sum of h not necessarily distinct elements from A. The smallest such h is called the *exact order* of A and is denoted by $g(A)$.

Let h be any positive integer. Then $A = \{1\} \cup \{nh \mid n \in \mathbb{N}\}$ is an exact asymptotic basis of order h. Let B be the set of all odd positive integers. Then B is not an exact asymptotic basis of any order h because the sum of h odd integers has the same parity as h. However, B with any additional even integer becomes an exact asymptotic basis of order 2.

As an example, we show that $A = \bigcup_{k=0}^{\infty} \left(2^{2k},\ 2^{2k+1}\right]$ is an exact asymptotic basis of order 3, where we use $(a, b]$ to denote the set of integers x with $a < x \leq b$. First, for a large k, $n = 2^{2k} + 1$ cannot be written as a sum of two elements from A. Otherwise, say $n = a + b$, then $a, b \leq 2^{2(k-1)+1}$. Hence $a + b \leq 2^{2k-1} + 2^{2k-1} \leq 2^{2k} < n$, a contradiction. We then show that every large integer n not in A can be written as a sum of two elements from A. First assume $n = 2^{2k+1} + 1$. If k is even, $2^k + 1 \in A$ and $2^{k+1} \in A$. Hence $n = (2^k + 1) + 2^{k+1} \in 2A$. If k is odd, then $n = 2^k + (2^{k+1} + 1) \in 2A$. Now assume $2^{2k+1} + 2 \leq n \leq 2^{2k+2}$ for some k. Then, we can rewrite $n = (2^{2k} + s) + (2^{2k} + t)$ with $1 \leq s < 2^{2k}$ and $1 \leq t \leq 2^{2k}$. Hence, $2^{2k} + s, 2^{2k} + t \in (2^{2k}, 2^{2k+1}] \subseteq A$, which implies that $n \in 2A$. Now pick a large positive integer n and write $n = m + 5$. If $m \notin A$, then, as shown earlier, m is a sum of two elements in A. Hence, n is a sum of three elements in A. If $m = 2^{2k} + 1$, then $n = 6 + 2^{2k-1} + 2^{2k-1} \in 3A$. If $m \in [2^{2k} + 2, 2^{2k+1}] \subseteq A$, then $n = 5 + (2^{2k-1} + s) + (2^{2k-1} + t) \in 3A$. Therefore, we proved that A is an exact asymptotic basis of order 3.

2 Subsets of Exact Asymptotic Bases

A subset of an exact asymptotic basis of order h may not be again an exact asymptotic basis of any order. For instance, $A = \{0\} \cup \{1, 3, 5, \ldots\}$ is an exact asymptotic basis of order 2. However, $A - \{0\}$, the set of all odd positive integers is not an exact asymptotic basis of any order. When a subset of an exact asymptotic basis of order h

is an exact asymptotic basis, its exact order can be larger. Nathanson defined the following function to study how large the exact order of subsets of an exact asymptotic basis of order h when finite.

Let A be an asymptotic basis of order h. Given a positive integer k, let

$$I_k(A) = \{F \subseteq A \mid |F| = k, \text{ and } A - F \text{ is an exact asymptotic basis}\}.$$

Define

$$G_k(h) = \max_{\substack{A \\ g(A) \le h}} \max_{F \in I_k(A)} g(A - F),$$

where the first maximum is taken over all exact asymptotic bases of order at most h. Nathanson [48] proved in 1984 the following theorem.

Theorem 1 (Nathanson [48], 1984). *For $h > k$,*

$$G_k(h) \ge \left(\left\lfloor \frac{h}{k+1} \right\rfloor + 1\right)^{k+1} - 1 \approx \left(\frac{h}{k+1}\right)^{k+1}.$$

In 1988, Jia [21] improved Nathanson's lower bound to:

$$G_k(h) \ge \frac{4}{3}\left(\frac{h}{k+1}\right)^{k+1} + O(h^k) \qquad \text{as} \quad h \to \infty.$$

Later this has been further improved [24] to

$$G_k(h) \ge (k+1)\left(\frac{k+1}{k+2}\right)^{k}\left(\frac{h}{k+1}\right)^{k+1} + O(h^k) \qquad \text{as} \quad h \to \infty.$$

Let A be a set of nonnegative integers. The *lower density* of A is defined by

$$d(A) = \liminf_{m \to \infty} \frac{A(m)}{m},$$

where $A(m) = |\{a \in A \mid 0 < a \le m\}|$. The following version of Kneser's Theorem [37] is useful in establishing upper bounds for $G_k(h)$.

Theorem 2 (Kneser [37], 1953). *Let $C = A_1 + \cdots + A_n$. Then either*

$$d(C) \ge d(A_1) + \cdots + d(A_n)$$

or C is equal to, with at most finitely many exceptions, a residue class modulo g for some positive integer g.

By using Kneser's Theorem for the upper bounds, Nash [45] proved in 1985 that

$$G_k(2) = 2k + 2 \qquad \text{for all} \quad k \ge 1.$$

In the general case, as $k \to \infty$ for any given $h \geq 1$, we proved [24] that

$$G_k(h) + 1 \geq 2\left(\frac{k}{h-1}\right)^{h-1} + (4h-5)\left(\frac{k}{h-1}\right)^{h-2} + O(k^{h-3}),$$

$$G_k(h) + 1 \leq \frac{2}{(h-1)!}k^{h-1} + \frac{h-1}{(h-2)!}k^{h-2} + O(k^{h-3}).$$

When $h = 2$, this is Nash's formula for $G_k(2)$. Farhi [7] studied the exact order $g(A-F)$ in terms of some parameters of the set F other than simply the cardinality of F. Define

$$d = \frac{\operatorname{diam}(F)}{\gcd\{x - y \mid x, y \in A\}}.$$

Farhi [7] proved in 2008 that

$$g(A-F) \leq \frac{h(h+3)}{2} + \frac{dh(h-1)(h+4)}{6}$$

for any exact asymptotic basis A of order h. In particular, if F is an arithmetic progression, then

$$g(A-F) \leq \frac{h(h+3)}{2} + \frac{(|F|-1)h(h-1)(h+4)}{6}.$$

These lower bounds are better in many cases.

For more references on the Postage stamp problem, see also [53, 54].

3 Exact Order of Asymptotic Bases

A set A of nonnegative integers is called an *asymptotic basis* of order h if every sufficiently large integer can be written as a sum of at most h not necessarily distinct elements from A. If we use notation

$$h^0 A = \bigcup_{s=1}^{h} sA,$$

then A is an asymptotic basis of order h if and only if $h^0 A$ contains all sufficient large integers.

The set of all positive odd integers is an asymptotic basis of order 2, while it is not an exact asymptotic basis of any order. The following theorem of Erdős and Graham provides a necessary and sufficient condition for an asymptotic basis to be an exact asymptotic basis.

Theorem 3 (Erdős and Graham [5], 1980). *Assume that A is an asymptotic basis. Then A is an exact asymptotic basis if and only if*

$$\gcd\{a - a' : a, a' \in A\} = 1.$$

A more general theorem is Nash and Nathanson [47] in 1985.

Theorem 4 (Nash and Nathanson [47], 1985). *If A is an asymptotic basis containing at most a finite number of negative terms such that*

$$\gcd\{a - a' : a, a' \in A\} = d,$$

then there exists an positive integer q such that every term of an infinite arithmetic progression with difference d can be written as a sum of exactly q elements in A.

Let $A = \bigcup_{k=0}^{\infty} \left(2^{2k},\ 2^{2k+1}\right]$. Since every large integer n not in A can be written as a sum of two elements from A, A is an asymptotic basis of order two. Note that A is an exact asymptotic basis of order 3. A natural question is how much bigger the exact order is when an asymptotic basis of order h is also an exact asymptotic basis.

Theorem 5 (Nathanson [48], 1984). *Let $h \geq 2$ be an integer. Then*

$$G_1(h) = \max_A\{g(A) \mid A \text{ is an asymptotic basis of order } h\} \tag{1}$$

Proof. Denote the right hand side of (1) by $t(h)$. We only need to prove $G_1(h) = t(h)$.

Assume that A is an exact asymptotic basis with $g(A) = h$ and $g(A - \{x\}) = G_1(h) = g$. Define $A_1 = \{a - x \mid a \in A \text{ and } a \neq x\}$. Let n be any large positive integer. Then

$$n + hx = a_1 + a_2 + \cdots + a_h, \qquad \text{where} \quad a_i \in A.$$

Then

$$n = (a_1 - x) + (a_2 - x) + \cdots + (a_h - x).$$

After deleting the "0" terms in the above summation, we see that n is a sum of at most h elements from A_1, i.e., A_1 is an asymptotic basis of order h. Since $g(A - \{x\}) = g$, we see that

$$n + gx = \sum_{i=1}^{g} a_i, \qquad \text{where} \quad a_i \in A - \{x\}.$$

Hence

$$n = \sum_{i=1}^{g}(a_i - x), \qquad \text{where} \quad a_i - x \in A_1.$$

Thus, A_1 is an exact asymptotic basis with $g(A_1) \leq g$. In fact, easy to see that $g(A_1) = g$. Therefore, $t(h) \geq g = G_1(h)$.

On the other hand, assume that A is an asymptotic basis of order h so that $g(A) = t(h)$. Since $t(h) > h$, $0 \notin A$. Define $A' = A \cup \{0\}$. Then A' is an exact asymptotic basis of order h, i.e., $g(A') = h$. Since $g(A' - \{0\}) = g(A) = t(h)$, we have that $g = G_1(h) \geq t(h)$. Therefore, $G_1(h) = t(h)$. □

Noting that $A = \bigcup_{k=0}^{\infty} \left(2^{2k},\ 2^{2k+1}\right]$ is an asymptotic basis of order 2 and its exact order is 3 as an exact asymptotic basis, we see that $G_1(2) \geq 3$. The following are the known exact values for the function:

$G_1(2) = 3$	(Erdős and Graham [5], 1980),
$G_1(3) = 7$	(Nash [45], 1985),
$G_1(4) = 10$	(Li [40], 1989, uncomfirmed),
$G_1(5) = 15$	(Li [40], 1989, uncomfirmed),
$G_2(3) = 13$	(Nash [45], 1985),
$G_k(2) = 2k + 2$	(Nash [45], 1985),

The following is a list of known estimates for $G_1(h)$ by various authors:

$$\frac{1}{4}h^2 + o(h^2) \leq G_1(h) \leq \frac{5}{4}h^2 + o(h^2) \qquad \text{(Erdős and Graham [5], 1980),}$$

$$\frac{1}{3}h^2 + O(h) \leq G_1(h) \leq h^2 + h \qquad \text{(Grekos [9], 1982),}$$

$$G_1(h) \leq \frac{1}{2}h^2 + h \qquad \text{(Nash [46], 1993).}$$

4 Postage Stamp Problem

Support an envelope has space for only up to h stamps, and $A = \{a_1 = 1, a_2, \ldots, a_k\}$ is the set of stamp face values. The postage stamp problem consists of computing the smallest postage value $n(h, A) + 1$ that cannot be stamped by using the given stamps. In other words, $n(h, A)$ is the largest integer such that every positive integer $\leq n(h, A)$ can be represented as a linear combination

$$\sum_{i=1}^{k} x_i a_i$$

with $x_i \geq 0$ and

$$\sum_{i=1}^{k} x_i \leq h.$$

Define, for any given positive integers h and k,

$$n(h,k) = \max_{|A|=k} n(h,A).$$

For convenience, a set $A \subseteq \{1,2,\ldots,n\}$ is called an *h-basis* for n if $\{1,2,\ldots,n\} \subseteq h^0 A$. $n(h,A)$ is called the *h-range* of A, and $n(h,k)$ is called the (h,k)-*range*. One central problem is to calculate the (h,k)-range $n(h,k)$.

The postage stamp problem has been around for a long time. However, it seems that the problem appeared only as recreational and entertaining mathematics (e.g., Sprague [62, Problem 18] and Legard [39]) until 1937 when Rohrbach [58] first formalized and analyzed the problem mathematically. Since then, there have been extensive research on the problem (see Guy [13], Hofmeister [16], Hofmeister et al. [17], Klotz [35], etc.). A similar and related problem is the *Frobenius Coin Problem* which asks the largest amount of postage that is impossible to pay by using a given set of stamps (sufficient supply). It turns out that this is an incredibly difficult problem to solve. We even do not know the answer with only three kinds of stamps! Computation of Frobenius problem is NP-complete (see an interesting article of Cipra [3] in the *Science* magazine). See a survey up to 1980 by Alter and Barnett [1], and [31] for recent developments on the postage stamp problem. See Selmer [60,61] for a comprehensive introduction to the problem [32,33,43,56,57].

Stohr [63,64] proved in 1955 that

$$n(h,2) = \left\lfloor \frac{h^2+6h+1}{4} \right\rfloor, \qquad h \geq 2. \tag{2}$$

Hofmeister [15] proved in 1968 that, for all integers $h \geq 23$,

$$n(h,3) = \frac{4}{81}h^3 + \frac{2}{3}h^2 + \alpha h + \beta, \tag{3}$$

where α and β are determined constants depending only on $h \pmod 9$.

The following theorem (see [23]) provides lower bounds for $G_k(h)$ by using the lower bounds for $n(h,k)$.

Theorem 6. *Let $h \geq 3$ and $k \geq 1$ be integers. Then*

(i) $G_k(h) \geq n(h-1,k+1)$;
(ii) $G_k(h) \geq 2n(h-1,k) + h$.

For convenience, let us define, for any given positive integer h,

$$\sigma_h = \limsup_{k\to\infty} \frac{n(h,k)}{k^h}.$$

Mrose [44] proved in 1979 that $\sigma_2 \geq \frac{2}{7} \approx 0.2857$. As for the upper bound, it is easy to see that $\sigma \leq 1/2$. This trivial bound has been improved several times:

$\sigma_2 \leq 0.4992$	(1937, Rohrbach [58])
$\sigma_2 \leq 0.4903$	(1960, Moser [41])
$\sigma_2 \leq 0.4867$	(1960, Riddell [55])
$\sigma_2 \leq 0.4847$	(1969, Moser, Pounder and Riddell [42])
$\sigma_2 \leq 0.4802$	(1969, Klotz [35, 36])
$\sigma_2 \leq 0.4789$	(2006, Güntürk and Nathanson [12])
$\sigma_2 \leq 0.4778$	(2007, Horváth [18])
$\sigma_2 \leq 0.4697$	(2009, Yu [67])

For $h = 3$, Mrose [44] proved in 1979 that

$$n(h, 3) \geq \frac{32}{27}\left(\frac{k}{3}\right)^3 + O(k^2).$$

This was improved by Windecker (unpublished) to

$$n(3, k) \geq \frac{4}{3}\left(\frac{k}{3}\right)^3 + O(k^2) \qquad \text{as} \quad k \to \infty. \tag{4}$$

Therefore, it follows from the lower bounds and (*ii*) in Theorem 6 that

$$G_k(3) \geq \frac{4}{7}k^2 + O(k) \qquad \text{as} \quad k \to \infty,$$

and

$$G_k(4) \geq \frac{8}{81}k^3 + O(k^2) \qquad \text{as} \quad k \to \infty.$$

Using the following recursive inequality for $n(h, k)$

$$n(h_1 + h_2, k_1 + k_2) \geq n(h_1, k_1)n(h_1, k_2), \tag{5}$$

One can prove that, for any fixed $h \geq 3$, as $k \to \infty$,

$$n(h, k) \geq c_h \left(\frac{4}{3}\right)^{\lfloor k/3 \rfloor} \left(\frac{k}{h}\right)^h,$$

where c_h are absolute constants depending only on h mod 3. It then follows from Theorem 6 (*ii*) that

$$G_k(h) \geq 2c_{h-1}\left(\frac{k}{h-1}\right)^{h-1}.$$

5 Extremal Bases for Finite Cyclic Groups

Let A be a set of k distinct integers. A is called an *h-basis* for $\mathbb{Z}_m$ if every element in $\mathbb{Z}_m$ can be written as a sum of at most h not necessarily distinct elements of A. In other words, A is an h-basis for $\mathbb{Z}_m$ if and only if $h^0 A = \mathbb{Z}_m$. Let $m(h, A)$ denote the largest positive integer m so that A is an h-basis for $\mathbb{Z}_m$. Given positive integers h and k, we define

$$m(h,k) = \max_{\substack{A \\ |A|=k}} m(h,A)$$

It is easy to see that $nm(h,k) \geq (h,k) + 1$ for all $h \geq 1$ and $h \geq 1$. Similar to the Theorem of Erdős and Graham, one can prove that a set $A = \{a_1, a_2, \ldots, a_k\}$ is an h-basis for $\mathbb{Z}_m$ for some integer h if and only if

$$\gcd\{m, a_1, a_2, \ldots, a_k\} = 1.$$

It is clear that an h-basis for n is always an h-basis for $\mathbb{Z}_{n+1}$. However, an h-basis for $\mathbb{Z}_m$ may not be an h-basis for $m-1$.

Extremal bases for finite cyclic groups are closely related to interconnection network designs. Extremal bases and related problems have been one of central focuses in the study of combinatorial networks, which emerges as a broad area of research. For more information, see, for instance, Graham and Sloane [8], Erdős and Hsu [6], Du and Hsu [4], Hsu and Jia [19, 20], and Jia [28, 29], etc.

Hsu and Jia [19] proved in 1994 that

$$m(h,2) = \left\lfloor \frac{h(h+4)}{3} \right\rfloor + 1 \qquad \text{for all} \quad h \geq 2. \tag{6}$$

For $k = 3$, it seems harder to handle bases for $\mathbb{Z}_n$ when compared with bases for $[1, n]$. The analog of Hofmeister's formula (3) for $n(h, 3)$ in the postage stamp problem, an exact formula for $m(h, 3)$, is yet to be found. Hsu and Jia [19] showed in 1994 that

$$m(h,3) \geq \frac{1}{16}h^3 + O(h^2) \approx 0.0625h^3 + O(h^2) \qquad \text{as} \quad h \to \infty.$$

It is easy to verify the following recursive addition inequality similar to (5)

$$m(h_1 + h_2, k_1 + k_2) \geq m(h_1, k_1)m(h_1, k_2). \tag{7}$$

Using this addition inequality, we can provide lower bounds for $m(h, k)$ by improving lower bounds for $m(h, k)$ with small ks.

Jia [21, 24, 25] showed in 1990 that, for fixed $k \geq 4$ as $h \to \infty$,

$$m(h,k) \geq \alpha_k \left(\frac{256}{125}\right)^{\lfloor k/4 \rfloor} \left(\frac{h}{k}\right)^k + O(h^{k-1}),$$

where $\alpha_k = 1, 1, 4/3$ and $27/16$ according as $k \equiv 0, 1, 2$ or $3 \pmod 4$. Chen and Gu [2] proved in 1992 that, for fixed k and $h \to \infty$,

$$m(h,k) \geq \beta_k \left(\frac{2{,}048}{625}\right)^{\lfloor k/4 \rfloor} \left(\frac{h}{k}\right)^k + O\left(h^{k-1}\right),$$

where $\beta_k = 1, 1, 4/3$, or $135/64$, according as $k = 0, 1, 2$, or 3 mod 4. In 1993, Su [65] constructed a new five-element h-basis which provides a new lower bound for $m(h,5)$, and hence a better lower bound in the general case:

$$\begin{aligned} m(h,k) &\geq \gamma_k \left(\frac{5^5 \cdot 7^4}{17^5}\right)^{\lfloor k/5 \rfloor} \left(\frac{h}{k}\right)^k + O(h^{k-1}) \\ &\approx \gamma_k (5.2844)^{\lfloor k/5 \rfloor} \left(\frac{h}{k}\right)^k + O(h^{k-1}), \end{aligned} \tag{8}$$

where

$$\gamma_k = \begin{cases} 1 & \text{if } k \equiv 0, 1 \pmod 5 \\ 4/3 & \text{if } k \equiv 2 \pmod 5 \\ \dfrac{4{,}752}{2{,}197} \approx 2.163 & \text{if } k \equiv 3 \pmod 5 \\ \dfrac{165{,}888}{50{,}625} = 3.2768 & \text{if } k \equiv 4 \pmod 5 \end{cases}$$

This implies the following lower bound for $G_k(h)$, which is best known at the time this article was written:

$$G_k(h) \geq \gamma_{k+1} \left(\frac{5^5 \cdot 7^4}{17^5}\right)^{\lfloor (k+1)/5 \rfloor} \left(\frac{h}{k+1}\right)^{k+1} + O(h^k) \qquad \text{as} \quad h \to \infty.$$

6 Remarks and Open Problems

1. Kirfel [34] proved in 1990 that the following limit

$$\tau_k = \lim_{h \to \infty} \frac{n(h,k)}{h^k}$$

exits for every $k \geq 1$. It is known that

$$\tau_1 = 1, \qquad \tau_2 = \frac{1}{4}, \qquad \text{and} \qquad \tau_3 = \frac{4}{81}.$$

It seems much harder to deal with $n(h,k)$ with given $h \geq 1$ as k approaches infinity. Current best known bounds for $n(2,k)$ are proved by Mrose [44] and Yu [67]:

$$0.2857 < \frac{n(2,k)}{k^2} < 0.4697.$$

2. It is natural to ask if any of the following limits exists:

$$\lim_{h\to\infty} \frac{G_k(h)}{h^{k+1}} \qquad \text{and} \qquad \lim_{k\to\infty} \frac{G_k(h)}{k^{h-1}}.$$

The only known nontrivial case is Nash's formula for $G_k(2) = 2k + 2$. We even do not know the answer when $k = 1$.

3. Similarly we may ask if the following limits exist:

$$\eta_k = \lim_{h\to\infty} \frac{m(h,k)}{h^k}.$$

The only known exact values are $\eta_1 = 1$ and $\eta_2 = 1/3$. One annoying fact is that we do not even have any nontrivial lower bound[1] or upper bound for $m(2,k)$.

4. A set A of nonnegative integers is called an *restricted exact asymptotic basis* of order h if every large positive integer can be written as a sum of h distinct elements from A. Similar questions can be asked for restricted exact asymptotic bases. We known little about restricted exact asymptotic bases. Not much is known for the restricted version for both $n(h,k)$ and $m(h,k)$, especially $m(h,k)$.

5. Let A be a finite set of integers. If A is a basis for $\mathbb{Z}_m$, then the average order of A for $\mathbb{Z}_m$ is defined by

$$\lambda(A, \mathbb{Z}_m) = \frac{1}{m} \sum_{t=0}^{m-1} h(t),$$

where $h(t)$ is the *length* of t by A, which is defined as the minimum number of elements (another problem if distinct elements required) of A with sum t. Similar functions can be defined:

$$m(\lambda, k) = \max\{m \mid \exists A \text{ with } |A| = k \text{ and } \lambda(A, \mathbb{Z}_m) \leq \lambda\},$$
$$n(\lambda, k) = \max\{n \mid \exists A \text{ with } |A| = k \text{ and } \lambda(A, [1,n]) \leq \lambda\}.$$

See [29–31] for some preliminary results in these cases.

6. Graham and Sloane [8] studied a set of four extremal functions related to additive bases. Those extremal functions can be generalized [20]. These functions are related to $G_k(h)$. There are still many open problems related these functions.

[1] The author does not know any lower bound for $m(2,k)$ other than $m(2,k) \geq \frac{2}{7}k^2 + O(k)$, which is obtained by $n(2,k) + 1 \leq m(2,k)$.

7. Extremal bases for finite cyclic groups are considered as good underlying topology for interconnection networks (see [6, 20]). This area has been studied extensively in recent years, see [31] for a more complete survey.
8. Wong and Coppersmith [66] discovered a geometric representation of bases for finite cyclic groups, which helps establish upper bounds for $m(k, h)$. With the help of Kneser's theorem, one might be able to utilize Wong-Coppersmith's representation to obtain upper bounds for $G_k(h)$. For more information on the geometric representation of bases, see Jia and Hsu [31].
9. Bases for finite groups have also been studied extensively in the past. Among many interesting problems in this area, Rohrbach's problem on bases for finite groups attracts a lot of attention. If A is a basis of order h for a finite group Γ with $|\Gamma| = m$, then
$$|A| \geq m^{1/h} - 1.$$
Rohrbach [58, 59] asked in 1937 the following question: Is it true that, for every positive integer h, there exists a constant $c = c(h) > 0$ such that every finite group Γ with $|\Gamma| = m$ contains a basis A of order h for Γ such that
$$|A| \leq cm^{1/h}?$$
This problem is related to short products of elements from a finite group [27]. The question is still largely open. See [22, 26, 31, 38] for recent developments.

Acknowledgements I like to thank Professor Mel Nathanson from whom I leaned combinatorial additive number theory, a wonderful and entertaining field of mathematics.

References

[1] R. Alter and J. A. Barnett. Research Problems: A Postage Stamp Problem. *Am. Math. Monthly*, 87(3):206–210, 1980. ISSN 0002-9890.
[2] S. Chen and W. Gu. Exact order of subsets of asymptotic bases. *J. Number Theory*, 41(1): 15–21, 1992. ISSN 0022-314X.
[3] B. Cipra. Exact-postage poser still not licked. *Science*, 319:898–899, 15 Febrary 2008.
[4] D.-Z. Du and D. F. Hsu, editors. *Combinatorial network theory*, volume 1 of *Applied Optimization*. Kluwer Academic Publishers, Dordrecht, 1996. ISBN 0-7923-3777-8.
[5] P. Erdős and R. L. Graham. On bases with an exact order. *Acta Arith.*, 37:201–207, 1980. ISSN 0065-1036.
[6] P. Erdős and D. F. Hsu. Distributed loop network with minimum transmission delay. *Theoret. Comput. Sci.*, 100(1):223–241, 1992. ISSN 0304-3975.
[7] B. Farhi. Upper bounds for the order of an additive basis obtained by removing a finite subset of a given basis. *J. Number Theory*, 128(8):2214–2230, 2008. ISSN 0022-314X.
[8] R. L. Graham and N. J. A. Sloane. On additive bases and harmonious graphs. *SIAM J. Algebraic Discrete Methods*, 1(4):382–404, 1980. ISSN 0196-5212.
[9] G. Grekos. *Quelques aspects de la théorie additive des nombres*. PhD thesis, Université de Bordeaux I, France, 1982.
[10] G. Grekos. On the order of a minimal additive basis. *J. Number Theory*, 71(2):307–311, 1998. ISSN 0022-314X.

[11] G. Grekos. Extremal problems about asymptotic bases: a survey. In *Combinatorial number theory*, pages 237–242. de Gruyter, Berlin, 2007.
[12] C. S. Güntürk and M. B. Nathanson. A new upper bound for finite additive bases. *Acta Arith.*, 124(3):235–255, 2006. ISSN 0065-1036.
[13] R. K. Guy. *Unsolved problems in number theory*. Problem Books in Mathematics. Springer-Verlag, New York, third edition, 2004. ISBN 0-387-20860-7.
[14] H. Halberstam and K. F. Roth. *Sequences*. Springer-Verlag, New York, second edition, 1983. ISBN 0-387-90801-3.
[15] G. Hofmeister. Asymptotische Abschätzungen für dreielementige Extremalbasen in natürlichen Zahlen. *J. Reine Angew. Math.*, 232:77–101, 1968. ISSN 0075-4102.
[16] G. Hofmeister. Thin bases of order two. *J. Number Theory*, 86(1):118–132, 2001. ISSN 0022-314X.
[17] G. Hofmeister, C. Kirfel, and H. Kolsdorf. *Extremale Reichweiten*, volume 60 of *Mathematics Department Research Reports*. University of Bergen, Bergen, Norway, 1991.
[18] G. Horváth. An improvement of an estimate for finite additive bases. *Acta Arith.*, 130(4): 369–380, 2007. ISSN 0065-1036.
[19] D. F. Hsu and X. D. Jia. Extremal problems in the construction of distributed loop networks. *SIAM J. Discrete Math.*, 7(1):57–71, 1994. ISSN 0895-4801.
[20] D. F. Hsu and X. Jia. Additive bases and extremal problems in groups, graphs and networks. *Util. Math.*, 66:61–91, 2004. ISSN 0315-3681.
[21] X. D. Jia. Exact order of subsets of asymptotic bases in additive number theory. *J. Number Theory*, 28(2):205–218, 1988. ISSN 0022-314X.
[22] X. D. Jia. Thin bases for finite abelian groups. *J. Number Theory*, 36(2):254–256, 1990a. ISSN 0022-314X.
[23] X. D. Jia. *Some Results in Additive Number Theory*. PhD thesis, The Graduate Center of the City University of New York, New York, New York, USA, 1990b.
[24] X. D. Jia. On the order of subsets of asymptotic bases. *J. Number Theory*, 37(1):37–46, 1991. ISSN 0022-314X.
[25] X. D. Jia. Extremal bases for finite cyclic groups. *J. Number Theory*, 41(1):116–127, 1992a. ISSN 0022-314X.
[26] X. D. Jia. Thin bases for finite nilpotent groups. *J. Number Theory*, 41(3):303–313, 1992b. ISSN 0022-314X.
[27] X. D. Jia. Representation of finite groups as short products of subsets. *Bull. Austral. Math. Soc.*, 49(3):463–467, 1994. ISSN 0004-9727.
[28] X. D. Jia. Extremal Cayley digraphs of finite cyclic groups. *SIAM J. Discrete Math.*, 8(1): 62–75, 1995a. ISSN 0895-4801.
[29] X. D. Jia. Cayley digraphs of finite cyclic groups with minimal average distance. In *Interconnection networks and mapping and scheduling parallel computations (New Brunswick, NJ, 1994)*, volume 21 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 229–250. Am. Math. Soc., Providence, RI, 1995b.
[30] X. Jia and W. Su. Triple loop networks with minimal transmission delay. *Int. J. Found. Comput. Sci.*, 8:305–328, 1997.
[31] X. D. Jia and D. F. Hsu. *Combinatorial Networks*. Under Preparation, 2010.
[32] C. Kirfel. *On Extremal Bases for the h-range Problem, I*, volume 53 of *Mathematics Department Research Reports*. University of Bergen, Bergen, Norway, 1989.
[33] C. Kirfel. *On Extremal Bases for the h-range Problem, II*, volume 53 of *Mathematics Department Research Reports*. University of Bergen, Bergen, Norway, 1990a.
[34] C. Kirfel. Golay-koden, kulepakkinger og nye simple grupper. *Normat*, 38(4):160–178, 192, 1990b. ISSN 0801-3500.
[35] W. Klotz. Eine obere Schranke für die Reichweite einer Extremalbasis zweiter Ordnung. *J. Reine Angew. Math.*, 238:161–168, 1969a. ISSN 0075-4102.
[36] W. Klotz. Extremalbasen mit fester Elementeanzahl. *J. Reine Angew. Math.*, 237:194–220, 1969b. ISSN 0075-4102.
[37] M. Kneser. Abschätzung der asymptotischen Dichte von Summenmengen. *Math. Z.*, 58: 459–484, 1953. ISSN 0025-5874.

[38] G. Kozma and A. Lev. On h-bases and h-decompositions of the finite solvable and alternating groups. *J. Number Theory*, 49(3):385–391, 1994. ISSN 0022-314X.
[39] A. Legard. Brain-teaser. *Sunday Times*, Dec. 23, 1962 and Jan. 20, 1963.
[40] Y.-F. Li. personal communication, 1989.
[41] L. Moser. On the representation of $1, 2, \cdots, n$ by sums. *Acta Arith.*, 6:11–13, 1960. ISSN 0065-1036.
[42] L. Moser, J. R. Pounder, and J. Riddell. On the cardinality of h-bases for n. *J. Lond. Math. Soc.*, 44:397–407, 1969. ISSN 0024-6107.
[43] A. Mrose. *Die Bestimmung der extremalen regulären Abschnittasen mit Hilfe einer Klasse von Kettenbruchdeterminnanten*. PhD thesis, Freie Universität Berlin, Berlin, 1969.
[44] A. Mrose. Untere Schranken für die Reichweiten von Extremalbasen fester Ordnung. *Abh. Math. Sem. Univ. Hamburg*, 48:118–124, 1979. ISSN 0025-5858.
[45] J. C. M. Nash. *Results on Bases in Additive Number Theorey*. PhD thesis, Rutgers University, New Jersey, USA, 1985.
[46] J. C. M. Nash. Some applications of a theorem of M. Kneser. *J. Number Theory*, 44(1):1–8, 1993. ISSN 0022-314X.
[47] J. C. M. Nash and M. B. Nathanson. Cofinite subsets of asymptotic bases for the positive integers. *J. Number Theory*, 20(3):363–372, 1985. ISSN 0022-314X.
[48] M. B. Nathanson. The exact order of subsets of additive bases. In *Number theory (New York, 1982)*, volume 1052 of *Lecture Notes in Math.*, pages 273–277. Springer, Berlin, 1984.
[49] M. B. Nathanson. On a problem of Rohrbach for finite groups. *J. Number Theory*, 41(1): 69–76, 1992. ISSN 0022-314X.
[50] M. B. Nathanson. *Additive number theory*, volume 164 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1996a. ISBN 0-387-94656-X. The classical bases.
[51] M. B. Nathanson. *Additive number theory*, volume 165 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1996b. ISBN 0-387-94655-1. Inverse problems and the geometry of sumsets.
[52] M. B. Nathanson. *Elementary methods in number theory*, volume 195 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. ISBN 0-387-98912-9.
[53] A. Plagne. Removing one element from an exact additive basis. *J. Number Theory*, 87(2): 306–314, 2001. ISSN 0022-314X.
[54] A. Plagne. À propos de la fonction X d'Erdős et Graham. *Ann. Inst. Fourier (Grenoble)*, 54 (6):1717–1767 (2005), 2004. ISSN 0373-0956.
[55] J. Riddell. *On bases for sets of integers.* Master's thesis, University of Alberta, Alberta, Canada, T6G 2R3, 1960.
[56] Ø. J. Rødseth. On h-bases for n. *Math. Scand.*, 48(2):165–183, 1981. ISSN 0025-5521.
[57] Ø. J. Rødseth. An upper bound for the h-range of the postage stamp problem. *Acta Arith.*, 54(4):301–306, 1990. ISSN 0065-1036.
[58] H. Rohrbach. Ein Beitrag zur additiven Zahlentheorie. *Math. Z.*, 42(1):1–30, 1937a. ISSN 0025-5874.
[59] H. Rohrbach. Anwendung eines Satzes der additiven Zahlentheorie auf eine gruppentheoretische Frage. *Math. Z.*, 42(1):538–542, 1937b. ISSN 0025-5874.
[60] E. S. Selmer. *The Local Postage Stamp Problem, Part I: General Theory*, volume 42 of *Mathematics Department Research Reports*. University of Bergen, Bergen, Norway, April 1986a.
[61] E. S. Selmer. *The Local Postage Stamp Problem, Part II: The Bases A_3 and A_4*, volume 44 of *Mathematics Department Research Reports*. University of Bergen, Bergen, Norway, September 1986b.
[62] R. Sprague. *Unterhaltsame Mathematik: Neue Probleme—überraschende Lösungen*. Friedr. Vieweg & Sohn, Braunschweig, 1961. (Translation in English by T. H. O'Beirne was published by Dover, New York in 1963.).
[63] A. Stöhr. Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe. I, II. *J. Reine Angew. Math.*, 194:40–65, 111–140, 1955a. ISSN 0075-4102.
[64] A. Stöhr. Bemerkungen zur additiven Zahlentheorie. III. Vereinfachter Beweis eines Satzes von A. Brauer. *J. Reine Angew. Math.*, 195:172–174 (1956), 1955b. ISSN 0075-4102.

[65] W. Su. *A combinatorial problem in the construction of distributed loop networks.* Master's thesis, Texas State University, Texas, USA, 1993.

[66] C. K. Wong and D. Coppersmith. A combinatorial problem related to multimodule memory organizations. *J. Assoc. Comput. Mach.*, 21:392–402, 1974. ISSN 0004-5411.

[67] G. Yu. Upper bounds for finite additive 2-bases. *Proc. Am. Math. Soc.*, 137(1):11–18, 2009. ISSN 0002-9939.

The Erdős–Turán Problem in Infinite Groups

Sergei V. Konyagin and Vsevolod F. Lev

Dedicated to Mel Nathanson on the occasion of his 60th birthday

Summary Let G be an infinite abelian group with $|2G| = |G|$. We show that if G is not the direct sum of a group of exponent 3 and the group of order 2, then G possesses a perfect additive basis; that is, there is a subset $S \subseteq G$ such that every element of G is uniquely representable as a sum of two elements of S. Moreover, if G *is* the direct sum of a group of exponent 3 and the group of order 2, then it does not have a perfect additive basis; however, in this case, there exists a basis $S \subseteq G$ such that every element of G has at most two representations (distinct under permuting the summands) as a sum of two elements of S. This solves completely the Erdős–Turán problem for infinite groups.

It is also shown that if G is an abelian group of exponent 2, then there is a subset $S \subseteq G$ such that every element of G has a representation as a sum of two elements of S, and the number of representations of nonzero elements is bounded by an absolute constant.

Keywords Additive basis · Erdős–Turán problem · Representation function

Mathematics Subject Classifications (2010). Primary: 11B13, Secondary: 05E15

The first author was supported by grants 08-01-00208 from the Russian Foundation for Basic Research and NSh-3233.2008.1 from the Program Supporting Leading Scientific Schools.

S.V. Konyagin
Steklov Mathematical Institute, 8 Gubkin Street, Moscow 119991, Russia
e-mail: konyagin@ok.ru

V.F. Lev
Department of Mathematics, The University of Haifa at Oranim, Tivon 36006, Israel
e-mail: seva@math.haifa.ac.il

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_14,

1 The Background

A subset of an abelian semigroup is called an *additive basis of order* 2, or *basis* for short, if every element of the semigroup is representable as a sum of two elements of the subset. We say that a basis is *perfect* if every element is represented uniquely, up to the order of summands. The representation function of a basis associates with each element the number of its (ordered) representations as a sum of two elements from the basis. If the semigroup can be embedded into an involution-free group, then for a basis to be perfect, it is necessary and sufficient that its representation function is bounded by 2.

A famous open conjecture of Erdős and Turán [ET41] is that every basis of the semigroup $\mathbb{N}_0$ of non-negative integers has unbounded representation function; that is, if $S \subseteq \mathbb{N}_0$ is a set such that each non-negative integer is representable as a sum of two elements of S, then there are integers with arbitrarily many representations.

Most investigations related to the Erdős–Turán conjecture study representation functions of bases of $\mathbb{N}_0$ (see [NS07] for a survey) or consider the analogous problem for infinite abelian semigroups other than $\mathbb{N}_0$, and also for infinite families of abelian semigroups (see [HH04]). In the present paper, we are concerned with the latter line of research.

There are several noticeable cases where bases with bounded representation functions are known to exist. As an example, Nathanson [N03] proved that the group of integers possesses a perfect basis. Ruzsa [R90] showed that if p is a prime with $(2/p) = -1$, then the group $\mathbb{F}_p \times \mathbb{F}_p$ possesses a basis such that every group element has at most 18 representations as a sum of two elements of this basis. (Here and below for a prime p, we denote by $\mathbb{F}_p$ the finite field with p elements. To simplify the notation, we occasionally identify a field with its additive group.) As a corollary, Ruzsa derived a result [R90, Theorem 1] which easily implies that every finite cyclic group has a basis whose representation function is bounded by an absolute constant, independent of the order of the group. The approach of [R90] was further developed by Haddad and Helou [HH04] to show that for any finite field $\mathbb{F}$ of odd characteristic, the group $\mathbb{F} \times \mathbb{F}$ has a basis whose representation function does not exceed 18. In the case where $\mathbb{F}$ is a finite field of characteristic 2, a basis in $\mathbb{F} \times \mathbb{F}$ with a bounded representation function was constructed in [GDT91, Lemma 1], though the property we are interested in has never been identified explicitly to our knowledge.

2 The Results

For a subset C of an abelian group and an integer $n \geq 1$, we write

$$nC := \{nc : c \in C\}.$$

Our main result is Theorem 1.

Theorem 1. *Let G be an infinite abelian group with $|2G| = |G|$.*

(i) If G is not the direct sum of a group of exponent 3 and the group of order 2, then G has a perfect basis.

(ii) If G is the direct sum of a group of exponent 3 and the group of order 2, then G does not have a perfect basis, but has a basis such that every element of G has at most two representations (distinct under permuting the summands) as a sum of two elements of the basis.

Clearly, if G is an infinite abelian group with $|2G| < |G|$, then for any basis S of G (and indeed, for any subset $S \subseteq G$ with $|S| = |G|$) there is an element of G having as many as $|G|$ representations of the form $2s$ with $s \in S$. In particular, this applies to infinite abelian groups of exponent 2. Similarly, for no infinite family of groups of exponent 2 can one find bases with uniformly bounded representation functions, even if the groups of the family are finite. We show that, nevertheless, efficient bases in such groups do exist if we exclude the zero element from consideration.

Theorem 2. *Each abelian group of exponent 2 possesses a basis such that every nonzero element of the group has at most 36 representations as a sum of two elements of this basis.*

Combined with the result of Haddad and Helou and the corollary of Ruzsa's result, mentioned in Sect. 1, Theorems 1 and 2 readily yield Corollary 1.

Corollary 1. *Let G be an abelian group. If G is infinite with $|2G| = |G|$, cyclic, or has prime exponent, then it possesses a basis with the representation function bounded by an absolute constant (independent of the group), except for the value of the function on the zero element in the case where G is of exponent 2.*

We notice that excluding the zero element for groups of exponent 2 is equivalent to disregarding representations with equal summands. To our present knowledge, a universal constant K may exist with the property that each abelian group possesses a basis such that every element of the group has at most K representations as a sum of two distinct elements of this basis.

3 The Proofs

In this section, we use the word "basis" both in the above-defined and linear-algebraic meaning, adding the attribute *linear* in the latter case to avoid confusion.

Our argument depends on the axiom of choice, which we assume for the rest of the paper.

To handle infinite groups of exponents 2 and 3, we need the following lemma.

Lemma 1. *If G is an infinite abelian group of prime exponent p, then there exists an algebraically closed field $\mathbb{F}$ of characteristic p such that $G \cong \mathbb{F} \times \mathbb{F}$.*

The proof uses several facts, well known from algebra and set theory; namely,

(i) Every vector space has a linear basis;
(ii) An infinite set can be partitioned into two disjoint subsets of equal cardinality;
(iii) An infinite vector space over a finite field has the same cardinality as any of its linear bases;
(iv) The field of rational functions over a finite field in the variables, indexed by the elements of an infinite set, has the same cardinality as this set;
(v) The algebraic closure of an infinite field has the same cardinality as the field itself.

We notice that (i) follows easily from Zorn's lemma, while (ii)–(v) are not difficult to derive from the basic set theory result saying that for any infinite cardinal m, a union of at most m sets, each of cardinality at most m, has cardinality at most m.

Proof of Lemma 1. Considering G as a vector space over the field $\mathbb{F}_p$, find a linear basis B of G and fix a partition $B = B_1 \cup B_2$, where B_1 and B_2 are disjoint subsets of equal cardinality. For $i \in \{1, 2\}$ denote by G_i the group of functions from B_i to $\mathbb{F}_p$ with a finite support; thus, $G_1 \cong G_2$, and since G is isomorphic to the group of functions from B to $\mathbb{F}_p$ with a finite support, we have $G \cong G_1 \times G_2$. Let $\mathbb{F}$ be the algebraic closure of the field of rational functions over $\mathbb{F}_p$ in the variables, indexed by the elements of G_1. By (iv) and (v), the cardinality of $\mathbb{F}$ is equal to the cardinality of G_1. From (iii), we conclude now that every linear basis of $\mathbb{F}$ has the same cardinality as B_1, which, we recall, is a linear basis of G_1. Any bijection from B_1 to a linear basis of $\mathbb{F}$ determines a group isomorphism between G_1 and the additive group of $\mathbb{F}$. As a result, we have $G_1 \cong \mathbb{F}$, and hence also $G_2 \cong \mathbb{F}$, implying the assertion. □

For an abelian group G, an integer $n \geq 1$, and subsets $A, B, C \subseteq G$, we write

$$G_n := \{g \in G : ng = 0\},$$
$$A \pm B := \{a \pm b : a \in A,\ b \in B\},$$

and

$$A + B - C := \{a + b - c : a \in A,\ b \in B,\ c \in C\}.$$

From $G/G_n \cong nG$ we conclude that if $|G|$ is infinite, then $\max\{|G_n|, |nG|\} = |G|$.

Yet another result used in the proof of Theorem 1 is Lemma 2.

Lemma 2. *Let G be an abelian group such that $2G$ is infinite. If $A, B \subseteq G$ satisfy*

$$\max\{|A|, |B|\} < \min\{|2G|, |3G|\},$$

then there exists an element $s \in G$ with $2s \notin A$ and $3s \notin B$.

Proof. Suppose for a contradiction that for every $s \in G$, we have either $2s \in A$ or $3s \in B$. Without loss of generality we assume $B \subseteq 3G$, and we find then a subset $U \subseteq G$ with $|U| = |B|$ and $B = \{3u: u \in U\}$.

Fix $w \in G$ with $3w \notin B$. For any $g \in G_3$, we have $3(w+g) = 3w \notin B$, whence $2(w+g) \in A$ and therefore, $2g \in -2w + A$. Now if $s \in G$ satisfies $3s \in B$, then $s = u + g$ with some $u \in U$ and $g \in G_3$, implying $2s = 2u + 2g \in 2u - 2w + A \subseteq -2w + 2U + A$. It follows that for every $s \in G$ we have either $2s \in -2w + 2U + A$ or $2s \in A$; this, however, is impossible as $|-2w + 2U + A| \leq |U||A| < |2G|$ and $|A| < |2G|$ by the assumptions. □

Eventually, we are ready to prove Theorem 1.

Proof of Theorem 1. We split the proof into three parts.

1. First, suppose that G is of exponent 3. By Lemma 1, we can assume $G = \mathbb{F} \times \mathbb{F}$, where $\mathbb{F}$ is an algebraically closed field of characteristic 3. Set

$$S := \{(x, x^2): x \in \mathbb{F}\}.$$

For each pair $(u, v) \in G$, the number of representations of (u, v) as a sum of two elements of S is the number of solutions of the equation

$$x^2 + (u - x)^2 = v; \quad x \in \mathbb{F},$$

which is either 1 or 2. The assertion follows.

We remark that this argument above actually goes through for any odd prime exponent; however, only the case of exponent 3 is not covered by the proof below.

2. Now suppose that $G = F \oplus \{0, h\}$, where F is of exponent 3 and h has order 2. As shown above, F has a perfect basis S, and it is immediate that $S \cup (h + S)$ is a basis of G such that every element of G has at most two representations (distinct under permuting the summands) as a sum of two elements of this basis.

Assuming, on the other hand, that G possesses a *perfect* basis, we write this basis as $T = T_0 \cup (h + T_1)$ with $T_0, T_1 \subseteq F$. Shifting T appropriately, we assume furthermore that $0 \in T_0$. To obtain a contradiction, we observe that the unique representation of h as a sum of two elements of T has the form $h = t_0 + (h + t_1)$ with $t_0 \in T_0$ and $t_1 \in T_1$; hence, $2(t_1 + h) = t_0 + 0$ gives two representations of t_0 as a sum of two elements of T.

3. Turning to the general case, we denote by μ the initial ordinal of the cardinal $|G|$ and consider a well-ordering $G = \{g_\iota: \iota < \mu\}$. Notice, that μ is a limit ordinal, and hence the successor of any ordinal, smaller than μ, is also smaller than μ.

We set $S_0 := \varnothing$ and construct a chain of subsets S_ι, for each ordinal $\iota \leq \mu$, so that

- $S_\iota \subseteq S_\rho$ whenever $\iota < \rho \leq \mu$;
- if ι is a finite ordinal, then S_ι is finite, and if $\iota \leq \mu$ is infinite, then $|S_\iota| \leq |\iota|$;
- $g_\iota \in S_\rho + S_\rho$ whenever $\iota < \rho \leq \mu$;
- for any ordinal $\iota \leq \mu$ and element $g \in G$ there is at most one representation of g as a sum of two elements of S_ι.

The proof is then completed by observing that S_μ is a perfect basis of G; hence, it suffices to show that the subsets S_ι can be constructed.

We use transfinite recursion, assuming that $\nu \le \mu$ and that S_ι has already been found for each ordinal $\iota < \nu$, and constructing S_ν. If ν is a limit ordinal, then we put $S_\nu := \cup_{\iota<\nu} S_\iota$. If ν is a successor ordinal and $g_{\nu-1} \in S_{\nu-1} + S_{\nu-1}$, then we put $S_\nu := S_{\nu-1}$. In the remaining case, where ν is a successor ordinal and $g_{\nu-1} \notin S_{\nu-1} + S_{\nu-1}$, we put $S_\nu := S_{\nu-1} \cup \{s, t\}$, where $s, t \in G$ with $s + t = g_{\nu-1}$ are chosen to satisfy the following conditions:

(a) if $|3G| < |G|$, then $s, t \notin S_{\nu-1} + 3G$;
(b) $s, t \notin S_{\nu-1} + S_{\nu-1} - S_{\nu-1}$;
(c) $2s, 2t \notin S_{\nu-1} + S_{\nu-1}$;
(d) $s - t \notin S_{\nu-1} - S_{\nu-1}$;
(e) $2s - t, 2t - s \notin S_{\nu-1}$.

Condition (a) is of technical nature and its exact purpose will be clarified in the following paragraph, while the last four conditions ensure that the unique representation property of $S_{\nu-1}$ is inherited by S_ν. Thus, to complete the proof it suffices to show that s and $t := g_{\nu-1} - s$ satisfying (a)–(e) can be found.

To this end we first observe that if $S_{\nu-1}$ is infinite, then condition (e) excludes at most $|S_{\nu-1}| \le |\nu - 1| < |\mu| = |G|$ options for $3s$, and similarly (a)–(d) together exclude fewer, than $|G|$ options for $2s$. Clearly, this conclusion remains valid if $S_{\nu-1}$ is finite. Therefore, in view of Lemma 2, we can assume that $|3G| < |G|$. Consequently, securing (a) at each step of the construction, we have ensured that all elements of $S_{\nu-1}$ fall into distinct cosets of $3G$, and in particular, each of $g_{\nu-1} + S_{\nu-1}$ and $2g_{\nu-1} - 2S_{\nu-1}$ contains at most one element from $3G$. Since (e) can be re-written as

$$3s \notin (g_{\nu-1} + S_{\nu-1}) \cup (2g_{\nu-1} - S_{\nu-1}),$$

if $|3G| \ge 3$, then there exists $g \in G$ such that every $s \in g + G_3$ satisfies (e). As remarked above, (a)–(d) reduce to forbidding fewer than $|G|$ values for $2s$; that is, forbidding fewer than $|G|$ cosets of G_2 for s. Since $|g + G_3| = |G|$ in view of $|3G| < |G|$, and every G_2-coset intersects $g + G_3$ by at most one element, there exists $s \in g + G_3$ with an admissible value of $2s$, proving the assertion.

Suppose, therefore, that $|3G| < 3$. If $|3G| = 1$, then G is of exponent 3, the case which has been addressed above. If $|3G| = 2$, then the identity $g = -2g + 3g$ shows that $G = G_3 + 3G$, and the sum is direct as if $g \in G_3 \cap 3G$, then $2g = 0$ (as $g \in 3G$ and $|3G| = 2$) and $3g = 0$ (as $g \in G_3$), implying $g = 0$. Consequently, G is the direct sum of a group of exponent 3 and the group of order 2. This completes the proof. □

Finally, we prove Theorem 2. As indicated in the introduction, the construction employed in the proof is adopted from [GDT91], where it is used (in the finite-dimensional case) to find small codes with covering radius 2.

Proof of Theorem 2. In view of Lemma 1, it suffices to show that if the field $\mathbb{F}$ of characteristic 2 is either finite or algebraically closed, then the group $\mathbb{F} \times \mathbb{F}$ has a basis with the representation function bounded by 18. Clearly, we can assume $|\mathbb{F}| > 2$.

We fix $d_1, d_2, d_3 \in \mathbb{F}^\times$ with $d_1 + d_2 + d_3 = 0$, write

$$S_i := \{(x, d_i/x) : x \in \mathbb{F}^\times\};\ i \in \{1, 2, 3\},$$

and put $S = S_1 \cup S_2 \cup S_3$. For $(u, v) \in \mathbb{F} \times \mathbb{F}$ let $r(u, v)$ denote the number of representations of (u, v) as a sum of two elements of S, and for $i, j \in \{1, 2, 3\}$ denote by $r_{ij}(u, v)$ the number of representations of (u, v) as a sum of an element of S_i and an element of S_j. Since the sets S_1, S_2, and S_3 are pairwise disjoint, we have

$$r(u, v) = \sum_{i,j=1}^{3} r_{ij}(u, v)$$

and furthermore,

$$r_{ij}(u, v) = \left|\{x \in \mathbb{F} \setminus \{0, u\} : d_i/x + d_j/(x + u) = v\}\right|; \quad i, j \in \{1, 2, 3\}.$$

The equation $d_i/x + d_j/(x + u) = v$ can be re-written as

$$vx^2 + (uv + d_i + d_j)x + d_i u = 0 \tag{1}$$

and since it has a nonzero coefficient unless $(u, v) \neq (0, 0)$, we have $r_{ij}(u, v) \leq 2$, except if $u = v = 0$. It follows that $r(u, v) \leq 18$ and to achieve our goal it suffices to show that for any $(u, v) \neq (0, 0)$ there are $i, j \in \{1, 2, 3\}$ with $r_{ij}(u, v) > 0$. We consider three cases.

If $u = 0$ and $v \neq 0$, then $r_{12}(u, v)$ is the number of solutions of $d_1/x + d_2/x = v$, which is 1.

If $u \neq 0$ and $v = 0$, then $r_{12}(u, v)$ is the number of solutions of $d_1/x = d_2/(x+u)$; this leads to a nondegenerate linear equation, the solution of which is distinct from both 0 and u.

Finally, suppose that $u \neq 0$ and $v \neq 0$. In this case for $i = j$ equation (1) takes the form

$$vx^2 + uvx + d_i u = 0, \tag{2}$$

and for $r_{ii}(u, v)$ to be nonzero it is necessary and sufficient that (2) has a solution (which automatically is then distinct from 0 and u). If $\mathbb{F}$ is algebraically closed, then we are done; suppose, therefore, that $\mathbb{F}$ is finite. Since (2) can be re-written as

$$(x/u)^2 + (x/u) = d_i/(uv),$$

it has a solution if and only if $d_i/(uv)$ belongs to the image of the linear transformation $x \mapsto x + x^2$ of the field $\mathbb{F}$ considered as a vector space over $\mathbb{F}_2$. The kernel of this transformation is a subspace of dimension 1; hence its image is a subspace of $\mathbb{F}$ of co-dimension 1. (This image actually is the set of all the elements of $\mathbb{F}$ with zero trace, but we do not use this fact.) Consequently,

$$d_1/(uv) + d_2/(uv) + d_3/(uv) = 0$$

implies that at least one of $d_1/(uv)$, $d_2/(uv)$, and $d_3/(uv)$ is an element of the image. □

Priority remark. As we learned when this paper was about to be published, there is some intersection between our results and the results of [HH08]. Namely, in [HH08, Theorem 5.1] perfect bases in infinite vector spaces over a field of characteristic, distinct from 2, are constructed, using essentially the same approach as in part 1 of the proof of Theorem 1.

References

[ET41] P. ERDŐS and P. TURÁN, On a problem of Sidon in additive number theory, and on some related problems, *J. Lond. Math. Soc.* **16** (1941), 212–215.

[GDT91] E. GABIDULIN, A. DAVYDOV, and L. TOMBAK, Linear codes with covering radius 2 and other new covering codes, *IEEE Trans. Inform. Theory* **37** (1) (1991), 219–224.

[HH04] L. HADDAD and C. HELOU, Bases in some additive groups and the Erdős-Turán conjecture, *J. Combin. Theory Ser. A* **108** (1) (2004), 147–153.

[HH08] L. HADDAD and C. HELOU, Additive bases representations in groups, *Integers* **8** (2) (2008), #A5.

[N03] M. NATHANSON, Unique representation bases for the integers, *Acta Arith.* **108** (1) (2003), 1–8.

[NS07] J. NEŠETŘIL and O. SERRA, On a conjecture of Erdős and Turán for additive bases, In: *Bib. de la Revista Matematica Iberoamericana, Proceedings of the Segundas Jornadas de Theoria de Numeros (Madrid 2007)*, 1–12.

[R90] I.Z. RUZSA, A just basis, *Monatsh. Math.* **109** (2) (1990), 145–151.

A Tiling Problem and the Frobenius Number

D. Labrousse and J.L. Ramírez Alfonsín

Summary In this paper, we investigate tilings of tori and rectangles with rectangular tiles. We present necessary and sufficient conditions for the existence of an integer C such that any torus, having dimensions greater than C, is tiled with two given rectangles (C depending on the dimensions of the tiles). We also give sufficient conditions to tile a *sufficiently* large n-dimensional rectangle with a set of (n-dimensional) rectangular tiles. We do this by combining the periodicity of some particular tilings and results concerning the so-called *Frobenius number*.

Keywords Frobenius number · Rectangle · Tiling · Torus

Mathematics Subject Classifications (2010). 20M99, 90C10

1 Introduction

Let a and b be positive integers. Let $R(a,b)$ be the 2-dimensional rectangle of sides a and b and let $T(a,b)$ be the 2-dimensional torus. We think of $T(a,b)$ as a rectangle where their parallel sides are identified in the usual way. We will say that a torus T (or a rectangle R) can be *tiled* with *tiles* (i.e., smaller 2-dimensional rectangles) $R_1,\ldots,R_k$ if T (or R) can be filled entirely with copies of R_i, $1 \leq i \leq k$ where rotations are not allowed.

D. Labrousse
Université Pierre et Marie Curie, Paris 6, Equipe Combinatoire et Optimisation,
Case 189 - 4 Place Jussieu 75252 Paris Cedex 05, France

J.L. Ramírez Alfonsín
Université Pierre et Marie Curie, Paris 6, Equipe Combinatoire et Optimisation,
Case 189 - 4 Place Jussieu 75252 Paris Cedex 05, France
and
Institut de Mathématiques et de Modelisation de Montpellier,
Université Montpellier 2, Place Eugène, 34095 Montpellier, France
e-mail: jramirez@math.univ-montp2.fr

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_15,

Question 1. Does there exist a function $C_T = C_T(x, y, u, v)$ (resp. $C_R = C_R(x, y, u, v)$) such that for all integers $a, b \geq C_T$ (resp. $a, b \geq C_R$) the torus $T(a, b)$ (resp. rectangle $R(a, b)$) can be tiled with copies of the rectangles $R(x, y)$ and $R(u, v)$ for given positive integers x, y, u and v?

The special case of Question 1 for $R(a, b)$ when $x = 4, y = 6, u = 5$ and $v = 7$ was posed in the 1991 William Lowell Putnam Examination (Problem B-3). In this case, Klosinski et al. [9] gave a lower bound for C_R. Their method was based on knowledge of the *Frobenius number*. The *Frobenius number*, denoted by $g(s_1, \ldots, s_n)$, of a set of relatively prime positive integers $s_1, \ldots, s_n$, is defined as the largest integer that is not representable as a nonnegative integer combination of $s_1, \ldots, s_n$. It is well known [15] that

$$g(s_1, s_2) = s_1 s_2 - s_1 - s_2. \tag{1}$$

It turns out that the computation of a similar (simple) formula when $n \geq 3$ is much more difficult. In fact, finding $g(s_1, \ldots, s_n)$, for general n, is a hard problem from the computational point of view (we refer the reader to [13] for a detailed discussion on the Frobenius number). Let us notice that equality (1) can be interpreted in terms of *1-dimensional tilings* as follows:

> all sufficiently large interval can be tiled by two given intervals whose lengths are relatively primes.

Klosinski et al. [9] used (1), with particular values for s_1 and s_2, to show that $R(a, b)$ can be tiled with $R(4, 6)$ and $R(5, 7)$ if $a, b \geq 2214$. We improve the latter by showing (see Remark 1) that if $a, b \geq 198$ then $R(a, b)$ can be tiled with $R(4, 6)$ and $R(5, 7)$. This lower bound is not optimal, Narayan and Schwenk [10] showed that it is enough to have $a_1, a_2 \geq 33$ by presenting tilings with more complicated patterns (allowing rotations of both tiles) which is not the case here. We also mention that Barnes [1, Theorem 2.1] used algebraic arguments to show the existence of C_R if some *complex set points* conditions are verified but explicit value for C_R was not given.

In the same spirit of the subjects treated in the volume *Unusual Applications of Number Theory* [11], we explore the connection between tilings and the Frobenius number. We show how plane *periodic* tilings can be *pertubed* with tilings, obtained via the Frobenius number, leading to a positive answer to Question 1. We hope these new methods will motivate further investigations.

The paper is organized as follows. In the next section, we shall give necessary and sufficient conditions on integers x, y, u, v for the existence of $C_T(x, y, u, v)$ (see Theorem 3). In Sect. 3, we give various results in relation with a generalization of C_R for n-dimensional rectangles (see Theorem 5). In particular, the knowledge of an upper bound for $g(s_1, \ldots, s_n)$ is used to show that an n-dimensional rectangle $R(a_1, \ldots, a_n)$ can be tiled with a given set of tiles if $a_j > r^{2n}$ for all $1 \leq j \leq n$ where r is the largest length among all the tiles (see Corollary 1). We finally give some results concerning the tilings of n-dimensional cubes.

2 Tiling Tori

It is known [5,8] that $R(a,b)$ can be tiled with $R(x,y)$ if and only if either x divides one side of R and y divides the other or xy divides one side of R and the other side can be expressed as a nonnegative integer combination of x and y. This shows that a rectangle $R(a,b)$ can be tiled with $R(1,n)$ if and only n divide either a or b. It is clear that this condition is also sufficient for tiling $T(a,b)$ (since a tiling of $R(a,b)$ is also a tiling of $T(a,b)$) but it is not necessary, see for instance Fig. 1.

Proposition 1. *Let n be a prime integer. Then, $T(a,b)$ can be tiled with $R(1,n)$ if and only if n divides either a or b.*

Proof. If n divides either a or b then there is a trivial tiling of $T(a,b)$. If $T(a,b)$ is tiled with $R(1,n)$ then n must divides ab and since it is prime then n must divides either a or b. □

In 1995, Fricke [6] gave the following characterization for tiling a rectangle with two squares.

Theorem 1. *Let a,b,x, and y be positive integers with $\gcd(x,y) = 1$ [6]. Then, $R(a,b)$ can be tiled with $R(x,x)$ and $R(y,y)$ if and only if either a and b are both multiple of x or a and b are both multiple of y or one of the numbers a,b is a multiple of xy and the other can be expressed as a nonnegative integer combination of x and y.*

The conditions of Theorem 1 are again sufficient for tiling $T(a,b)$ but they are not necessary, that is, there are tilings of $T(a,b)$ with $R(x,x)$ and $R(y,y)$ not verifying the above conditions (and thus not tiling $R(a,b)$), see for instance Fig. 2.

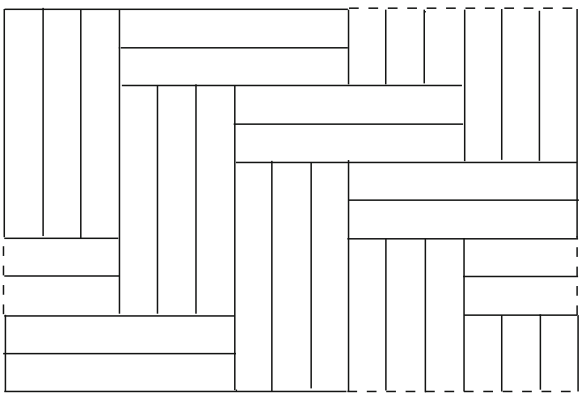


Figure 1 A tiling of $T(15,10)$ with $R(1,6)$ and $R(6,1)$

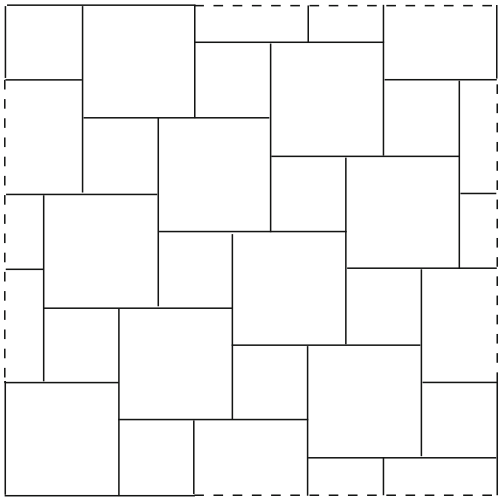


Figure 2 A tiling of $T(13, 13)$ with $R(2, 2)$ and $R(3, 3)$

Remila [14] studied tilings of $T(a, b)$ with two *bars* (that is, when the rectangles are of the form $R(1, y)$ and $R(u, 1)$) where rotations are not allowed. In [14, Sect. 8] the problem of investigating tilings of tori with two general rectangles (not necessarily bars) was posed. By using the algebraic approach (via polynomials and ideals) first introduced by Barnes [1,2], Clivio found [4, Theorem 6.2] the existence of a value C such that for any n-dimensional torus T, having dimensions at least C, there exist necessary and sufficient conditions for T to be tiled with two given n-dimensional rectangles. In particular, for the 2-dimensional case, Clivio's result reads as follows.

Theorem 2. *[4, Theorem 6.2] For arbitrary rectangles $R(x, y)$ and $R(u, v)$ there exists integer C such that for every $T(a, b)$ with $a, b \geq C$, $T(a, b)$ can be tiled with $R(x, y)$ and $R(u, v)$ if and only if* $\gcd\left(\frac{uv}{\gcd(u,a)\gcd(v,b)}, \frac{xy}{\gcd(x,a)\gcd(y,b)}\right) = 1$.

Theorem 2 gives a characterization of *sufficiently large* tori to be tiled with two given rectangles. An estimation of value C was not given in [4] (even for $n = 2$). Clivio remarked that if the volumes of the two given rectangles $R(x, y)$ and $R(u, v)$ (and, in general, the two given n-dimensional rectangles) are relatively primes, that is, if $\gcd(xy, uv) = 1$, then the condition of Theorem 2 always holds.

Proposition 2. *[4, Proposition 6.1, Step 2] Let u, v, x, y and s be positive integers with $\gcd(xy, uv) = \gcd(s, xy) = \gcd(s, uv) = 1$ and such that $T(s, s)$ is tiled with $R(xy, xy)$ and $R(uv, uv)$. Then, $T(a, b)$ can be tiled with $R(x, y)$ and $R(u, v)$ if $a, b \geq s(xy)(uv)$.*

This yield to the following lower bound (by taking $s = xy + uv$)

$$C_T \geq (xy + uv)xyuv. \tag{2}$$

We might improve the latter by using a complete different technique.

Theorem 3. *Let u, v, x, and y be positive integers such that* $\gcd(xy, uv) = 1$. *Then, $T(a, b)$ can be tiled with $R(x, y)$ and $R(u, v)$ if*

$$a, b \geq \min\{n_1(uv + xy) + 1, n_2(uv + xy) + 1\}$$

where $n_1 = \max\{ux, vy\}$ and $n_2 = \max\{vx, uy\}$.

We notice that the above lower bound improves the one given in (2) by a factor of $\max\{ux, vy\}$. For instance, if we take $R(3, 5)$ and $R(4, 2)$ then $n_1 = 12$, $n_2 = 20$ and Theorem 3 gives $C_T \geq 12(15 + 8) + 1 = 277$ while (2) gives $C_T \geq (15 + 8)(15)(8) = 2760$. The latter lower bound can be improved since, by Proposition 2, $C_T \geq 120s$ where $\gcd(s, 15) = \gcd(s, 8) = 1$ and such that $T(s, s)$ is tiled with $R(15, 15)$ and $R(8, 8)$. It is clear that such integer s must be at least 11 and thus obtaining $C_T \geq 1320$ (which still worst than our lower bound).

Theorem 3 implies the following characterization.

Theorem 4. *Let u, v, x, and y be positive integers. Then, there exists $C_T(x, y, u, v)$ such that any $T(a, b)$ with $a, b \geq C_T$ can be tiled with $R(x, y)$ and $R(u, v)$ if and only if* $\gcd(xy, uv) = 1$.

Proof. The sufficiency follows from Theorem 3. For the necessity, suppose, by contradiction, that $\gcd(xy, uv) = d > 1$. Since $T(a, b)$ can be tiled with $R(x, y)$ and $R(u, v)$, then $ab = l_1(xy) + l_2(uv)$ for some nonnegative integers l_1, l_2 and any $a, b \geq C_T$. Since $\gcd(xy, uv) = d$, then d divides ab for any $a, b \geq C_T$. In particular, d divides pq for any pair of primes $p, q > C_T$, which is a contradiction. □

In order to prove Theorem 3, we may consider a special Euclidean plane tiling $\mathcal{T}^*$ formed with two rectangles $R(x, y)$ and $R(u, v)$ with sides parallel to the real axes, as shown in Fig. 3 (we always suppose that the sides u and x are horizontal and the sides y and v are vertical).

Let u and v be positive integers. A plane tiling $\mathcal{T}$ is said to be *horizontally periodic* with *horizontal period*, denoted by $h_{\mathcal{T}}$, equals to u (resp. *vertically periodic* with *vertical period*, denoted by $v_{\mathcal{T}}$, equals to v) if $\mathcal{T} + (u, 0)$ (resp. $\mathcal{T} + (0, v)$) is a congruent transform mapping $\mathcal{T}$ into itself. A tiling $\mathcal{T}$ is *periodic*, if it is both horizontally and vertically periodic.

Lemma 1. *Let $\mathcal{T}^*$ be the plane tiling given in Fig. 3 with $R(x, y)$ and $R(u, v)$. Then, $\mathcal{T}^*$ is periodic with $h_{\mathcal{T}^*} = v_{\mathcal{T}^*} = uv + xy$.*

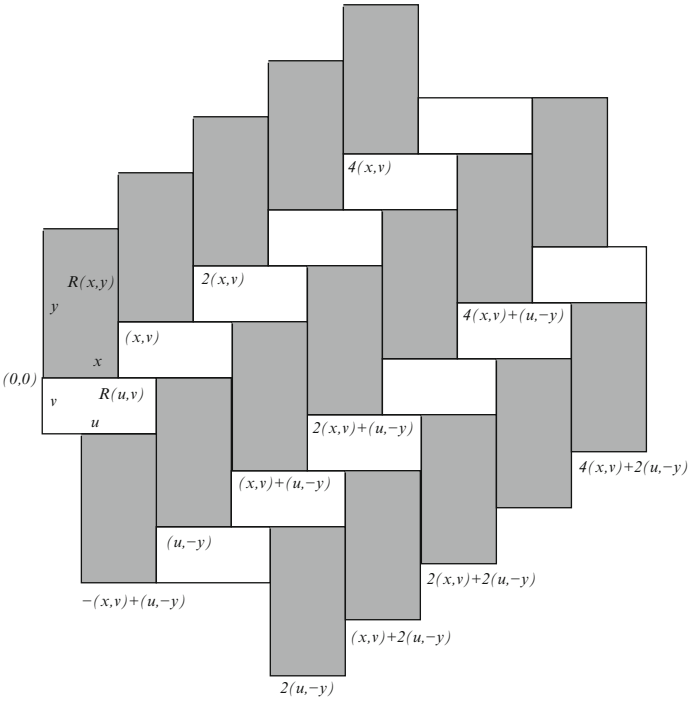


Figure 3 Tiling T^* of the plane

Proof. Without loss of generality, we assume that the lower leftmost corner of one copy of $R(x, y)$ is placed at $(0, 0)$. It is clear that the coordinate of the lower leftmost corner of any other copy of $R(x, y)$ is given by $p(x, v) + q(u, -y)$ with $p, q \in \mathbf{Z}$. And thus, the translation $\mathscr{T}^* + (px + qu, pv - qy)$ is a congruent transform mapping $\mathscr{T}^*$ into itself. In particular, by taking $p = y$ and $q = v$ (resp. by taking $p = u$ and $q = -x$) we have that $\mathscr{T}^* + (vu + yx, 0)$ (resp. $\mathscr{T}^* + (0, xy + uv)$) is a congruent transform mapping $\mathscr{T}^*$ into itself. Therefore, $\mathscr{T}^*$ is periodic with $h_{\mathscr{T}^*} = v_{\mathscr{T}^*} = uv + xy$. □

Proposition 3. *Let $p, q \geq 1$ be integers. Then, $T(ph_{\mathscr{T}^*}, qv_{\mathscr{T}^*})$ can be tiled with $R(x, y)$ and $R(u, v)$.*

Proof. Without loss of generality, we assume that the lower leftmost corner of one copy of $R(x, y)$ is placed at $(0, 0)$. Let B be the rectangle formed by lines $x_1 = 0$, $x_2 = ph_{\mathscr{T}^*}$, $y_1 = 0$ and $y_2 = qv_{\mathscr{T}^*}$. By definition of horizontally period, if a rectangle R is split by a line x_1 into two parts, r_1 (the part lying inside B) and r_2 (the part lying outside B) then the corresponding translated rectangle is also split by line x_2 into two parts r_1' (the part lying outside B) and r_2' (the part lying inside B) where r_1 is congruent to r_1' and r_2 is congruent to r_2' (similarly for the

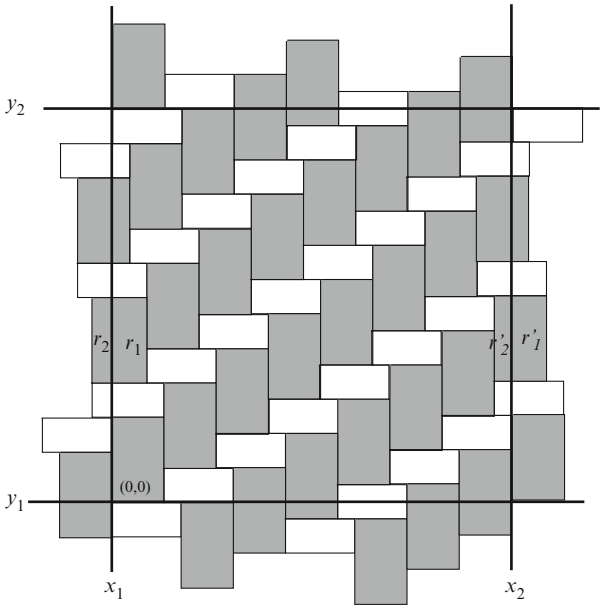


Figure 4 Rectangle B formed with $R(3,5)$ and $R(4,2)$

split rectangles by lines y_1 and y_2), this is illustrated in Fig. 4 when $p = q = 1$ and $x = 3, y = 5, u = 4, v = 2$. Thus, the tiling induced by the copies inside B where their opposites sides are identified gives the desired tiling of $T(ph_{\mathscr{T}^*}, qv_{\mathscr{T}^*})$. □

Proposition 4. *Let x, y, u, and v be positive integers. Then, $T(ph_{\mathscr{T}^*} + sux, qv_{\mathscr{T}^*} + tvy)$ can be tiled with $R(x, y)$ and $R(u, v)$ for all integers $p, q \geq 1$ and $s, t \geq 0$.*

Proof. Let E_1 (resp. E_2) be the row formed by sticking together u (resp. x) copies of $R(x, y)$ (resp. $R(u, v)$) and let F_1 (resp. F_2) be the column formed by sticking together v (resp. y) copies of $R(x, y)$ (resp. $R(u, v)$), see Fig. 5.

Given the constructed rectangle B as in Proposition 3 (that induces a tiling of $T(ph_{\mathscr{T}^*}, qv_{\mathscr{T}^*})$), we shall construct a rectangle B' that will induce a tiling of $T(ph_{\mathscr{T}^*} + sux, qv_{\mathscr{T}^*} + tvy)$. We will do this as follows (each step of the construction is illustrated with the case when $p = q = s = t = 1, x = 3, y = 5, u = 4, v = 2$).

Let E (resp. F) be the set of rectangles R of $\mathscr{T}^*$ such that either R shares its left-hand side border (resp. its bottom border) with the right-hand side border (resp. the top border) of B or R is cut by the right-hand side border (resp. the top border) of B. Let $\bar{B}$ be the union of the rectangles inside B together with sets E and F. We place $\bar{B}$ in the plane such that the leftmost bottom corner of one copy of $R(x, y)$ is placed at $(0, 0)$. Figure 6 illustrates the construction of $\bar{B}$.

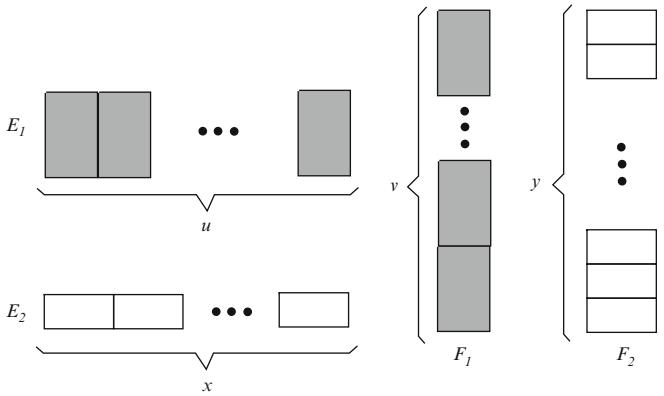


Figure 5 Blocks of rows and columns

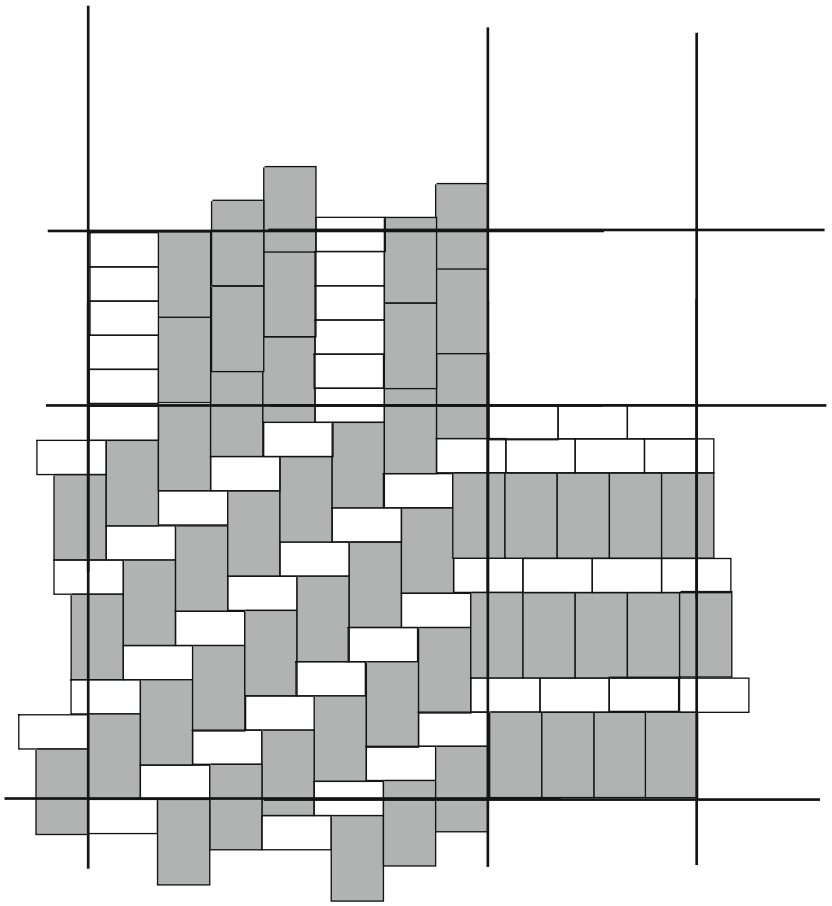


Figure 6 Rectangle $\bar{B}$

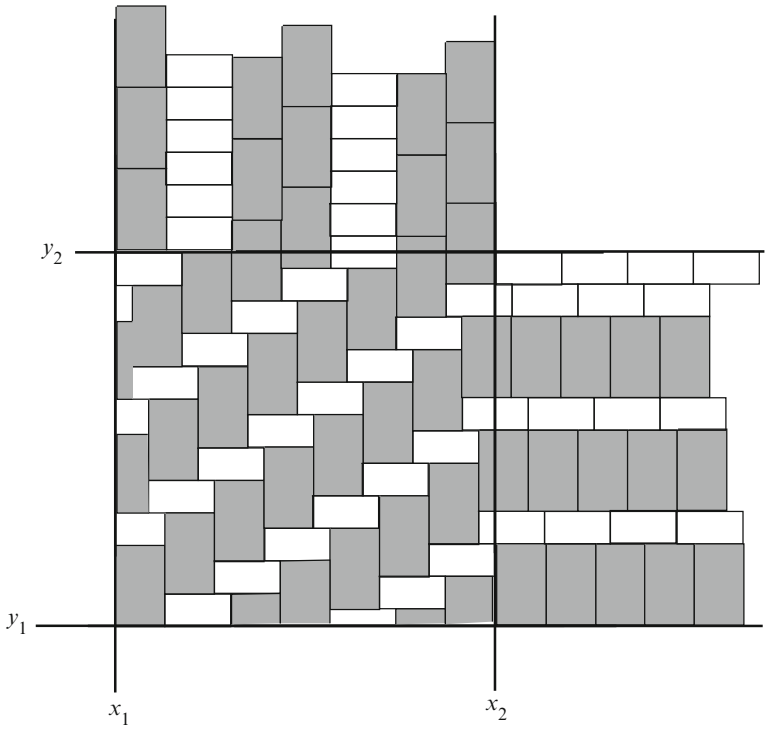


Figure 7 Extended tiling

Now, for each rectangle R of E, we stick s copies of E_1 if $R = R(x, y)$ (or s copies of E_2 if $R = R(u, v)$) to the right-hand side of R. And, analogously, for each rectangle R of F, we stick t copies of F_1 if $R = R(x, y)$ (or t copies of F_2 if $R = R(u, v)$) above R, we do this in Fig. 7.

Let B' be the rectangle formed by lines $x_1 = 0, y_1 = 0, x_3 = ph_{\mathscr{T}^*} + sux$ and $y_3 = qv_{\mathscr{T}^*} + tvy$ (notice that if $s = t = 0$ then $x_3 = x_2$ and $y_3 = y_2$). The rectangle formed by lines x_2, x_3, y_2 and y_3 (lying inside B' in its rightmost top corner) is of size $(sux) \times (tvy)$ and it can be tiled by placing tv, rows each formed by sticking together su copies of E_1, this is done in Fig. 8.

We have the following two observations concerning B'.

(a) By definition of horizontal (resp. vertical) periodicity of $\mathscr{T}^*$, the intersection of x_1 and y_3 (resp. of y_1 and x_3) is a leftmost bottom corner of a copy of $R(x, y)$.
(b) If a rectangle is split by line x_1 into two parts r_1 (part lying inside B') and r_2 (part lying outside B') then the corresponding translated rectangle is split by line x_3 into two parts, r_1' (part lying outside B') and r_2' (part lying inside B') where r_1 is congruent to r_1' and r_2 is congruent to r_2' (similarly for the split rectangles by lines y_1 and y_3).

Therefore, by the above observations, the desired tiling of $T(ph_{\mathscr{T}^*} + sux, qv_{\mathscr{T}^*} + tvy)$ is obtained by identifying opposite sides of B'. □

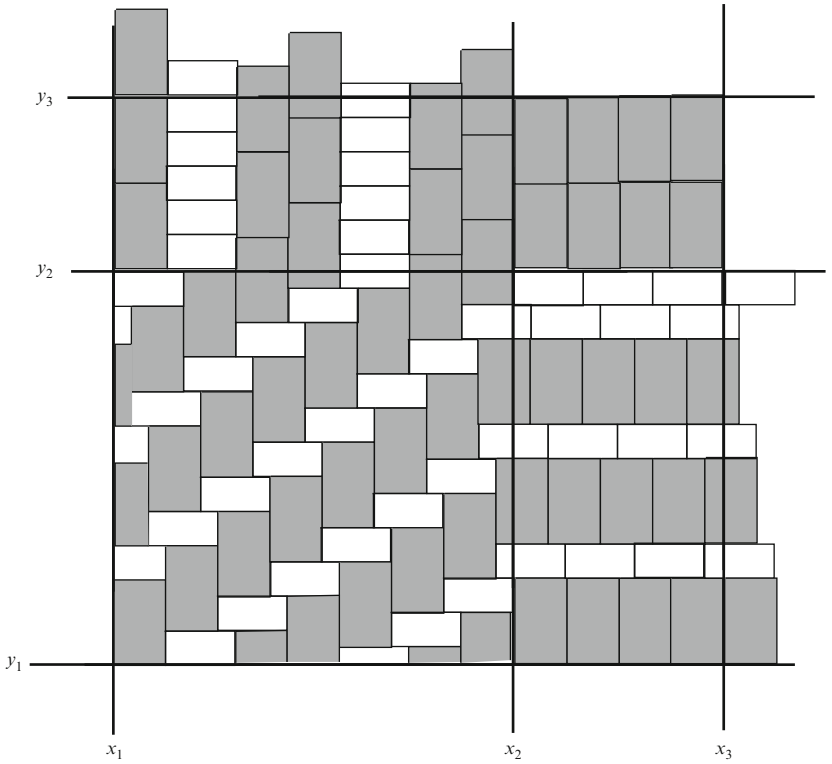


Figure 8 Rectangle B' inducing a tiling of $T^*(33, 35)$ with $R(3, 5)$ and $R(4, 2)$

Proposition 5. *Let $x, y, u,$ and v be positive integers such that* $\gcd(xy, uv) = 1$. *Then,* $\gcd(xy + uv, vx) = \gcd(xy + uv, uy) = 1$.

Proof. We first show that if $\gcd(xy, uv) = 1$, then $\gcd(u, x) = \gcd(u, y) = \gcd(v, x) = \gcd(v, y) = 1$. Indeed, if $\gcd(x, u) = d > 1$, then there exists an integer $k > 1$ with $k|d$. So, k divides both x and u and thus $k|\gcd(xy, uv)$ implying that $\gcd(xy, uv) > 1$ which is a contradiction (similar for the other cases).

We shall now show that $\gcd(xy + uv, vx) = 1$ (the case $\gcd(uv + xy, uy) = 1$ can be done similarly). Let us suppose that $\gcd(uv + xy, vx) = k > 1$ and thus k divides both $uv + xy$ and vx. Let $p > 1$ be a prime such that p divides k. Then p also divides both $uv + xy$ and vx, and since p is prime then we have that p divides either v or x.

Case 1. If p divides v, then $p|uv$ and since $p|(uv + xy)$, then either $p|x$ (but since $p|v$ then $p|\gcd(x, v)$ implying that $\gcd(x, v) > 1$ which is a contradiction) or $p|y$ (but since $p|v$ then $p|\gcd(y, v)$ implying that $\gcd(y, v) > 1$ which is a contradiction).

Case 2. If p divides x, then $p|xy$ and since $p|(uv + xy)$, then either $p|u$ (but since $p|x$ then $p|\gcd(x,u)$ implying that $\gcd(x,u) > 1$ which is a contradiction) or $p|v$ (but since $p|x$ then $p|\gcd(x,v)$ implying that $\gcd(x,v) > 1$ which is a contradiction). □

We may now prove Theorem 3.

Proof of Theorem 3. Let $C = \max\{g(h_{\mathscr{T}^*}, ux) + h_{\mathscr{T}^*} + 1, g(v_{\mathscr{T}^*}, vy) + v_{\mathscr{T}^*} + 1\}$ (notice that the Frobenius numbers are well defined by Proposition 5). Let us suppose that $C = g(h_{\mathscr{T}^*}, ux) + h_{\mathscr{T}^*} + 1$ (similarly, in the case $C = g(v_{\mathscr{T}^*}, uy) + v_{\mathscr{T}^*} + 1$). Then, by definition of the Frobenius number, there exist integers $p, s \geq 0$ such that $N = ph_{\mathscr{T}^*} + sux$ for any integer $N \geq g(h_{\mathscr{T}^*}, ux) + 1$. Thus there exist integers $p \geq 1$ and $s \geq 0$ such that $N = ph_{\mathscr{T}^*} + sux$ for any integer $N \geq g(h_{\mathscr{T}^*}, vx) + h_{\mathscr{T}^*} + 1$. So, since $p \geq 1$ then, by Proposition 4, $T(a,b)$ can be tiled with $R(x,y)$ and $R(u,v)$ if $a, b \geq \max\{g(h_{\mathscr{T}^*}, ux) + h_{\mathscr{T}^*} + 1, g(v_{\mathscr{T}^*}, vy) + v_{\mathscr{T}^*} + 1\}$ or equivalently, by (1), if $a, b \geq \max\{vx(uv + xy) + 1, uy(uv + xy) + 1\}$.

We finally observe that in the construction of $\mathscr{T}^*$ we assume that the sides v and y are vertical and the sides u and x are horizontal but we could construct a similar tiling with the sides u and y vertical and the sides v and x horizontal. In this case, by applying the same arguments as above, we obtain that $T(a,b)$ can be tiled with $R(x,y)$ and $R(v,u)$ if $a, b \geq \max\{ux(uv + xy) + 1, vy(uv + xy) + 1\}$, and the result follows. □

3 Tiling Rectangles

Let $a_1, \ldots, a_n$ be positive integers. We denote by $R = R(a_1 \ldots, a_n)$ the n-dimensional rectangle of sides a_i, that is, $R = \{(x_1, \ldots, x_n) \in \mathbf{R}^n | 0 \leq x_i \leq a_i, i = 1, \ldots, n\}$. A n-dimensional rectangle R is said to be *tiled* with *tiles* (n-dimensional rectangles) $R_1, \ldots, R_k$ if R can be filled entirely with copies of R_i, $1 \leq i \leq k$ (rotations are not allowed).

Our main result in this section is given by Theorem 5 (below) stating that a *sufficiently* large n-dimensional rectangle can be tiled with a set of $n + k - 1$ tiles if any k-subset of the set of 1-coordinates (set of the first lengths) of the tiles are relatively primes and the set of j-coordinates (set of the j^{th} lengths) of the tiles are pairwise relatively prime for each $j = 2, \ldots, n$. We shall use again the Frobenius number and for, we need the following result.

Proposition 6. *Let $a_1, \ldots, a_n$ be positive integers such that $\gcd(a_i, a_j) = 1$, for all $1 \leq i \neq j \leq n$. Then,*

$$\gcd\left(\frac{a_{i_1} \cdots a_{i_\ell}}{a_{i_\ell}}, \ldots, \frac{a_{i_1} \cdots a_{i_\ell}}{a_{i_1}}\right) = 1$$

for any $\{i_1 < \cdots < i_\ell\} \subseteq \{1, \ldots, n\}$.

We leave the reader to prove this proposition by induction on ℓ.

Theorem 5. *Let $k \geq 2$ and $n \geq 1$ be integers. Let $R_i(x_1^i, \ldots, x_n^i)$, $i = 1, \ldots, n + k - 1$ be rectangles formed with integers $x_j^i \geq 2$ such that*

(a) $\gcd(x_1^{i_1}, \ldots, x_1^{i_k}) = 1$ *for any* $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n + k - 1\}$ *and*
(b) $\gcd(x_j^{i_1}, x_j^{i_2}) = 1$ *for any* $\{i_1, i_2\} \subset \{1, \ldots, n + k - 1\}$ *and any* $j = 2, \ldots, n$.

Let $g_1 = \max\{g(x_{i_1}^1, \ldots, x_{i_k}^1) | \{i_1, \ldots, i_k\} \subset \{1, \ldots, n + k - 1\}\}$ *and*

$$g_\ell = \max\left\{ g\left(\frac{x_\ell^{i_1} \cdots x_\ell^{i_{\ell+k-2}}}{x_\ell^{i_{\ell+k-2}}}, \ldots, \frac{x_\ell^{i_1} \cdots x_\ell^{i_{\ell+k-2}}}{x_\ell^{i_1}} \right) \mid \{i_1, \ldots, i_{\ell+k-2}\} \subseteq \{1, \ldots, n + k - 1\} \right\}$$

for each $\ell = 2, \ldots, n$. *Then,*

$R(a_1, \ldots, a_n)$ *can be tiled with tiles* $R_1, \ldots, R_{n+k-1}$ *if* $a_j > \max\limits_{1 \leq \ell \leq n} \{g_\ell\}$ *for all* j.

Notice that when $k = 2$, the number of tiles is $n + 1$ which is the minimum required since, by Theorem 1, two square tiles do not suffice to tile all sufficiently large rectangles. Also, notice that if $k = 2$ condition (a) becomes condition (b) with $j = 1$ and when $k > 2$ the number of tiles is increased but condition (a) is less restrictive that condition (b), we justify this below (see second paragraph after Corollary 1). We finally remark that the Frobenius numbers g_i used in Theorem 5 are well defined by Proposition 6.

In order to understand how the Frobenius number is used, we show how the constructive proof proceeds in the special case when $n = 2$ and $k = 2$ (the complete proof, given below, will be done by induction on n). Let us consider a rectangle $R(a_1, a_2)$ and tiles $R_i(x_1^i, x_2^i)$ with $i = 1, \ldots, 3$. Since $\gcd(x_1^i, x_1^j) = 1$ then if $a_1 > g_1$, we have $a_1 = ux_1^i + vx_1^j$ for all $1 \leq i \neq j \leq 3$. So, we can form a rectangle $R_{ij} = R(a_1, x_2^i x_2^j)$ by sticking together u copies of R_i and v copies of R_j along the first coordinate, and then by replacing each R_i (resp. R_j) by a column of x_2^j (resp. of x_2^i) copies of R_i (resp. R_j). Now, since $\gcd(x_2^i, x_2^j) = 1$ for all $1 \leq i \neq j \leq 3$ then, by Proposition 6, $\gcd(x_2^1 x_2^2, x_2^1 x_2^3, x_2^2 x_2^3) = 1$. So, if $a_2 > g(x_2^1 x_2^2, x_2^1 x_2^3, x_2^2 x_2^3)$, we have $a_2 = ux_2^1 x_2^2 + vx_2^1 x_2^3 + wx_2^2 x_2^3$. Therefore, $R(a_1, a_2)$ can be tiled with R_1, R_2, R_3 by sticking together u copies of R_{12}, v copies of R_{13} and w copies of R_{23} along the second coordinate.

Remark 1. $R(a, b)$ can be tiled with $R(4, 6)$ and $R(5, 7)$ if $a, b > 197$.

Proof. We apply the above argument with $R_1(6, 4), R_2(5, 7)$ and $R_3(7, 5)$ obtaining that $g_1 = \max\{g(6, 5), g(6, 7), g(5, 7)\} = \max\{19, 29, 23\} = 29$ and $g_2 = \max\{g(28, 20, 35)\} = 197$. □

We denote by $(R;q)$ the rectangle obtained from $R(x_1,\dots,x_n)$ by sticking together q copies of R along the n^{th}-axis, that is, $(R;q) = R(x_1,\dots,x_{n-1},qx_n)$. We also denote by $\bar{R}$ the $(n-1)$-dimensional rectangle obtained from $R(x_1,\dots,x_n)$ by setting $x_n = 0$, that is, $\bar{R} = \bar{R}(x_1,\dots,x_{n-1})$.

Proof of Theorem 5. We shall use induction on the dimension n with a fixed $k \geq 2$. For $n = 1$, we have that $\gcd(x_1^{i_1},\dots,x_1^{i_k}) = 1$ for any $\{i_1,\dots,i_k\} \subset \{1,\dots,n+k-1\}$. Since $a_1 > g_1$ then, by definition of the Frobenius number, any integer $a_1 > g(x_1^{i_1},\dots,x_1^{i_k})$ is of the form $a_1 = \sum_{j=1}^{k} u_j x_1^{i_j}$ where u_j is a nonnegative integer. Thus, the 1-dimensional rectangle $R(a_1)$ (that is, the interval $[0,a_1]$) can be tiled by sticking together tiles $(R^{i_1};u_1),\dots,(R^{i_k};u_k)$ (that is, by sticking together intervals $[0,u_1x_1^{i_1}],\dots,[0,u_kx_1^{i_k}]$).

Let us suppose that it is true for $n-1 \geq 1$ and we prove it for n. Let x_j^i be a positive integer for each $j = 1,\dots,n$ and each $i = 1,\dots,n+k-1$ with $\gcd(x_1^{i_1},\dots,x_1^{i_k}) = 1$ for any $\{i_1,\dots,i_k\} \subset \{1,\dots,n+k-1\}$ and $\gcd(x_j^{i_1},x_j^{i_2}) = 1$ for any $\{i_1,i_2\} \subset \{1,\dots,n+k-1\}$ and any $j = 2,\dots,n$. Let $R_i = R_i(x_1^i,\dots,x_n^i)$, $i = 1,\dots,n+k-1$ and $a_j > \max\{g_1,g_2,\dots,g_n\}$. By induction, $R(a_1,\dots,a_{n-1})$ can be tiled with tiles $\bar{R}_{i_1},\dots,\bar{R}_{i_{n+k-2}}$ for any $\{i_1 < \cdots < i_{n+k-2}\} \subset \{1,\dots,n+k-1\}$ since $a_j > \max\{g_1,g_2,\dots,g_{n-1}\}$ for any $1 \leq j \leq n-1$.

We claim that $R(a_1,\dots,a_{n-1},x_n^{i_1}\cdots x_n^{i_{n+k-2}})$ can be tiled with tiles $R_{i_1},\dots,R_{i_{n+k-2}}$ for any $\{i_1 < \cdots < i_{n+k-2}\} \subset \{1,\dots,n+k-1\}$. Indeed, if we consider the rectangle $R(a_1,\dots,a_{n-1})$ embedded in $\mathbf{R}^n$ with $x_n = 0$ then, by replacing each tile $\bar{R}_{i_j}$ (used in the tiling of $R(a_1,\dots,a_{n-1})$) by

$$\left(R_{i_j};\frac{x_n^{i_1}\cdots x_n^{i_n}}{x_n^{i_j}}\right)$$

we obtain a tiling of $R(a_1,\dots,a_{n-1},x_n^{i_1}\cdots x_n^{i_{n+k-2}})$ with tiles $R_{i_1},\dots,R_{i_{n+k-2}}$.

Now, since $a_n > g_n$ then

$$a_n = w_{n+k-1}\left(\frac{x_n^1\cdots x_n^{n+k-1}}{x_n^{n+k-1}}\right)+\cdots+w_1\left(\frac{x_n^1\cdots x_n^{n+k-1}}{x_n^1}\right)$$

where each w_i is a nonnegative integer. By the above claim, rectangle

$$R'_j = \left(a_1,\dots,a_{n-1},\frac{x_n^1\cdots x_n^{n+k-1}}{x_n^j}\right)$$

can be tiled with tiles $\{R_1,\dots,R_{n+k-1}\}/R_j$ for each $j = 1,\dots,n+k-1$. Thus, $R(a_1,\dots,a_{n-1},a_n)$ can be tiled with $R_1,\dots,R_{n+k-1}$ by sticking together tiles $(R'_1;w_1),\dots,(R'_{n+k-1};w_{n+k-1})$ along the n^{th}-axis. □

Example 1. Let $R_1 = (22, 3, 3)$, $R_2 = (14, 5, 5)$, $R_3 = (21, 2, 2)$, $R_4 = (15, 7, 7)$ and $R_5 = (55, 11, 11)$. In this case, we have $k = n = 3$.

$$\begin{aligned} g_1 &= \max\{g(22,14,21), g(22,14,15), g(22,14,55), g(22,21,15), g(22,21,55),\\ &\qquad g(22,15,55), g(14,21,15), g(14,21,55), g(14,15,55), g(21,15,55)\}\\ &= \max\{139, 91, 173, 181, 243, 97, 288, 151, 179\} = 288. \end{aligned}$$

With $\ell = 2$ we obtain

$$\begin{aligned} g_2 &= \max\{g(3\cdot 5, 3\cdot 2, 5\cdot 2), g(3\cdot 5, 3\cdot 7, 5\cdot 7), g(3\cdot 5, 3\cdot 11, 5\cdot 11),\\ &\qquad g(3\cdot 2, 3\cdot 7, 2\cdot 7), g(3\cdot 2, 3\cdot 11, 2\cdot 11), g(3\cdot 7, 3\cdot 11, 7\cdot 11),\\ &\qquad g(5\cdot 2, 5\cdot 7, 2\cdot 7), g(5\cdot 2, 5\cdot 11, 2\cdot 11), g(5\cdot 7, 5\cdot 11, 7\cdot 11),\\ &\qquad g(2\cdot 7, 2\cdot 11, 7\cdot 11)\}\\ &= \max\{g(15,6,10), g(15,21,35), g(15,33,55), g(6,21,14), g(6,33,22),\\ &\qquad g(21,33,77), g(10,35,14), g(10,55,22), g(35,55,77), g(14,22,77)\}\\ &= \max\{29, 139, 227, 43, 71, 331, 81, 133, 603, 195\} = 603. \end{aligned}$$

And with $\ell = 3$

$$\begin{aligned} g_3 &= \max\{g(3\cdot 5\cdot 2, 3\cdot 5\cdot 7, 3\cdot 2\cdot 7, 5\cdot 2\cdot 7), g(3\cdot 5\cdot 2, 3\cdot 5\cdot 11, 3\cdot 2\cdot 11, 5\cdot 2\cdot 11),\\ &\qquad g(3\cdot 5\cdot 7, 3\cdot 5\cdot 11, 3\cdot 7\cdot 11, 5\cdot 7\cdot 11),\\ &\qquad g(5\cdot 2\cdot 7, 5\cdot 2\cdot 11, 5\cdot 7\cdot 11, 2\cdot 7\cdot 11)\}\\ &= \max\{g(30,105,42,70), g(30,165,66,110), g(105,165,231,385),\\ &\qquad g(70,110,385,154)\}\\ &= \max\{383, 619, 2579, 1591\} = 2579. \end{aligned}$$

Therefore, Theorem 5 implies that $R(a_1, a_2, a_3)$ can be tiled with tiles $R_1, \ldots, R_5$ if $a_1, a_2, a_3 > \max\{g_1, g_2, g_3\} = \{288, 603, 2579\} = 2579$.

Corollary 1. *Let $k \geq 2$ and $n \geq 1$ be integers and let $R_i(x_1^i, \ldots, x_n^i)$, $i = 1, \ldots, n + k - 1$ be rectangles formed with integers $x_j^i \geq 2$ verifying conditions (a) and (b) of Theorem 5. Then,*

$$R(a_1, \ldots, a_n) \text{ can be tiled with tiles } R_1, \ldots, R_{n+k-1} \text{ if } a_j > r^{2n} \text{ for all } j$$

where r is the largest length among all the tiles R_i.

Proof. The following upper bound for the Frobenius number, due to Wilf [13, Theorem 3.1.9], states that

$$g(b_1, \ldots, b_n) \leq b_n^2 \tag{3}$$

where $b_1 < \cdots < b_n$ are relatively prime integers. In our case, this gives

$$g_\ell \leq (z_\ell^\ell)^2$$

where $z_\ell = \max\{x_\ell^1, \ldots, x_\ell^{n+k-1}\}$ for each $\ell = 1, \ldots, n$. Therefore, by Theorem 5, we have that $R(a_1, \ldots, a_n)$ can be tiled with tiles $R_1, \ldots, R_{n+k-1}$ if

$$a_i > r^{2n} \geq \max_{1 \leq \ell \leq n} \{z_\ell^{2\ell}\} \geq \max_{1 \leq \ell \leq n} \{g_\ell\}, \tag{4}$$

where r is the largest length among tiles $R_1, \ldots, R_{n+k-1}$. □

Notice that the lower bound given in the above corollary depends on the lower bound given by (3) and thus it is not necessary optimal. For instance, in the above example, Corollary 1 would give $a_1, a_2, a_3 > 55^6$ while $a_1, a_2, a_3 > 2579$ is sufficient as shown in the example.

In [3, Theorem 3], it was announced (without proof) Theorem 5 for the case when $k = 2$, that is, when each set consisting of the j^{th} lengths of the tiles, are pairwise relatively prime. The latter is sometimes restrictive, for instance, the above example cannot be considered under these conditions. Indeed, any permutation of the coordinates (lengths) of tiles in this case will give a pair of j^{th}-coordinates not relatively primes for some $1 \leq j \leq 3$.

Katona and Szász [7] also investigated conditions for tiling n-dimensional rectangles by applying a generalization of the well-known Marriage theorem. They showed [7, Theorems 2 and 3] that $R(a_1, \ldots, a_n)$ can be tiled with tiles $R_1, \ldots, R_m$ if

$$a_j > 3^{km2^{mk}} r^{2^{kn}+2} \text{ for all } j$$

where r is the largest length among all the tiles and $k \geq 1$ is the cardinality of *special* sets constructed from the lengths of the tiles. In particular, when $k = 1$ (the smallest cardinality possible) the above inequality gives

$$a_j > 3^{m2^m} a^{2^n+2}. \tag{5}$$

It is clear that this lower bound is exponentially worst than the one given by Corollary 1.

3.1 Cube Tiles

Theorem 6. *[3, Theorem 4] All sufficiently large n-dimensional rectangle R can be tiled by any given set of $n+1$ cubes with pairwise relatively prime edge lengths.*

We notice that this theorem is a particular case of Theorem 5 by taking $k = 2$ and $x_j^i = a_i$ for each $i = 1, \ldots, n+1$ and all $1 \leq j \leq n$ where $1 < a_1 < a_2 < \cdots < a_{n+1}$ are pairwise relatively prime integers, $n \geq 1$. Moreover, Theorem 5 implies that $R(\underbrace{a, \ldots, a}_{n})$ can be tiled with $R(\underbrace{a_1, \ldots, a_1}_{n}), \ldots, R(\underbrace{a_{n+1}, \ldots, a_{n+1}}_{n})$ if

$$a > g(A_1, \ldots, A_{n+1}) \tag{6}$$

where $A_i = P/a_i$ with $P = \prod_{j=1}^{n+1} a_j$. It turns out that the above lower bound can be stated explicitly by using the following formula, due to Tripathi [16], when $1 < a_1 < a_2 < \cdots < a_{n+1}$ are pairwise relatively prime integers,

$$g(A_1, \ldots, A_{n+1}) = nP - \sum_{i=1}^{n+1} A_i. \tag{7}$$

The above lower bound is not optimal in general. For instance, by combining (6) and (7), we obtain that $R(a, a)$ can be tiled with $R(2, 2)$, $R(3, 3)$ and $R(p, p)$ if $a \geq 7p - 6$ where p is an odd integer with $3 \nmid p$. The following result improves the latter.

Theorem 7. *Let $p > 4$ be an odd integer with $3 \nmid p$. Then, $R(a, a)$ can be tiled with $R(2, 2)$, $R(3, 3)$ and $R(p, p)$ if $a \geq 3p + 2$.*

We refer the reader to [12] where a collection of some unpublished work, due to D.A. Klarner, in relation with Theorem 7 can be found.

Proposition 7. *Let L, a, b, c and r be positive integers with $b|r$ and such that $r = x_1 a + x_2 c$ for some integers $x_1, x_2 \geq 0$ and $Lc = y_1 a + y_2 b$ for some integers $y_1, y_2 \geq 0$. Then, $R(r + ac, r + ac)$ and $R(Lc + kab, Lc + kab)$ can be tiled with $R(a, a)$, $R(b, b)$ and $R(c, c)$ for any integer $k \geq 1$.*

Proof. Suppose that $b|r$. By Theorem 1, we have that

- $R(r, r)$ can be tiled with $R(b, b)$
- $R(ac, ac)$ can be tiled with $R(a, a)$,
- $R(ac, r)$ can be tiled with $R(a, a)$ and $R(c, c)$,
- $R(Lc, Lc)$ can be tiled with $R(c, c)$,
- $R(Lc, kab)$ can be tiled with $R(a, a)$ and $R(b, b)$ and
- $R(kab, kab)$ can be tiled with $R(a, a)$ (or with $R(b, b)$).

The results follow by sticking together copies of the tilings of the above rectangles as shown in Fig. 9. □

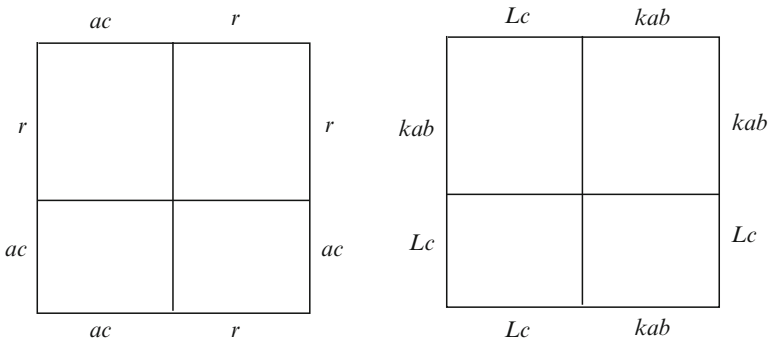


Figure 9 Compositions of tilings

Proof of Theorem 7. By Theorem 1, $R(f, f)$ can be tiled with $R(2,2)$ and $R(3,3)$ if $f \equiv 0 \bmod 2$ or $f \equiv 0 \bmod 3$. So, we only need to show that $R(f, f)$ can be tiled with $R(2,2)$, $R(3,3)$ and $R(p, p)$ if $f \geq 3p+2$ when f is odd and $f \equiv 1$ or $2 \bmod 3$. Since $3 \nmid p$ then $p \equiv i \bmod 3$ with $i = 1$ or 2.

Let $s = p - i + 3t \geq p+1$ for any integer $t \geq 1$. Since $s > g(2, p) = p-2$, then there exist nonnegative integers u and v such that $s = 2u + pv$. So, by Proposition 7 (with $a = 2, b = 3, r = s$ and $c = p$), we have that $R(s+2p, s+2p) = R(3(p+t) - i, 3(p+t) - i)$ can be tiled with $R(2,2)$, $R(3,3)$ and $R(p, p)$ for any integer $t \geq 1$. Or equivalently, $R(f, f)$ can be tiled with $R(2,2)$, $R(3,3)$ and $R(p, p)$ for any integer $f \geq 3p+1$ with $f \equiv -i \bmod 3$.

Also, since $p = 3t + i$ with $i = 1$ or 2 for some integer $t \geq 1$ then for $p > 3$ we have that $p = (t-1)3 + 2(2)$ and so, by Proposition 7 (with $a = 2, b = 3, r = s$, $c = p$ and $L = 1$), we have that, $R(p+6k, p+6k) = R(3(t+2k) + i, 3(t + 2k) + i)$ can be tiled with $R(2,2)$, $R(3,3)$ and $R(p, p)$ for any integer $k \geq 1$. Or equivalently, $R(f, f)$ can be tiled with $R(2,2)$, $R(3,3)$ and $R(p, p)$ for any odd integer $f \geq p+6$ with $f \equiv i \bmod 3$. □

Corollary 2. *$R(a, a)$ can be tiled with (a) $R(2,2)$, $R(3,3)$, and $R(5,5)$ if and only if $a \neq 1, 7$ and with (b) $R(2,2)$, $R(3,3)$, and $R(7,7)$ if and only if $a \neq 1, 5, 11$.*

Proof. (a) It is clear that $R(1,1)$ and $R(7,7)$ cannot be tiled with $R(2,2)$, $R(3,3)$, and $R(5,5)$. By Theorem 7, we have that $R(a, a)$ can be tiled with $R(2,2)$, $R(3,3)$, and $R(5,5)$ if $a \geq 3p+2 = 17$ and, by Theorem 1, $R(a, a)$ can be tiled with $R(2,2)$ and $R(3,3)$ if $a \equiv 0 \bmod 2$ or $a \equiv 0 \bmod 3$. These leave us with the cases when $a = 5, 11$ and 13. The case $a = 5$ is trivial. $R(11, 11)$ can be tiled with $R(2,2)$, $R(3,3)$, and $R(5,5)$ since, by Theorem 7, the result is true for any odd integer $a \geq p+6 = 11$ and $a \equiv 2 \bmod 3$. Finally, $R(13, 13)$ can be tiled as it is illustrated in Fig. 10.

(b) It is clear that $R(1,1)$, $R(5,5)$, and $R(11,11)$ cannot be tiled with $R(2,2)$, $R(3,3)$, and $R(7,7)$. By Theorem 7, we have that $R(a, a)$ can be tiled with $R(2,2)$, $R(3,3)$, and $R(7,7)$ if $a \geq 3p+2 = 23$ and, by Theorem 1, $R(a, a)$ can be tiled with $R(2,2)$ and $R(3,3)$ if $a \equiv 0 \bmod 2$ or $a \equiv 0 \bmod 3$. These leave us with the cases when $a = 7, 13, 17$ and 19. The case $a = 7$ is trivial.

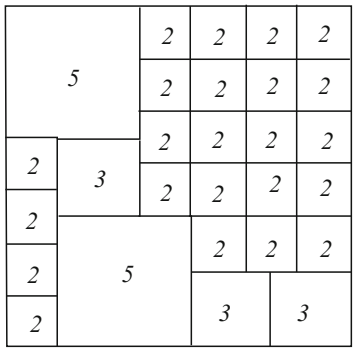


Figure 10 Tiling $R(13, 13)$ with $R(2,2)$, $R(3,3)$, and $R(5,5)$

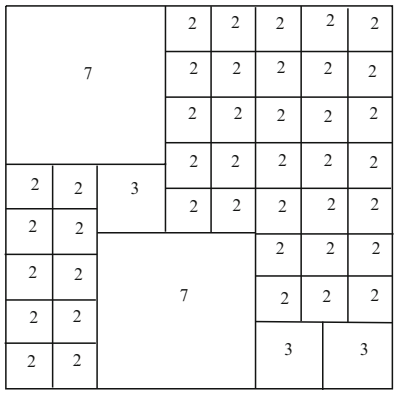


Figure 11 Tiling $R(17, 17)$ with $R(2, 2)$, $R(3, 3)$, and $R(7, 7)$

$R(13, 13)$ and $R(19, 19)$ both can be tiled since, by Theorem 7, the result is true for any odd integer $a \geq p + 6 = 13$ with $a \equiv 1 \bmod 3$. Finally, $R(17, 17)$ can be tiled as it is illustrated in Fig. 11. □

References

1. F.W. Barnes, Algebraic theory of bricks packing I, *Discrete Math.* **42** (1982), 7–26.
2. F.W. Barnes, Algebraic theory of bricks packing II, *Discrete Math.* **42** (1982), 129–144.
3. R.J. Bower and T.S. Michael, Packing boxes with bricks, *Mathematics Magazine* **79**(1) (2006), 14–30.
4. A. Clivio, Tilings of a torus with rectangular boxes, *Discrete Math.* **91** (1991), 121–139.
5. N.G. de Bruijn, Filling boxes with bricks, *Am. Math. Monthly* **76** (1969), 37–40.
6. J. Fricke, Quadratzerlegung eines Rechtecks, *Math. Semesterber.* **42** (1995), 53–62.
7. G. Katona and D. Szász, Matching problems, *J. Comb. Th.* Ser B **10** (1971), 60–92.
8. D.A. Klarner, Packing a rectangle with congruent n-ominoes, *J. Comb. Theory* **7** (1969), 107–115.
9. L.F. Klosinski, G.L. Alexanderson and L.C. Larson, The Fifty-Second William Lowell Putman mathematical competition, *Am. Math. Month.* **9** (1992), 715–724.
10. D.A. Narayan and A.J. Schwenk, Tiling large rectangles *Math. Mag.* **75**(5) (2002), 372–380.
11. M.B. Nathanson, Unusual Applications of Number Theory, *DIMACS, Series in Discrete Mathematics and Theoretical Computer Sciences* **64**, American Mathematical Society, Providence, RI (2004).
12. T. Plambeck (March, 2001). Web site http://www.plambeck.org/oldhtml/mathematics/klarner/
13. J.L. Ramírez Alfonsín, The Diophantine Frobenius Problem, *Oxford Lectures Series in Mathematics and its Applications* **30**, Oxford University Press, Oxford (2005).
14. E. Remila, On the tiling of a torus with two bars, *Theoretical Computer Sciences* **134** (1994), 415–426.
15. J.J. Sylvester, Problem 7382, *Educational Times* **37** (1884), 26; reprinted in: Mathematical questions with their solution, *Educational Times* (with additional papers and solutions) **41** (1884), 21.
16. A. Tripathi, On a linear diophantine problem of Frobenius, *INTEGERS* **6** (2006), # A14.

Sumsets and the Convex Hull

Máté Matolcsi and Imre Z. Ruzsa

Summary We extend Freiman's inequality on the cardinality of the sumset of a d-dimensional set. We consider different sets related by an inclusion of their convex hull, and one of them added possibly several times.

Mathematics Subject Classifications (2000). 11B50, 11B75, 11P70

Keywords Multidimensional sumsets

1 Introduction

The aim of this paper is to give a lower estimate for the cardinality of certain sumsets in $\mathbb{R}^d$.

We say that a set in $\mathbb{R}^d$ is *proper d-dimensional* if it is not contained in any affine hyperplane.

Our starting point is the following classical theorem of Freiman.

Theorem 1.1 (Freiman [1], Lemma 1.14). *Let $A \subset \mathbb{R}^d$ be a finite set, $|A| = m$. Assume that A is proper d-dimensional. Then,*

$$|A + A| \geq m(d + 1) - \frac{d(d + 1)}{2}.$$

Supported by Hungarian National Foundation for Scientific Research (OTKA), Grants No. PF-64061, T-049301, T-047276.
Supported by Hungarian National Foundation for Scientific Research (OTKA), Grants No. K 61908, K 72731.

M. Matolcsi
Alfréd Rényi Institute of Mathematics, Budapest, Pf. 127, H-1364 Hungary
(also at BME Department of Analysis, Budapest, H-1111, Egry J. u. 1)
e-mail: matomate@renyi.hu

I.Z. Ruzsa
Alfréd Rényi Institute of Mathematics, Budapest, Pf. 127, H-1364 Hungary
e-mail: ruzsa@renyi.hu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_16,

We will show that to get this inequality it is sufficient to use the vertices (extremal points) of A.

Definition 1.2. *We say that a point $a \in A$ is a* vertex *of a set $A \subset \mathbb{R}^d$ if it is not in the convex hull of $A/\{a\}$. The set of vertices will be denoted by* vert A.

The convex hull of a set A will be denoted by conv A.

Theorem 1.3. *Let $A \subset \mathbb{R}^d$ be a finite set, $|A| = m$. Assume that A is proper d-dimensional, and let $A' =$ vert A, We have*

$$|A + A'| \geq m(d+1) - \frac{d(d+1)}{2}.$$

This can be extended to different summands as follows.

Theorem 1.4. *Let $A, B \subset \mathbb{R}^d$ be finite sets, $|A| = m$. Assume that B is proper d-dimensional and $A \subset$ conv B. We have*

$$|A + B| \geq m(d+1) - \frac{d(d+1)}{2}.$$

Finally, we extend it to several summands as follows. We use $kB = B + \cdots + B$ to denote repeated addition. As far as we know, even the case of $A = B$ seems to be new here.

Theorem 1.5. *Let $A, B \subset \mathbb{R}^d$ be finite sets, $|A| = m$. Assume that B is proper d-dimensional and $A \subset$ conv B. Let k be a positive integer. We have*

$$|A + kB| \geq m\binom{d+k}{k} - k\binom{d+k}{k+1} = \left(m - \frac{kd}{k+1}\right)\binom{d+k}{k}. \tag{1.1}$$

The case $d = 1$ of the above theorems is quite obvious. In [2], we gave a less obvious result which compares a complete sum and its subsums, which sounds as follows.

Theorem 1.6. *Let $A_1, \ldots, A_k$ be finite, nonempty sets of integers. Let A_i' be the set consisting of the smallest and the largest elements of A_i (so that $1 \leq |A_i'| \leq 2$). Put*

$$\begin{aligned} S &= A_1 + \cdots + A_k, \\ S_i &= A_1 + \cdots + A_{i-1} + A_{i+1} + \cdots + A_k, \\ S_i' &= A_1 + \cdots + A_{i-1} + A_i' + A_{i+1} + \cdots + A_k, \\ S' &= \bigcup_{i=1}^{k} S_i'. \end{aligned}$$

We have

$$|S| \geq |S'| \geq \frac{1}{k-1}\sum_{i=1}^{k} |S_i| - \frac{1}{k-1}. \tag{1.2}$$

Problem 1.7. Generalize Theorem 1.6 to multidimensional sets. A proper generalization should give the correct order of magnitude, hence the analog of (1.2) could be of the form

$$|S| \geq |S'| \geq \left(\frac{k^{d-1}}{(k-1)^d} - \varepsilon \right) \sum_{i=1}^{k} |S_i|$$

if all sets are sufficiently large.

Problem 1.8. Let $A, B_1, \ldots, B_k \subset \mathbb{R}^d$ such that the B_i are proper d-dimensional and

$$A \subset \operatorname{conv} B_1 \subset \operatorname{conv} B_2 \subset \cdots \subset \operatorname{conv} B_k .$$

Does the estimate given in (1.1) also hold for $A + B_1 + \cdots + B_k$?

This is easy for $d = 1$.

2 A Simplicial Decomposition

We will need a result about simplicial decomposition.

By a *simplex* in $\mathbb{R}^d$, we mean a proper d-dimensional compact set, which is the convex hull of $d + 1$ points.

Definition 2.1. *Let $S_1, S_2 \subset \mathbb{R}^d$ be simplices, $B_i = \operatorname{vert} S_i$. We say that they are in* regular position*, if*

$$S_1 \cap S_2 = \operatorname{conv}(B_1 \cap B_2),$$

that is, they meet in a common k-dimensional face for some $k \leq d$. (This does not exclude the extremal cases when they are disjoint or they coincide.) We say that a collection of simplices is in regular position if any two of them are.

Lemma 2.2. *Let $B \subset \mathbb{R}^d$ be a proper d dimensional finite set, $S = \operatorname{conv} B$. There is a sequence $S_1, S_2, \ldots, S_n$ of distinct simplices in regular position with the following properties.*

(a) $S = \bigcup S_i$.
(b) $B_i = \operatorname{vert} S_i = S_i \cap B$.
(c) Each S_i, $2 \leq i \leq n$ meets at least one of $S_1, \ldots, S_{i-1}$ in a $(d-1)$ dimensional face.

We mentioned this lemma to several geometers and all answered "of course" and offered a proof immediately, but none could name a reference with this formulation, so we include a proof for completeness. This proof was communicated to us by prof. Károly Böröczki.

Proof. We use induction on $|B|$. The case $|B| = 2$ is clear. Let $|B| = k$, and assume we know it for smaller sets (in any possible dimension).

Let b be a vertex of B and apply it for the set $B' = B/\{b\}$. This set may be d or $d-1$ dimensional.

First case: B' is d dimensional. With the natural notation let

$$S' = \bigcup_{i=1}^{n'} S'_i$$

be the prescribed decomposition of $S' = \operatorname{conv} B'$. We start the decomposition of S with these, and add some more as follows.

We say that a point x of S' is *visible* from b, if x is the only point of the segment joining x and b in S'. Some of the simplices S'_i have (one or more) $d-1$ dimensional faces that are completely visible from b. Now if F is such a face, then we add the simplex

$$\operatorname{conv}(F \cup \{b\})$$

to our list.

Second case: B' is $d-1$ dimensional. Again we start with the decomposition of S', just in this case the sets S'_i will be $d-1$ dimensional simplices. Now the decomposition of S will simply consist of

$$S_i = \operatorname{conv}(S'_i \cup \{b\}),\ n = n'.$$

□

The construction above immediately gave property (c). We note that it is not really an extra requirement, every decomposition has it after a suitable rearrangement. This just means that the graph obtained by using our simplices as vertices and connecting two of them if they share a $d-1$ dimensional face is connected. Now take two simplices, say S_i and S_j. Take an inner point in each and connect them by a segment. For a generic choice of these points, this segment will not meet any of the $\le d-2$ dimensional faces of any S_k. Now as we walk along this segment and go from one simplex into another, this gives a path in our graph between the vertices corresponding to S_i and S_j.

3 The Case of a Simplex

Here we prove Theorem 1.5 for the case $|B| = d+1$.

Lemma 3.1. *Let $A, B \subset \mathbb{R}^d$ be finite sets, $|A| = m$, $|B| = d+1$. Assume that B is proper d-dimensional and $A \subset \operatorname{conv} B$. Let k be a positive integer. Write $|A \cap B| = m_1$. We have*

$$|A + kB| = (m - m_1)\binom{d+k}{k} + \binom{d+k+1}{k+1} - \binom{d-m_1+k+1}{k+1}. \quad (3.3)$$

In particular, if $|A \cap B| \leq 1$, *then*

$$|A + kB| = m\binom{d+k}{k}. \tag{3.4}$$

We have always

$$|A + kB| \geq m\binom{d+k}{k} - k\binom{d+k}{k+1} = \left(m - \frac{kd}{k+1}\right)\binom{d+k}{k}. \tag{3.5}$$

Proof. Put $A_1 = A \cap B$, $A_2 = A/B$. Write $B = \{b_0, \ldots, b_d\}$, arranged in such a way that

$$A_1 = A \cap B = \{b_0, \ldots, b_{m_1-1}\}.$$

The elements of kB are the points of the form

$$s = \sum_{i=0}^{d} x_i b_i,\ x_i \in \mathbb{Z}\ , x_i \geq 0,\ \sum x_i = k,$$

and this representation is unique. Clearly

$$|kB| = \binom{d+k}{k}.$$

Each element of A has a unique representation of the form

$$a = \sum_{i=0}^{k} \alpha_i d_i,\ \alpha_i \in \mathbb{R}\ , \alpha_i \geq 0,\ \sum \alpha_i = 1,$$

$$a = \sum_{i=0}^{d} \alpha_i b_i,\ \alpha_i \in \mathbb{R}\ , \alpha_i \geq 0,\ \sum \alpha_i = 1,$$

and if $a \in A_1$, then some $\alpha_i = 1$ and the others are equal to 0, while if $a \in A_2$, then at least two α_i's are positive.

Assume now that $a + s = a' + s'$ with certain $a, a' \in A$, $s, s' \in kB$. By substituting the above representations, we obtain

$$\sum(\alpha_i + x_i)b_i = \sum(\alpha_i' + x_i')b_i,\ \sum(\alpha_i + x_i) = \sum(\alpha_i' + x_i') = k + 1,$$

hence $\alpha_i + x_i = \alpha_i' + x_i'$ for all i. By looking at the integral and fractional parts, we see that this is possible only if $\alpha_i = \alpha_i'$, or one of them is 1 and the other is 0. If the second possibility never happens, then $a = a'$. If it happens, say $\alpha_i = 1, \alpha_i' = 0$ for some i, then $\alpha_j = 0$ for all $j \neq i$ and then each a_j' must also be 0 or 1, that is, $a, a' \in A_1$.

The previous discussion shows that $(A_1 + kB) \cap (A_2 + kB) = \emptyset$ and the sets $a + kB$, $a \in A_2$ are disjoint, hence

$$|A + kB| = |A_1 + kB| + |A_2 + kB|$$

and

$$|A_2 + kB| = |A_2|\,|kB| = (m - m_1)\binom{d+k}{k}. \tag{3.6}$$

Now we calculate $|A_1 + kB|$. The elements of this set are of the form

$$\sum_{i=0}^{d} x_i b_i,\ x_i \in \mathbb{Z}\ , x_i \geq 0,\ \sum x_i = k + 1,$$

with the additional requirement that there is at least one subscript i, $i \leq m_1 - 1$ with $x_i \geq 1$. Without this requirement the number would be the same as

$$|(k+1)B| = \binom{d+k+1}{k+1}.$$

The vectors $(x_0, \ldots, x_d)$ that violate this requirement are those that use only the last $d - m_1$ coordinates, hence their number is

$$\binom{d - m_1 + k + 1}{k+1}.$$

We obtain that

$$|A_1 + kB| = \binom{d+k+1}{k+1} - \binom{d - m_1 + k + 1}{k+1}.$$

Adding this formula to (3.6) we get (3.3).

If $m_1 = 0$ or 1, this formula reduces to the one given in (3.4).

To show inequality (3.5), observe that this formula is a decreasing function of m_1, hence the minimal value is at $m_1 = d + 1$, which after an elementary transformation corresponds to the right side of (3.5). Naturally this is attained only if $m \geq d + 1$, and for small values of m the right side of (3.5) may even be negative. □

4 The General Case

Proof (Proof of Theorem 1.5). We apply Lemma 2.2 to our set B. This decomposition induces a decomposition of A as follows. We put

$$A_1 = A \cap S_1, A_2 = A \cap (S_2/S_1), \ldots, A_n = A \cap \left(\frac{S_n}{(S_1 \cup S_2 \cup \cdots \cup S_{n-1})}\right).$$

Clearly the sets A_i are disjoint and their union is A. Recall the notation $B_i = \operatorname{vert} S_i$.

We claim that the sets $A_i + kB_i$ are also disjoint.

Indeed, suppose that $a + s = a' + s'$ with $a \in A_i$, $a' \in A_j$, $s \in kB_i$, $s' \in kB_j$, $i < j$. We have

$$\frac{a+s}{k+1} \in S_i, \ \frac{a'+s'}{k+1} \in S_j,$$

and these points are equal, so they are in

$$S_i \cap S_j = \mathrm{conv}(B_i \cap B_j).$$

This means that in the unique convex representation of $(a'+s')/(k+1)$ by points of B_j only elements of $B_i \cap B_j$ are used. However, we can obtain this representation via using the representation of a' and the components of s', hence we must have $a' \in \mathrm{conv}(B_i \cap B_k) \subset S_i$, a contradiction.

This disjointness yields

$$|A + kB| \geq \sum |A_i + kB_i|.$$

We estimate the summands using Lemma 3.1.

If $i > 1$, then $|A_i \cap B_i| \leq 1$. Indeed, there is a $j < i$ such that S_j has a common $d-1$ dimensional face with S_i, and then the d vertices of this face are excluded from A_i by definition. So in this case, (3.4) gives

$$|A_i + kB_i| = |A_i| \binom{d+k}{k}.$$

For $i = 1$, we can only use the weaker estimate (3.5):

$$|A_1 + kB_1| \geq |A_1| \binom{d+k}{k} - k\binom{d+k}{k+1}.$$

Summing these equations we obtain (1.1). □

Acknowledgement The authors profited much from discussions with Katalin Gyarmati and Károly Böröczky.

References

1. G. Freiman, *Foundations of a structural theory of set addition*, Am. Math. Soc., 1973.
2. K. Gyarmati, M. Matolcsi, and I. Z. Ruzsa, *A superadditivity and submultiplicativity property for cardinalities of sumsets*, Combinatorica, to appear.

Explicit Constructions of Infinite Families of MSTD Sets

Steven J. Miller and Daniel Scheinerman

Summary Given a finite set of integers A, we may consider its sumset $A + A$ and its difference set $A - A$. As addition is commutative and subtraction is not, it was initially believed that as $r \to \infty$, almost all of the 2^r subsets of $\{1, \dots, r\}$ would have $|A - A| > |A + A|$; if $|A + A| > |A - A|$, we say A is an MSTD (more sums than differences) set. While Martin and O'Bryant [MO06] disproved this conjecture by showing that a small but positive percentage of such sets are MSTD, previous explicit constructions only found families of size $f(r)2^{r/2}$ for some polynomial $f(r)$.

Below we present a new construction that yields a family of MSTD sets in $\{1, \dots, r\}$ of size $C2^r/r^4$ for a fixed, non-zero constant C; thus our family is significantly denser than previous constructions. Our method has been generalized further with Brooke Orosz to handle certain ternary combinations; the details below are adapted from our paper [MOS09].

We conclude with an appendix on a special case of a result of Hegarty and Miller [HM07], which supports the intuition behind the false conjecture. Specifically, if $p(r)$ is a monotonically decreasing function tending to 0, and for each r every element in $\{1, \dots, r\}$ is in a subset A with probability $p(r)$, then as $r \to \infty$ almost no subsets (with respect to this probability) are MSTD.

Keywords More sum than difference sets

Mathematics Subject Classifications (2010). 11P99.

S.J. Miller
Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA
e-mail: Steven.J.Miller@williams.edu

D. Scheinerman
Department of Mathematics, Brown University, Providence, RI 02912, USA
e-mail: Daniel_Scheinerman@brown.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_17,

1 Introduction

Given a finite set of integers A, we define its sumset $A + A$ and difference set $A - A$ by

$$A + A = \{a_i + a_j : a_i, a_j \in A\}$$
$$A - A = \{a_i - a_j : a_i, a_j \in A\}, \tag{1}$$

and let $|X|$ denote the cardinality of X. If $|A + A| > |A - A|$, then, following Nathanson, we call A an MSTD (more sums than differences) set. As addition is commutative while subtraction is not, we expect that for a 'generic' set A we have $|A-A| > |A+A|$, as a typical pair (x, y) contributes one sum and two differences; thus we expect MSTD sets to be rare.

Martin and O'Bryant [MO06] proved that, in some sense, this intuition is wrong. They considered the uniform model[1] for choosing a subset A of $\{1, \ldots, n\}$, and showed that there is a positive probability that a random subset A is an MSTD set (though, not surprisingly, the probability is quite small). However, the answer is very different for other ways of choosing subsets randomly, and if we decrease slightly the probability that an element is chosen, then our intuition is correct. Specifically, consider the binomial model with parameter $p(n)$, with $\lim_{n\to\infty} p(n) = 0$ and $n^{-1} = o(p(n))$ (so $p(n)$ doesn't tend to zero so rapidly that the sets are too sparse).[2] Hegarty and Miller [HM07] recently proved that, in the limit as $n \to 0$, the percentage of subsets of $\{1, \ldots, n\}$ that are MSTD sets tends to zero in this model. See Appendix 2 for full statements and a self-contained proof when $p(n) = o(n^{-1/2})$.

Though MSTD sets are rare, they do exist (and, in the uniform model, are somewhat abundant by the work of Martin and O'Bryant). Examples go back to the 1960s. Conway is said to have discovered $\{0, 2, 3, 4, 7, 11, 12, 14\}$, while Marica [Ma69] gave $\{0, 1, 2, 4, 7, 8, 12, 14, 15\}$ in 1969 and Freiman and Pigarev [FP73] found $\{0, 1, 2, 4, 5, 9, 12, 13, 14, 16, 17, 21, 24, 25, 26, 28, 29\}$ in 1973. Recent work includes infinite families constructed by Hegarty [He07] and Nathanson [Na07], as well as existence proofs by Ruzsa [Ru76, Ru84, Ru92].

Most of the previous constructions[3] of infinite families of MSTD sets start with a symmetric set, which is then 'perturbed' slightly through the careful addition of a few elements that increase the number of sums more than the number of differences; see [He, Na07] for a description of some previous constructions and methods.

[1] This means each of the 2^n subsets of $\{1, \ldots, n\}$ are equally likely to be chosen, or, equivalently, that the probability any $k \in \{1, \ldots, n\}$ is in A is just $1/2$.

[2] This model means that the probability $k \in \{1, \ldots, n\}$ is in A is $p(n)$.

[3] An alternate method constructs an infinite family from a given MSTD set A by considering $A_t = \{\sum_{i=1}^{t} a_i m^{i-1} : a_i \in A\}$. For m sufficiently large, these will be MSTD sets; this is called the base expansion method. Note, however, that these will be very sparse. See [He07] for more details.

In many cases, these symmetric sets are arithmetic progressions; such sets are natural starting points because if A is an arithmetic progression, then $|A+A| = |A-A|$.[4]

In this work, we present a new method which takes an MSTD set satisfying certain conditions and constructs an infinite family of MSTD sets. While these families are not dense enough to prove that a positive percentage of subsets of $\{1, \dots, r\}$ are MSTD sets, we are able to elementarily show that the percentage is at least C/r^4 for some constant C. Thus, our families are far denser than those in [He, Na07]; trivial counting[5] shows all of their infinite families give at most $f(r)2^{r/2}$ of the subsets of $\{1, \dots, r\}$ (for some polynomial $f(r)$) are MSTD sets, implying a percentage of at most $f(r)/2^{r/2}$.

We first introduce some notation:

- We let $[a, b]$ denote all integers from a to b; thus $[a, b] = \{n \in \mathbf{Z} : a \le n \le b\}$.
- We say a set of integers A have the property P_n (or is a P_n-set) if both its sumset and its difference set contain all but the first and last n possible elements (and of course, it may or may not contain some of these fringe elements).[6] Explicitly, let $a = \min A$ and $b = \max A$. Then, A is a P_n-set if

$$[2a + n,\ 2b - n] \subset A + A \tag{2}$$

and

$$[-(b - a) + n,\ (b - a) - n] \subset A - A. \tag{3}$$

We can now state our construction and main result.

Theorem 1. *Let $A = L \cup R$ be a P_n, MSTD set where $L \subset [1, n]$, $R \subset [n + 1, 2n]$, and $1, 2n \in A$;[7] see Remark 2 for an example of such an A. Fix a $k \ge n$ and let m be arbitrary. Let M be any subset of $[n+k+1, n+k+m]$ with the property that it does*

[4] As $|A + A|$ and $|A - A|$ are not changed by mapping each $x \in A$ to $\alpha x + \beta$ for any fixed α and β, we may assume our arithmetic progression is just $\{0, \dots, n\}$, and thus the cardinality of each set is $2n + 1$.

[5] For example, consider the following construction of MSTD sets from [Na07]: let $m, d, k \in \mathbf{N}$ with $m \ge 4$, $1 \le d \le m - 1$, $d \ne m/2$, $k \ge 3$ if $d < m/2$ else $k \ge 4$. Set $B = [0, m - 1]\backslash\{d\}$, $L = \{m - d, 2m - d, \dots, km - d\}$, $a^* = (k + 1)m - 2d$ and $A = B \cup L \cup (a^* - B) \cup \{m\}$. Then A is an MSTD set. The width of such a set is of the order km. Thus, if we look at all triples (m, d, k) with $km \le r$ satisfying the above conditions, these generate on the order of at most $\sum_{k \le r} \sum_{m \le r/k} \sum_{d \le m} 1 \ll r^2$, and there are of the order 2^r possible subsets of $\{0, \dots, r\}$; thus this construction generates a negligible number of MSTD sets. Though we write $f(r)/2^{r/2}$ to bound the percentage from other methods, a more careful analysis shows it is significantly less; we prefer this easier bound as it is already significantly less than our method. See for example Theorem 2 of [He07] for a denser example.

[6] It is not hard to show that for fixed $0 < \alpha \le 1$ a random set drawn from $[1, n]$ in the uniform model is a $P_{\lfloor \alpha n \rfloor}$-set with probability approaching 1 as $n \to \infty$.

[7] Requiring $1, 2n \in A$ is quite mild; we do this so that we know the first and last elements of A.

not have a run of more than k missing elements (i.e., for all $\ell \in [n+k+1, n+m+1]$ *there is a* $j \in [\ell, \ell+k-1]$ *such that* $j \in M$*). Assume further that* $n+k+1 \notin M$ *and set* $A(M;k) = L \cup O_1 \cup M \cup O_2 \cup R'$*, where* $O_1 = [n+1, n+k]$*,* $O_2 = [n+k+m+1, n+2k+m]$ *(thus the* O_i*'s are just sets of k consecutive integers), and* $R' = R + 2k + m$*. Then*

(1) $A(M;k)$ *is an MSTD set, and thus we obtain an infinite family of distinct MSTD sets as* M *varies;*
(2) There is a constant $C > 0$ *such that as* $r \to \infty$ *the percentage of subsets of* $\{1, \dots, r\}$ *that are in this family (and thus are MSTD sets) is at least* C/r^4*.*

Remark 1. We quickly highlight the main idea of the construction, referring to Sect. 2 for details. The idea is to take an MSTD set A and augment it to a new set A' such that the number of sums added ($|A'+A'| - |A+A|$) equals the number of differences added ($|A'-A'| - |A-A|$). This is accomplished by having the two blocks O_1, O_2 of consecutive elements and then making sure that we always take at least one out of every k elements between O_1 and O_2. Counting arguments then show that every possible new difference and new sum is included.

Remark 2. In order to show that our theorem is not trivial, we must of course exhibit at least one P_n, MSTD set A satisfying all our requirements (else our family is empty!). We may take the set[8] $A = \{1, 2, 3, 5, 8, 9, 13, 15, 16\}$; it is an MSTD set as

$$\begin{aligned} A+A = \{&2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, \\ &22, 23, 24, 25, 26, 28, 29, 30, 31, 32\} \\ A-A = \{&-15, -14, -13, -12, -11, -10, -8, -7, -6, -5, -4, -3, -2, -1, \\ &0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15\} \end{aligned} \tag{4}$$

(so $|A+A| = 30 > 29 = |A-A|$). A is also a P_n-set, as (2) is satisfied since $[10, 24] \subset A+A$ and (3) is satisfied since $[-7, 7] \subset A-A$.

For the uniform model, a subset of $[1, 2n]$ is a P_n-set with high probability as $n \to \infty$, and thus examples of this nature are plentiful. For example, of the 1,748 MSTD sets with minimum 1 and maximum 24,1008 are P_n-sets.

Unlike other estimates on the percentage of MSTD sets, our arguments are not probabilistic, and rely on explicitly constructing large families of MSTD sets. Our arguments share some similarities with the methods in [He07] (see for example, Case I of Theorem 8) and [MO06]. There the fringe elements of the set were also chosen first. A random set was then added in the middle, and the authors argued that with high probability the resulting set is an MSTD set. We can almost add a random set in the middle; the reason we do not obtain a positive percentage is that we have

[8] This A is trivially modified from [Ma69] by adding 1 to each element, as we start our sets with 1 while other authors start with 0. We chose this set as our example as it has several additional nice properties that were needed in earlier versions of our construction which required us to assume slightly more about A.

the restriction that there can be no consecutive block of size k of numbers in the middle that are not chosen to be in $A(M;k)$. This is easily satisfied by requiring us to choose at least one number in consecutive blocks of size $k/2$, and this is what leads to the loss of a positive percentage (though we do obtain sets that are known to be MSTD sets, and not just highly likely to be MSTD sets).

The paper is organized as follows. We describe our construction in Sect. 2, and prove our claimed lower bounds for the percentage of sets that are MSTD sets in Sect. 3. We end with some concluding remarks and suggestions for future research in Sect. 4.

On a personal note, the first named author would like to thank Mel for introducing him to much of additive number theory, ranging from his accessible books to numerous conversations over the years. This paper (as well as the paper by Hegarty and Miller [HM07]) is an outgrowth of a talk Mel gave at Brown in 2007 on MSTD sets as well as conversations at CANT 2007, and it is a pleasure to thank him for an introduction to such a fascinating subject.

2 Construction of Infinite Families of MSTD Sets

Let $A \subset [1,2n]$. We can write this set as $A = L \cup R$ where $L \subset [1,n]$ and $R \subset [n+1,2n]$. We have

$$A + A = [L + L] \cup [L + R] \cup [R + R] \tag{5}$$

where $L + L \subset [2,2n]$, $L + R \subset [n+2,3n]$, and $R + R \subset [2n+2,4n]$, and

$$A - A = [L - R] \cup [L - L] \cup [R - R] \cup [R - L] \tag{6}$$

where $L-R \subset [-1,-2n+1]$, $L-L \subset [-(n-1),n-1]$, $R-R \subset [-(n-1),n-1]$ and $R - L \subset [1,2n-1]$.

A typical subset A of $\{1,\dots,2n\}$ (chosen from the uniform model, see Footnote 1) will be a P_n-set (see Footnote 6). It is thus the interaction of the "fringe" elements that largely determines whether a given set is an MSTD set. Our construction begins with a set A that is both an MSTD set and a P_n-set. We construct a family of P_n, MSTD sets by inserting elements into the middle in such a way that the new set is a P_n-set, and the number of added sums is equal to the number of added differences. Thus, the new set is also an MSTD set.

In creating MSTD sets, it is very useful to know that we have a P_n-set. The reason is that we have all but the "fringe" possible sums and differences, and are thus reduced to studying the extreme sums and differences. The following lemma shows that if A is a P_n, MSTD set and a certain extension of A is a P_n-set, then this extension is also an MSTD set. The difficult step in our construction is determining a large class of extensions which lead to P_n-sets; we will do this in Lemma 2.

Lemma 1. *Let $A = L \cup R$ be a P_n-set where $L \subset [1, n]$ and $R \subset [n+1, 2n]$. Form $A' = L \cup M \cup R'$ where $M \subset [n+1, n+m]$ and $R' = R + m$. If A' is a P_n-set, then $|A' + A'| - |A + A| = |A' - A'| - |A - A| = 2m$ (i.e., the number of added sums is equal to the number of added differences). In particular, if A is an MSTD set, then so is A'.*

Proof. We first count the number of added sums. In the interval $[2, n+1]$, both $A + A$ and $A' + A'$ are identical, as any sum can come only from terms in $L + L$. Similarly, we can pair the sums of $A + A$ in the region $[3n+1, 4n]$ with the sums of $A' + A'$ in the region $[3n + 2m + 1, 4n + 2m]$, as these can come only from $R + R$ and $(R + m) + (R + m)$, respectively. Since we have accounted for the n smallest and largest terms in both $A + A$ and $A' + A'$, and as both are P_n-sets, the number of added sums is just $(3n + 2m + 1) - (3n + 1) = 2m$.

Similarly, differences in the interval $[1 - 2n, -n]$ that come from $L - R$ can be paired with the corresponding terms from $L-(R+m)$, and differences in the interval $[n, 2n-1]$ from $R - L$ can be paired with differences coming from $(R + m) - L$. Thus the size of the middle grows from the interval $[-n+1, n-1]$ to the interval $[-n-m+1, n+m-1]$. Thus we have added $(2n + 2m + 3) - (2n + 3) = 2m$ sums. Thus, $|A' + A'| - |A + A| = |A' - A'| - |A - A| = 2m$ as desired. □

The above lemma is not surprising, as in it we assume A' is a P_n-set; the difficulty in our construction is showing that our new set $A(M; k)$ is also a P_n-set for suitably chosen M. This requirement forces us to introduce the sets O_i (which are blocks of k consecutive integers), as well as requiring M to have at least one of every k consecutive integers.

We are now ready to prove the first part of Theorem 1 by constructing an infinite family of distinct P_n, MSTD sets. We take a P_n, MSTD set and insert a set in such a way that it remains a P_n-set; thus by Lemma 1, we see that this new set is an MSTD set.

Lemma 2. *Let $A = L \cup R$ be a P_n-set where $L \subset [1, n]$, $R \subset [n+1, 2n]$, and $1, 2n \in A$. Fix a $k \geq n$ and let m be arbitrary. Choose any $M \subset [n+k+1, n+k+m]$ with the property that M does not have a run of more than k missing elements, and form $A(M; k) = L \cup O_1 \cup M \cup O_2 \cup R'$ where $O_1 = [n+1, n+k]$, $O_2 = [n+k+m+1, n+2k+m]$, and $R' = R+2k+m$. Then $A(M; k)$ is a P_n-set.*

Proof. For notational convenience, denote $A(M; k)$ by A'. Note $A' + A' \subset [2, 4n + 4k + 2m]$. We begin by showing that there are no missing sums from $n + 2$ to $3n + 4k + 2m$; proving an analogous statement for $A' - A'$ shows A' is a P_n-set. By symmetry[9], we only have to show that there are no missing sums in $[n+2, 2n + 2k + m]$. We consider various ranges in turn.

We observe that $[n + 2, n + k + 1] \subset A' + A'$ because we have $1 \in L$ and these sums result from $1 + O_1$. Additionally, $O_1 + O_1 = [2n + 2, 2n + 2k] \subset A' + A'$. Since $n \leq k$ we have $n + k + 1 \geq 2n + 1$, these two regions are contiguous and overlap and thus $[n + 2, 2n + 2k] \subset A' + A'$.

[9] Apply the arguments below to the set $2n + 2k + m - A'$, noting that $1, 2n + 2k + m \in A'$.

Now consider $O_1 + M$. Since M does not have a run of more than k missing elements, the worst case scenario for us for elements in the sumset is that the smallest element of M is $n+2k$ and that the largest element is $n+m+1$ (and, of course, we still have at least one out of every k consecutive integers is in M). If this is the case, then we still have $O_1 + M \supset [(n+1)+(n+2k), (n+k)+(n+m+1)] = [2n+2k+1, 2n+k+m+1]$. We had already shown that $A' + A'$ has all sums up to $2n+2k$; this extends the sumset to all sums up to $2n+k+m+1$.

All that remains is to show we have all sums in $[2n+k+m+2, 2n+2k+m]$. This follows immediately from $O_1+O_2 = [2n+k+m+2, 2n+3k+m] \subset A'+A'$. This extends our sumset to include all sums up to $2n+3k+m$, which is well past our halfway mark of $2n+2k+m$; the remaining sums follow from a similar argument. Thus, we have shown that $A' + A' \supset [n+2, 3n+4k+2m+1]$.

We now do a similar calculation for the difference set, which is contained in $[-(2n+2k+m)+1, (2n+2k+m)-1]$. As we have already analyzed the sumset, all that remains to prove A is a P_n-set is to show that $A'-A' \supset [-n-2k-m+1, n+2k+m-1]$. As all difference sets are symmetric about and contain 0, it suffices to show the positive elements are present, i.e., that $A' - A' \supset [1, n+2k+m-1]$.

We easily see $[1, k-1] \subset A' - A'$ as $[0, k-1] \subset O_1 - O_1$. Now consider $M - O_1$. Again the worst case scenario for us is that the least element of M is $n+2k$ and the greatest is $n+m+1$. With this in mind, we see that $M - O_1 \supset [(n+2k)-(n+k), (n+m+1)-(n+1)] = [k, m]$. Now $O_2 - O_1 \supset [(n+k+m+1)-(n+k), (n+2k+m)-(n+1)] = [m+1, 2k+m-1]$, and we therefore have all differences up to $2k+m-1$.

Since $2n \in A$ we have $2n+2k+m \in A'$. Consider $(2n+2k+m) - O_1 = [n+k+m, n+2k+m-1]$. Since $k \geq n$ we see that $n+k+m \leq 2k+m$; this implies that we have all differences up to $n+2k+m-1$ (this is because we already have all differences up to $2k+m-1$, and $n+k+m$ is either less than $2k+m-1$, or at most one larger). □

Proof of Theorem 1(1). The proof of the first part of Theorem 1 follows immediately. By Lemma 2, our new sets $A(M;k)$ are P_n-sets, and by Lemma 1, they are also MSTD. All that remains is to show that the sets are distinct; this is done by requiring $n+k+1$ is not in our set (for a fixed k, these sets have elements $n+1, \ldots, n+k$ but not $n+k+1$; thus different k yield distinct sets).

3 Lower Bounds for the Percentage of MSTDs

To finish the proof of Theorem 1, for a fixed n, we need to count how many sets $\widetilde{M}$ of the form $O_1 \cup M \cup O_2$ (see Theorem 1 for a description of these sets) of width $r = 2k+m$ can be inserted into a P_n, MSTD set A of width $2n$. As O_1 and O_2 are just intervals of k consecutive ones, the flexibility in choosing them comes solely from the freedom to choose their length k (so long as $k \geq n$). There is far more freedom to choose M.

There are two issues we must address. First, we must determine how many ways there are to fill the elements of M such that there are no runs of k missing elements. Second, we must show that the sets generated by this method are distinct. We saw in the proof of Theorem 1(1) that the latter is easily handled by giving $A(M;k)$ (through our choice of M) slightly more structure. Assume that the element $n+k+1$ is *not* in M (and thus not in A). Then for a fixed width $r = 2k + m$, each value of k gives rise to necessarily distinct sets, since the set contains $[n + 1, n + k]$ but not $n+k+1$. In our arguments below, we assume our initial P_n, MSTD set A is fixed; we could easily increase the number of generated MSTD sets by varying A over certain MSTD sets of size $2n$. We choose not to do this as n is fixed, and thus varying over such A will only change the percentages by a constant independent of k and m.

Fix n and let r tend to infinity. We count how many $\widetilde{M}$'s there are of width r such that in M, there is at least one element chosen in any consecutive block of k integers. One way to ensure this is to divide M into consecutive, non-overlapping blocks of size $k/2$, and choose at least one element in each block. There are $2^{k/2}$ subsets of a block of size $k/2$, and all but one have at least one element. Thus, there are $2^{k/2} - 1 = 2^{k/2}(1 - 2^{-k/2})$ valid choices for each block of size $k/2$. As the width of M is $r - 2k$, there are $\lceil \frac{r-2k}{k/2} \rceil \le \frac{r}{k/2} - 3$ blocks (the last block may have length less than $k/2$, in which case any configuration will suffice to ensure there is not a consecutive string of k omitted elements in M because there will be at least one element chosen in the previous block). We see that the number of valid M's of width $r - 2k$ is at least $2^{r-2k}\left(1 - 2^{-k/2}\right)^{\frac{r}{k/2}-3}$. As O_1 and O_2 are two sets of k consecutive 1's, there is only one way to choose either.

We therefore see that, for a fixed k, of the $2^r = 2^{m+2k}$ possible subsets of r consecutive integers, we have at least $2^{r-2k}\left(1 - 2^{-k/2}\right)^{\frac{r}{k/2}-3}$ are permissible to insert into A. To ensure that all of the sets are distinct, we require $n + k + 1 \notin M$; the effect of this is to eliminate one degree of freedom in choosing an element in the first block of M, and this will only change the proportionality constants in the percentage calculation (and *not* the r or k dependencies). Thus, if we vary k from n to $r/4$ (we could go a little higher, but once k is as large as a constant times r the number of generated sets of width r is negligible), we have at least some fixed constant times

$$2^r \sum_{k=n}^{r/4} \frac{1}{2^{2k}} \left(1 - 2^{-k/2}\right)^{\frac{r}{k/2}-3}$$

MSTD sets; equivalently, the percentage of sets $O_1 \cup M \cup O_2$ with O_i of width $k \in \{n, \ldots, r/4\}$ and M of width $r - 2k$ that we may add is at least this divided by 2^r, or some universal constant times

$$\sum_{k=n}^{r/4} \frac{1}{2^{2k}} \left(1 - \frac{1}{2^{k/2}}\right)^{\frac{r}{k/2}} \tag{7}$$

(as $k \ge n$ and n is fixed, we may remove the -3 in the exponent by changing the universal constant).

We now determine the asymptotic behavior of this sum. More generally, we can consider sums of the form

$$S(a,b,c;r) = \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{8}$$

For our purposes, we take $a = 2$ and $b = c = 1/2$; we consider this more general sum so that any improvements in our method can readily be translated into improvements in counting MSTD sets. While we know (from the work of Martin and O'Bryant [MO06]) that a positive percentage of such subsets are MSTD sets, our analysis of this sum yields slightly weaker results. The approach in [MO06] is probabilistic, obtained by fixing the fringes of our subsets to ensure certain sums and differences are in (or not in) the sum- and difference sets. While our approach also fixes the fringes, we have far more possible fringe choices than in [MO06] (though we do not exploit this). While we cannot prove a positive percentage of subsets are MSTD sets, our arguments are far more elementary.

The proof of Theorem 1(2) is clearly reduced to proving the following lemma, and then setting $a = 2$ and $b = c = 1/2$.

Lemma 3. *Let*

$$S(a,b,c;r) = \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{9}$$

Then for any $\varepsilon > 0$, we have

$$\frac{1}{r^{a/b}} \ll S(a,b,c;r) \ll \frac{(\log r)^{2a+\epsilon}}{r^{a/b}}. \tag{10}$$

Proof. We constantly use $(1 - 1/x)^x$ is an increasing function in x. We first prove the lower bound. For $k \geq (\log_2 r)/b$ and r large, we have

$$\left(1 - \frac{1}{2^{bk}}\right)^{r/ck} = \left(1 - \frac{1}{2^{bk}}\right)^{2^{bk}\frac{r}{ck2^{bk}}} \geq \left(1 - \frac{1}{r}\right)^{r\cdot\frac{b}{c\log_2 r}} \geq \frac{1}{2} \tag{11}$$

(in fact, for r large the last bound is almost exactly 1). Thus, we trivially have

$$S(a,b,c;r) \geq \sum_{k=(\log_2 r)/b}^{r/4} \frac{1}{2^{ak}} \cdot \frac{1}{2} \gg \frac{1}{r^{a/b}}. \tag{12}$$

For the upper bound, we divide the k-sum into two ranges: (1) $bn \le bk \le \log_2 r - \log_2(\log r)^\delta$; (2) $\log_2 r - \log_2(\log r)^\delta \le bk \le br/4$. In the first range, we have

$$\begin{aligned}\left(1-\frac{1}{2^{bk}}\right)^{r/ck} &\le \left(1-\frac{(\log r)^\delta}{r}\right)^{r/ck} \\ &\ll \exp\left(-\frac{b(\log r)^\delta}{c\log_2 r}\right) \\ &\le \exp\left(-\frac{b\log 2}{c}\cdot(\log r)^{\delta-1}\right). \end{aligned} \tag{13}$$

If $\delta > 2$, then this factor is dominated by $r^{-\frac{b\log 2}{c}\cdot(\log r)^{\delta-2}} \ll r^{-A}$ for any A for r sufficiently large. Thus there is negligible contribution from k in range (1) if we take $\delta = 2 + \epsilon/a$ for any $\epsilon > 0$.

For k in the second range, we trivially bound the factors $\left(1-1/2^{bk}\right)^{r/ck}$ by 1. We are left with

$$\sum_{k\ge \frac{\log_2 r}{b} - \frac{\log_2(\log r)^\delta}{b}} \frac{1}{2^{ak}}\cdot 1 \le \frac{(\log r)^{a\delta}}{r^{a/b}}\sum_{\ell=0}^{\infty}\frac{1}{2^{a\ell}} \ll \frac{(\log r)^{a\delta}}{r^{a/b}}. \tag{14}$$

Combining the bounds for the two ranges with $\delta = 2 + \epsilon/a$ completes the proof. □

Remark 3. The upper and lower bounds in Lemma 3 are quite close, differing by a few powers of $\log r$. The true value will be at least $(\log r/r)^{a/b}$; we sketch the proof in Appendix 1.

Remark 4. We could attempt to increase our lower bound for the percentage of subsets that are MSTD sets by summing r from R_0 to R (as we have fixed r above, we are only counting MSTD sets of width $2n + r$ where 1 and $2n + r$ are in the set. Unfortunately, at best we can change the universal constant; our bound will still be of the order $1/R^4$. To see this, note the number of such MSTD sets is at least a constant times $\sum_{r=R_0}^{R} 2^r/r^4$ (to get the percentage, we divide this by 2^R). If $r \le R/2$, then there are exponentially few sets. If $r \ge R/2$, then $r^{-4} \in [1/R^4, 16/R^4]$. Thus, the percentage of such subsets is still only at least of order $1/R^4$.

4 Concluding Remarks and Future Research

We observed earlier (Footnote 6) that for a constant $0 < \alpha \le 1$, a set randomly chosen from $[1, 2n]$ is a $P_{\lfloor \alpha n\rfloor}$-set with probability approaching 1 as $n \to \infty$. MSTD sets are of course not random, but it seems logical to suppose that this pattern continues.

Conjecture 1. Fix a constant $0 < \alpha \leq 1/2$. Then as $n \to \infty$, the probability that a randomly chosen MSTD set in $[1, 2n]$ containing 1 and $2n$ is a $P_{\lfloor \alpha n \rfloor}$-set goes to 1.

In our construction and that of [MO06], a collection of MSTD sets is formed by fixing the fringe elements and letting the middle vary. The intuition behind both is that the fringe elements matter most and the middle elements least. Motivated by this, it is interesting to look at all MSTD sets in $[1, n]$ and ask with what frequency a given element is in these sets. That is, what is

$$\gamma(k; n) = \frac{\#\{A : k \in A \text{ and } A \text{ is an MSTD set}\}}{\#\{A : A \text{ is an MSTD set}\}} \tag{15}$$

as $n \to \infty$? We can get a sense of what these probabilities might be from Figure 1.

Note that as the graph suggests, γ is symmetric about $(n + 1)/2$, that is, $\gamma(k, n) = \gamma(n + 1 - k, n)$. This follows from the fact that the cardinalities of the sumset and difference set are unaffected by sending $x \to \alpha x + \beta$ for any α, β. Thus, for each MSTD set A we get a distinct MSTD set $n + 1 - A$ showing that our function γ is symmetric. These sets are distinct since if $A = n + 1 - A$, then A is sum-difference balanced.[10]

From [MO06] we know that a positive percentage of sets are MSTD sets. By the central limit theorem, we then get that the average size of an MSTD set chosen from

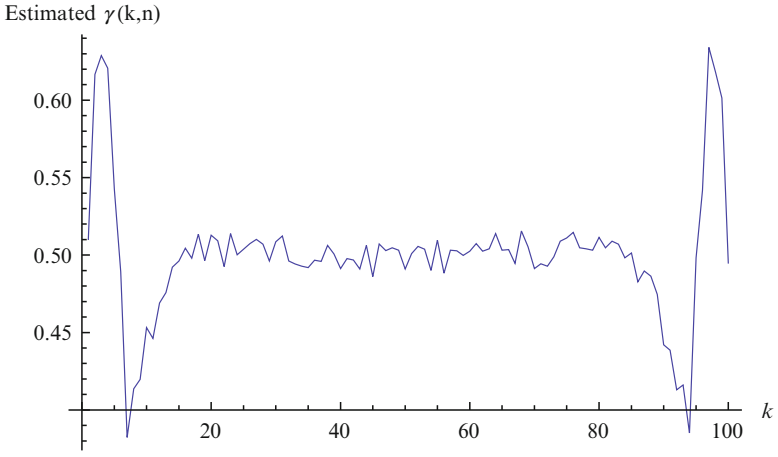


Figure 1 Estimation of $\gamma(k, 100)$ as k varies from 1 to 100 from a random sample of 4458 MSTD sets. The sample was obtained by choosing sets from the uniform model (i.e., for each $A \subset \{1, \ldots, n\}$ the probability $k \in A$ is 1/2)

[10] The following proof is standard (see, for instance, [Na07]). If $A = n + 1 - A$ then

$$|A + A| = |A + (n + 1 - A)| = |n + 1 + (A - A)| = |A - A|. \tag{16}$$

$[1, n]$ is about $n/2$. This tells us that on average $\gamma(k, n)$ is about $1/2$. The graph above suggests that the frequency goes to $1/2$ in the center. This leads us to the following conjecture:

Conjecture 2. Fix a constant $0 < \alpha < 1/2$. Then $\lim_{n\to\infty} \gamma(k, n) = 1/2$ for $\lfloor \alpha n \rfloor \le k \le n - \lfloor \alpha n \rfloor$.

Remark 5. More generally, we could ask which non-decreasing functions $f(n)$ have $f(n) \to \infty$, $n - f(n) \to \infty$ and $\lim_{n\to\infty} \gamma(k, n) = 1/2$ for all k such that $\lfloor f(n) \rfloor \le k \le n - \lfloor f(n) \rfloor$.

Appendix 1: Size of $S(a, b, c; r)$

We sketch the proof that the sum

$$S(a, b, c; r) = \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck} \tag{17}$$

is at least $(\log r/r)^{a/b}$. We determine the maximum value of the summands

$$f(a, b, c; k, r) = \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{18}$$

Clearly $f(a, b, c; k, r)$ is very small if k is small due to the second factor; similarly it is small if k is large because of the first factor. Thus, the maximum value of $f(a, b, c; k, r)$ will arise not from an endpoint but from a critical point.

It is convenient to change variables to simplify the differentiation. Let $u = 2^k$ (so $k = \log u/\log 2$). Then,

$$g(a, b, c; u, r) = f(a, b, c; k, r) = u^{-a} \left(1 - \frac{1}{u^b}\right)^{u^b \cdot \frac{m \log 2}{cu^b \log u}}. \tag{19}$$

Thus,

$$g(a, b, c; u, r) \approx u^{-a} \exp\left(-\frac{r \log 2}{cu^b \log u}\right). \tag{20}$$

Maximizing this is the same as minimizing $h(a, b, c; u, r) = 1/g(a, b, c; u, r)$. After some algebra, we find

$$h'(a, b, c; u, r) = \frac{h(a, b, c; u, r)}{cu \log^2 u} \left(acu^b \log^2 u - r \log 2 \cdot (b \log u + 1)\right). \tag{21}$$

Setting the derivative equal to zero yields

$$acu^b \log^2 u = r \log 2 \cdot (b \log u + 1)\,. \tag{22}$$

As we know u must be large, looking at just the main term from the right hand side yields

$$acu^b \log u \approx rb \log 2, \tag{23}$$

or

$$u^b \log u \approx Cr, \quad C = \frac{b \log 2}{ac}. \tag{24}$$

To first order, we see the solution is

$$u_{\max} = \left(\frac{(Cr)}{\frac{\log(Cr)}{b}}\right)^{\frac{1}{b}} \approx C' \left(\frac{r}{\log r}\right)^{\frac{1}{b}}. \tag{25}$$

Straightforward algebra shows that the maximum value of our summands is approximately

$$(C' \mathrm{e}^{1/b})^{-a} \left(\frac{\log r}{r}\right)^{a/b}.$$

Appendix 2: When Almost ALL Sets are not MSTD Sets

Peter Hegarty and Steven J. Miller

In [Na06], Nathanson remarked: *Even though there exist sets A that have more sums than differences, such sets should be rare, and it must be true with the right way of counting that the vast majority of sets satisfies $|A - A| > |A + A|$.* While we now know (thanks to the work of Martin and O'Bryant [MO06]) that a positive percentage of all subsets of $\{1, \dots, N\}$ are MSTD sets, the answer is markedly different when we consider instead a binomial model with parameter decreasing to zero as $N \to \infty$. In [HM07], it is shown that Nathanson's intuition is correct for such a model.

Theorem 2. *Let $p : \mathbf{N} \to (0, 1)$ be any function such that*

$$N^{-1} = o(p(N)) \quad \text{and} \quad p(N) = o(1). \tag{26}$$

For each $N \in \mathbf{N}$, let A be a random subset of $\{1, \dots, N\}$ chosen according to a binomial distribution with parameter $p(N)$. Then, as $N \to \infty$, the probability that A is difference dominated tends to one.

More precisely, let $\mathscr{S}, \mathscr{D}$ denote respectively the random variables $|A + A|$ and $|A - A|$. Then the following three situations arise:

(i) $p(N) = o(N^{-1/2})$*: Then,*

$$\mathscr{S} \sim \frac{(N \cdot p(N))^2}{2} \quad \text{and} \quad \mathscr{D} \sim 2\mathscr{S} \sim (N \cdot p(N))^2. \tag{27}$$

(ii) $p(N) = c \cdot N^{-1/2}$ *for some* $c \in (0, \infty)$*: Define the function* $g : (0, \infty) \to (0, 2)$ *by*

$$g(x) := 2\left(\frac{e^{-x} - (1 - x)}{x}\right). \tag{28}$$

Then,

$$\mathscr{S} \sim g\left(\frac{c^2}{2}\right) N \quad \text{and} \quad \mathscr{D} \sim g(c^2)N. \tag{29}$$

(iii) $N^{-1/2} = o(p(N))$*: Let* $\mathscr{S}^c := (2N + 1) - \mathscr{S}$, $\mathscr{D}^c := (2N + 1) - \mathscr{D}$. *Then*

$$\mathscr{S}^c \sim 2 \cdot \mathscr{D}^c \sim \frac{4}{p(N)^2}. \tag{30}$$

Parts (i) and (ii) of the theorem can be proven by elementary means; a standard second moment analysis (Chebyshev's inequality applied to a sum of indicator random variables) suffices to prove strong concentration of the variables $\mathscr{S}$ and $\mathscr{D}$, while in part (ii) an additional inclusion-exclusion type argument is used to obtain the correct form of the function g. Our proof of part (iii) requires different and more sophisticated concentration machinery recently developed by Kim and Vu [KV00, Vu00, Vu02]. For the benefit of the reader not familiar with probabilistic techniques, we present below an entirely self-contained proof of a more explicit form of the simplest case of our theorem, namely part (i). See [HM07] for proofs of the other cases, as well as generalizations to comparing arbitrary binary forms.[11]

We prove the following special case of Theorem 2.

Theorem 3. *Let* $p(N) := cN^{-\delta}$ *for some* $c > 0$, $\delta \in (1/2, 1)$. *Set* $C := \max(1, c)$, $f(\delta) := \min\{\frac{1}{2}, \frac{3\delta - 1}{2}\}$ *and let* $g(\delta)$ *be any function such that* $0 < g(\delta) < f(\delta)$ *for all* $\delta \in (1/2, 1)$. *Set* $P_1(N) := \frac{4}{c}N^{-(1-\delta)}$ *and* $P_2(N) := N^{-(f(\delta)-g(\delta))}$. *For any subset chosen with respect to the binomial model with parameter* $p = p(N)$*, with probability at least* $1 - P_1(N) - P_2(N)$ *the ratio of the cardinality of its difference*

[11] Let u_1, u_2, v_1, v_2 be fixed integers, and define two binary forms $f(x, y) = u_1x + v_1y$ and $g(x, y) = u_2x + v_2y$. By $f(A)$ we mean $\{f(a_1, a_2) : a_i \in A\}$ (and similarly for $g(A)$). Theorem 2 can be generalized to analyze how often $|f(A)| > |g(A)|$ when A is drawn from $\{1, \ldots, N\}$ from a binomial model with parameter $p(N)$.

set to the cardinality of its sumset is $2 + O_C(N^{-g(\delta)})$. *Thus the probability a subset chosen with respect to the binomial model is not difference dominated is at most* $P_1(N) + P_2(N)$, *which tends to zero rapidly with* N *for* $\delta \in (1/2, 1)$.

We first establish some notation, and then prove a sequence of lemmas from which Theorem 3 immediately follows. Our goal is to provide explicit bounds which decay like N to a power.

Let $I_N = \{1, \dots, N\}$ and let $X_{n;N}$ denote the binary indicator variable for n being in a subset (it is thus 1 with probability $cN^{-\delta}$ and 0 otherwise), and let X be the random variable denoting the cardinality of a subset (thus $X = \sum_n X_{n;N}$). For two pairs of ordered elements (m, n) and (m', n') in $I_N \times I_N$ ($m < n$, $m' < n'$), let $Y_{m,n,m',n'} = 1$ if $n - m = n' - m'$, and 0 otherwise.

Lemma 4. *With probability at least* $1 - P_1(N)$,

$$X \in \left[\frac{1}{2}\, cN^{1-\delta},\ \ \frac{3}{2}\, cN^{1-\delta}\right]. \tag{31}$$

Let $\mathscr{O}$ *denote the number of ordered pairs* (m, n) *(with* $m < n$*) in a subset of* I_N *chosen with respect to the binomial model. Then with probability at least*[12] $1 - P_1(N)$ *we have*

$$\frac{\frac{1}{2}cN^{1-\delta}\left(\frac{1}{2}cN^{1-\delta} - 1\right)}{2} \le \mathscr{O} \le \frac{\frac{3}{2}cN^{1-\delta}\left(\frac{3}{2}cN^{1-\delta} - 1\right)}{2}. \tag{32}$$

Proof. We have $\mathbf{E}[X] = \sum_n \mathbf{E}[X_{n;N}] = cN^{1-\delta}$. As the $X_{n;N}$ are independent,

$$\sigma_X^2 = \sum_n \sigma_{X_{n;N}}^2 = N\left(cN^{-\delta} - c^2N^{-2\delta}\right). \tag{33}$$

Thus,

$$\sigma_X \le \sqrt{c} \cdot N^{\frac{1-\delta}{2}}. \tag{34}$$

By Chebyshev's inequality,

$$\text{Prob}(|X - cN^{1-\delta}| \le k\sigma_X) \ge 1 - \frac{1}{k^2}. \tag{35}$$

For $X \in \left[\frac{1}{2}\, cN^{1-\delta},\ \ \frac{3}{2}\, cN^{1-\delta}\right]$, we choose k so that

$$k\sigma_X = \frac{1}{2}\, cN^{1-\delta} \le k\sqrt{c}N^{\frac{1-\delta}{2}}. \tag{36}$$

[12] By using the Central Limit Theorem instead of Chebyshev's inequality we could obtain a better estimate on the probability of X lying in the desired interval.

Thus $k \geq \frac{1}{2}\sqrt{c}N^{(1-\delta)/2}$, and the probability that X lies in the stated interval is at least $1-(cN^{1-\delta}/4)^{-1}$. The second claim follows from the fact that there are $\binom{r}{2}$ ways to choose two distinct objects from r objects. □

Proof of Theorem 3. By Lemma 4, (32) holds with probability at least $1-P_1(N)$. The main contribution to the cardinalities of the sumset and the difference set is from ordered pairs (m,n) with $m<n$. With probability at least $1-P_1(N)$, there are on the order $N^{2-2\delta}$ such pairs, which are much larger than the order $N^{1-\delta}$ pairs with $m=n$. The proof is completed by showing that almost all of the ordered pairs yield distinct sums (and differences). Explicitly, we shall show that for a subset chosen from I_N with respect to the binomial model, if $\mathscr{O}$ is the number of ordered pairs (which is of size $N^{2-2\delta}$ with high probability), then with high probability the cardinality of its difference set is $2\mathscr{O}+O_C(N^{3-4\delta})$ while the cardinality of its sumset is $\mathscr{O}+O_C(N^{3-4\delta})$. This argument crucially uses $\delta>1/2$ (if $\delta=1/2$, then the error term is the same size as the main term, and a more delicate argument is needed). We shall show that almost all of the ordered pairs generate distinct differences; the argument for the sums follows similarly.

Each ordered pair (m,n) yields two differences: $m-n$ and $n-m$. The problem is that two different ordered pairs could generate the same differences. To calculate the size of the difference set, we need to control how often two different pairs give the same differences. Consider two distinct ordered pairs (m,n) and (m',n') with $m<n$ and $m'<n'$ (as the $N^{1-\delta} \ll N^{2-2\delta}$ 'diagonal' pairs (n,n) yield the same difference, namely 0, it suffices to study the case of ordered pairs with distinct elements). Without loss of generality, we may assume $m \leq m'$. If $n-m=n'-m'$, then these two pairs contribute the same differences. There are two possibilities: (1) all four indices are distinct; (2) $n=m'$.

We calculate the expected number of pairs of non-diagonal ordered pairs with the same difference by using our binary indicator random variables $Y_{m,n,m',n'}$. Set

$$Y = \sum_{1\leq m\leq m'\leq N} \sum_{m'<n'\leq N} \sum_{\substack{m<n\leq N \\ n'-m'=n-m}} Y_{m,n,m',n'}. \tag{37}$$

If the four indices are distinct, then $\mathbf{E}[Y_{m,n,m',n'}] = c^4N^{-4\delta}$; if $n=m'$, then $\mathbf{E}[Y_{m,n,m',n'}] = c^3N^{-3\delta}$.

The number of tuples (m,n,m',n') of distinct integers satisfying our conditions is bounded by N^3 (once m, n and m' are chosen there is at most one choice for $n' \in \{m+1,\ldots,N\}$ with $n'-m'=n-m$)[13]. If instead $n=m'$ then there are at most N^2 tuples satisfying our conditions (once m and n are chosen, m' and n' are uniquely determined, though they may not satisfy our conditions). Therefore,

$$\mathbf{E}[Y] \leq N^3\cdot c^4N^{-4\delta} + N^2\cdot c^2N^{-3\delta} \leq 2C^4N^{3-4\delta} \tag{38}$$

as $\delta \in (1/2,1)$.

[13] Although we do not need the actual value, simple algebra yields the number of tuples is $N^3/6+O(N^2)$.

As $N^{3-4\delta}$ is much smaller than $N^{2-2\delta}$ for $\delta > 1/2$, most of the differences are distinct. To complete the proof, we need some control on the variance of Y. In Lemma 5, we show that

$$\sigma_Y \le 7C^4 N^{r(\delta)}, \tag{39}$$

where

$$2r(\delta) = \max\{3-4\delta, 5-7\delta\}. \tag{40}$$

While we cannot use the Central Limit Theorem (the $Y_{m,n,m',n'}$ are not independent), we may use Chebyshev's inequality to bound the probability that Y is close to its mean (recall the mean is at most $2C^4N^{3-4\delta}$). We have

$$\text{Prob}(|Y - \mathbf{E}[Y]| \le k\sigma_Y) \ge 1 - \frac{1}{k^2}. \tag{41}$$

Simple algebra shows that if we take $k = N^{2-2\delta-r(\delta)-g(\delta)}$, then with probability at least $1 - N^{-(f(\delta)-g(\delta))}$ we have $Y \le 9C^4N^{2-2\delta-g(\delta)}$, which is a positive power of N less than $N^{2-2\delta}$. Thus an at most negligible amount of the differences are repeated.

The argument for two ordered pairs yielding the same sum proceeds similarly: if $\mu + \nu = \mu' + \nu'$, then $\nu - \mu' = \nu' - \mu$.

For our ratio to be $2 + O_C(N^{-g(\delta)})$, two events must happen. As the probability the first does not occur is at most $P_1(N)$ and the probability the second does not occur is at most $P_2(N)$, the probability that the two desired events happen is at least $1 - P_1(N) - P_2(N)$.

Except for the claimed estimate on σ_Y, the above completes the proof of Theorem 3. We now prove our bound for σ_Y.

Lemma 5. *Let the notation be as in Theorem 3 and (A.10). We have*

$$\sigma_Y \le 7C^4 N^{r(\delta)}. \tag{42}$$

Proof. If U and V are two random variables, then

$$\text{Var}(U+V) = \text{Var}(U) + \text{Var}(V) + 2\text{CoVar}(U,V). \tag{43}$$

By the Cauchy-Schwartz inequality, $\text{CoVar}(U,V) \le \sqrt{\text{Var}(U)\text{Var}(V)}$. Thus

$$\text{Var}(U+V) \le 3\text{Var}(U) + 3\text{Var}(V). \tag{44}$$

We may therefore write

$$\sum Y_{m,n,m',n'} = \sum U_{m,n,m',n'} + \sum V_{m,n,n'} = U + V, \tag{45}$$

where in the U-sum all four indices are distinct (with $1 \le m < m' \le N$, $m < n \le N$, $m' < n' \le N$ and $n-m = n'-m'$) and in the V-sum all three indices are distinct (with $1 \le m < n < n' \le N$ and $n-m = n'-n$). As $\mathrm{Var}(Y) \le 3\mathrm{Var}(U)+3\mathrm{Var}(V)$, we are reduced to bounding the variances of U and V.

We first bound $\mathrm{Var}(U)$. Standard algebra yields

$$\begin{aligned}\mathrm{Var}(U) &= \mathrm{Var}\left(\sum U_{m,n,m',n'}\right)\\ &= \sum \mathrm{Var}(U_{m,n,m',n'}) + 2 \sum_{(m,n,m',n')\neq(\widetilde{m},\widetilde{n},\widetilde{m'},\widetilde{n'})} \mathrm{CoVar}(U_{m,n,m',n'}, U_{\widetilde{m},\widetilde{n},\widetilde{m'},\widetilde{n'}}).\end{aligned} \tag{46}$$

As $\mathrm{Var}(U_{m,n,m',n'}) = c^4N^{-4\delta} - c^8N^{-8\delta}$ and there are at most N^3 ordered tuples (m,n,m',n') of distinct integers with $n-m = m'-n'$, the $\mathrm{Var}(U_{m,n,m',n'})$ term is bounded by $c^4N^{3-4\delta}$.

For the covariance piece, if all eight indices $(m,n,m',n',\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$ are distinct, then $U_{m,n,m',n'}$ and $U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}$ are independent and thus the covariance is zero. There are four cases; in each case, there are always at most N^3 choices for the tuple (m,n,m',n'), but often there will be significantly fewer choices for the tuple $(\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$. We only provide complete details for the first and third cases, as the other cases follow similarly.

- Seven distinct indices: There are at most N^2 choices for $(\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$. The covariance of each such term is bounded by $c^7N^{-7\delta}$. To see this, note

$$\begin{aligned}&\mathrm{CoVar}(U_{m,n,m',n'}, U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'})\\ &\quad= \mathbf{E}[U_{m,n,m',n'}U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}] - \mathbf{E}[U_{m,n,m',n'}]\mathbf{E}[U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}].\end{aligned} \tag{47}$$

  The product of the expected values is $c^8N^{-8\delta}$, while the expected value of the product is $c^7N^{-7\delta}$. Thus the covariances of these terms contribute at most $c^7N^{5-7\delta}$.
- Six distinct indices: The covariances of these terms contribute at most $c^6N^{4-6\delta}$.
- Five distinct indices: The covariances of these terms contribute at most $c^5N^{3-5\delta}$ (once three of the $\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'$ have been determined, the fourth is uniquely determined; thus there are at most N^3 choices for the first tuple and at most 1 choice for the second).
- Four distinct indices: The covariances of these terms contribute at most $c^4N^{3-4\delta}$.

The N-dependence from the case of seven distinct indices is greater than the N-dependence of the other cases (except for the case of four distinct indices if $\delta > 2/3$). We also only increase the contributions if we replace c with $C = \max(c,1)$. We therefore find

$$\begin{aligned}\mathrm{Var}(U) &\le C^4N^{3-4\delta} + 2\left(C^7N^{5-7\delta} + C^6N^{4-6\delta} + C^5N^{3-5\delta} + C^4N^{3-4\delta}\right)\\ &= 3C^4N^{3-4\delta} + 6C^7N^{5-7\delta}.\end{aligned} \tag{48}$$

Similarly we have,

$$\begin{aligned}\mathrm{Var}(V) &= \mathrm{Var}\left(\sum V_{m,n,n'}\right)\\ &= \sum \mathrm{Var}(V_{m,n,n'}) + 2 \sum_{(m,n,n')\neq(\widetilde{m},\widetilde{n},\widetilde{n'})} \mathrm{CoVar}(V_{m,n,n'}, V_{\widetilde{m},\widetilde{n},\widetilde{n'}}). \end{aligned} \tag{49}$$

The $\mathrm{Var}(V_{m,n,n'})$ piece is bounded by $N^2 \cdot c^3 N^{-3\delta}$ (as there are at most N^2 tuples with $n' - n = n - m$). The covariance terms vanish if the six indices are distinct. A similar argument as before yields bounds of $c^5 N^{3-5\delta}$ for five distinct indices, $c^4 N^{2-4\delta}$ for four distinct indices, and $c^3 N^{2-3\delta}$ for three distinct indices. The largest N-dependence is from the $c^3 N^{2-3\delta}$ term (as $\delta > 1/2$). Arguing as before and replacing c with C yields

$$\mathrm{Var}(V) \le C^3 N^{2-3\delta} + 2 \cdot 3C^3 N^{2-3\delta} \le 7C^3 N^{2-3\delta}. \tag{50}$$

As $\delta < 1$, $2 - 3\delta < 3 - 4\delta$. Therefore,

$$\begin{aligned}\mathrm{Var}(Y) &\le 3 \cdot \left(3C^4 N^{3-4\delta} + 6C^7 N^{5-7\delta}\right) + 3 \cdot 7C^3 N^{2-3\delta}\\ &\le 30C^4 N^{3-4\delta} + 18C^7 N^{5-7\delta} \le 49C^8 N^{2r(\delta)}, \end{aligned} \tag{51}$$

which yields

$$\sigma_Y \le 7C^4 N^{r(\delta)}. \tag{52}$$

□

Remark 6. An extreme choice of g would be to choose $g(\delta) = \varepsilon$, for some small positive constant ε. Since $f(\delta) \ge 1/4$ for all $\delta \in (1/2, 1)$, we then obtain a bound of $2 + O_C(N^{-\varepsilon})$ for the ratio of the cardinality of the difference set to the sumset with probability $1 - O_C(N^{-\min\{1-\delta, \frac{1}{4}-\varepsilon\}})$.

References

[FP73] Freiman, G.A. and Pigarev, V.P.: The relation between the invariants R and T. In: Number theoretic studies in the Markov spectrum and in the structural theory of set addition (Russian), pp. 172–174. Kalinin. Gos. Univ., Moscow (1973).

[He07] Hegarty, P.V.: Some explicit constructions of sets with more sums than differences. Acta Arithmetica **130**, no. 1, 61–77 (2007).

[HM07] Hegarty, P.V. and Miller, S.J.: When almost all sets are difference dominated. Random Struct. Algorithm. **35**, no. 1, 118–136 (2009).

[KV00] Kim, J.H. and Vu, V.H.: Concentration of multivariate polynomials and its applications. Combinatorica **20**, 417–434 (2000).

[Ma69] Marica, J.: On a conjecture of Conway. Canad. Math. Bull. **12**, 233–234 (1969).

[MO06] Martin, G. and O'Bryant, K.: Many sets have more sums than differences. In: Additive Combinatorics, in: CRM Proc. Lecture Notes, vol. 43, Am. Math. Soc., Providence, RI, 2007, pp. 287–305.

[MOS09] Miller, S.J., Orosz, B. and Scheinerman, D.: Explicit constructions of infinite families of MSTD sets. To appear in: Journal of Number Theory (2010), doi:10.1016/j.jnt.2009.09.003.

[Na06] Nathanson, M.B.: Problems in additive number theory, I. In: Additive Combinatorics, in: CRM Proc. Lecture Notes, vol. 43, Am. Math. Soc., Providence, RI, 2007, pp. 263–270.

[Na07] Nathanson, M.B.: Sets with more sums than differences. Integers: Electron. J. Combinatorial Number Theory **7**, Paper A5, 24 (2007).

[Ru76] Ruzsa, I.Z.: On the cardinality of $A+A$ and $A-A$. In: Combinatorics year (Keszthely, 1976), vol. 18, Coll. Math. Soc. J. Bolyai, North-Holland-Bolyai Tàrsulat, pp. 933–938 (1978).

[Ru84] Ruzsa, I.Z.: Sets of sums and differences. In: Séminaire de Théorie des Nombres de Paris 1982–1983, pp. 267–273. Birkhäuser, Boston (1984).

[Ru92] Ruzsa, I.Z.: On the number of sums and differences. Acta Math. Sci. Hungar. **59**, 439–447 (1992).

[Vu00] Vu, V.H.: New bounds on nearly perfect matchings of hypergraphs: Higher codegrees do help. Random Struct. Algorithm. **17**, 29–63 (2000).

[Vu02] Vu, V.H.: Concentration of non-Lipschitz functions and Applications. Random Struct. Algorithm. **20**, no. 3, 262–316 (2002).

An Inverse Problem in Number Theory and Geometric Group Theory

Melvyn B. Nathanson*

Summary This paper describes a new link between combinatorial number theory and geometry. The main result states that A is a finite set of relatively prime positive integers if and only if $A = (K - K) \cap \mathbf{N}$, where K is a compact set of real numbers such that for every $x \in \mathbf{R}$ there exists $y \in K$ with $x \equiv y \pmod 1$. In one direction, given a finite set A of relatively prime positive integers, the proof constructs an appropriate compact set K such that $A = (K - K) \cap \mathbf{N}$. In the other direction, a strong form of a fundamental result in geometric group theory is applied to prove that $(K - K) \cap \mathbf{N}$ is a finite set of relatively prime positive integers if K satisfies the appropriate geometrical conditions. Some related results and open problems are also discussed.

Keywords Relatively prime integers · Combinatorial number theory · Additive number theory · Geometric group theory

Mathematics Subject Classifications (2010). Primary 11A05, 11B75, 11P21, 20F65

1 From Compact Sets to Integers

The object of this note is to describe a new connection between number theory and geometry. Let $\mathbf{R}$, $\mathbf{Z}$, and $\mathbf{N}$ denote the real numbers, integers, and positive integers, respectively. For every $x \in \mathbf{R}$, let $[x] \in \mathbf{Z}$ and $(x) \in [0, 1)$ denote the integer part and fractional part of x. Let $\mathbf{Z}^n$ denote the additive group of n-dimensional lattice points in the Euclidean space $\mathbf{R}^n$.

*Supported in part by a grant from the PSC-CUNY Research Award Program. This article was written while the author was a visiting fellow in the mathematics department at Princeton University.

M.B. Nathanson
Department of Mathematics, Lehman College (CUNY), Bronx, New York 10468
and
CUNY Graduate Center, New York, New York 10016
e-mail: melvyn.nathanson@lehman.cuny.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_18,

We recall the following definitions. The set A of integers is *relatively prime*, denoted $\gcd(A) = 1$, if A is nonempty and the elements of A have no common factor greater than 1. Equivalently, A is relatively prime if A generates the additive group $\mathbf{Z}$. The set A of n-dimensional lattice points is relatively prime if the elements of A generate the additive group $\mathbf{Z}^n$.

Let H be a subgroup of a multiplicative group G, and let x and y be elements of G. We say that x and y are *congruent modulo* H, denoted $x \equiv y \pmod{H}$, if $xy^{-1} \in H$. If the group G is additive, then $x \equiv y \pmod{H}$ if $x - y \in H$. For example, let $G = \mathbf{R}$ and $H = \mathbf{Z}$. The real numbers x and y are congruent modulo $\mathbf{Z}$, that is, $x \equiv y \pmod{\mathbf{Z}}$ or, in more traditional notation, $x \equiv y \pmod{1}$, if and only if they have the same fractional part.

In a multiplicative group G, the *difference set* of a subset K of G is

$$KK^{-1} = \{xy^{-1} : x, y \in K\}.$$

In an additive abelian group G, the *difference set* of a subset K of G is

$$K - K = \{x - y : x, y \in K\}.$$

Note that a difference set is symmetric: $z \in KK^{-1}$ if and only if $z^{-1} \in KK^{-1}$ (respectively, $z \in K - K$ if and only if $-z \in K - K$).

We are interested in sets of integers contained in difference sets of sets of real numbers. Our main theorem gives a geometric condition for a finite set of positive integers to be relatively prime. The geometry uses the concept of an $\mathcal{N}$-set, which is a compact subset K of $\mathbf{R}^n$ such that for every $x \in \mathbf{R}^n$ there exists $y \in K$ with $x \equiv y \pmod{\mathbf{Z}^n}$.

Theorem 1. *Let A be a finite set of positive integers. The set A is relatively prime if and only if there exists an $\mathcal{N}$-set K in $\mathbf{R}$ such that $A = (K - K) \cap \mathbf{N}$.*

In Sect. 2, we solve the inverse problem: Given a finite set of relatively prime integers, we construct an $\mathcal{N}$-set K in $\mathbf{R}$ such that $A = (K - K) \cap \mathbf{N}$. In Sect. 3 we apply the "fundamental observation of geometric group theory" to relatively prime sets of integers. An explanation of the "fundamental observation of geometric group theory" appears in Appendix A.

Ideas from geometric group theory have been used recently to obtain new results in number theory (e.g., Nathanson [4–6]), and should continue to be useful. The book of de la Harpe, *Topics in Geometric Group Theory* [1], is an excellent survey of this subject. Theorem 5 was discovered and proved independently by Efremovič [2], Švarc [7], and Milnor [3].

2 The Inverse Problem

In this section we prove that every finite set of relatively prime positive integers can be realized as the difference set of an $\mathcal{N}$-set. The construction depends on the following simple observation.

Lemma 1. *Let K be a set of real numbers, and let $a \in \mathbf{Z} \setminus \{0\}$. Then $a \in K - K$ if and only if there is a two-element subset $\{x, y\}$ of K such that $(x) = (y)$ and $a = [x] - [y]$.*

Proof. For any nonzero integer a, we have $a \in K - K$ if and only if there exist $x, y \in K$ such that $x \neq y$ and

$$a = x - y = [x] - [y] + (x) - (y).$$

Because $[x] - [y] \in \mathbf{Z}$ and $-1 < (x) - (y) < 1$, it follows that $(x) - (y) = 0$ and $a = [x] - [y]$. The set $\{x, y\}$ satisfies the conditions of the Lemma.

Here are three examples. We associate the set $A_1 = \{2, 5\}$ with the $\mathcal{N}$-set

$$K(A_1) = [0, 1/3] \cup [2 + 1/3, 2 + 2/3] \cup [4 + 2/3, 5].$$

There are only three two-element subsets $\{x, y\}$ of $K(A_1)$ such that x and y have the same fractional part: $\{1/3, 2 + 1/3\}$, $\{2 + 2/3, 4 + 2/3\}$, and $\{0, 5\}$, and

$$A_1 = (K(A_1) - K(A_1)) \cap \mathbf{N}.$$

The set $A_2 = \{6, 10, 15\}$ arises from the $\mathcal{N}$-set

$$K(A_2) = [0, 1/3] \cup [9 + 2/3, 10] \cup [15 + 1/3, 15 + 2/3].$$

The complete list of the two-element subsets $\{x, y\}$ of $K(A_2)$ such that x and y have the same fractional part is: $\{1/3, 15 + 1/3\}$, $\{15 + 2/3, 9 + 2/3\}$, and $\{0, 10\}$.

For the set $A_3 = \{18, 28, 63\}$, the $\mathcal{N}$-set

$$K(A_3) = \bigcup_{i=0}^{9} [-18i + i/13, -18i + (i + 1)/13] \cup [-99 + 10/13, -99 + 11/13]$$
$$\cup [-36 + 11/13, -36 + 12/13] \cup [27 + 12/13, 28]$$

satisfies

$$A_3 = (K(A_3) - K(A_3)) \cap \mathbf{N}.$$

There are exactly 13 two-element subsets $\{x, y\}$ of $K(A_3)$ such that x and y have the same fractional part.

In the following Lemma we construct an important example of an $\mathcal{N}$-set on the real line, and its associated difference set of integers.

Lemma 2. *For the positive integer w, let*

$$\lambda_0 < \lambda_1 < \cdots < \lambda_{w-1} < \lambda_w$$

be a strictly increasing sequence of real numbers such that

$$\lambda_w = \lambda_0 + 1$$

and let $b_0, b_1, \ldots, b_{w-1}$ *be a sequence of integers such that*

$$b_{k-1} \neq b_k \text{ for } k = 1, \ldots, w-1$$

and

$$1 + b_{w-1} \neq b_0.$$

The set

$$K' = \bigcup_{k=0}^{w-1} [b_k + \lambda_k, b_k + \lambda_{k+1}]$$

is an $\mathcal{N}$*-set, and*

$$(K' - K') \cap \mathbf{N} = \{|b_k - b_{k-1}| : k = 1, \ldots, w-1\} \cup \{|1 + b_{w-1} - b_0|\}$$

is a finite set of relatively prime positive integers.

Proof. The set K' is compact because it is a finite union of closed intervals, and an $\mathcal{N}$-set because

$$\bigcup_{k=0}^{w-1} [\lambda_k, \lambda_{k+1}] = [\lambda_0, \lambda_w] = [\lambda_0, \lambda_0 + 1].$$

Let A be the finite set of positive integers contained in the difference set $K' - K'$. As

$$\{\{b_{k-1} + \lambda_k, b_k + \lambda_k\} : k = 1, \ldots, w-1\} \cup \{\{b_0 + \lambda_0, b_{w-1} + \lambda_w\}\}$$

is the set of all two-element subsets $\{x, y\}$ of K' with $(x) = (y)$, it follows that

$$A = (K' - K') \cap \mathbf{N} = \{|b_k - b_{k-1}| : k = 1, \ldots, w-1\} \cup \{|1 + b_{w-1} - b_0|\}.$$

Choose $\varepsilon_k \in \{1, -1\}$ such that

$$|b_k - b_{k-1}| = \varepsilon_k (b_k - b_{k-1})$$

for $k = 1, \ldots, w-1$, and $\varepsilon_w \in \{1, -1\}$ such that

$$|1 + b_{w-1} - b_0| = \varepsilon_w (1 + b_{w-1} - b_0).$$

As

$$1 = \varepsilon_w |1 + b_{w-1} - b_0| - \sum_{k=1}^{w-1} \varepsilon_k |b_k - b_{k-1}|$$

it follows that A is a finite set of relatively prime positive integers.

Theorem 2. *If A is a finite set of relatively prime positive integers then there is an $\mathcal{N}$-set K such that $A = (K - K) \cap \mathbf{N}$.*

Proof. As the elements of A are relatively prime, we can write 1 as an integral linear combination of elements of A. Thus, there exist pairwise distinct integers $a_1, \ldots, a_h$ in A, positive integers $w_1, \ldots, w_h$, and $\varepsilon_1, \ldots, \varepsilon_h \in \{1, -1\}$ such that

$$\sum_{i=1}^{h} \varepsilon_i w_i a_i = 1. \tag{1}$$

Let $w_0 = 0$ and $w = \sum_{i=1}^{h} w_i$. For $j = 1, 2, \ldots, w$, we define integers $\tilde{a}_j$ as follows: If

$$w_1 + \cdots + w_{i-1} + 1 \leq j \leq w_1 + \cdots + w_{i-1} + w_i$$

then

$$\tilde{a}_j = -\varepsilon_i a_i .$$

It follows that

$$1 + \sum_{j=1}^{w} \tilde{a}_j = 1 + \sum_{i=1}^{h} w_i (-\varepsilon_i a_i) = 0.$$

For $k = 0, 1, \ldots, w$, we consider the integers

$$b_k = \sum_{j=1}^{k} \tilde{a}_j \tag{2}$$

and real numbers

$$\lambda_k = \frac{k}{w}.$$

Then $b_0 = 0$,

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_w = 1$$

and, for $k = 1, \ldots, w$,

$$b_k - b_{k-1} = \tilde{a}_k \neq 0.$$

It follows from (1) and (2) that

$$\begin{aligned} 1 + b_{w-1} &= 1 + \sum_{j=1}^{w-1} \tilde{a}_j = 1 + \sum_{j=1}^{w} \tilde{a}_j - \tilde{a}_w \\ &= -\tilde{a}_w = \varepsilon_h a_h \neq 0 = b_0. \end{aligned}$$

Construct the $\mathcal{N}$-set

$$K' = \bigcup_{k=0}^{w-1} [b_k + \lambda_k, b_k + \lambda_{k+1}].$$

By Lemma 2,

$$\begin{aligned}(K' - K') \cap \mathbf{N} &= \{|b_k - b_{k-1}| : k = 1, \ldots, w-1\} \cup \{|1 + b_{w-1} - b_0|\} \\ &= \{|\tilde{a}_k| : k = 1, \ldots, w-1\} \cup \{|-\tilde{a}_w|\} \\ &= \{a_i : i = 1, \ldots, h\}.\end{aligned}$$

Let $\mathrm{card}(A) = \ell$. If $\ell = h$ then $A = \{a_1, \ldots, a_h\}$ and we set $K = K'$.

Suppose that $\ell > h$ and $A \setminus \{a_1, \ldots, a_h\} = \{a_{h+1}, \ldots, a_\ell\} \neq \emptyset$. Because

$$0 < \frac{1}{w(\ell - h + 1)} < \frac{2}{w(\ell - h + 1)} < \cdots < \frac{\ell - h}{w(\ell - h + 1)} < \frac{1}{w}$$

and

$$\left[0, \frac{1}{w}\right] = [b_0 + \lambda_0, b_0 + \lambda_1] \subseteq K'$$

it follows that

$$K = K' \cup \left\{ a_{h+i} + \frac{i}{w(\ell - h + 1)} : i = 1, 2, \ldots, \ell - h \right\}$$

is an $\mathcal{N}$-set such that $A = (K - K) \cap \mathbf{N}$. This completes the proof.

Let A be a finite set of relatively prime positive integers. We define the weight of a representation of 1 in the form (1) by:

$$\sum_{i=1}^{h} w_i + \mathrm{card}(A) - h.$$

We define the *additive weight* of A, denoted $\mathrm{Add}(A)$ as the smallest weight of a representation of 1 in the form (1) with integers $a_i \in A$. Note that $\mathrm{Add}(A) \geq \mathrm{card}(A)$ for all A, and $\mathrm{Add}(A) = \mathrm{card}(A)$ if and only if there exist distinct integers $a_1, \ldots, a_h \in A$ and $\varepsilon_1, \ldots, \varepsilon_h \in \{1, -1\}$ such that $\sum_{i=1}^{h} \varepsilon_i a_i = 1$.

We define the weight of an $\mathcal{N}$-set K as the number of connected components of K, and the *geometric weight* of A, denoted $\mathrm{Geo}(A)$ as the smallest weight of an $\mathcal{N}$-set K such that $A = (K - K) \cap \mathbf{N}$.

The following result follows immediately from the proof of Theorem 2.

Corollary 1. *Let A be a finite set of relatively prime positive integers. Then*

$$\mathrm{Geo}(A) \leq \mathrm{Add}(A).$$

There exist sets A such that $\mathrm{Geo}(A) < \mathrm{Add}(A)$. For example, if $A = \{1, 2, 3, \ldots, n\}$ then $K = [0, n]$ is an $\mathcal{N}$-set of weight 1 such that $(K-K) \cap \mathbf{N} = A$, and so $\mathrm{Geo}(A) = 1 < n = \mathrm{Add}(A)$.

3 Relatively Prime Sets of Lattice Points

In this section we obtain the converse of Theorem 2.

Theorem 3. *If K is an $\mathcal{N}$-set in $\mathbf{R}$ then $A = (K-K) \cap \mathbf{N}$ is a finite set of relatively prime positive integers.*

We prove this result in n dimensions.

Theorem 4. *If K is an $\mathcal{N}$-set in $\mathbf{R}^n$ then $A = (K-K) \cap \mathbf{Z}^n$ is a finite set of relatively prime lattice points.*

Note the necessity of the compactness condition. For $n \geq 1$, the noncompact set $K = [0, 1)^n$ has the property that for all $x \in \mathbf{R}^n$ there exists $y \in K$ with $x \equiv y \pmod{\mathbf{Z}^n}$, but $(K-K) \cap \mathbf{Z}^n = \{0\}$.

Proof. The proof uses a theorem called "the fundamental observation of geometric group theory." We discuss this result in Appendix.

The additive group $\mathbf{Z}^n$ acts isometrically and properly discontinuously on $\mathbf{R}^n$ by translation: $(g, x) \mapsto g + x$ for $g \in \mathbf{Z}^n$ and $x \in \mathbf{R}^n$. The quotient space $\mathbf{Z}^n \setminus \mathbf{R}^n$ is the n-dimensional torus, which is compact, and so the group action $\mathbf{Z}^n \curvearrowright \mathbf{R}^n$ is cocompact. Let $\pi : \mathbf{R}^n \to \mathbf{Z}^n \setminus \mathbf{R}^n$ be the quotient map. Then $\pi(x) = \langle x \rangle$ is the orbit of x for all $x \in \mathbf{R}^n$. If K is an $\mathcal{N}$-set in $\mathbf{R}^n$ then K is compact, and for every $x \in \mathbf{R}^n$ there exists $y \in K$ such that $x \equiv y \pmod{\mathbf{Z}^n}$. This means that $\pi(y) = \langle x \rangle$, and so $\pi(K) = \mathbf{Z}^n \setminus \mathbf{R}^n$. Applying Theorem 5 to the set K, we conclude that the set

$$A = \{a \in \mathbf{Z}^n : K \cap (a + K) \neq \emptyset\}$$

is a finite set of generators for $\mathbf{Z}^n$. Moreover, $a \in A$ if and only if $a \in \mathbf{Z}^n$ and there exists $x \in K$ such that $x \in a + K$, that is, $x = a + y$ for some $y \in K$. Equivalently, $a \in A$ if and only if $a = x - y \in (K-K) \cap \mathbf{Z}^n$. This proves Theorem 4.

The symmetry of the difference set immediately implies Theorem 3.

We can state the following general inverse problem in geometric group theory: If A is a finite set of generators for a group G such that A is symmetric and contains the identity of G, does there exist a geometric action of G on a metric space X such that $A = (K-K) \cap G$ for some compact set K with $\pi(K) = G \backslash X$? In this article we proved that the answer is "yes" for $G = \mathbf{Z}$, but the answer is not known for other groups. In particular, it is not known for lattice points. Equivalently, does every finite symmetric set A of relatively prime n-dimensional lattice points with $0 \in A$ come from an $\mathcal{N}$-set K in $\mathbf{R}^n$, in the sense that $A = (K-K) \cap \mathbf{Z}^n$? This is not known even in dimension 2.

Appendix: The Fundamental Observation of Geometric Group Theory

The proof of Theorem 4 is an application of what is often called the "fundamental observation of geometric group theory" [1, Chap. IV, pp. 87–88]. We shall describe this result, which is not well known to number theorists.

We begin by introducing the class of boundedly compact geodesic metric spaces. The Heine-Borel theorem states that, in Euclidean space $\mathbf{R}^n$ with the usual metric, a closed and bounded set is compact. We shall call a metric space (X, d) *boundedly compact* if every closed and bounded subset of X is compact. Equivalently, X is boundedly compact if every closed ball

$$B^*(x_0, r) = \{x \in X : d(x_0, x) \leq r\}$$

is compact for all $x_0 \in X$ and $r \geq 0$. Boundedly, compact metric spaces are also called *proper* metric spaces.

A metric space (X, d) is *geodesic* if, for all points $x_0, x_1 \in X$ with $x_0 \neq x_1$, there is an isometry γ from an interval $[a, b] \subseteq \mathbf{R}$ into X such that $\gamma(a) = x_0$ and $\gamma(b) = x_1$. Thus, if $t, t' \in [a, b]$ then $d(\gamma(t), \gamma(t')) = |t - t'|$. In particular, $d(x_0, x_1) = d(\gamma(a), \gamma(b)) = b - a$. For example, let $x_0, x_1 \in \mathbf{R}^n$ with $|x_1 - x_0| = T$. Define $\gamma : [0, T] \to \mathbf{R}^n$ by:

$$\gamma(t) = x_0 + \frac{t}{T}(x_1 - x_0).$$

Then $\gamma(0) = x_0$, $\gamma(T) = x_1$, and

$$\begin{aligned} |\gamma(t) - \gamma(t')| &= \left| \left(x_0 + \frac{t}{T}(x_1 - x_0) \right) - \left(x_0 + \frac{t'}{T}(x_1 - x_0) \right) \right| \\ &= \left| \left(\frac{t - t'}{T} \right) (x_1 - x_0) \right| = |t - t'|. \end{aligned}$$

Thus, $\mathbf{R}^n$ is a boundedly compact geodesic metric space.

Let G be a group that acts on a topological space X. We denote the group action by $G \curvearrowright X$. For every $g \in G$ we define the function $\alpha_g : X \to X$ by $\alpha_g(x) = g \cdot x$. If the function $\alpha_g : X \to X$ is continuous for all $g \in G$ then $\alpha_{g^{-1}} = \alpha_g^{-1}$ implies that α_g is a homeomorphism for all $g \in G$. We say that the group G *acts isometrically* on a metric space (X, d) if the function $x \mapsto gx$ is an isometry for every $g \in G$.

The action of a group G on a topological space X is called *properly discontinuous* if, for every compact subset K of X, there are only finitely many $a \in G$ such that $K \cap aK \neq \emptyset$. Let $A = \{a \in G : K \cap aK \neq \emptyset\}$. Then $A \neq \emptyset$ because $e \in A$. As

$$K \cap a^{-1}K = a^{-1}(K \cap aK)$$

it follows that $A^{-1} = A$.

For every element $x_0 \in X$, the *orbit* of x_0 is the set

$$\langle x_0 \rangle = \{gx_0 : g \in G\} = Gx_0.$$

The orbits of elements of X partition the set X. Let $G\backslash X$ denote the set of orbits of the group action, and define the function $\pi : X \to G\backslash X$ by $\pi(x) = \langle x \rangle$. We call $G \setminus X$ the *quotient space* of X by G, and we call π the *quotient map* of X onto $G\backslash X$. Note that every orbit $\langle x \rangle$ is a subset of the set X and a point in the quotient space $G\backslash X$.

We define the quotient topology on $G\backslash X$ as follows: A set V in $G\backslash X$ is open if and only if $\pi^{-1}(V)$ is open in X. This is the largest topology on the quotient space $G\backslash X$ such that the quotient map π is continuous. We call the group action $G \curvearrowright X$ *co-compact* if the quotient space $G\backslash X$ is compact. An isometric, properly discontinuous, co-compact action of a group G on a boundedly compact geodesic metric space is called a *geometric action*.

We now state the "fundamental observation of geometric group theory."

Theorem 5. *Let (X, d) be a boundedly compact geodesic metric space and let G be a group that acts isometrically on X. Suppose that the group action $G \curvearrowright X$ is properly discontinuous and co-compact. Let $\pi : X \to G\backslash X$ be the quotient map, and let K be a compact subset of X such that $\pi(K) = G\backslash X$. Then*

$$A = \{a \in G : K \cap aK \neq \emptyset\}$$

is a finite set of generators for G.

For example, the additive group $\mathbf{Z}^n$ of n-dimensional lattice points acts on Euclidean space $\mathbf{R}^n$ by translation: $\alpha_g(x) = g + x$ for $g \in \mathbf{Z}^n$ and $x \in \mathbf{Z}^n$. The group $\mathbf{Z}^n$ acts isometrically on $\mathbf{R}^n$ because

$$|\alpha_g(x) - \alpha_g(y)| = |(g + x) - (g + y)| = |x - y|.$$

Let K be a compact subset of $\mathbf{R}^n$. Then K is bounded and there is a number $r > 0$ such that $|x| < r$ for all $x \in K$. If $g \in \mathbf{Z}^n$ and $K \cap (g + K) \neq \emptyset$ then there exists $x \in K$ such that $g + x \in K$. Therefore,

$$|g| - r < |g| - |x| \leq |g + x| < r$$

and $|g| < 2r$. There exist only finitely many lattice points in $\mathbf{Z}^n$ of length less than $2r$, and so the action on $\mathbf{Z}^n$ on $\mathbf{R}^n$ is properly discontinuous.

We shall prove that this group action $\mathbf{Z}^n \curvearrowright \mathbf{R}^n$ is co-compact. Let $\pi : \mathbf{R}^n \to \mathbf{Z}^n \setminus \mathbf{R}^n$ be the quotient map. The quotient space $\mathbf{T}^n = \mathbf{Z}^n \setminus \mathbf{R}^n$ is called the *n-dimensional torus*. Let $\{W_i\}_{i \in I}$ be an open cover of $\mathbf{T}^n$, and define $V_i = \pi^{-1}(W_i)$ for all $i \in I$. Then $\{V_i\}_{i \in I}$ is an open cover of $\mathbf{R}^n$. The *unit cube*

$$K = [0, 1]^n = \{x = (x_1, \ldots, x_n) \in \mathbf{R}^n : 0 \leq x_i \leq 1 \text{ for all } i = 1, \ldots, n\}$$

is a compact subset of $\mathbf{R}^n$, and $\pi(K) = \mathbf{T}^n$. As $\{V_i\}_{i \in I}$ is an open cover of K, it follows that there is a finite subset J of I such that $K \subseteq \bigcup_{j \in J} V_j$, and so

$$\mathbf{T}^n = \pi(K) \subseteq \bigcup_{j \in J} \pi(V_j) = \bigcup_{j \in J} W_j.$$

Therefore, $\mathbf{T}^n$ is compact and the group action $\mathbf{Z}^n \curvearrowright \mathbf{R}^n$ is co-compact.

References

1. P. de la Harpe, *Topics in Geometric Group Theory*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 2000.
2. V. A. Efremovič, *The proximity geometry of Riemannian manifolds*, Uspekhi Math. Nauk **8** (1953), 189.
3. J. Milnor, *A note on curvature and fundamental group*, J. Diff. Geom. **2** (1968), 1–7.
4. M. B. Nathanson, *Phase transitions in infinitely generated groups, and related problems in additive number theory*, arXiv: 0811.3990, 2008. Integers, to appear.
5. M. B. Nathanson, *Nets in groups, minimum length g-adic representations, and minimal additive complements*, arXiv: 0812.0560, 2008.
6. M. B. Nathanson, *Bi-Lipschitz equivalent metrics on groups, and a problem in additive number theory*, arXiv: 0902.3254, 2009.
7. A. S. Švarc, *A volume invariant of coverings*, Dokl. Akad. Nauk SSSR (N.S.) **105** (1955), 32–34.

Cassels Bases

Melvyn B. Nathanson*

Summary This paper describes several classical constructions of thin bases of finite order in additive number theory, and, in particular, gives a complete presentation of a beautiful construction of Cassels of a class of polynomially asymptotic bases. Some open problems are also discussed.

Keywords Additive basis · Sumset · Thin basis · Polynomially asymptotic basis · Cassels basis · Raikov-Stöhr basis · Jia-Nathanson basis · Additive number theory

Mathematics Subject Classifications (2010). 11B13, 11B75, 11P70,11P99

1 Additive Bases of Finite Order

The fundamental object in additive number theory is the *sumset*. If $h \geq 2$ and $A_1, \ldots, A_h$ are sets of integers, then we define the sumset

$$A_1 + \cdots + A_h = \{a_1 + \cdots + a_h : a_i \in A_i \text{ for } i = 1, \ldots, h\}. \tag{1}$$

If $A_1 = A_2 = \cdots = A_h = A$, then the sumset

$$hA = \underbrace{A + A + \cdots + A}_{h \text{ summands}} \tag{2}$$

* This work was supported in part by the PSC-CUNY Research Award Program.

M.B. Nathanson
Department of Mathematics, Lehman College (CUNY), Bronx, New York 10468
and
CUNY Graduate Center, New York, New York 10016
e-mail: melvyn.nathanson@lehman.cuny.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_19,

is called the *h-fold sumset of A*. If $0 \in A$, then

$$A \subseteq 2A \subseteq \cdots \subseteq hA \subseteq (h+1)A \subseteq \cdots$$

For example,

$$\{0, 1, 4, 5\} + \{0, 2, 8, 10\} = [0, 15]$$

and

$$\begin{aligned} &\{3, 5, 7, 11\} + \{3, 5, 7, 11, 13, 17, 19\} \\ &\quad = \{6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30\}. \end{aligned}$$

The set A is called a *basis of order h* for the set B if every element of B can be represented as the sum of exactly h not necessarily distinct elements of A, or, equivalently, if $B \subseteq hA$. The set A is an *asymptotic basis of order h* for B if the sumset hA contains all but finitely many elements of B, that is, if $\text{card}(B \backslash hA) < \infty$. The set A is a basis (respectively, asymptotic basis) of finite order for B if A is a basis (respectively, asymptotic basis) of order h for B for some positive integer h. The set A of non-negative integers is a basis of finite order for the non-negative integers only if $0, 1 \in A$.

Many classical results and conjectures in additive number theory state that some "interesting" or "natural" set of non-negative integers is a basis or asymptotic basis of finite order. For example, the Goldbach conjecture asserts that the set of odd prime numbers is a basis of order 2 for the even integers greater than 4. Lagrange's theorem states the set of squares is a basis of order 4 for the non-negative integers $\mathbf{N}_0$. Wieferich proved that the set of non-negative cubes is a basis of order 9 for $\mathbf{N}_0$, and Linnik proved that the set of non-negative cubes is an asymptotic basis of order 7 for $\mathbf{N}_0$. More generally, for any integer $k \geq 2$, Waring's problem, proved by Hilbert in 1909, states that the set of non-negative k-th powers is a basis of finite order for $\mathbf{N}_0$. Vinogradov proved that the set of odd prime numbers is an asymptotic basis of order 3 for the odd positive integers. Nathanson [11] contains complete proofs of all of these results.

Notation: Let $\mathbf{N}$, $\mathbf{N}_0$, and $\mathbf{Z}$ denote the sets of positive integers, non-negative integers, and integers, respectively. For real numbers x and y, we define the intervals of integers $[x, y] = \{n \in \mathbf{Z} : x \leq n \leq y\}$, $(x, y] = \{n \in \mathbf{Z} : x < n \leq y\}$, and $[x, y) = \{n \in \mathbf{Z} : x \leq n < y\}$. For any sets A and A' of integers and any integer c, we define the *difference set*

$$A - A' = \{a - a' : a \in A \text{ and } a' \in A'\}$$

and the *dilation* by c of the set A

$$c * A = \{ca : a \in A\}.$$

Thus, $2 * \mathbf{N}$ is the set of positive even integers, and $2 * \mathbf{N} - \{0, 1\} = \mathbf{N}$.

Denote the cardinality of the set X by $|X|$.

Let f be a complex-valued function on the domain Ω and let g be a positive function on the domain Ω. Usually, Ω is the set of positive integers or the set of all real numbers $x \geq x_0$. We write $f \ll g$ or $f = O(g)$ if there is a number $c > 0$ such that $|f(x)| \leq cg(x)$ for all $x \in \Omega$. We write $f \gg g$ if there is a number $c > 0$ such that $|f(x)| \geq cg(x)$ for all $x \in \Omega$. We write $f = o(g)$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

2 A Lower Bound for Bases of Finite Order

For any set A of integers, the *counting function* of A, denoted $A(x)$, counts the number of positive integers in A not exceeding x, that is,

$$A(x) = \sum_{\substack{a\in A \\ 1\leq a\leq x}} 1 = |A \cap [1, x]| .$$

Theorem 1. *Let $h \geq 2$ and let $A = \{a_k\}_{k=1}^{\infty}$ be a set of non-negative integers with $a_k < a_{k+1}$ for all $k \geq 1$. If A is an asymptotic basis of order h then*

$$A(x) \gg x^{1/h} \tag{3}$$

for all sufficiently large real numbers x and

$$a_k \ll k^h \tag{4}$$

for all positive integers k. If A is a basis of order h, then inequality (3) holds for all real numbers $x \geq 1$.

Proof. If A is an asymptotic basis of order h, then there exists an integer n_0 such that every integer $m \geq n_0$ can be represented as the sum of h elements of A. Let $x \geq x_0$ and let n be the integer part of x. Then $A(x) = A(n)$. There are $n - n_0 + 1$ integers m such that

$$n_0 \leq m \leq n.$$

As the elements of A are non-negative integers, it follows that if

$$m = a'_1 + \cdots + a'_h \qquad \text{with } a'_k \in A \text{ for } k = 1, \ldots, h$$

then

$$0 \leq a'_k \leq m \leq n \qquad \text{for } k = 1, \ldots, h.$$

The set A contains exactly $A(n)$ positive integers not exceeding n, and A might also contain 0, hence A contains at most $A(n) + 1$ non-negative integers not

exceeding n. As the number of ways to choose h elements with repetitions from a set of cardinality $A(n)+1$ is $\binom{A(n)+h}{h}$, it follows that

$$n+1-n_0 \leq \binom{A(n)+h}{h} < \frac{(A(n)+h)^h}{h!}$$

and so

$$A(x) = A(n) > (h!(n+1-n_0))^{1/h} - h \gg n^{1/h} \gg x^{1/h}$$

for all sufficiently large x. We have $A(a_k) = k$ if $a_1 \geq 1$ and $A(a_k) = k-1$ if $a_1 = 0$, hence

$$k \geq A(a_k) \gg a_k^{1/h}$$

or, equivalently,

$$a_k \ll k^h$$

for all sufficiently large integers k, hence for all positive integers k.

If A is a basis of order h then $1 \in A$ and so $A(n)/n > 0$ for all $n \geq 1$. Therefore, $A(x) \gg x^{1/h}$ for all $x \geq 1$. This completes the proof. □

Let A be a set of non-negative integers. By Theorem 1, if A is an asymptotic basis of order h then $A(x) \gg x^{1/h}$. If A is an asymptotic basis of order h such that

$$A(x) \ll x^{1/h}$$

then A is called a *thin asymptotic basis* of order h. If $hA = \mathbf{N}_0$ and $A(x) \ll x^{1/h}$ then A is called a *thin basis* of order h. In the next section we construct examples of thin bases.

3 Raikov-Stöhr Bases

In 1937 Raikov and Stöhr independently published the first examples of thin bases for the natural numbers. Their construction is based on the fact that every non-negative integer can be written uniquely as the sum of pairwise distinct powers of 2. The sets constructed in the following theorem will be called *Raikov-Stöhr bases*.

Theorem 2 (Raikov-Stöhr). *Let $h \geq 2$. For $i = 0, 1, \ldots, h-1$, let $W_i = \{i, h+i, 2h+i, \ldots\}$ denote the set of all non-negative integers that are congruent to i modulo h, and let $\mathcal{F}(W_i)$ be the set of all finite subsets of W_i. Let*

$$A_i = \left\{ \sum_{f \in F} 2^f : F \in \mathcal{F}(W_i) \right\}$$

and

$$A = A_0 \cup A_1 \cup \cdots \cup A_{h-1}.$$

Then A is a thin basis of order h.

Proof. Note that for all $i = 0, 1, \ldots, h-1$ we have $0 \in A_i$ since $\emptyset \in \mathcal{F}(W_i)$ and $\sum_{f \in \emptyset} 2^f = 0$. This implies that

$$A_0 + A_1 + \cdots + A_{h-1} \subseteq h\left(\bigcup_{i=0}^{h-1} A_i\right) = hA$$

Moreover, $A_i \cap A_j = \{0\}$ if $0 \leq i < j \leq h-1$.

First we show that A is a basis of order h. Every positive integer n is uniquely the sum of distinct powers of two, so we can write

$$n = \sum_{j=0}^{\infty} \varepsilon_j 2^j,$$

where the sequence $\{\varepsilon_j\}_{j=0}^{\infty}$ satisfies $\varepsilon_j \in \{0, 1\}$ for all $j \in \mathbf{N}_0$ and $\varepsilon_j = 0$ for all sufficiently large j. Because

$$\sum_{\substack{j=0 \\ j \equiv i \pmod h}}^{\infty} \varepsilon_j 2^j \in A_i,$$

it follows that

$$\begin{aligned} n &= \sum_{j=0}^{\infty} \varepsilon_j 2^j \\ &= \sum_{i=0}^{h-1} \left(\sum_{\substack{j=0 \\ j \equiv i \pmod h}}^{\infty} \varepsilon_j 2^j \right) \\ &\in A_0 + A_1 + \cdots + A_{h-1} \\ &\subseteq hA \end{aligned}$$

and so A is a basis of order h.

We shall compute the counting functions of the sets A_i and A. Let $x \geq 2^{h-1}$. For every $i \in \{0, 1, \ldots, h-1\}$, there is a unique positive integer r such that

$$2^{(r-1)h+i} \leq x < 2^{rh+i}.$$

If $a_i \in A_i$ and $a_i \leq x$ then there is a set

$$F \subseteq \{i, h+i, \ldots, (r-1)h+i\}$$

such that

$$a_i = \sum_{f \in F} 2^f.$$

The number of such sets F is exactly 2^r. Because $0 \in A_i$, we have

$$A_i(x) \leq 2^r - 1 < 2^r \leq 2^{1-i/h} x^{1/h}$$

and so

$$\begin{aligned} A(x) &= A_0(x) + A_1(x) + \cdots + A_{h-1}(x) \\ &< \left(\sum_{i=0}^{h-1} 2^{1-i/h} \right) x^{1/h} \\ &= \left(\frac{1}{1 - 2^{-1/h}} \right) x^{1/h}. \end{aligned}$$

Thus, A is a thin basis of order h. This completes the proof. □

For $h = 2$, the Raikov-Stöhr construction produces the thin basis $A = A_0 \cup A_1$ of order 2, where

$$A_0 = \{0, 1, 4, 5, 16, 17, 20, 21, 64, 65, 68, 69, 80, 81, 84, 85, 256, \ldots\}$$

is the set of all finite sums of even powers of 2, and

$$A_1 = \{0, 2, 8, 10, 32, 34, 40, 42, 128, 130, 136, 138, 160, 162, 168, 170, 512, \ldots\}$$

is the set of all finite sums of odd powers of 2.

4 Construction of Thin g-adic Bases of Order h

Lemma 1. *Let $g \geq 2$. Let W be a nonempty set of non-negative integers such that*

$$W(x) = \theta x + O(1)$$

for some $\theta \geq 0$ and all $x \geq 1$. Let $\mathcal{F}(W)$ be the set of all finite subsets of W. Let $A(W)$ be the set consisting of all integers of the form

$$a = \sum_{w \in F} e_w g^w, \tag{5}$$

where $F \in \mathcal{F}(W)$ *and* $e_w \in \{0, 1, \ldots, g-1\}$ *for all* $w \in F$. *Then*

$$x^\theta \ll A(W)(x) \ll x^\theta$$

for all sufficiently large x.

Proof. The nonempty set W is finite if and only if $\theta = 0$, and in this case $A(W)$ is also nonempty and finite, or, equivalently, $1 \ll A(W)(x) \ll 1$.

Suppose that $\theta > 0$ and the set W is infinite. Let $W = \{w_i\}_{i=1}^\infty$, where $0 \le w_1 < w_2 < w_3 < \cdots$. Let $\delta = 0$ if $w_1 \ge 1$ and $\delta = 1$ if $w_1 = 0$. For $x \ge g^{w_1}$, we choose the positive integer k so that

$$g^{w_k} \le x < g^{w_k+1}.$$

Then

$$w_k \le \frac{\log x}{\log g} < w_{k+1}$$

and

$$k = W\left(\frac{\log x}{\log g}\right) + \delta = \frac{\theta \log x}{\log g} + O(1),$$

where $W(x)$ is the counting function of the set W.

If $a \in A(W)$ and $a \le x$, then every power of g that appears with a nonzero coefficient in the g-adic representation (5) of a does not exceed g^{w_k}, and so a can be written in the form:

$$a = \sum_{i=1}^{k} e_{w_i} g^{w_i}, \qquad \text{where } e_{w_i} \in \{0, 1, \ldots, g-1\}.$$

There are exactly g^k integers of this form, and so

$$A(W)(x) \le g^k = g^{\frac{\theta \log x}{\log g} + O(1)} \ll x^\theta.$$

Similarly, if a is one of the $g^{k-1} - 1$ positive integers that can be represented in the form:

$$a = \sum_{i=0}^{k-1} e_{w_i} g^{w_i}$$

then

$$a \le \sum_{i=0}^{k-1} (g-1) g^{w_i} \le \sum_{j=0}^{w_{k-1}} (g-1) g^j < g^{w_{k-1}+1} \le g^{w_k} \le x$$

and so

$$A(W)(x) \geq g^{k-1} - 1 \gg x^{\theta}.$$

This completes the proof. □

Theorem 3 (Jia-Nathanson). *Let $g \geq 2$ and $h \geq 2$. Let $W_0, W_1, \ldots, W_{h-1}$ be nonempty sets of non-negative integers such that*

$$\mathbf{N}_0 = W_0 \cup W_1 \cup \cdots \cup W_{h-1}$$

and

$$W_i(x) = \theta_i x + O(1),$$

where $0 \leq \theta_i \leq 1$ for $i = 0, 1, \ldots, h-1$. Let

$$\theta = \max(\theta_0, \theta_1, \ldots, \theta_{h-1}).$$

Let $A(W_0), A(W_1), \ldots, A(W_{h-1})$ be the sets of non-negative integers constructed in Lemma 1. The set

$$A = A(W_0) \cup A(W_1) \cup \cdots \cup A(W_{h-1})$$

is a basis of order h, and

$$A(x) = O\left(x^{\theta}\right).$$

In particular, if

$$W_i(x) = \frac{x}{h} + O(1)$$

for $i = 0, 1, \ldots, h-1$ then $A = A(W_0) \cup A(W_1) \cup \cdots \cup A(W_{h-1})$ is a thin basis of order h.

Note that it is not necessary to assume that the sets $W_0, W_1, \ldots, W_{h-1}$ are pairwise disjoint.

Proof. Every non-negative integer n has a g-adic representation of the form

$$n = \sum_{w=0}^{t} e_w g^w,$$

where $t \geq 0$ and $e_w \in \{0, 1, \ldots, g-1\}$ for $w = 0, 1, \ldots, t$. We define the sets

$$\begin{aligned}
F_0 &= \{w \in \{0, 1, \ldots, t\} : w \in W_0\} \\
F_1 &= \{w \in \{0, 1, \ldots, t\} : w \in W_1 \setminus W_0\} \\
F_2 &= \{w \in \{0, 1, \ldots, t\} : w \in W_2 \setminus (W_0 \cup W_1)\} \\
&\vdots \\
F_{h-1} &= \{w \in \{0, 1, \ldots, t\} : w \in W_{h-1} \setminus (W_0 \cup \cdots \cup W_{h-2})\}.
\end{aligned}$$

Then $F_i \in \mathcal{F}(W_i)$ for all $i = 0, 1, \ldots, h-1$. Since $0 \in A(W_i)$ for $i = 0, 1, \ldots, h-1$, we have

$$n = \sum_{w=0}^{t} e_w g^w = \sum_{i=0}^{h-1} \sum_{w \in F_i} e_w g^w \in A(W_0) + \cdots + A(W_{h-1}) \in hA.$$

Thus, A is a basis of order h.

By Lemma 1,

$$A(W_i)(x) = O\left(x^{\theta_i}\right) = O\left(x^{\theta}\right)$$

for all $i = 0, 1, \ldots, h-1$, and so

$$A(W)(x) \leq \sum_{i=0}^{h-1} A(W_i)(x) = O\left(x^{\theta}\right).$$

If $\theta_i = 1/h$ for all i then $\theta = 1/h$ and A is a thin basis. This completes the proof. □

Consider the case when $W_i = \{w \in \mathbf{N}_0 : w \equiv i \pmod{h}\}$ for $i = 0, 1, \ldots, h-1$. We shall compute an upper bound for the counting functions $A_i(x)$ and $A(x)$. For each i and $x \geq g^i$, choose the positive integer r such that

$$g^{(r-1)h+i} \leq x < g^{rh+i}.$$

Then

$$A_i(x) \leq g^r - 1 < g^r \leq g^{1-(i/h)} x^{1/h}$$

and so

$$A(x) = \sum_{i=0}^{h-1} A_i(x) < \sum_{i=0}^{h-1} g^{1-(i/h)} x^{1/h} = \frac{g-1}{1-g^{-1/h}} x^{1/h}.$$

Applying the mean value theorem to the function $f(x) = x^{1/h}$, we obtain $A(x) < ghx^{1/h}$. In particular, if $g = 2$, we obtain $A(x) < \frac{1}{1-2^{-1/h}} x^{1/h} < 2hx^{1/h}$. This special case is the Raikov-Stöhr construction. For $h = 2$ the Raikov-Stöhr basis $A = \{a_k\}_{k=1}^{\infty}$ with $a_k < a_{k+1}$ for $k \geq 1$ satisfies

$$\frac{A(x)}{\sqrt{x}} < 2 + \sqrt{2} = 3.4142\ldots.$$

Letting $x = a_k$, we obtain

$$\frac{a_k}{k^2} > \frac{3 - 2\sqrt{2}}{2} = 0.0857\ldots.$$

If A is a basis of order h then the order of magnitude of the counting function $A(x)$ must be at least $x^{1/h}$, and there exist thin bases, such as the Raikov-Stöhr

bases and the Jia-Nathanson bases, with exactly this order of magnitude. Two natural constants associated with thin bases of order h are

$$\alpha_h = \inf_{\substack{A \subseteq \mathbf{N}_0 \\ hA = \mathbf{N}_0}} \liminf_{x \to \infty} \frac{A(x)}{x^{1/h}}$$

and

$$\beta_h = \inf_{\substack{A \subseteq \mathbf{N}_0 \\ hA = \mathbf{N}_0}} \limsup_{x \to \infty} \frac{A(x)}{x^{1/h}}$$

Stöhr [16] proved the following lower bound for β_h.

Theorem 4 (Stöhr).

$$\beta_h \geq \frac{\sqrt[h]{h!}}{\Gamma(1 + 1/h)},$$

where $\Gamma(x)$ is the Gamma function.

In particular, $\limsup_{x \to \infty} A(x)/\sqrt{x} \geq \sqrt{8/\pi}$ for every basis A of order 2.

Open Problem 1 *Compute the numbers α_h and β_h for all $h \geq 2$.*

This is an old unsolved problem in additive number theory. Even the numbers α_2 and β_2 are unknown.

5 Asymptotically Polynomial Bases

Let $h \geq 2$, and let $A = \{a_k\}_{k=1}^{\infty}$ be a set of non-negative integers with $a_1 = 0$ and $a_k < a_{k+1}$ for all $k \geq 1$. If A is a basis of order h, then there is a real number λ_2 such that $a_k \leq \lambda_2 k^h$ for all k (Theorem 1). The basis A is called *thin* if there is also a number $\lambda_1 > 0$ such that $a_k \geq \lambda_1 k^h$ for all k. Thus, if A is a thin basis of order h, then there exist positive real numbers λ_1 and λ_2 such that

$$\lambda_1 \leq \frac{a_k}{k^h} \leq \lambda_2$$

for all k. In Theorems 2 and 3 we constructed examples of thin bases of order h for all $h \geq 2$.

The sequence $A = \{a_k\}_{k=0}^{\infty}$ is called *asymptotically polynomial of degree d* if there is a real number $\lambda > 0$ such that $a_k \sim \lambda k^d$ as $k \to \infty$. If A is a basis of order h and if A is also asymptotically polynomial of degree d, then $d \leq h$. We shall describe a beautiful construction of Cassels of a family of additive bases of order h that are asymptotically polynomial of degree h. The key to the construction is the following result, which allows us to embed a sequence of non-negative integers with regular growth into a sequence of non-negative integers with asymptotically polynomial growth.

Theorem 5. *Let $h \geq 2$ and let $A = \{a_k\}_{k=1}^{\infty}$ be a sequence of non-negative integers such that*

$$\liminf_{k\to\infty} \frac{a_{k+1} - a_k}{a_k^{(h-1)/h}} = \alpha > 0.$$

For every real number γ with $0 < \gamma < \alpha$, there exists a sequence $C = \{c_k\}_{k=0}^{\infty}$ of non-negative integers such that C is a supersequence of A and

$$c_k = \left(\frac{\gamma k}{h}\right)^h + O\left(k^{h-1}\right).$$

Proof. Let $B = \{b_k\}_{k=1}^{\infty}$ be a strictly increasing sequence of non-negative integers such that

$$b_k = \left(\frac{\gamma k}{h}\right)^h + O(k^{h-2}).$$

Because $h \geq 2$ and $b_k = (\gamma k/h)^h \left(1 + O\left(k^{-2}\right)\right)$, we have

$$\begin{aligned}
\frac{b_{k+1} - b_k}{b_k^{(h-1)/h}} &= \frac{\left(\frac{\gamma}{h}\right)^h \left((k+1)^h - k^h + O(k^{h-2})\right)}{\left(\frac{\gamma k}{h}\right)^{h-1} \left(1 + O\left(k^{-2}\right)\right)^{(h-1)/h}} \\
&= \frac{\gamma \left(hk^{h-1} + O\left(k^{h-2}\right)\right)}{hk^{h-1} \left(1 + O\left(k^{-2}\right)\right)^{(h-1)/h}} \\
&= \frac{\gamma \left(1 + O\left(k^{-1}\right)\right)}{\left(1 + O\left(k^{-2}\right)\right)^{(h-1)/h}} \\
&= \gamma(1 + o(1))
\end{aligned}$$

and so

$$\lim_{k\to\infty} \frac{b_{k+1} - b_k}{b_k^{(h-1)/h}} = \gamma.$$

Suppose there exist infinitely many k such that, for some integer $m = m(k)$,

$$b_k < a_m < a_{m+1} \leq b_{k+1}.$$

The inequality

$$\frac{b_{k+1} - b_k}{b_k^{(h-1)/h}} > \frac{a_{m+1} - a_m}{b_k^{(h-1)/h}} > \frac{a_{m+1} - a_m}{a_m^{(h-1)/h}}$$

implies that

$$\gamma = \lim_{k\to\infty} \frac{b_{k+1} - b_k}{b_k^{(h-1)/h}} \geq \liminf_{m\to\infty} \frac{a_{m+1} - a_m}{a_m^{(h-1)/h}} \geq \alpha > \gamma$$

which is impossible. Therefore, there exists an integer K such that, for every integer $k \geq K$, the interval $(b_k, b_{k+1}]$ contains at most one element of A.

Choose the integer L such that

$$a_L \leq b_K < a_{L+1}.$$

We define the sequence $C = \{c_k\}_{k=0}^{\infty}$ as follows: Let $c_k = a_k$ for $k = 1, 2, \ldots, L$. For $i \geq 1$, we choose $c_{L+i} \in (b_{K+i-1}, b_{K+i}]$ as follows: If the interval $(b_{K+i-1}, b_{K+i}]$ contains the element a_ℓ from the sequence A then $c_{L+i} = a_\ell$. Otherwise, let $c_{L+i} = b_{K+i}$. As the interval $(b_{K+i-1}, b_{K+i}]$ contains at most one element of A for all $i \geq 1$, and since every element a_k of A with $k > L$ is contained in some interval of the form $(b_{K+i-1}, b_{K+i}]$ with $i \geq 1$, it follows that A is a subsequence of C. Moreover, for every $k \geq L + 1$,

$$b_{k-L+K-1} < c_k \leq b_{k-L+K}.$$

As

$$b_{k-L+K} = \left(\frac{\gamma}{h}\right)^h (k - L + K)^h + O(k^{h-2}) = \left(\frac{\gamma k}{h}\right)^h + O\left(k^{h-1}\right)$$

and, similarly, $b_{k-L+K-1} = (\gamma k/h)^h + O\left(k^{h-1}\right)$, it follows that

$$c_k = \left(\frac{\gamma k}{h}\right)^h + O\left(k^{h-1}\right).$$

This completes the proof. □

6 Bases of Order 2

In this section we describe Cassels' construction in the case $h = 2$. We need the following convergence result.

Lemma 2. *Let $0 < \alpha < 1$. If $\{q_k\}_{k=1}^{\infty}$ is a sequence of positive integers such that*

$$\lim_{k\to\infty} \frac{q_{k-1}}{q_k} = \alpha$$

then

$$\lim_{k\to\infty} \frac{q_1 + q_2 + \cdots + q_k}{q_k} = \frac{1}{1-\alpha}.$$

Proof. For every non-negative integer j we have

$$\lim_{k\to\infty} \frac{q_{k-j}}{q_k} = \lim_{k\to\infty} \prod_{i=0}^{j-1} \frac{q_{k-i-1}}{q_{k-i}} = \alpha^j. \tag{6}$$

Let β be a real number such that $\alpha < \beta < 1$. For every $\varepsilon > 0$ there exists a number $K = K(\beta, \varepsilon)$ such that

$$\frac{q_{k-1}}{q_k} < \beta \qquad \text{for all } k \geq K \tag{7}$$

and

$$\beta^K < \frac{(1-\beta)\varepsilon}{4}. \tag{8}$$

If $k \geq K$ and $k - K = r$ then

$$q_k > \beta^{-1} q_{k-1} > \beta^{-2} q_{k-2} > \cdots > \beta^{-r} q_{k-r} = \beta^{K-k} q_K = c\beta^{-k},$$

where $c = \beta^K q_K > 0$, and so

$$\lim_{k\to\infty} q_k = \infty. \tag{9}$$

If $0 \leq j \leq k - K + 1$, then inequality (7) implies

$$\frac{q_{k-j}}{q_k} = \prod_{i=0}^{j-1} \frac{q_{k-i-1}}{q_{k-i}} < \beta^j.$$

For $k \geq 2K$ we obtain

$$\begin{aligned}
\left| \frac{q_1 + q_2 + \cdots + q_k}{q_k} - \frac{1}{1-\alpha} \right| &= \left| \sum_{j=0}^{k-1} \frac{q_{k-j}}{q_k} - \sum_{j=0}^{\infty} \alpha^j \right| \\
&\leq \sum_{j=0}^{K-1} \left| \frac{q_{k-j}}{q_k} - \alpha^j \right| + \sum_{j=K}^{k-K+1} \frac{q_{k-j}}{q_k} + \sum_{j=k-K+2}^{k-1} \frac{q_{k-j}}{q_k} + \sum_{j=K}^{\infty} \alpha^j \\
&< \sum_{j=0}^{K-1} \left| \frac{q_{k-j}}{q_k} - \alpha^j \right| + \sum_{j=K}^{k-K+1} \beta^j + \sum_{j=1}^{K-2} \frac{q_j}{q_k} + \sum_{j=K}^{\infty} \beta^j \\
&< \sum_{j=0}^{K-1} \left| \frac{q_{k-j}}{q_k} - \alpha^j \right| + \sum_{j=1}^{K-2} \frac{q_j}{q_k} + \frac{2\beta^K}{1-\beta}
\end{aligned}$$

It follows from (6), (9), and (8) that for $j = 0, 1, \ldots, K-1$ and all sufficiently large k

$$\left| \frac{q_{k-j}}{q_k} - \alpha^j \right| < \frac{\varepsilon}{4K}$$

and

$$\frac{q_j}{q_k} < \frac{\varepsilon}{4K}$$

and so

$$\left| \frac{q_1 + q_2 + \cdots + q_k}{q_k} - \frac{1}{1-\alpha} \right| < \varepsilon.$$

This completes the proof. □

Theorem 6. *Let $\{q_i\}_{i=1}^{\infty}$ and $\{m_i\}_{i=1}^{\infty}$ be sequences of positive integers such that*

$$q_1 = 1 \tag{10}$$

and, for all $i \geq 2$,

$$(q_{i-1}, q_i) = (q_{i-1}, q_{i+1}) = 1 \tag{11}$$

$$m_{i-1} \geq q_i + q_{i+1} - 2 \tag{12}$$

and

$$m_{i+1} q_{i+1} \geq m_i q_i + m_{i-1} q_{i-1}. \tag{13}$$

Define the sequences $\{Q_k\}_{k=1}^{\infty}$ of non-negative integers and $\{A_k\}_{k=1}^{\infty}$ of finite arithmetic progressions of non-negative integers by:

$$Q_k = \sum_{i=1}^{k-1} m_i q_i$$

and

$$A_k = Q_k + q_k * [0, m_k].$$

Let

$$A = \bigcup_{k=1}^{\infty} A_k = \{a_n\}_{n=0}^{\infty}$$

where $a_0 = 0 < a_1 < a_2 < \cdots$. Then A is a basis of order 2, and, for every positive integer K, the set $\bigcup_{k=K}^{\infty} A_k$ is an asymptotic basis of order 2.

Let $A(x)$ be the counting function of the set A, and let $M_k = \sum_{i=1}^{k-1} m_i$ for $k \geq 1$. If $M_k \leq n \leq M_{k+1}$, then

$$a_n = Q_k + (n - M_k) q_k. \tag{14}$$

If $Q_k \le x \le Q_{k+1}$, *then*

$$A(x) = M_k + \left[\frac{x - Q_k}{q_k}\right]. \tag{15}$$

Proof. As $Q_{k+1} - Q_k = m_k q_k$, it follows that

$$\{Q_k, Q_{k+1}\} \subseteq A_k \subseteq [Q_k, Q_{k+1}]$$

and

$$A_k = Q_{k+1} - q_k * [0, m_k].$$

Also, $Q_1 = 0$, $Q_2 = m_1 q_1 = m_1$, and $A_1 = [0, m_1]$, hence

$$[2Q_1, 2Q_2] = [0, 2m_1] = 2A_1.$$

We shall prove that

$$[2Q_k, 2Q_{k+1}] \subseteq A_{k-1} + (A_k \cup A_{k+1}) \subseteq 2\,(A_{k-1} \cup A_k \cup A_{k+1}) \tag{16}$$

for all $k \ge 2$.

Let $n \in [2Q_k, 2Q_{k+1}]$. There are two cases. In the first case we have

$$2Q_k \le n \le Q_k + Q_{k+1} - (q_k - 1)q_{k-1}. \tag{17}$$

Because $(q_k, q_{k-1}) = 1$, there is a unique integer r such that

$$n \equiv 2Q_k - rq_{k-1} \pmod{q_k}$$

and, by (12),

$$0 \le r \le q_k - 1 \le m_{k-1}. \tag{18}$$

Then $Q_k - rq_{k-1} \in A_{k-1}$. There is a unique integer s such that

$$sq_k = n - 2Q_k + rq_{k-1}.$$

It follows from (17) and (18) that

$$0 \le n - 2Q_k + rq_{k-1} \le Q_{k+1} - Q_k = m_k q_k,$$

and so

$$0 \le s \le m_k.$$

Therefore, $Q_k + sq_k \in A_k$ and

$$n = (Q_k - rq_{k-1}) + (Q_k + sq_k) \in A_{k-1} + A_k.$$

In the second case we have

$$Q_k + Q_{k+1} - (q_k - 1)q_{k-1} + 1 \leq n \leq 2Q_{k+1}. \tag{19}$$

The set $R = [q_k - 1, q_k + q_{k+1} - 2]$ is a complete set of representatives of the congruence classes modulo q_{k+1}. Because $(q_{k-1}, q_{k+1}) = 1$, it follows that there is a unique integer $r \in R$ such that

$$n \equiv Q_k + Q_{k+1} - rq_{k-1} \pmod{q_{k+1}}.$$

Inequality (12) implies that

$$0 \leq q_k - 1 \leq r \leq q_k + q_{k+1} - 2 \leq m_{k-1} \tag{20}$$

and so $Q_k - rq_{k-1} \in A_{k-1}$. There is a unique integer t such that

$$tq_{k+1} = n - Q_k - Q_{k+1} + rq_{k-1},$$

Inequalities (19), (20), and (13) imply that

$$tq_{k+1} \geq (r - q_k + 1)q_{k-1} + 1 \geq 1$$

and

$$tq_{k+1} \leq Q_{k+1} - Q_k + rq_{k-1} \leq m_k q_k + m_{k-1}q_{k-1} \leq m_{k+1}q_{k+1},$$

and so

$$1 \leq t \leq m_{k+1}.$$

Therefore, $Q_{k+1} + tq_{k+1} \in A_{k+1}$ and

$$n = (Q_k - rq_{k-1}) + (Q_{k+1} + tq_{k+1}) \in A_{k-1} + A_{k+1}.$$

This proves (16). It follows that $\bigcup_{k=1}^{\infty} A_k$ is a basis of order 2. Moreover, for every positive integer K,

$$[2Q_{K+1}, \infty) \subseteq 2\left(\bigcup_{k=K}^{\infty} A_k\right)$$

and so $\bigcup_{k=K}^{\infty} A_k$ is an asymptotic basis of order 2.

Let $A = \{a_n\}_{n=0}^{\infty}$, where $a_0 = 0 < a_1 < a_2 < \cdots$, and let $A(x)$ be the counting function of the set A. Formulas (14) and (15) are immediate consequences of the construction of the set A. This completes the proof. □

Theorem 7. *Let $0 < \alpha < 1$ and let $\{q_i\}_{i=1}^{\infty}$ be a sequence of positive integers with $q_1 = 1$ such that, for all $i \geq 2$,*

$$(q_{i-1}, q_i) = (q_{i-1}, q_{i+1}) = 1 \tag{21}$$

$$q_{i+1}(q_{i+2} + q_{i+3}) \geq q_i(q_{i+1} + q_{i+2}) + q_{i-1}(q_i + q_{i+1}) \tag{22}$$

and

$$\lim_{i\to\infty} \frac{q_{i-1}}{q_i} = \alpha. \tag{23}$$

Define the sequences $\{Q_k\}_{k=1}^{\infty}$ of non-negative integers and $\{A_k\}_{k=1}^{\infty}$ of finite arithmetic progressions of non-negative integers by:

$$Q_k = \sum_{i=1}^{k-1} q_i(q_{i+1} + q_{i+2})$$

and

$$A_k = Q_k + q_k * [0, q_{k+1} + q_{k+2}].$$

Let

$$A = \bigcup_{k=1}^{\infty} A_k = \{a_n\}_{n=0}^{\infty},$$

where $a_0 = 0 < a_1 < a_2 < \cdots$. Then A is a basis of order 2 such that

$$\liminf_{k\to\infty} \frac{a_{n+1} - a_n}{n} \geq \frac{\alpha^2(1-\alpha)}{1+\alpha} > 0.$$

Note that the sequence $\{q_i\}_{i=1}^{\infty}$ of Fibonacci numbers defined by $q_1 = q_2 = 1$ and $q_{i+2} = q_{i+1} + q_i$ for $i \geq 1$ satisfies the conditions of Theorem 7 with $\alpha = (\sqrt{5} - 1)/2$.

Proof. For every integer $i \geq 1$ we define the positive integer $m_i = q_{i+1} + q_{i+2}$. Inequality (22) implies that the sequence $\{m_i\}_{i=1}^{\infty}$ satisfies the hypotheses of Theorem 6, and so A is a basis of order 2. For $k \geq 1$ we define:

$$M_k = \sum_{i=1}^{k-1} m_i = \sum_{i=1}^{k-1} (q_{i+1} + q_{i+2}).$$

Then $\{M_k\}_{k=1}^{\infty}$ is a strictly increasing sequence of positive integers. For every positive integer n there is a unique integer k such that

$$M_k \leq n < M_{k+1}.$$

By (14) we have

$$a_n = Q_k + (n - M_k)q_k$$

and so

$$a_{n+1} - a_n = q_k,$$

hence

$$\frac{a_{n+1} - a_n}{n} = \frac{q_k}{n} > \frac{q_k}{M_{k+1}}.$$

Condition (23) implies that $\lim_{k\to\infty} q_k = \infty$. As

$$\begin{aligned}\frac{M_{k+1}}{q_k} &= \sum_{i=1}^{k} \frac{q_{i+1} + q_{i+2}}{q_k} \\ &= \sum_{i=2}^{k+1} \frac{q_i}{q_k} + \sum_{i=3}^{k+2} \frac{q_i}{q_k} \\ &= 2\sum_{i=1}^{k} \frac{q_i}{q_k} + 2\frac{q_{k+1}}{q_k} + \frac{q_{k+2}}{q_k} - 2\frac{q_1}{q_k} - \frac{q_2}{q_k},\end{aligned}$$

it follows from Lemma 2 that

$$\lim_{k\to\infty} \frac{M_{k+1}}{q_k} = \frac{2}{1-\alpha} + \frac{2}{\alpha} + \frac{1}{\alpha^2} = \frac{1+\alpha}{\alpha^2(1-\alpha)}.$$

Therefore,

$$\liminf_{k\to\infty} \frac{a_{n+1} - a_n}{n} \geq \lim_{k\to\infty} \frac{q_k}{M_{k+1}} = \frac{\alpha^2(1-\alpha)}{1+\alpha} > 0.$$

This completes the proof. □

Theorem 8 (Cassels). *There exist a basis $C = \{c_n\}_{n=0}^{\infty}$ of order 2 and a real number $\lambda > 0$ such that $c_n = \lambda n^2 + O(n)$.*

Proof. By Theorem 7, there exists a basis $A = \{a_n\}_{n=0}^{\infty}$ of order 2 such that $\liminf_{n\to\infty}(a_{n+1} - a_n)/n > 0$. Applying Theorem 1 with $h = 2$, we see that $a_n \ll n^2$ and so $\liminf_{n\to\infty}(a_{n+1} - a_n)/a_n^{1/2} > 0$. Applying Theorem 5 with $h = 2$, we obtain a sequence $C = \{c_n\}_{n=0}^{\infty}$ of non-negative integers and a positive real number λ such that C is a supersequence of A and $c_n = \gamma n^2 + O(n)$. This completes the proof. □

7 Bases of Order $h \geq 3$

We start with Cassels' construction of a finite set C of integers such that the elements of C are widely spaced and C is a basis of order h for a long interval of integers. The construction uses a perturbation of the g-adic representation.

Lemma 3. *Let $h \geq 3$. Let v and L be positive integers with $L \geq h$. Define*

$$g = 2^{h+1}v.$$

Let $C = C(v, L)$ denote the finite set consisting of the following integers:

$$\begin{array}{ll} g^h + eg^{h-1} + 2vg^{h-2} + e & \text{for } 0 \leq e < g, \\ (i+1)g^h + eg^{h-1} + eg^i & \text{for } 0 \leq i \leq h-3 \text{ and } 0 \leq e < g, \\ (h-1)g^h + (4vq + r)g^{h-1} + (4vq + r)g^{h-2} & \text{for } 0 \leq q < 2^{h-1} \text{ and } 0 \leq r < 2v, \\ hg^h + \ell g^{h-1} & \text{for } 0 \leq \ell < Lg. \end{array}$$

Then

(i) The h-fold sumset hC contains every integer n in the interval

$$\left[\left(\frac{h^2+3h-2}{2}\right)g^h, \left(\frac{h(h+1)}{2} + L\right)g^h\right).$$

(ii) If $c \in C$ then

$$g^h \leq c < (h+L)g^h.$$

If $c \geq hg^h$, then $c \equiv 0 \pmod{g^{h-1}}$.

(iii) If $c, c' \in C$ and $c \neq c'$, then

$$|c - c'| \geq vg^{h-2} - g.$$

(iv) If $c \in C$ and y is any integer such that

$$y \equiv -vg^{h-2} \pmod{4vg^{h-2}}$$

then

$$|c - y| \geq vg^{h-2} - g.$$

Proof. (i) Every non-negative integer n has a unique g-adic representation in the form:

$$n = e_{h-1}g^{h-1} + e_{h-2}g^{h-2} + \cdots + e_1 g + e_0, \tag{24}$$

where $e_{h-1} \geq 0$ and

$$0 \leq e_j < g \qquad \text{for } j = 0, 1, \ldots, h-2.$$

If n satisfies the inequality

$$\left(\frac{h^2+3h-2}{2}\right)g^h \leq n < \left(\frac{h(h+1)}{2}+L\right)g^h$$

then e_{h-1} satisfies the inequality

$$\left(\frac{h^2+3h-2}{2}\right)g \leq e_{h-1} < \left(\frac{h(h+1)}{2}+L\right)g. \tag{25}$$

The digit e_{h-2} satisfies the inequality $0 \leq e_{h-2} < g = 4v2^{h-1}$. There are two cases, which depend on the remainder of e_{h-2} when divided by $4v$.

In the first case, we have

$$e_{h-2} = 4vq + r \qquad \text{with } 0 \leq q < 2^{h-1} \text{ and } 0 \leq r < 2v.$$

Rearranging the g-adic representation (24), we obtain

$$\begin{aligned} n = &\left((h-1)g^h + (4vq+r)g^{h-1} + (4vq+r)g^{h-2}\right) \\ &+ \sum_{i=0}^{h-3}\left((i+1)g^h + e_i g^{h-1} + e_i g^i\right) + \left(hg^h + \ell g^{h-1}\right), \end{aligned} \tag{26}$$

where

$$\ell = e_{h-1} - \sum_{i=0}^{h-2} e_i - \frac{h(h+1)g}{2}.$$

Inequality (25) implies that

$$\ell \geq \left(\frac{h^2+3h-2}{2}\right)g - (h-1)(g-1) - \frac{h(h+1)g}{2} = h-1 > 0$$

and

$$\ell < \left(\frac{h(h+1)}{2}+L\right)g - \frac{h(h+1)g}{2} = Lg$$

and so $hg^h + \ell g^{h-1} \in C$. Thus, (26) is a representation of n as the sum of h elements of C, that is, $n \in hC$.

In the second case, we have

$$e_{h-2} = 4vq + r + 2v \qquad \text{with } 0 \leq q < 2^{h-1} \text{ and } 0 \leq r < 2v.$$

From the g-adic representation (24), we obtain

$$\begin{aligned} n = &\left((h-1)g^h + (4vq+r)g^{h-1} + (4vq+r)g^{h-2}\right) \\ &+ \sum_{i=1}^{h-3}\left((i+1)g^h + e_i g^{h-1} + e_i g^i\right) \\ &+ \left(g^h + e_0 g^{h-1} + 2vg^{h-2} + e_0\right) + \left(hg^h + \ell g^{h-1}\right), \end{aligned} \tag{27}$$

where

$$\ell = e_{h-1} - (e_{h-2} - 2v) - \sum_{i=0}^{h-3} e_i - \left(\frac{h(h+1)}{2}\right) g.$$

As in the first case, inequality (25) implies that $0 < h-1 \le \ell < Lg$ and so $hg^h + \ell g^{h-1} \in C$. Thus, (27) is a representation of n as the sum of h elements of C, that is, $n \in hC$. This proves (i).

To prove (ii), we observe that the smallest element of C is g^h and the largest is $hg^h + (Lg-1)g^{h-1} < (h+L)g^h$. If $c \in C$ and $c \ge hg^h$ then $c = hg^h + \ell g^{h-1}$ for some non-negative integer $\ell < Lg$, hence $c \equiv 0 \pmod{g^{h-1}}$.

To prove (iii), we assert that every integer $c \in C$ satisfies an inequality of the form:

$$4svg^{h-2} \le c < (4s+2)vg^{h-2} + g \tag{28}$$

for some non-negative integer s. There are four cases to check.

If $c = g^h + eg^{h-1} + 2vg^{h-2} + e$ with $0 \le e < g$ then we choose $s = 2^{h-1}(g+e)$. Because

$$4svg^{h-2} = g^h + eg^{h-1}$$

and

$$(4s+2)vg^{h-2} + g = g^h + eg^{h-1} + 2vg^{h-2} + g$$

it follows that c satisfies (28).

If $c = (i+1)g^h + eg^{h-1} + eg^i$ with $0 \le e < g$ and $0 \le i \le h-3$ then c satisfies (28) with $s = 2^{h-1}((i+1)g + e)$.

If $c = (h-1)g^h + (4vq+r)g^{h-1} + (4vq+r)g^{h-2}$ with $0 \le q < 2^{h-1}$ and $0 \le r < 2v$ then c satisfies (28) with $s = 2^{h-1}((h-1)g + 4vq + r) + q$.

If $c = hg^h + \ell g^{h-1}$ with $0 \le \ell < Lg$ then c satisfies (28) with $s = 2^{h-1}(hg + \ell)$.

This proves (28). It follows that the distance between elements of C that satisfy inequality (28) for different values of s is at least $2vg^{h-2} - g$. If c and c' are distinct elements of C that satisfy inequality (28) for the same value of s, and if $c' < c$, then we must have

$$0 < c - c' < 2vg^{h-2} + g.$$

This can happen only if $c = g^h + eg^{h-1} + 2vg^{h-2} + e$ and $c' = g^h + eg^{h-1} + e$ with $0 \le e < g$, and so $c - c' = 2vg^{h-2}$. This proves (iii).

Finally, to prove (iv), we observe that if $y \equiv -vg^{h-2} \pmod{4vg^{h-2}}$ then $y = 4s'vg^{h-2} - vg^{h-2}$ for some integer s', and the distance between y and any integer satisfying an inequality of the form (28) is at least $vg^{h-2} - g$. This completes the proof of the Lemma. □

Lemma 4. *For $h \geq 3$, let $v_i = 2^i$ and $g_i = 2^{h+1}v_i = 2^{i+h+1}$ for $i = 0, 1, 2, \ldots$. Then*

$$p_j = \sum_{i=0}^{j} v_i g_i^{h-2} < g_j^h.$$

Proof. We compute p_j explicitly as follows:

$$\begin{aligned} p_j &= \sum_{i=0}^{j} v_i g_i^{h-2} = \sum_{i=0}^{j} 2^i \left(2^{i+h+1}\right)^{h-2} = 2^{(h-2)(h+1)} \sum_{i=0}^{j} 2^{(h-1)i} \\ &= 2^{(h-2)(h+1)} \left(\frac{2^{(h-1)(j+1)} - 1}{2^{h-1} - 1} \right) = \frac{2^{h^2+hj-j-3} - 2^{h^2-h-2}}{2^{h-1} - 1} \\ &< 2^{h(j+h+1)} = g_j^h \end{aligned}$$

because, for $h \geq 3$,

$$\begin{aligned} 2^{h^2+hj-j-3} + 2^{h^2+hj+h} &< 2^{h^2+hj+h+1} < 2^{h^2+hj+2h-1} \\ &< 2^{h^2+hj+2h-1} + 2^{h^2-h-2}. \end{aligned}$$

□

Theorem 9. *Let $h \geq 3$. There exists a strictly increasing sequence $A = \{a_k\}_{k=1}^{\infty}$ of non-negative integers such that A is a basis of order h and*

$$\liminf_{k\to\infty} \frac{a_{k+1} - a_k}{a_k^{(h-1)/h}} \geq \frac{1}{2^{3h-1}}.$$

Proof. Let

$$A(-1) = \left[0, 2^{h^2+2h}\right].$$

We define

$$L = 2^{2h} - h - 1$$

and, for $i = 0, 1, 2, \ldots$,

$$\begin{aligned} v_i &= 2^i \\ g_i &= 2^{h+1}v_i = 2^{i+h+1} \end{aligned}$$

and

$$p_j = \sum_{i=0}^{j} v_i g_i^{h-2}.$$

For $j = 0, 1, 2, \ldots$, let

$$A(j) = p_j + C(v_j, L),$$

where $C(v_j, L)$ is the finite set of positive integers constructed in Lemma 3. We begin by proving that

$$A = \bigcup_{j=-1}^{\infty} A(j)$$

is a basis of order h.

First, we observe that

$$I(-1) = \left[0, h2^{h^2+2h}\right] = hA(-1) \subseteq hA$$

and, by Lemma 3,

$$I(j) = \left[hp_j + \left(\frac{h^2+3h-2)}{2}\right) g_j^h, hp_j + \left(\frac{h(h+1)}{2} + L\right) g_j^h\right) \subseteq hA(j)$$

for $j = 0, 1, 2, \ldots$. As $h^2 + 3h - 2 \leq 2^{h+1}$ for $h \geq 3$, it follows that

$$\begin{aligned} hp_0 + \left(\frac{h^2+3h-2)}{2}\right) g_0^h &= h2^{(h+1)(h-2)} + \left(\frac{h^2+3h-2}{2}\right) 2^{h(h+1)} \\ &\leq h2^{h^2-h-2} + 2^{h^2+2h} \\ &\leq h2^{h^2+2h} \end{aligned}$$

and so the intervals $I(-1)$ and $I(0)$ overlap. Similarly, for $j \geq 0$ the intervals $I(j)$ and $I(j+1)$ overlap if

$$hp_{j+1} + \left(\frac{h^2+3h-2}{2}\right) g_{j+1}^h \leq hp_j + \left(\frac{h(h+1)}{2} + L\right) g_j^h. \tag{29}$$

Because $v_{j+1} = 2v_j$ and $g_{j+1} = 2g_j$, we have

$$p_{j+1} - p_j = v_{j+1} g_{j+1}^{h-2} = 2^{h+j-1} g_j^{h-2} = \frac{g_j^h}{2^{h+j+3}}.$$

Rearranging inequality (29) and dividing by g_j^h, we see that it suffices to prove that

$$\frac{h}{2^{h+j+3}} + \left(\frac{h^2+3h-2}{2}\right) 2^h \leq \frac{(h-2)(h+1)}{2} + 2^{2h}.$$

This follows immediately from the inequalities $h^2 + 3h - 2 \leq 2^{h+1}$ and

$$\frac{h}{2^{h+j+3}} \leq 2 \leq \frac{(h-2)(h+1)}{2}$$

for $j \geq 0$ and $h \geq 3$. Thus, the set A is a basis of order h.

Next, we show that the elements of A are widely spaced. Let $a, a' \in A$ with $a' \neq a$ and $a \in A(j)$ and $a' \in A(j')$ for $j, j' \geq 0$. We shall prove that

$$|a - a'| \geq v_j g_j^{h-2} - g_j.$$

Suppose not. If $j = j'$ then there exist $c, c' \in C(v_j, L)$ with $c \neq c'$ such that $a = p_j + c$ and $a' = p_j + c'$. By Lemma 3 (iii) we have $|a - a'| = |c - c'| \geq v_j g_j^{h-2} - g_j$. Thus, if $|a - a'| < v_j g_j^{h-2} - g_j$ then $j \neq j'$.

The sequences $\{p_j\}_{j=0}^{\infty}$ and $\{g_j\}_{j=0}^{\infty}$ are strictly increasing sequences of positive integers. If $j < j'$, then $v_j g_j^{h-3} < v_{j'} g_{j'}^{h-3}$ and so

$$v_j g_j^{h-2} - g_j = (v_j g_j^{h-3} - 1) g_j < (v_{j'} g_{j'}^{h-3} - 1) g_{j'} = v_{j'} g_{j'}^{h-2} - g_{j'}.$$

Thus, if $j < j'$ and $|a - a'| < v_j g_j^{h-2} - g_j$ then also $|a - a'| < v_{j'} g_{j'}^{h-2} - g_{j'}$. Therefore, without loss of generality, we can assume that $j' < j$.

By Lemma 3 (ii) we have $a \geq p_j + g_j^h$ and $a' < p_{j'} + (h+L) g_{j'}^h$. The inequality $|a - a'| < v_j g_j^{h-2} - g_j$ implies that

$$\begin{aligned} a' &> a - v_j g_j^{h-2} + g_j > p_j + g_j^h - v_j g_j^{h-2} \\ &= p_{j-1} + g_j^h = p_{j-1} + 2^h g_{j-1}^h > p_{j-1} + h g_{j-1}^h. \end{aligned}$$

Combining the upper bound in Lemma 3 (ii) with Lemma 4, we get

$$a' < p_{j'} + (h+L) g_{j'}^h < (h+1+L) g_{j'}^h = 2^{2h} g_{j'}^h = 2^h g_{j'+1}^h = g_{j'+2}^h.$$

As $g_j^h < a' < g_{j'+2}^h$, we see that $j' < j < j' + 2$ and so $j = j' + 1$ and $a' = p_{j-1} + c'$ for some $c' \in C(v_{j-1}, L)$ with $c' \geq h g_{j-1}^h$. By Lemma 3 (ii), we have $c' \equiv 0 \pmod{g_{j-1}^{h-1}}$ and so

$$a' = p_{j-1} + c' \equiv p_{j-1} = p_j - v_j g_j^{h-2} \pmod{g_{j-1}^{h-1}}.$$

Since

$$g_{j-1}^{h-1} = 2^{h+j} g_{j-1}^{h-2} = 4 v_j 2^{h-2} g_{j-1}^{h-2} = 4 v_j g_j^{h-2}$$

it follows that

$$y = a' - p_j \equiv -v_j g_j^{h-2} \pmod{4 v_j g_j^{h-2}}.$$

There exists $c \in C(v_j, L)$ such that $a = p_j + c$. Lemma 3 (iv) implies that

$$|a - a'| = |c - (a' - p_j)| = |c - y| \geq v_j g_j^{h-2} - g_j$$

which is a contradiction. This proves that if $a, a' \in A \setminus A(-1)$ with $a \neq a'$ and $a \in A(j)$, then $|a - a'| \geq v_j g_j^{h-2} - g_j$.

From Lemmas 3 (ii) and 4 we also have

$$a = p_j + c < g_j^h + (h + L)g_j^h = 2^{2h} g_j^h = (4g_j)^h$$

and so $a^{(h-1)/h} < (4g_j)^{h-1}$ and

$$\frac{|a - a'|}{a^{(h-1)/h}} > \frac{v_j g_j^{h-2} - g_j}{(4g_j)^{h-1}} = \frac{v_j}{4^{h-1} g_j} - \frac{1}{4^{h-1} g_j^{h-2}} = \frac{1}{2^{3h-1}} - \frac{1}{4^{h-1} g_j^{h-2}}.$$

Writing A as a strictly increasing sequence $A = \{a_k\}_{k=1}^{\infty}$ of non-negative integers, we obtain

$$\begin{aligned} \liminf_{k \to \infty} \frac{a_{k+1} - a_k}{a_k^{(h-1)/h}} &\geq \liminf_{\substack{a, a' \in A \setminus A(-1) \\ a \neq a'}} \frac{|a - a'|}{a^{(h-1)/h}} \\ &\geq \liminf_{j \to \infty} \left(\frac{1}{2^{3h-1}} - \frac{1}{4^{h-1} g_j^{h-2}} \right) \\ &= \frac{1}{2^{3h-1}}. \end{aligned}$$

This completes the proof. □

Theorem 10 (Cassels). *For every integer $h \geq 3$ there exist a basis $C = \{c_n\}_{n=0}^{\infty}$ of order h and real number $\lambda > 0$ such that $c_n = \lambda n^h + O\left(n^{h-1}\right)$.*

Proof. This follows immediately from Theorems 9 and 5. □

Open Problem 2 *Let $h \geq 2$. Does there exist a basis $C = \{c_n\}_{n=0}^{\infty}$ of order h such that $c_n = \gamma n^h + o(n^{h-1})$ for some $\gamma > 0$?*

Open Problem 3 *Let $h \geq 2$. Does there exist a basis $C = \{c_n\}_{n=0}^{\infty}$ of order h such that $c_n = \gamma n^h + O(n^{h-2})$ for some $\gamma > 0$?*

Open Problem 4 *Let $h \geq 2$. Compute or estimate*

$$\sup\{\lambda > 0 : \text{ there exists a basis } C = \{c_n\}_{n=0}^{\infty} \text{ of order } h \text{ such that } c_n \sim \lambda n^h\}.$$

8 Notes

Raikov [13] and Stöhr [15] independently constructed the first examples of thin bases of order h. Another early, almost forgotten construction of thin bases is due to Chartrovsky [3]. The g-adic generalization of the Raikov-Stöhr construction appears in work of Jia and Nathanson [8, 9] on minimal asymptotic bases. The currently "thinest" bases of finite order appear in recent articles by Hofmeister [7] and Blomer [1]. An old but still valuable survey of combinatorial problems in additive number theory is Stöhr [16].

The classical bases in additive number theory are the squares, cubes, and, for every integer $k \geq 4$, the kth powers of non-negative integers, and also the sets of polygonal numbers and of prime numbers. Using probability arguments, one can prove that all of the classical bases contain thin subsets that are bases of order h for sufficiently large h (Choi-Erdős-Nathanson [4], Erdős-Nathanson [5], Nathanson [10], Wirsing[18], and Vu [17]).

The construction in this paper of polynomially asymptotic thin bases of order h appeared in the classic article of Cassels [2] in 1957. There is a recent quantitative improvement by Schmitt [14], and also related work on Cassels bases by Grekos, Haddad, Helou, and Pihko [6] and Nathanson [12].

References

1. V. Blomer, *Thin bases of order h*, J. Number Theory **98** (2003), 34–46.
2. J. W. S. Cassels, *Über Basen der natürlichen Zahlenreihe*, Abhandlungen Math. Seminar. Univ. Hamburg **21** (1975), 247–257.
3. L. Chatrovsky, *Sur les bases minimales de la suite des nombres naturels*, Bull. Acad. Sci. URSS. Sér. Math. [Izvestia Akad. Nauk SSSR] **4** (1940), 335–340.
4. S. L. G. Choi, P. Erdős, and M. B. Nathanson, *Lagrange's theorem with $N^{1/3}$ squares*, Proc. Am. Math. Soc. **79** (1980), 203–205.
5. P. Erdős and M. B. Nathanson, *Lagrange's theorem and thin subsequences of squares*, Contributions to probability, Academic Press, New York, 1981, pp. 3–9.
6. G. Grekos, L. Haddad, C. Helou, and J. Pihko, *Variations on a theme of Cassels for additive bases*, Int. J. Number Theory **2** (2006), no. 2, 249–265.
7. G. Hofmeister, *Thin bases of order two*, J. Number Theory **86** (2001), no. 1, 118–132.
8. X.-D. Jia, *Minimal bases and g-adic representations of integers*, Number Theory: New York Seminar 1991-1995 (New York), Springer-Verlag, New York, 1996, pp. 201–209.
9. X.-D. Jia and M. B. Nathanson, *A simple construction of minimal asymptotic bases*, Acta Arith. **52** (1989), no. 2, 95–101.
10. M. B. Nathanson, *Waring's problem for sets of density zero*, Analytic number theory (Philadelphia, PA, 1980), Lecture Notes in Math., vol. 899, Springer, Berlin, 1981, pp. 301–310.
11. M. B. Nathanson, *Additive Number Theory: The Classical Bases*, Graduate Texts in Mathematics, vol. 164, Springer-Verlag, New York, 1996.
12. M. B. Nathanson, *Supersequences, rearrangements of sequences, and the spectrum of bases in additive number theory*, J. Number Theory **129** (2009), 1608–1621.
13. D. Raikov, *Über die Basen der natürlichen Zahlentreihe*, Mat. Sbornik N. S. 2 **44** (1937), 595–597.

14. C. Schmitt, *Uniformly thin bases of order two*, Acta Arith. **124** (2006), no. 1, 17–26.
15. A. Stöhr, *Eine Basis h-Ordnung für die Menge aller natürlichen Zahlen*, Math. Zeit. **42** (1937), 739–743.
16. A. Stöhr, *Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe. I, II*, J. Reine Angew. Math. **194** (1955), 40–65, 111–140.
17. Van H. Vu, *On a refinement of Waring's problem*, Duke Math. J. **105** (2000), no. 1, 107–134.
18. E. Wirsing, *Thin subbases*, Analysis **6** (1986), 285–308.

Asymptotics of Weighted Lattice Point Counts Inside Dilating Polygons

Marina Nechayeva and Burton Randol

Summary We study the family of normalized discrete measures induced on the unit circle by radially projecting onto the circle the integral lattice points contained in dilations of a fixed polygon satisfying certain algebraic properties. We examine the asymptotic effect of such measures on a function f on S^1 by weighting the lattice points and their projections by a homogeneous extension of f to R^2. We then derive an almost everywhere result for almost all rotations of the polygon.

Keywords Lattice point asymptotics · Polygons

Mathematics Subject Classifications (2010). Primary 11P21, 11K60

1 Introduction

A generalization of the classical lattice point problem, which has found recent applications in string theory [1], as well as being of considerable intrinsic mathematical interest in its own right, occurs when the lattice points in a dilating domain ρD in R^n are weighted with a homogeneous function which is not necessarily constant. The homogeneity requirement on a weighting function is quite natural, since such functions are exceptionally well adapted to the discussion of dilations, and in our discussions of aspects of this subject, we will require that the weighting function be of homogeneous weight $\alpha \geq 0$. Such a weighted lattice point count is obviously equal to the total measure on S^{n-1} produced by the weighted radial projections onto S^{n-1} of lattice points contained in ρD. The case $\alpha = 0$ leads to the consideration of

M. Nechayeva
LaGuardia Community College, 31-10 Thomson Avenue Long Island City, NY 11101, USA
e-mail: mnechayeva@lagcc.cuny.edu

B. Randol
Graduate Center of CUNY, 365 Fifth Avenue, New York, NY 10016, USA
e-mail: brandol@gc.cuny.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_20,

the integral of a function f on S^{n-1} with respect to the discrete measure on S^{n-1} given by radial projections of the lattice points in ρD, which clearly corresponds to the sum of the lattice points in ρD, weighted by the 0-homogeneous extension of f to R^n.

It can be shown (cf. [4]), that this quantity will, under suitable restrictions on D, and after division by ρ^n, tend to

$$\frac{1}{n}\int_{S^{n-1}} f(\theta)m(\theta)\,\mathrm{d}\theta\,,$$

where $\mathrm{d}\theta$ is Lebesgue measure on S^{n-1}, and $m(\theta)$ is defined by the requirement that the boundary ∂D of D is given in polar coordinates by $r = (m(\theta))^{1/n}$. Put another way, suitably normalized discrete measures produced by radial projection onto S^{n-1} of atomic measures at the lattice points in ρD, converge weakly to the measure $\frac{1}{n}\,m(\theta)\,\mathrm{d}\theta$ on S^{n-1}. This fact can serve as the basis for a discrete approximation to the effect of the measure $m(\theta)\mathrm{d}\theta$ on a test function $f(\theta)$, the efficiency of which, over a suitable suite of test functions, depends on the rapidity of convergence of this discrete process to the above integral. It is of course clear that the rate at which the process converges will depend on the function m, or equivalently, on the geometric nature of the boundary of D as well as on the class of test functions.

In this paper, we will employ a general method of analysis, developed in [4–7] (cf. also [1], where the case in which ∂D has positive curvature is discussed using somewhat different techniques), to study this question in the case in which the region being dilated is a polygon. In greater detail, we will discuss the relevant asymptotics when D is a polygon satisfying a natural algebraic condition. As we mention at an appropriate point in the paper, the algebraic hypothesis could be replaced by a Diophantine condition which subsumes it, but we carry out the analysis in the algebraic case, because of its exceptional importance and because it generically illustrates the techniques.

In higher dimensions, the general approach we employ applies to polyhedra, but the combinatorics become extremely formidable, and we defer their investigation for now. For treatments of some aspects of the constant-weight higher-dimensional polyhedral case, see, for example, [8, 11, 12].

2 The Algebraic Case

We will now take up the case of an algebraic polygon, and will initially assume that the weighting function has a high degree of homogeneity, since it will be convenient for it to have a certain degree of smoothness at the origin. We will then use a standard Stieltjes argument to generalize our result to an arbitrary weight, in particular, to weight zero.

We will also require that the polygon contains the origin, and that the normals to its sides be poorly approximable in the sense of Diophantine approximation. That is,

roughly speaking, the lines they determine cannot come too close to integral lattice points, when the approach is measured as a function of distance from the origin. In other terms, this condition is intended to ensure, in a sense to be made precise later, that the slope of a normal to a side cannot be approximated very well by rationals.

This is the case, for example, for polygons for which the normals to the sides have algebraic slopes; a fact that is a direct consequence of Roth's Theorem [9], which asserts that if γ is real algebraic, then for any $\varepsilon > 0$, the number of solutions of

$$\left|\gamma - \frac{p}{q}\right| \leq q^{-(2+\varepsilon)}$$

is finite. We will refer to such polygons as algebraic.

Remark: In the following discussion, the critical hypothesis will be that the normals to the sides of the polygon are poorly approximable, as in Roth's theorem. There are, of course, many numbers besides algebraic ones which satisfy such a condition, but as previously mentioned, we have chosen to illustrate the argument in the algebraic case, because of its exceptional interest and generic character.

Now let $f(\theta)$ be a smooth function on S^1, and let F be the weight-α homogeneous extension of f to R^2, given in polar coordinates by $\mathrm{F}(r, \theta) = r^\alpha f(\theta)$. Assume for now that α is large enough to ensure that F is sufficiently smooth at the origin to justify subsequent arguments. We will study the asymptotic behavior of the F-weighted lattice point count over dilates ρD of the polygon D.

Let χ_ρ denote the indicator function of the dilated polygon and let $F_\rho = \mathrm{F} \cdot \chi_\rho$ (we will write F in place of F_1.) The F-weighted lattice-point count, $\mathfrak{N}(\rho)$, inside the dilated polygon is given by

$$\mathfrak{N}(\rho) = \sum_{N \in \mathbb{Z}^2} F_\rho(N) = \int_{R^2} F_\rho + R(\rho) = \rho^{2+\alpha} \int_D F + R(\rho), \tag{1}$$

Our goal is to estimate the magnitude of $R(\rho)$ as $\rho \to \infty$. The form of (1) suggests the use of the Poisson Summation Formula, however F_ρ is not a smooth function. To overcome this problem, we will use convolution to create a parameterized family of C^∞ functions that are suitably close to F_ρ and whose sum over lattice points will approximate $\mathfrak{N}(\rho)$.

We begin with δ, a non-negative C^∞ radial function supported on the unit disc and such that $\int_{R^2} \delta = 1$, and define a family of functions $F_\rho * \delta_\varepsilon$, where

$$\delta_\varepsilon(X) = \frac{1}{\varepsilon^2} \delta\left(\frac{X}{\varepsilon}\right).$$

Summing $F_\rho * \delta_\varepsilon$ over lattice points, we get a modified lattice-point count

$$\mathfrak{N}_\varepsilon(\rho) = \sum_{N \in \mathbb{Z}^2} F_\rho * \delta_\varepsilon(N).$$

Now, since $F_\rho * \delta_\varepsilon$ is C^∞, the Poisson summation formula can be applied to estimate the convergent sum above as

$$\mathfrak{N}_\varepsilon(\rho) = \sum_{N\in\mathbb{Z}^2} \widehat{F}_\rho(N)\widehat{\delta}_\varepsilon(N) = \int F_\rho + R_\varepsilon(\rho), \tag{2}$$

where the principal term $\int F_\rho = \rho^{2+\alpha}\int_D F$ comes from the lattice point at the origin and, since $\widehat{F}_\rho(N) = \rho^{\alpha+2}\widehat{F}(\rho N)$ and $\widehat{\delta}_\varepsilon(N) = \widehat{\delta}(\varepsilon N)$, the sum of the remaining terms may be expressed as

$$R_\varepsilon(\rho) = \rho^{2+\alpha} \sum_{N\in\mathbb{Z}^2\setminus(0,0)} \widehat{F}(\rho N)\widehat{\delta}(\varepsilon N). \tag{3}$$

We will later show exactly how $\mathfrak{N}_\varepsilon(\rho)$ serves as an estimate for $\mathfrak{N}(\rho)$. At the moment, comparison of (1) and (2) suggests that $R_\varepsilon(\rho)$ will be useful in the study of $\mathrm{R}(\rho)$.

To obtain an estimate for $R_\varepsilon(T)$, we need to address the asymptotics of $\widehat{F}$.

Using the divergence theorem and a lemma in [5, pp. 260–261], we get

$$\widehat{F}(Y) = \int_D \mathrm{e}^{2\pi i(X,Y)} F(X)\mathrm{d}X = \frac{1}{2\pi i|Y|}\int_{\partial D} \mathrm{e}^{2\pi i(X,Y)}(n(X), G(X))\,\mathrm{d}S_X,$$

where ∂D is the boundary of the polygon D, $n(X)$ is the exterior normal to ∂D at X, and $G(X)$ is a vector field on R^2, such that

$$\mathbf{div}[(2\pi i|Y|)^{-1}\mathrm{e}^{2\pi i(X,Y)}G(X)] = \mathrm{e}^{2\pi i(X,Y)}F(X)\,.$$

Now, since $n(X)$, restricted to any side S of the polygon is a constant, and since the lemma from [5] ensures that derivatives of the components of $G(X)$ can, up to a level depending on the smoothness of F, be bounded on the boundary of D independently of Y, we conclude that $\widehat{F}(Y)$ can be expressed as a finite linear combination of terms like

$$H(Y) = \frac{1}{|Y|}\int_S g(X)\,\mathrm{e}^{2\pi i(X,Y)}\mathrm{d}S_X,$$

where the derivatives of $g(X)$ are uniformly controlled in Y.

From this and (3), it follows that

$$R_\varepsilon(\rho) \ll \rho^{2+\alpha} \sum_{N\in\mathbb{Z}^2\setminus(0,0)} H(\rho N)\widehat{\delta}(\varepsilon N). \tag{4}$$

We will now proceed to obtain an estimate of H(Y) that will be useful when ε is small.

Let ψ denote the acute positive angle formed by Y and the line through the normal to the side S. Since translation and rescaling of a side contribute a constant multiple to the line integral above, and since inner product is rotation invariant, we get:

$$H(Y) = \frac{1}{|Y|}\int_{-1}^{1} h(x)\, \mathrm{e}^{i|Y|x\sin\psi}\,\mathrm{d}x \tag{5}$$

where h corresponds to g in the obvious way, and straightforward integration by parts leads to an estimate

$$H(Y) \ll \frac{1}{|Y|^2\psi}, \tag{6}$$

uniformly in Y.

This estimate will not quite be enough to cover the whole range of summation in (4), since our approach to estimating (4) will require replacing part of the infinite sum with a corresponding convergent integral, and $1/\psi$ is not an L^1 function in a neighborhood of $\psi = 0$. If, however, we are interested in an estimate in the complement of a band or strip centered around the y-axis, we can effect a kind of trade-off, accepting a worse estimate for $|Y|$ in exchange for a better estimate for ψ. In more detail, if $|Y|$ lies in the complement of such a strip, then at worst, on the boundary of the strip, ψ is of the order of $1/|Y|$, and becomes larger as we move away from the boundary. Thus, for example, if, for small $\gamma > 0$, in (6) we replace one of the $|Y|$ factors by $|Y|^{1-\gamma}$, and the ψ factor by $\psi^{1-\gamma}$, we find that in the complement of the strip,

$$H(Y) \ll \frac{1}{|Y|^{2-\gamma}\psi^{1-\gamma}}\,. \tag{7}$$

It is, of course, possible to use other tradeoffs between $|Y|$ and ψ in order to prove an estimate for $H(Y)$ in the complement of a band of the described type. For example, in a subsequent section, we will use the estimate

$$H(Y) \ll \frac{\log^{1+\gamma}|Y|}{(|Y|^2\psi)(\log^{1+\gamma}(1/\psi))}\,. \tag{8}$$

In the following, we will combine (6) and (7) with (4) to derive an estimate for $R_\varepsilon(\rho)$.

As we have mentioned, the form of (4) suggests the possibility of deriving an estimate for $R_\varepsilon(\rho)$ by replacing part of the sum with a corresponding convergent integral, and it is with that in mind that we have produced an L^1 estimate for $H(Y)$.

However, the desired substitution must take place in a region that can be covered by suitable neighborhoods of lattice points, so that the value of $H(Y)$ at any lattice point in the region is uniformly comparable with the value of $H(Y)$ integrated over

the neighborhood of that point. This presents a problem, since $H(Y)$ explodes as Y gets close to the normal to a side of the polygon, and there can be lattice points lying very close to these "bad" directions.

With this in view, we will partition the plane into a "good" region G, within which $H(Y)$ is such that the summation can be replaced with an integral, and a "bad" region B, where another method of summation will have to be employed.

A partition corresponding to a side S over which the Fourier transform is computed is defined as follows. Take B to be a closed strip of width $2 + \sqrt{2}$ whose middle line runs through the origin and is perpendicular to the side S of the polygon, and let G be the complement of B. Using our previous notation, we have:

$$G = \left\{ Y \ : \ |Y| \sin\psi \geq 1 + \frac{\sqrt{2}}{2} \right\} \text{ and } B = R^2 \backslash G \cup \{(0,0)\}. \tag{9}$$

The purpose of the following lemma is to ascertain that G is indeed a "good" region in the above sense, where suitable neighborhoods for lattice points in G are closed unit squares centered at those points and with sides parallel to the coordinate axes.

Lemma 1. *Let $H(Y)$ be as above. Let $N \in Z^2 \cap G$. Let Q be a closed unit square centered at N with sides parallel to the coordinate axes.*

Claim:

$$H(N) \ll \int_Q \frac{1}{|Y|^{2-\gamma} \psi^{1-\gamma}}.$$

By (7), this will allow us to replace the sum over G by an integral.

Proof. It is easily checked that the quotient of any two values of

$$\frac{1}{|Y|^{2-\gamma} \psi^{1-\gamma}}$$

within a Q of the described type is bounded independently of $N \in G$, which establishes the lemma. □

Thus, we have established that G is a "good" region. We can therefore estimate $R_\varepsilon^G(\rho)$, the contribution to $R_\varepsilon(\rho)$ which comes from lattice points located in G, as proposed above.

We start with subdividing G into a bounded and an unbounded part, namely $G_1 = \{Y \in G : |Y| < \frac{1}{\varepsilon}\}$ and $G_2 = G \setminus G_1$ and defining, for $i = 1, 2$

$$S_i = \rho^2 \sum_{N \in G_i} H(\rho N) \widehat{\delta_\varepsilon}(N).$$

We will estimate each S_i separately, using estimate (7) and replacing sums with appropriate integrals. First, since $\widehat{\delta_\varepsilon}(N)$ is bounded, we get:

$$S_1 \ll \rho^2 \sum_{N\in G_1} H(\rho N) \ll \rho^2 \sum_{N\in G_1} \frac{1}{(\rho|N|)^{2-\gamma}} \frac{1}{\psi^{1-\gamma}}$$

$$\ll \rho^\gamma \int_0^{2\pi}\int_1^{1/\varepsilon} \frac{1}{r^{2-\gamma}} \frac{1}{\psi^{1-\gamma}} r \,\mathrm{d}r\,\mathrm{d}\psi \ll \rho^\gamma r^\gamma]_1^{1/\varepsilon} \ll \left(\frac{\rho}{\varepsilon}\right)^\gamma.$$

Now, the contribution of the $\widehat{\delta_\varepsilon}(N)$ factor, immaterial in case of S_1, becomes essential for the convergence of S_2. Since $\delta_\varepsilon(N)$ is C^∞, $\widehat{\delta_\varepsilon}(N)$ deteriorates very rapidly as $N \to \infty$, but we only need to use $\widehat{\delta_\varepsilon}(N) \ll \frac{1}{\varepsilon|N|}$ to obtain:

$$S_2 \ll \rho^2 \sum_{N\varepsilon G_2} \frac{1}{(\rho|N|)^{2-\gamma}} \frac{1}{\psi^{1-\gamma}} \frac{1}{\varepsilon|N|}$$

$$\ll \frac{\rho^\gamma}{\varepsilon} \int_0^{2\pi}\int_{1/\varepsilon}^{\infty} \frac{1}{r^{3-\gamma}} \frac{1}{\psi^{1-\gamma}} r \,\mathrm{d}r\,\mathrm{d}\psi \ll \frac{\rho^\gamma}{\varepsilon} r^{\gamma-1}]_{1/\varepsilon}^{\infty} \ll \left(\frac{\rho}{\varepsilon}\right)^\gamma.$$

Thus, we now have the following result:

$$R_\varepsilon^G(\rho) \ll \rho^{2+\alpha} \sum_{N\in G} H(\rho N)\widehat{\delta_\varepsilon}(N) \ll \rho^\alpha(S_1+S_2) \ll \rho^\alpha \left(\frac{\rho}{\varepsilon}\right)^\gamma. \tag{10}$$

Our next task will be to estimate $R_\varepsilon^B(\rho)$, the contribution to $R_\varepsilon(\rho)$ which comes from lattice points located in the strip B.

Now, since $|N|\sin\psi$, the distance from the lattice point N to the vector normal to the face, can be arbitrarily small within B, and hence $F(N)$ can get uncontrollably large, we will need a method different from the one just employed.

At this point, the so far unused requirement that the normal to a side of the polygon, or equivalently the slope λ of the side, be poorly approximable (for example, algebraic), is called upon. In the case of an algebraic polygon, Roth's Theorem applies and from it we derive that, given any $\delta > 0$, there is a constant c, which depends only on δ, such that for any integer q,

$$\langle q\lambda\rangle > \frac{c}{q^{1+\delta}},$$

where $\langle q\lambda\rangle$ stands for the distance from $q\lambda$ to the nearest integer.

This poor approximability of λ will play a crucial part in the estimate we are after, since for any lattice point $N = (p,q)$ we have $|N|\sin\psi \gg |q||\lambda - p/q|$, and so $|N|^2\sin\psi \gg |q|\langle q\lambda\rangle$, which produces, (using the previously derived estimate $H(Y) \ll \frac{1}{|Y|^2\sin\psi}$),

$$H(\rho N) \ll \frac{1}{\rho^2|N|^2\sin\psi} \ll \frac{1}{\rho^2|q|\langle q\lambda\rangle}.$$

This, combined with the fact that for all integers q, there exists a constant c such that $|\{p \in \mathbb{Z} : (p,q) \in B\}| < c$, and the earlier estimate $\widehat{\delta_\varepsilon}(N) \ll \frac{1}{1+\varepsilon|N|} \ll \frac{1}{1+\varepsilon|q|}$, yields

$$R_\varepsilon^B(\rho) \ll \rho^{2+\alpha} \sum_{N\in B} H(\rho N)\widehat{\delta_\varepsilon}(N) \ll \rho^\alpha \sum_{q=1}^{\infty} \frac{1}{q\langle q\lambda\rangle}\frac{1}{1+\varepsilon q}\,.$$

As before, we will divide the sum above into two parts, to be estimated separately, namely

$$S_1 = \sum_{q=1}^{m} \frac{1}{q\langle q\lambda\rangle}\frac{1}{1+\varepsilon q} \ll \sum_{q=1}^{m} \frac{1}{q\langle q\lambda\rangle} \tag{11}$$

and

$$S_2 = \sum_{q=m+1}^{\infty} \frac{1}{q\langle q\lambda\rangle}\frac{1}{1+\varepsilon q} \ll \frac{1}{\varepsilon}\lim_{M\to\infty}\sum_{q=m+1}^{M} \frac{1}{q^2\langle q\lambda\rangle}, \tag{12}$$

where m is the largest integer smaller then $\frac{1}{\varepsilon}$.

Lemma 2. *Let* $s_k = \sum_{q=1}^{k} \frac{1}{\langle q\lambda\rangle}$. *Suppose* $\langle q\lambda\rangle > \frac{c}{q^{1+\delta}}$, *for some* $c, \delta > 0$.

Claim $s_k \ll k^{1+2\delta}$.

Proof. This is an immediate consequence of the penultimate line in the proof of Lemma 3.3 of [3], (p. 123). □

By Roth's theorem, the hypothesis of the lemma above holds for any $\delta > 0$. And so, we conclude that for any $\gamma > 0, s_k \ll k^{1+\gamma}$.

Now, applying partial summation and the above result to (11) we arrive at

$$S_1 \ll \left(\sum_{k=1}^{m} \frac{s_k}{k(k+1)}\right) + \frac{s_m}{m+1} \ll \sum_{k=1}^{m} \frac{k^\gamma}{k} + m^\gamma \ll \int_1^{\frac{1}{\varepsilon}} \frac{x^\gamma}{x}\mathrm{d}x + \left(\frac{1}{\varepsilon}\right)^\gamma \ll \left(\frac{1}{\varepsilon}\right)^\gamma.$$

We now proceed to estimate (12). Using the same tools as above, we first obtain

$$S_2^M = \sum_{q=m+1}^{M} \frac{1}{q^2\langle q\lambda\rangle} = \left(\sum_{q=m+1}^{M} \frac{2k+1}{k^2(k+1)^2} s_k\right) + \frac{s_M}{(M+1)^2}$$

$$\ll \sum_{k=m+1}^{M} \frac{k^\gamma}{k^2} + \frac{M^\gamma}{M} \ll \int_{\frac{1}{\varepsilon}}^{M} \frac{x^\gamma}{x^2}\mathrm{d}x + \frac{M^\gamma}{M}$$

Then, taking the limit as M approaches ∞ we get

$$S_2 \ll \frac{1}{\varepsilon}\int_{\frac{1}{\varepsilon}}^{\infty} \frac{x^\gamma}{x^2}\mathrm{d}x \ll \left(\frac{1}{\varepsilon}\right)^\gamma.$$

Putting it all together, we get an estimate for the contribution to the $R_\varepsilon(\rho)$ coming from lattice points in the "bad" region B.

$$R_\varepsilon^B(\rho) \ll \rho^\alpha(S_1 + S_2) \ll \rho^\alpha \left(\frac{1}{\varepsilon}\right)^\gamma. \tag{13}$$

And finally, combining (10) with (13), we derive

$$R_\varepsilon(\rho) \ll \rho^\alpha \left(\frac{\rho}{\varepsilon}\right)^\gamma.$$

and from (2) obtain for any $\gamma > 0$

$$\mathfrak{N}_\varepsilon(\rho) = \rho^{2+\alpha} \int_D F + O\left(\rho^\alpha \left(\frac{\rho}{\varepsilon}\right)^\gamma\right). \tag{14}$$

We are now ready to derive an estimate for $\mathfrak{N}(\rho)$, the F-weighted lattice point count within ρD.

Observe that for large enough ρ, there exists a constant c, independent of ρ and ε, such that $(\rho+c\varepsilon)D$ contains $\rho\mathrm{D} + B_\varepsilon$ and $(\rho-c\varepsilon)D + B_\varepsilon$ is contained in ρD, where B_ε is a disc of radius ε. (In fact, for any $c > 1/|z|$, where z is a point on ∂D closest to the origin, the above condition is satisfied.)

In the case where F is a constant function, we then have, for all $N\varepsilon R^2$, $F_{\rho-c\varepsilon} * \delta_\varepsilon(N) \le F_\rho(N) \le F_{\rho+c\varepsilon} * \delta_\varepsilon(N)$, and summing over lattice points, we obtain

$$\mathfrak{N}_\varepsilon(\rho - c\varepsilon) \le \mathfrak{N}(\rho) \le \mathfrak{N}_\varepsilon(\rho + c\varepsilon).$$

The above inequality does not necessarily hold for non-constant F. But we can modify it to fit the general case as follows. Note that any Y in ρD we have $F_{\rho+c\varepsilon} * \delta_\varepsilon(Y) = F_{\rho+c\varepsilon}(Z)$ for some $Z \in B_{\varepsilon(Y)}$ and so $|F_\rho(Y) - F_{\rho+c\varepsilon} * \delta_\varepsilon(Y)|$ is bounded above by the oscillation of F over $B_\varepsilon(Y)$. The latter is in turn bounded by a product of 2ε with the maximum absolute value of the derivative of F on $B_\varepsilon(Y)$, and since the derivative of F has weight $\alpha - 1$, we obtain

$$|F_\rho(Y) - F_{\rho+c\varepsilon} * \delta_\varepsilon(Y)| \ll \varepsilon\rho^{\alpha-1}.$$

Thus, for every $N\ \varepsilon\ \rho D$, we get

$$F_{\rho-c\varepsilon} * \delta_\varepsilon(N) + O\left(\varepsilon\rho^{\alpha-1}\right) \le F_\rho(N) \le F_{\rho+c\varepsilon} * \delta_\varepsilon(N) + O\left(\varepsilon\rho^{\alpha-1}\right),$$

where the adjusting constants do not depend on N. Now, summing over lattice points, we arrive at

$$\mathfrak{N}_\varepsilon(\rho-c\varepsilon) + O\left(\varepsilon\rho^{\alpha-1}\right)\sum \chi_\rho(N) \le \mathfrak{N}(\rho) \le \mathfrak{N}_\varepsilon(\rho+c\varepsilon) + O\left(\varepsilon\rho^{\alpha-1}\right)\sum \chi_\rho(N),$$

and using the estimate $\sum \chi_\rho(N) \ll \rho^2$, we end up with

$$\mathfrak{N}_\varepsilon(\rho - c\varepsilon) + O\left(\varepsilon\rho^{\alpha+1}\right) \leq \mathfrak{N}(\rho) \leq \mathfrak{N}_\varepsilon\left(\rho + c\varepsilon\right) + O\left(\varepsilon\rho^{\alpha+1}\right).$$

Combining the last inequality with (1) and (14) and subtracting the principal term $\rho^{2+\alpha}\int_D F$, we obtain the following estimate:

$$R(\rho) \ll (\rho + c\varepsilon)^{2+\alpha} - \rho^{2+\alpha} + \rho^\alpha\left(\frac{\rho}{\varepsilon}\right)^\gamma + \varepsilon\rho^{\alpha+1}.$$

Now, using the binomial expansion, we derive

$$R(\rho) \ll \rho^\alpha\left(\rho\varepsilon + \left(\frac{\rho}{\varepsilon}\right)^\gamma\right), \tag{15}$$

and setting $\varepsilon = \rho^{-1}$ we finally produce the estimate

$$R(\rho) \ll \rho^{\alpha+2\gamma}$$

for any $\gamma > 0$, which amounts to the same thing as

$$R(\rho) \ll \rho^{\alpha+\gamma},$$

for any $\gamma > 0$. (We have not insisted on choosing ε in (15) to precisely balance the two terms, since to do so would not change the final result).

We now conclude that the F-weighted lattice-point count inside an algebraic polygon D dilated by ρ is given by

$$\mathfrak{N}(\rho) = \sum_{N\in\mathbb{Z}^2} F_\rho(N) = \int F_\rho + R(\rho) = \rho^{2+\alpha}\int_D F + O(\rho^{\alpha+\gamma}), \tag{16}$$

This estimate, obtained under the assumption that α is sufficiently large, can be used to obtain an estimate for an arbitrary weight β, in particular, for $\beta = 0$. In more detail, as noted in [1] and elsewhere, there is a standard Stieltjes integral technique for obtaining estimates for the asymptotic growth of a measure, weighted in various ways, from a single such estimate. In the case at hand, the α-weighted lattice point count can be regarded as the integral from 0 to ρ of an atomic measure $\mathrm{d}\mu$, concentrated on the values of ρ for which the dilation of ∂D by ρ contains lattice-points, with each lattice-point weighted by the corresponding value of the weight-α density F. The weight-β lattice point count is then given by

$$\begin{aligned}\int_0^\rho t^{\beta-\alpha}\,\mathrm{d}\mu(t) &= \rho^{2+\beta}\int_D F(x)\,\mathrm{d}x + \int_0^\rho \mathrm{d}O(t^{\beta+\gamma}) \\ &= \rho^{2+\beta}\int_D F(x)\,\mathrm{d}x + O(\rho^{\beta+\gamma}).\end{aligned}$$

In particular, in the case where the weighting function is of weight zero, we have the F-weighted lattice point count within ρD given by

$$\mathfrak{N}_D(\rho) = \rho^2 \int_D F + O(\rho^\gamma) \tag{17}$$

As we mentioned before, this case is of particular significance since the above equation can be modified to serve as a numerical integration scheme.

In fact, given a measure on S^1 with the positive density $m(\theta)$ with respect to Lebesgue measure, and for which the equation $r = (m(\theta))^{1/2}$ defines the boundary of a polygon D, and a smooth function f on S^1 with a weight zero homogeneous extension F, we have, as a consequence of equation (8) of [4]

$$\int_{S^1} f(\theta)m(\theta)\mathrm{d}\theta = 2\int_D F(x)\mathrm{d}x. \tag{18}$$

This, combined with [17] above, yields

$$\int_{S^1} f(\theta)m(\theta)\mathrm{d}\theta = (2/\rho^2)\mathfrak{N}_D(\rho) + O(\rho^{\gamma-2}). \tag{19}$$

The first term on the right side of the last equation provides the principal term in the estimate for the integral on the left. The number of points used in the estimate corresponding to the parameter ρ is $\mathrm{N} = O(\rho^2)$, (a rough estimate for the number of lattice points within ρD). Therefore, the error of the method is of the order

$$O\left(\rho^{\gamma-2}\right) = O(N^{-1+\gamma}),$$

where γ is an arbitrarily small positive constant.

We observe that this method of numerical integration requires the polygon whose boundary is given by $r = (m(\theta))^{1/2}$ to have normals with poorly approximable slopes, as is the case with an algebraic polygon considered above. If, for example, a polygon has a side, and thus a normal, with a rational slope, the error of the estimate would be significantly worse, since the error term in the integral lattice point count inside the dilated polygon would grow linearly with the dilation parameter. However, such "bad" polygons are rare exceptions. In the following section, we will show that for almost all rotations of a polygon, the slopes of the normal to the sides are in fact poorly approximable in the sense needed to carry out the derivation of the estimate.

3 An Almost Everywhere Result

In this section, we will derive an a.e. estimate for the error of a weighted lattice point count for a polygon, specifically that for almost all rotations θ of the polygon,

$$R(\rho, \theta) \ll \rho^\alpha \log^{3+\gamma}(\rho).$$

It is plausible that this can be improved to $\rho^{\alpha}\log^{1+\gamma}(\rho)$, by analogy with Khinchine's result [2] for the constant density case, a possibility which we plan to investigate.

Since the arguments are so similar to those of the last section, our description of them will be terser and more merely indicative than was the case with our previous exposition.

As in the algebraic case, we consider separately the contributions to the error term coming from the "good" and the "bad" region, defined as before, relative to the angle θ by which the polygon is rotated. That is, the bad regions will consist as before of bands surrounding the normals to the sides of the rotated polygon, and the good regions will be complementary to these. The estimate for the good region contribution does not depend on θ and is obtained in much the same way as the corresponding estimate in the case of an algebraic polygon, using the "give-and-take" observation in the form given by (8), and, as in the treatment of the algebraic case, replacing the sum with a corresponding integral. The calculations are very similar to those involved in the estimate for the corresponding sum in the algebraic case, and yield

$$R^{G}(\rho,\theta) \ll \rho^{\alpha}\log^{2+\gamma}(\rho/\varepsilon).$$

Now, to get an a.e. (in terms of θ) estimate for the bad region's contribution, we invoke a well-known metrical theorem of Khinchine, the relevant part of which states that

If $\sum \frac{1}{qf(q)} < \infty$ *for some monotonic function f(x), then for almost all values of* λ *the inequality* $\langle q\lambda\rangle < \frac{1}{qf(q)}$ *has only finitely many solutions, or equivalently for almost all values of* λ*, there exist a constant* c_λ *s.t.* $\langle q\lambda\rangle > \frac{c_\lambda}{qf(q)}$.

In particular, $f(x) = \log^{1+\gamma}(x)$ satisfies the hypothesis above. Hence, setting $\lambda = \tan(\theta)$, and noting that it is sufficient to consider a single normal, we have that for almost all rotations of the polygon,

$$\langle q\lambda\rangle > \frac{c_\lambda}{q\log^{1+\gamma}(q)}.$$

As in the last section, it now follows immediately from the penultimate line in the proof of Lemma 3.3 of [3], (p. 123), that $s_k = \sum_{q=1}^{k}\frac{1}{\langle q\lambda\rangle} \ll k\log^{2+\gamma}k$.

Now, as before, the bad region's contribution to the error term in the modified lattice point count is

$$R_{\varepsilon}^{B}(\rho) \ll \rho^{\alpha}\sum_{N\in B} H(N)\widehat{\delta_{\varepsilon}}(N) \ll \rho^{\alpha}(S_1+S_2) \tag{20}$$

where

$$S_1 = \sum_{q=1}^{m}\frac{1}{q\langle q\lambda\rangle}\frac{1}{1+\varepsilon q} \ll \sum_{q=1}^{m}\frac{1}{q\langle q\lambda\rangle}$$

and

$$S_2 = \sum_{q=m+1}^{\infty} \frac{1}{q\langle q\lambda\rangle} \frac{1}{1+\varepsilon q} \ll \frac{1}{\varepsilon} \lim_{M\to\infty} \sum_{q=m+1}^{M} \frac{1}{q^2\langle q\lambda\rangle}.$$

Applying partial summation and then replacing the sum with a corresponding integral as before, we obtain

$$S_1 \ll \left(\sum_{k=1}^{m} \frac{s_k}{k(k+1)}\right) + \frac{s_m}{m+1} \ll \int_1^{\frac{1}{\varepsilon}} \frac{\log^{2+\gamma} x}{x} \mathrm{d}x + \log^{2+\gamma}\left(\frac{1}{\varepsilon}\right) \ll \log^{3+\gamma}\left(\frac{1}{\varepsilon}\right)$$

and

$$S_2 \ll \frac{1}{\varepsilon} \sum_{q=m+1}^{\infty} \frac{2k+1}{k^2(k+1)^2} s_k + \frac{1}{\varepsilon} \lim_{M\to\infty} \frac{s_M}{(M+1)^2}$$

$$\ll \frac{1}{\varepsilon} \int_{\frac{1}{\varepsilon}}^{\infty} \frac{\log^{2+\gamma} x}{x^2} \mathrm{d}x + \frac{1}{\varepsilon} \lim_{M\to\infty} \frac{\log^{2+\gamma} M}{M}.$$

Integrating by parts, we get

$$\frac{1}{\varepsilon} \int_{\frac{1}{\varepsilon}}^{\infty} \frac{\log^{2+\gamma} x}{x^2} dx \ll \log^{2+\gamma}\left(\frac{1}{\varepsilon}\right).$$

Combined with (20) this leads to the following estimate, which holds for almost all angles of rotation of the polygon:

$$R_{\varepsilon}^{B}(\rho) \ll \rho^{\alpha} \log^{3+\gamma}\left(\frac{1}{\varepsilon}\right).$$

Putting this together with the good region contribution, setting $\varepsilon = 1/\rho$ as before, we get

$$R_{\varepsilon}(\rho, \theta) \ll \rho^{\alpha} \log^{3+\gamma} \rho$$

for almost all rotations θ of the polygon. We use this a.e. estimate of the error in the modified lattice point count to get an a.e. estimate of the error for the true lattice point count

$$R(\rho, \theta) \ll \rho^{\alpha} \log^{3+\gamma}(\rho).$$

The steps of the argument leading from $R_{\varepsilon}(\rho)$ to $R(\rho)$ are essentially the same as those in the previous section and we will omit them here.

4 Concluding Remarks

We have described a general method for describing the accuracy with which a class of measures on S^1 can be approximated by a naturally associated family of discrete measures.

In the classical constant-density lattice point problem, if at least one of the normals to a side of a polygon has rational slope, there are an infinite number of $\rho_i \to \infty$ from which an infinitesimal displacement results in a modification of the lattice-point count of order ρ, so in this circumstance the error estimate is of true order ρ. Since the simple estimate of Gauss shows that the error term is always $\ll \rho$, polygons can be worst possible cases for lattice-point error asymptotics, though paradoxically, this situation is not generic for polygons, as was noticed by Khintchine [2] in the constant density case.

In the case in which the polygon is algebraic, the constant density case has previously been discussed in [8] and in [10].

Our approach can be synopsized by noting that in the presence of adequate information about the Fourier transform, the lattice points on the Fourier transform side of the Poisson summation formula are split into two groups: those in finite-width bands surrounding the "bad" normal vectors of D, and all the rest. The contribution from the lattice points exterior to the bands can be estimated by comparison with an integral, while the series arising from the contributions from lattice points within bands is estimated by using Diophantine properties of the slopes of the corresponding normals. Since the relevant estimates for the Fourier transform of D are singular at these directions, the poor approximability of the slopes, which in the algebraic case is a consequence of Roth's Theorem, is crucial (cf. e.g., [7], p. 858 for a similar argument).

Our developing experience with variable density lattice point asymptotics suggests a kind of meta-conclusion, that in general, the derivable asymptotics associated with such problems coincide, after suitable weighting, with the corresponding results for the classical constant-density case, and that the key is that the relevant Fourier transform asymptotics are effectively identical.

References

1. Michael R. Douglas, Bernard Shiffman, and Steve Zelditch. Critical points and supersymmetric vacua, iii: string/m models. *Commun. Math. Phys.*, 265(3):617–671, 2006
2. A. Khintchine. Ein Satz über Kettenbrüche, mit arithmetischen Anwendungen. *Math. Zeitschrift*, 18:289–306, 1923
3. L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences*. Wiley, New York, 1974
4. Marina Nechayeva and Burton Randol. Approximation of measures on S^n by discrete measures. *arXiv:math/0601230v1*, 2006
5. Burton Randol. A lattice-point problem. *Trans. Am. Math. Soc.*, 121(1):257–268, 1966
6. Burton Randol. On the Fourier transform of the indicator function of a planar set. *Trans. Am. Math. Soc.*, 139:271–276, 1969

7. Burton Randol. The behavior under projection of dilating sets in a covering space. *Trans. Am. Math. Soc.*, 285:855–859, 1984
8. Burton Randol. On the number of integral lattice-points in dilations of polyhedra. *Int. Math. Res. Not.*, (6):259–270, 1997
9. K.F. Roth. Rational approximation to algebraic numbers. *Mathematika*, 2:1–20, 1955
10. Maxim Skriganov. On integer points in polygons. *Annales de l'Institut Fourier*, 43:313–323, 1993
11. Maxim Skriganov. Ergodic theorems on $SL(n)$, Diophantine approximations and anomalies in the lattice-point problem. *Invent. Math.*, 132(1):1–72, 1998
12. Maxim Skriganov and A.N. Starkov. On logarithmically small errors in the lattice point problem. *Ergodic Theory Dynam. Systems*, 20(5):1469–1476, 2000

Support Bases of Solutions of a Functional Equation Arising From Multiplication of Quantum Integers and the Twin Primes Conjecture

Lan Nguyen

Summary Let P be the support base of a solution Γ, with the field of coefficients of characteristic zero, of the functional equations arising from the multiplication of quantum integers discussed in [A. Borisov, M. Nathanson, Y. Wang, Quantum integers and cyclotomy, J. Number Theory (to appear); M. Nathanson, A functional equation arising from multiplication of quantum integers, J. Number Theory, 103(2), 214–233 (2003)]. It is known from the work of Nathanson as well as our work that there is a close relationship between P and the constructibility of Γ from quantum integers. In this paper, we prove that if the Twin Primes conjecture holds, then Γ is constructible from quantum integers if P contains infinitely many pairs of twin primes. This shows, in particular, that if the Twin Primes conjecture holds, then Γ is constructible from quantum integers whenever P has finite complement.

Keywords Polynomial functional equations · Q-series · Quantum integers · Quantum algebra

Mathematics Subject Classifications (2010). 11B13, 11C08, 11P99, 11T20

1 Introduction

Motivated by the study of the q-analogues of zeta and multiple zeta functions, Nathanson studies quantum integers and the functional equations arising from quantum addition and multiplication. Nathanson shows in [2] that there is a strong relationship between the support base P of a normalized sequence Γ of polynomials satisfying the functional equations arising from multiplication of quantum integers and the constructibility of such sequence by quantum integers. In particular, he shows that under some condition involving the degree of the non-zero polynomials in Γ, Γ is constructible in a unique way by quantum integers if P contains 2 and

L. Nguyen
Mathematics Department, The University of Michigan, Ann Arbor, MI, USA
e-mail: ltng@umich.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_21,

some odd prime. It is known from [1,2,4] that if the field of coefficients of Γ is $\mathbb{Q}$, then Γ is constructible from quantum integers if the support base P contains at least two primes. In [4,9], we also show that if the field of coefficients of Γ is of characteristic zero, then Γ is constructible from quantum integers if P contains all primes. Furthermore, it can be seen from [6,11] that the larger the cardinality of P is, the more likely that Γ is constructible from quantum integers. On the other hand, we show in [5,7] that if P has finite cardinality, then there always exists at least one such sequence Γ of polynomials, with support base P, which is not constructible from quantum integers. Therefore, if P is a set of primes such that all sequences Γ with support base P is constructible from quantum integers, then P must have infinite cardinality. In this paper, we also explore the relationship between the support base P of Γ and the constructibility of Γ from quantum integers for all the sequences Γ with fields of coefficients of characteristic zero. In particular, we seek to weaken the condition that P has to contain all primes for Γ to be constructible from quantum integers. One way to accomplish this is to make use of the well-known Twin Primes conjecture which we will recall below.

Twin Primes Conjecture. There are infinitely many pairs of primes of the form p and $p+2$.

Before studying the consequence of this conjecture on the support base P of a solution Γ of the functional equation arising from the multiplication of quantum integers discussed in [1,2] and below, let us give some basic background and some results from [1,4,5], which are relevant to this paper, concerning quantum integers and the functional equation arising from the multiplication of these integers.

Definition 1. A quantum integer is a polynomial in q of the form

$$[n]_q := q^{n-1} + \ldots + q + 1 = \frac{q^n - 1}{q - 1} \tag{1}$$

where n is any natural number.

From [1], the multiplication operation for quantum integers is defined by the following rule:

$$[m]_q \star [n]_q := [mn]_q = [m]_q \cdot [n]_{q^m} = [n]_q \cdot [m]_{q^n} \tag{2}$$

where $\star$ denotes the multiplication operation for quantum integers, and . denotes the usual multiplication of polynomials. Equation (1.2) is just the q-series expansion of the sumsets

$$\begin{aligned} \{0, 1, \ldots, mn-1\} &= \{0, 1, \ldots, m-1\} + \{0, m, \ldots, (n-1)m\} \\ &= \{0, 1, \ldots, n-1\} + \{0, m, \ldots, (m-1)n\} \end{aligned}$$

Motivated by (1.2), let $\Gamma = \{f_n(q)|n = 1, \ldots, \infty\}$ be a sequence of polynomials in q, with coefficients contained in some field, satisfying the following functional equations:

$$f_m(q)\, f_n(q^m) \overset{(1)}{=} f_n(q)\, f_m(q^n) \overset{(2)}{=} f_{mn}(q) \tag{3}$$

for all $m, n \in \mathbb{N}$. We refer to the first equality in the above functional equation as Functional Equation (1) and the second equality as Functional Equation (2). A sequence of polynomials which satisfies Functional Equation (2) automatically satisfies Functional Equation (1) but not vice versa (see [4] for an example). Functional Equation (1) induces an interesting equivalent relation which we explore in detail in [12].

The support of Γ, denoted by $supp\{\Gamma\}$, is the set of integers n in $\mathbb{N}$ where $f_n(q) \neq 0$. If P is a set of rational primes and A_P consists of 1 and all natural numbers such that all their prime factors come from P, then A_P is a multiplicative semigroup which is called a prime multiplicative semigroup associated to P.

Theorem 1 ([1]). *Let $\Gamma = \{f_n(q)\}$ be a sequence of polynomials satisfying Functional Equation* (2). *Then $supp\{\Gamma\}$ is of the form A_P for some set of primes P, and Γ is completely determined by the collection of polynomials:*

$$\{f_p(q)|p \in P\}.$$

Definition 2. Let P be the collection of primes associated to the support A_P, in the sense of Theorem 1, of a sequence of polynomials Γ satisfying Functional Equation (2). Then P is called the support base of Γ.

In the reverse direction, if P is a set of primes in $\mathbb{N}$ then there is at least one sequence Γ satisfying Functional Equation (2) with $supp\{\Gamma\} = A_P$. One such sequence can be defined as the set of polynomials:

$$f_m(q) = \begin{cases} [m]_q & \text{if } m \in A_P; \\ 0 & \text{otherwise.} \end{cases}$$

Note that the coefficients of $f_m(q)$ are properly contained in $\mathbb{Q}$.

We say that a sequence Γ is non-zero if $supp\{\Gamma\} \neq \emptyset$. If Γ satisfies Functional Equation (2), then Γ is non-zero if and only if $f_1(q) = 1$ (see [1]).

The degree of each polynomial $f_n(q) \in \Gamma$ is denoted by $deg(f_n(q))$. From [1], it is known that there exists a rational number t_Γ such that:

$$deg(f_n(q)) = t_\Gamma(n-1)$$

for all n in $supp\{\Gamma\}$. This number t_Γ is not necessarily an integer (see [4] for an example of such a sequence). We show in [2] and [4] that t_Γ can only be non-integral when the set of primes P associated to the support of Γ has the form $P = \{p\}$ for some prime p.

Theorem 2 ([1]). *Let P be a set of primes. Let $\Gamma' = \{f'_p(q)|p \in P\}$ be a collection of polynomials such that:*

$$f'_{p_1}(q) \cdot f'_{p_2}(q^{p_1}) = f'_{p_2}(q) \cdot f'_{p_1}(q^{p_2})$$

for all $p_i \in P$ (i.e, satisfying Functional Equation (1)). Then there exists a unique sequence $\Gamma = \{f_n(q)|n \in \mathbb{N}\}$ of polynomials satisfying Functional Equation (2) such that $f_p(q) = f'_p(q)$ for all primes $p \in P$.

Theorem 3 ([4]). *Let $\Gamma = \{f_n(q)|n \in \mathbb{N}\}$ be a non-zero sequence of polynomials satisfying Functional Equation (2) with support A_P for some set of primes P. Then there exists a unique completely multiplicative arithmetic function $\psi(n)$, a unique non-negative rational number t, and a unique sequence $\Sigma = \{g_n(q)\}$ satisfying Functional Equation (2) with*

$$Supp\{\psi\} = Supp\{\Sigma\} = Supp\{\Gamma\} = A_P$$

such that

$$f_n(q) = \psi(n)q^{t(n-1)}g_n(q)$$

where $g_n(q)$ is a monic polynomial with $g_n(0) \neq 0$ for all $n \in A_P$.

As a result, in the rest of this paper, unless otherwise stated, all sequences of polynomials which we consider are normalized so that each polynomial is monic and having non-zero constant terms. Such sequences are called normalized sequences.

For a sequence Γ of polynomials satisfying Functional Equation (2), the smallest field K which contains all the coefficients of all the polynomials in Γ is called *The Field of Coefficients of Γ*. We are only concerned with sequences of polynomials whose fields of coefficients K are of characteristic zero. The case of positive characteristic fields of coefficients will be reserved for our future papers. Unless stated otherwise, we always view Γ as a sequence of polynomials with coefficients in a fixed separable closure $\overline{K}$ of K which is embedded in $\mathbb{C}$ via a fixed embedding $\iota : \overline{K} \hookrightarrow \mathbb{C}$. Thus every element $f(q)$ of Γ can be viewed as a polynomial in $\mathbb{C}[q]$. We frequently view polynomials $f(q)$'s in Γ as elements of the ring $\mathbb{C}[q]$ throughout this paper. Thus whenever that is necessary, it is implicitly assumed.

In [3,8,10], we classify all normalized sequences when their fields of coefficients are of characteristic zero.

Theorem 4 ([4]). *Let $\Gamma = \{f_n(q)|n \in \mathbb{N}\}$ be a sequence of polynomials satisfying Functional Equation (2) and whose field of coefficients is of characteristic zero.*

(1) Field of coefficients is $\mathbb{Q}$: Suppose that $deg(f_p(q)) = t_\Gamma(p-1)$ with $t_\Gamma \geq 1$ for at least two distinct primes p and r, which means that the set P associated to the support A_P of Γ contains p and r and the elements $f_p(q)$ and $f_r(q)$

of Γ are non-constant polynomials. Then there exist ordered pairs of integers $\{u_i, t_i\}_i$ with $i = 1, \ldots, s$ such that $t_\Gamma = \sum_{i=1,\ldots,s} u_i t_i$ and

$$f_n(q) = \prod_{i=1}^{s} \left([n]_{q^{u_i}}\right)^{t_i}$$

for all n in $\mathbb{N}$.

(2) Field of coefficients strictly contains $\mathbb{Q}$: There is no sequence of polynomials Γ, with field of coefficients strictly containing $\mathbb{Q}$, satisfying Functional Equation (2) *and the condition $deg(f_p(q)) = t_\Gamma(p-1)$ with integral $t_\Gamma \geq 1$ for all primes p. The latter condition means that the set P associated to the support A_P of Γ contains all prime numbers and the correspondent elements $f_p(q)$ of Γ are non-constant polynomials.*

The decomposition of $f_n(q)$ into a product of quantum integers as above is unique in the sense that if $\{a_j, b_j\}$ is another set of integers such that $t_\Gamma = \sum_{j=1,\ldots,h} a_j b_j$ and

$$f_n(q) = \prod_{j=1}^{h} \left([n]_{q^{a_j}}\right)^{b_j}$$

for all $n \in supp\{\Gamma\}$, then for each u_i, there exists at least one a_j such that $u_i = a_j$. Moreover, if $I \subseteq \{1, \ldots, s\}$ and $J \subseteq \{1, \ldots, h\}$ are two collections of indexes such that $u_i = a_j$ exactly for all i in I and j in J and nowhere else, then

$$\sum_{i \in I} t_i = \sum_{j \in J} b_j,$$

and the above relation between any such set of integers $\{a_j, b_j\}_j$ and the set $\{u_i, t_i\}_i$ is an equivalent relation. However, if the condition $deg(f_p(q)) = t_\Gamma(p-1)$ with integral $t_\Gamma \geq 1$ for all primes p is not imposed on Γ, then there exist sequences Γ's of polynomials with fields of coefficients strictly greater than $\mathbb{Q}$ satisfying Functional Equation (2).

Definition 3. Let $\Gamma = \{f_n(q) | n \in \mathbb{N}\}$ be a sequence of polynomials satisfying Functional Equation (2). Then Γ is said to be constructible from quantum integers if there exist ordered pairs of integers $\{(u_i, t_i) \mid u_i \geq 1; i = 1, \ldots, s\}$ such that $t_\Gamma = \sum_{i=1,\ldots,s} u_i t_i$ and

$$f_n(q) = \prod_{i=1}^{s} \left([n]_{q^{u_i}}\right)^{t_i}$$

for all n in $supp\{\Gamma\}$.

Remark 1. By part (2) of Theorem 4, if Γ is constructible from quantum integers, then it is so in a unique way. Also, if Γ is constructible from quantum integers, then it is necessary that Γ is a normalized sequence.

2 Main Results

Assuming the Twin Primes conjecture, we prove some consequences of this conjecture on the support base P of a sequence Γ of polynomials satisfying Functional Equation (2).

Let Γ be a sequence of polynomials satisfying Functional Equation (2) and P be its support base. From Theorem 9 of [1], if P contains 2 and some odd prime p, then Γ is constructible from quantum integers if $t_\Gamma = 1$. However, when $t_\Gamma > 1$, there is no similar result (see [6]). Our first goal is to show that the Twin Primes conjecture gives an interesting analogue of this result for the case where $t_\Gamma > 1$.

From part (2) of Theorem 4, we know that there is no sequence Γ of polynomials satisfying Functional Equation (2) with support base P consisting of all primes and field of coefficients of characteristic zero strictly containing $\mathbb{Q}$. Therefore, every sequence Γ of polynomials satisfying Functional Equation (2) with support base P consisting of all primes must be constructible from quantum integers by part (1) of Theorem 4. In the hypothesis of Theorem 4, the complement of the support base of Γ, the collection of prime numbers not included in P, is empty. Our second goal is to show that Theorem 4 can be strengthened by showing that this complement can be enlarged as a consequence of the Twin Primes conjecture.

Our main results in this paper can be summarized as follows:

Theorem 5. *Let Γ be a sequence of polynomials satisfying Functional Equation* (2) *and let P be its support base. Assuming the Twin Primes conjecture, then there exist ordered pairs of integers $\{u_i, t_i\}_i$ with $i = 1, \ldots, s$ such that $t_\Gamma = \sum_{i=1,\ldots,s} u_i t_i$ and*

$$f_n(q) = \prod_{i=1}^{s} \left([n]_{q^{u_i}}\right)^{t_i}$$

for all n in supp$\{\Gamma\}$ if P contains infinitely many pairs of twin primes.

Corollary 1. *Assuming the Twin Primes conjecture, then every sequence Γ of polynomials satisfying Functional Equation* (2) *is constructible from quantum integers if its support base P has finite complement.*

Remark 2. We will generalize Corollary 1 in our future paper [3], a project which is currently in progress.

3 Proof of Main Results

Proof. (proof of Theorem 5)

Suppose the Twin Primes conjecture holds. Then there are infinitely many pairs of twin primes. Let

$$\Gamma := \{f_n(q) | n \in \mathbb{N}\}$$

be a non-zero sequence of polynomials satisfying Functional Equation (2) and let P be its support base containing infinitely many pairs of twin primes (not necessarily all twin primes). Suppose that Γ cannot be constructible from quantum integers. Then the field of coefficients of Γ strictly contains $\mathbb{Q}$ by part (1) of Theorem 4. Since P is contained in the support of Γ by definition, Γ contains a subsequence of the form $\{f_p(q)|p \in P\}$.

By Theorem 1.4, to prove that such a sequence Γ does not exist, it is sufficient for us to prove that the subsequence

$$\{f_p(q)|p \in P\}$$

of Γ does not exist.

Let u be any positive integer and p be any prime number. The polynomial denoted by $P_{u,p}(q)$ or $P_{up}(q)$ is the irreducible cyclotomic polynomial in $\mathbb{Q}[q]$ whose roots are all primitive up-roots of unity. $P_{u,p}(q)$ is sometimes denoted by $P_{up}(q)$ or $P_v(q)$ where $v = up$. For a primitive n-root of unity α in $\mathbb{C}$, which can be written in the form $\alpha = e^{(2\pi iw)/n}$ for some primitive residue class w modulo n, we always identify α, via the Chinese Remainder Theorem, with the tuples $(u_i)_i$ where $\prod_i (p_i)^{m_i}$ is the prime factorization of n and $u_i \in (\mathbb{Z}/p_i^{m_i}\mathbb{Z})^*$ for each i such that

$$u_i \equiv w \pmod{p_i^{m_i}}.$$

From [4], it is known that if Γ is a non-trivial sequence of polynomials satisfying Functional Equation (2) with support base P containing at least two primes and p is a prime in $supp\{\Gamma\}$, then all roots of $f_p(q)$ are roots of unity of orders divisible by p.

Let p and r be any distinct primes in the support of Γ. Define $f_{u_p,p}(q)$ to be the factor of $f_p(q)$ such that its roots consist of all the roots of $f_p(q)$ with multiplicities which are primitive pu_p-roots of unity. Then,

$$f_p(q) = \prod_{u_{p,j} > u_{p,j+1}} f_{u_{p,j},p}(q)$$

in the ring $\mathbb{C}[q]$. Similarly,

$$f_r(q) = \prod_{u_{r,i} > u_{r,i+1}} f_{u_{r,i},r}(q).$$

We call j (respectively, i) *j-level* (respectively, *i-level*) of $f_p(q)$ (respectively, $f_r(q)$) if $f_{u_{p,j}}(q)$ (respectively, $f_{u_{r,i}}(q)$) is a non-trivial factor of $f_p(q)$ (respectively, $f_r(q)$). We refer to $u_{p,j}$ (resp. $u_{r,i}$) the value of the level j (respectively, i) of $f_p(q)$ (respectively, $f_r(q)$). Define $V := \{v_{p,r,k}|v_{p,r,k} > v_{p,r,k+1}\} := \{u_{p,j}\}_j \cup \{u_{r,i}\}_i$. We refer to k as the *k-bi-level* with respect to p and r of $f_p(q)$ and $f_r(q)$ and $v_{p,r,k}$ as its value. Note that level i of $f_p(q)$ or $f_r(q)$ is not necessarily equal to

the bi-level i of $f_p(q)$ and $f_r(q)$. Using V and these product decompositions, we write Functional Equation (1) with respect to $f_p(q)$ and $f_r(q)$ as:

$$\begin{array}{ccc}
f_{v_{p,r,1},p}(q)^{s_{v_p,1}} f_{v_{p,r,1},r}(q^p)^{s_{v_r,1}} & \overset{(1)}{\longleftrightarrow} & f_{v_{p,r,1},r}(q)^{s_{v_r,1}} f_{v_{p,r,1},p}(q^r)^{s_{v_p,1}} \\
\cdots & \cdots & \cdots \\
f_{v_{p,r,k},p}(q)^{s_{v_p,k}} f_{v_{p,r,k},r}(q^p)^{s_{v_r,k}} & \overset{(k)}{\longleftrightarrow} & f_{v_{p,r,k},r}(q)^{s_{v_r,k}} f_{v_{p,r,k},p}(q^r)^{s_{v_p,k}} \\
\cdots & \cdots & \cdots \\
f_p(q) f_r(q^p) & = & f_r(q) f_p(q^r)
\end{array}$$

where:

- $s_{p,k} = 1$ if $f_{v_{p,r,k},p}(q)$ non-trivially divides $f_p(q)$ (i.e., $f_{v_{p,r,k},p}(q) = f_{u_i,p}(q)$ for some u_i) and 0 otherwise.
- $s_{r,k} = 1$ if $f_{v_{p,r,k},r}(q)$ non-trivially divides $f_r(q)$ (i.e., $f_{v_{p,r,k},r}(q) = f_{u_i,r}(q)$ for some u_i) and 0 otherwise.
- $\prod_k f_{v_{p,r,k},p}(q)^{s_{v_p,k}} f_{v_{p,r,k}}(q^p)^{s_{v_r,k}} = f_p(q) f_r(q^p)$.
- $\prod_j f_{v_{p,r,k},r}(q)^{s_{v_r,k}} f_{v_{p,r,k}}(q^r)^{s_{v_p,j}} = f_r(q) f_p(q^r)$.
- The symbol $\overset{(j)}{\longleftrightarrow}$ indicates the functional equation (1) at the bi-level j. Note that the polynomial expressions on the left hand side and the right hand side of $\longleftrightarrow$ at each bi-level are not necessarily equal.

Note that for every bi-level k where $v_{p,r,k}$ appears in the equation above, either $s_{p,k} = 1$ or $s_{r,k} = 1$.

The above version of Functional Equation (1) is called the *Expanded Functional Equation* (1) with respect to p and r, denoted by EFE(1). The EFE(1) above is said to be in *reduced form* if at each bi-level k where pr does not divide $v_{p,r,k}$, the line

$$f_{v_{p,r,k},p}(q)^{s_{v_p,k}} f_{v_{p,r,k},r}(q^p)^{s_{v_r,k}} \overset{(k)}{\longleftrightarrow} f_{v_{p,r,k},r}(q)^{s_{v_r,k}} f_{v_{p,r,k},p}(q^r)^{s_{v_p,k}}$$

in EFE (1) is replaced by

(i) $f_{v_{p,r,k},r}(q^p)^{s_{v_r,k}} \overset{(k)}{\longleftrightarrow} f_{v_{p,r,k},r}(q)^{s_{v_r,k}} \frac{f_{v_{p,r,k},p}(q^r)^{s_{v_p,k}}}{f_{v_{p,r,k},p}(q)^{s_{v_p,k}}}$ if $(r, v_{p,r,k}) = 1$.

(ii) $f_{v_{p,r,k},p}(q)^{s_{v_p,k}} \frac{f_{v_{p,r,k},r}(q^p)^{s_{v_r,k}}}{f_{v_{p,r,k},r}(q)^{s_{v_r,k}}} \overset{(k)}{\longleftrightarrow} f_{v_{p,r,k},p}(q^r)^{s_{v_p,k}}$ if $(p, v_{p,r,k}) = 1$, or

(iii) $\frac{f_{v_{p,r,k},p}(q^r)^{s_{v_p,k}}}{f_{v_{p,r,k},p}(q)^{s_{v_p,k}}} \overset{(k)}{\longleftrightarrow} \frac{f_{v_{p,r,k},r}(q^p)^{s_{v_r,k}}}{f_{v_{p,r,k},r}(q)^{s_{v_r,k}}}$ if $(pr, v_{p,r,k}) = 1$.

(iv) The line $f_p(q) f_r(q^p) = f_r(q) f_p(q^r)$ is replaced by $Q_{p,r}(q) = Q_{p,r}(q)$ where $Q_{p,r}(q)$ is the product of all expressions of the left-hand column (or the right-hand column) after (i), (ii), (iii) have taken place, i.e.,

$$\begin{aligned}
Q_{p,r}(q) &= \frac{f_p(q) f_r(q^p)}{\prod_i f_{v_{p,r,i},r}(q)^{s_{r,i}(1-\delta_{p,i})} f_{v_{p,r,i},p}(q)^{s_{p,i}(1-\delta_{r,i})}} \\
&= \frac{f_r(q) f_p(q^r)}{\prod_i f_{v_{p,r,i},r}(q)^{s_{r,i}(1-\delta_{p,i})} f_{v_{p,r,i},p}(q)^{s_{p,i}(1-\delta_{r,i})}}
\end{aligned}$$

where

$$\delta_{p,i} = \begin{cases} 1 & \text{if } p \text{ divides } v_{p,r,i}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\delta_{r,i} = \begin{cases} 1 & \text{if } r \text{ divides } v_{p,r,i}, \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3. (1) An EFE(1) with respect to p and r can be transformed into its reduced form by dividing both polynomials $f_p(q) f_r(q^p)$ and $f_r(q) f_p(q^r)$ by

$$\prod_i f_{v_{p,r,i},r}(q)^{s_{r,i}(1-\delta_{p,i})} f_{v_{p,r,i},p}(q)^{s_{p,i}(1-\delta_{r,i})};$$

(2) The product of all the rational expressions in the left-hand column and the product of those in the right-hand column of the reduced form of the EFE(1) are equal, and thus can be denoted by the same polynomial $Q_{p,r}(q)$; (3) For each line (i), the product of all expressions on both sides of $\longleftrightarrow$ remains equal after (i), (ii) or (iii) have taken place. It is shown in [2] that all the rational expressions above are actually polynomials when they occur, and that for each of these rational expressions, its roots are primitive roots of unity of the same order.

Definition 4. (1) Let $P_{u,p}(q)$ and $P_{u,r}(q)$ be the cyclotomic polynomials with coefficients in $\mathbb{Q}$ of orders up and ur respectively. Let $F_{u,p}(q)$ and $F_{u,r}(q)$ be two polynomials dividing $P_{u,p}(q)$ and $P_{u,r}(q)$ respectively. If $F_{u,p}(q)$ and $F_{u,r}(q)$ satisfy the condition that for each primitive residue class w modulo u, all the roots of $P_{u,p}(q)$ represented by the collection of tuples $\{(\gamma_p, (w_{p_j})_j) | \gamma_p = 1, \ldots, p-1\}$ if p does not divide u (respectively by the collection $\{(w_p + t(p^l), (w_{p_j})_{j,p_j \neq p}) | t = 0, \ldots, p-1\}$ if $p^l \| u$ for some positive integer $l \geq 1$) are roots $F_{u,p}(q)$ if and only if all the roots of $P_{u,r}(q)$ represented by the collection $\{\gamma_r, (w_{p_j})_j | \gamma_r = 1, \ldots, r-1\}$ if r does not divide u (respectively, by the collection $\{w_r + s(r^h), (w_{p_j})_{j,p_j \neq r} | s = 0, \ldots, r-1\}$ if $r^h \| u$ for some positive integer $h \geq 1$) are roots $F_{u,r}(q)$, then we will say that $F_{u,p}(q)$ and $F_{u,r}(q)$ are *compatible*. For example, $P_{u,p}(q)$ and $P_{u,r}(q)$ are compatible for any positive integer u, primes p and r, a fact which is proven in [4] for the case where pr does not divide u as well as when either p or r divides u.

(2) Two polynomials $f_{u,p}(q)$ and $f_{u,r}(q)$ are said to be *super-compatible* if $f_{u,p}(q) = \prod_i (F^{(i)}_{u,p}(q))^{n_i}$ and $f_{u,r}(q) = \prod_i (F^{(i)}_{u,r}(q))^{n_i}$ where $F^{(i)}_{u,p}(q)$ and $F^{(i)}_{u,r}(q)$ are polynomials which are compatible for all i. In particular, $P_{u,p}(q)^n$ and $P_{u,r}(q)^n$ are super-compatible for any non-negative integer n. Thus, compatibility is a special case of super-compatibility.

Remark 4. To understand the rationality of this definition, the readers can consult [4]. The polynomials $F^{(i)}_{u,\square}(q)$'s in the definition of super-compatible might not be unique for any i, where $\square$ denotes either p or r.

Let p and r be two primes in P and let $f_p(q)$ and $f_r(q)$ be the elements of Γ corresponding to p and r. As above, we can write

$$f_p(q) = \prod_i f_{u_{p,i},p}(q)$$

and

$$f_r(q) = \prod_j f_{u_{r,j},r}(q),$$

where $f_{u_{p,i},p}(q)$ (respectively, $f_{u_{r,j},r}(q)$) is the factor of $f_p(q)$ (respectively, $f_r(q)$) whose roots are all roots of $f_p(q)$ (respectively, $f_r(q)$) which are primitive $u_{p,i}\,p$ (respectively, $u_{r,j}\,r$) roots of unity for some integer $u_{p,i}$ (respectively, $u_{r,j}$).

Let $V_{p,r} := \{v_{p,r,l} | v_{p,r,l} > v_{p,r,l+1}\} := \{u_{p,i}\}_i \cup \{u_{r,j}\}_j$ and let

$$\begin{array}{ccc}
f_{v_{p,r,1},p}(q)^{s_{v_{p,1}}} f_{v_{p,r,1},r}(q^p)^{s_{v_{r,1}}} & \overset{(1)}{\longleftrightarrow} & f_{v_{p,r,1},r}(q)^{s_{v_{r,1}}} f_{v_{p,r,1},p}(q^r)^{s_{v_{p,1}}} \\
\cdots & \cdots & \cdots \\
f_{v_{p,r,k},p}(q)^{s_{v_{p,k}}} f_{v_{p,r,k},r}(q^p)^{s_{v_{r,k}}} & \overset{(k)}{\longleftrightarrow} & f_{v_{p,r,k},r}(q)^{s_{v_{r,k}}} f_{v_{p,r,k},p}(q^r)^{s_{v_{p,k}}} \\
\cdots & \cdots & \cdots \\
f_p(q) f_r(q^p) & = & f_r(q) f_p(q^r)
\end{array}$$

be EFE(1) with respect to p and r.

Proposition 1. *Define*

$$V := \bigcup_{p,r \in P} V_{p,r}.$$

Let $\mathscr{U}$ be the collection of all prime factors of every element of V. Then

$$|\mathscr{U}| < \infty.$$

Proof. From part (1) of Key Proposition $1'$ of [4], $v_{p,r,1}$ is independent of p and r. Moreover, for any pair of primes p and r,

$$v_{p,r,1} \geq v_{p,r,l}$$

for all bi-levels $l \geq 1$. Therefore,

$$s \leq v_{p,r,1}$$

for all primes in $\mathscr{U}$. Thus the result follows (see part (1) of Key Proposition 3 of [4] for more details). □

As a result of Proposition 1, there exists a natural number N such that if $s > N$ is a prime, then s is greater than every prime in $\mathscr{U}$ and thus is not in $\mathscr{U}$. Since P contains infinitely many pairs of twin primes, there exist infinitely many pairs of twin primes in P which are greater than N and thus are not in $\mathscr{U}$.

Now let $p > N$ and $r > N$ be two primes in P. Then p and s are not in $\mathscr{U}$. We have the following result.

Proposition 2. *The coefficients of $f_p(q)$ and $f_r(q)$ are not properly contained in $\mathbb{Q}$. In particular, there exist levels i and j such that the coefficients of $f_{u_{p,i},p}(q)$ and $f_{u_{r,j},r}(q)$ are not properly contained in $\mathbb{Q}$.*

Proof. Since p and r are greater than N, p and r are greater than all primes in $\mathscr{U}$. Then this proposition follows from the proof of Key Proposition 3 of [4]. □

Let l_p and l_r be the minimal indexes such that the coefficients of $f_{u_{p,l_p},p}(q)$ and $f_{u_{r,l_r},r}(q)$ are not properly contained in $\mathbb{Q}$.

Proposition 3. *There exists a bi-level $k_{p,r}$ such that $v_{p,r,k_{p,r}} = u_{p,l_p} = u_{r,l_r}$, where $v_{p,r,k_{p,r}}$ is the value of the $k_{p,r}$ bi-level of EFE(1) with respect to p and r.*

Proof. Since $p > N$, p is greater than every prime in $\mathscr{U}$. Therefore, the result follows from part (2) of Key Proposition 3 of [4]. □

From Proposition 3, we have

$$f_{v_{p,r,k_{p,r}},p}(q) = f_{u_{p,l_p},p}(q)$$

and

$$f_{v_{p,r,k_{p,r}},r}(q) = f_{u_{r,l_r},r}(q).$$

Hence, $f_{v_{p,r,k_{p,r}},p}(q) \neq 1$ and $f_{v_{p,r,k_{p,r}},r}(q) \neq 1$. Therefore,

$$s_{p,k_{p,r}} = s_{r,k_{p,r}} = 1.$$

Proposition 4. *Let r' be another prime in P such that $r' > N$. Then*

$$v_{p,r,k_{p,r}} = v_{p,r',k_{p,r'}}$$

where $v_{p,r',k_{p,r'}}$ is the analogue $v_{p,r,k_{p,r}}$ with r replaced by r'.

Proof. See part (3) of Key Proposition 3 of [4]. □

Therefore, there exists a natural number L such that $L = v_{p,s,k_{p,s}}$, for any prime $s > N$ in P.

Proposition 5. *Let $p > N$ and $s > N$ be primes in P. Then*

$$L > 1.$$

Proof. Since $p > N$, p is strictly greater than any prime in $\mathscr{U}$. Then the same argument as in Key Proposition 5 of [4] implies the result. □

Let us recall from [4] the following definition:

Definition 5. Let Γ be a sequence of polynomials satisfying Functional Equation (1) with support base P containing at least two primes. Suppose that the field of coefficients of Γ is a field of characteristic zero which strictly contains $\mathbb{Q}$. Let p and r be two primes in P and let $f_p(q)$ and $f_r(q)$ be the corresponding elements of Γ. Let $k_{p,r}$ be the smallest bi-level of EFE(1) with respect to p and r such that the coefficients $f_{v_{p,r,k_{p,r}},p}(q)$ and $f_{v_{p,r,k_{p,r}},r}(q)$ are not properly contained in $\mathbb{Q}$. Let $\square$ denote either p or r. Suppose that $\square$ does not divide $v_{p,r,k_{p,r}}$. Then a prime p_0 is call an essential prime of Γ if the following conditions are satisfied:

1. p_0 divides $v_{p,r,k_{p,r}}$.
2. The collection of tuples of integers representing the roots of $f_{v_{p,r,k_{p,r}},\square}(q)$ is a sub-collection of a collection of tuples of integers of the form

$$\prod_{p_i|\square v_{p,r,k_{p,r}};i\geq 1} A_{p_i^{n_i}} \times A_{p_0^{n_0}}$$

where:

- $\square v_{p,r,k_{p,r}} = \prod_{i\geq 0} p_i^{n_i}$ is the prime factorization of $\square v_{p,r,k_{p,r}}$.
- $A_{p_0^{n_0}}$ is a proper subset of $(\mathbb{Z}/p_0^{n_0})^*$.
- For each $p_i \neq p_0$ in the factorization of , $A_{p_i^{n_i}}$ is equal to $((\mathbb{Z}/p_i^{n_i})^*)^{m_i}$ for some natural number m_i.

As in [4], we may assume without loss of generality that L is square-free and all prime divisors of L are essential primes since all other cases are either irrelevant to or automatically satisfy what needs to be shown.

Let p be as above (i.e., $>N$) and let $s > N$ be a prime in P. Then s is also not in $\mathscr{U}$. Hence p and s do not divide L. As a result, it can be verified that line $(k_{p,s})$ of the reduced form of EFE(1) with respect to p and s has the form

$$\frac{f_{v_{p,s,k_{p,s}},p}(q^s)}{f_{v_{p,s,k_{p,s}},p}(q)} \overset{(k_{p,s})}{\longleftrightarrow} \frac{f_{v_{p,s,k_{p,s}},s}(q^p)}{f_{v_{p,s,k_{p,s}},s}(q)}$$

since $s_{p,k_{p,s}} = s_{r,k_{p,s}} = 1$. Moreover, it can also be verified that line $(k_{p,s})$ of the $(k_{p,s})$-super-reduced form of EFE(1) with respect to p and s has the form

$$f_{v_{p,s,c_{k_{p,s}}},s}(q)^{s_{s,c_{k_{p,s}}}\delta_{p,c_{k_{p,s}}}} \frac{f_{v_{p,s,k_{p,s}},p}(q^s)}{f_{v_{p,s,k_{p,s}},p}(q)}$$

$$\overset{(k_{p,s})}{=} f_{v_{p,s,d_{k_{p,s}}},p}(q)^{s_{p,d_{k_{p,s}}}\delta_{s,d_{k_{p,s}}}} \frac{f_{v_{p,s,k_{p,s}},s}(q^p)}{f_{v_{p,s,k_{p,s}},s}(q)}$$

where:

- $s_{s,c_{k_{p,s}}} = 1$ if there exists a bi-level $c_{k_{p,s}}$ such that

$$v_{p,s,c_{k_{p,s}}} = v_{p,r,k_{p,s}} p$$

and $s_{s,c_{k_{p,s}}} = 0$ otherwise. Hence $\delta_{p,c_{k_{p,s}}} = 1$ if $s_{s,c_{k_{p,s}}} = 1$.

- $s_{s,d_{k_{p,s}}} = 1$ if there exists a bi-level $d_{k_{p,s}}$ such that

$$v_{p,s,d_{k_{p,s}}} = v_{p,r,k_{p,s}} s$$

and $s_{s,d_{k_{p,s}}} = 0$ otherwise. Hence $\delta_{s,d_{k_{p,s}}} = 1$ if $s_{s,d_{k_{p,s}}} = 1$.

By Key Proposition 1′ of [4], the rational expressions

$$\frac{f_{v_{p,s,k_{p,s}},p}(q^s)}{f_{v_{p,s,k_{p,s}},p}(q)}$$

and

$$\frac{f_{v_{p,s,k_{p,s}},s}(q^p)}{f_{v_{p,s,k_{p,s}},s}(q)}$$

are polynomials. Since the coefficients of $f_{v_{p,s,k_{p,s}},p}(q)$ and $f_{v_{p,s,k_{p,s}},s}(q)$ are not properly contained in $\mathbb{Q}$ by definition of $k_{p,s}$, this implies that s and p are congruent to 1 modulo L by part (ii) of Key Proposition 1′ of [4] and by the assumption that L is square-free and all the prime divisors of L are essential primes. In other words, L divides $p-1$ and $s-1$. Since P contains infinitely many pairs of twin primes, it contains infinitely many pairs of twin primes which are greater that N. As a result, we may assume that $s = p+2$. This means that L must divide $s-p=2$.

As a result of Proposition 5,

$$L = v_{p,s,k_{p,s}} = 2.$$

By Key Proposition 2 of [4], we may assume that $k_{p,s} = 1$. Hence $f_{v_{p,s,k_{p,s}},p}(q) = f_{2,p}(q)$ and $f_{v_{p,s,k_{p,s}},s}(q) = f_{2,r}(q)$ are super-compatible by Key Proposition 1′ of [4]. Therefore, there exists a subset $\mathscr{A}_2$ of $((\mathbb{Z}/2\mathbb{Z})^*)^T$, for some positive integer T, such that roots of $f_{2,p}(q)$ and $f_{2,r}(q)$ are represented by the collection of tuples

$$\bigcup_{\alpha\in\mathscr{A}_2} \{\alpha\} \times (\mathbb{Z}/p\mathbb{Z})^*$$

and

$$\bigcup_{\alpha\in\mathscr{A}_2} \{\alpha\} \times (\mathbb{Z}/r\mathbb{Z})^*$$

respectively.

Since $(\mathbb{Z}/2\mathbb{Z})^* = \{1\}$,

$$\mathscr{A}_2 = ((\mathbb{Z}/2\mathbb{Z})^*)^{T'}$$

for some integer $T' \leq T$. It can be verified that the monic polynomial whose roots are primitive $2p$-roots of unity and the monic polynomial whose roots are primitive $2r$-roots of unity represented by the collection of tuples

$$\bigcup_{\alpha \in (\mathbb{Z}/2\mathbb{Z})^*} \{\alpha\} \times (\mathbb{Z}/p\mathbb{Z})^*$$

and

$$\bigcup_{\alpha \in (\mathbb{Z}/2\mathbb{Z})^*} \{\alpha\} \times (\mathbb{Z}/r\mathbb{Z})^*$$

respectively are the cyclotomic polynomial, with coefficients in $\mathbb{Q}$ of order $2p$, $P_{2p}(q)$ and the cyclotomic polynomial, with coefficients in $\mathbb{Q}$ of order $2r$, $P_{2r}(q)$. As a result,

$$f_{2,p}(q) = P_{2p}(q)^{T'}$$

and

$$f_{2,r}(q) = P_{2r}(q)^{T'}.$$

Therefore, the coefficients of $f_{2,p}(q)$ and $f_{2,r}(q)$ are properly contained in $\mathbb{Q}$. Since the coefficients of the polynomials $f_{v_{p,s,k_{p,s}},p}(q)$ and $f_{v_{p,s,k_{p,s}},s}(q)$ are not properly contained in $\mathbb{Q}$ by definition of $k_{p,r}$, this is a contradiction. Therefore, the field of coefficients of Γ must be $\mathbb{Q}$, and thus Γ is constructible from quantum integers by Theorem 4. □

Proof. (proof of Corollary 1)

Suppose that the Twin Primes conjecture holds. Let Γ be a sequence of polynomials satisfying Functional Equation (2) and let P be its support base. Suppose that P has finite complement, i.e., P contains all but a finite number of primes. Then P must contain infinitely many pairs of twin primes. Therefore, Γ is constructible from quantum integers by Theorem 5. □

References

1. Borisov, A., Nathanson, M., Wang, Y.: Quantum Integers and Cyclotomy, Journal of Number Theory, Volume 109, Issue 1, 120–135 (November 2004)
2. Nathanson, M.: A Functional Equation Arising From Multiplication of Quantum Integers, Journal of Number Theory, Volume 103, No. 2, 214–233 (2003)
3. Nguyen, L.: On the Classification of Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
4. Nguyen, L.: On the Solutions of a Functional Equation Arising from Multiplication of Quantum Integers. (To appear in Journal of Number Theory)
5. Nguyen, L.: On the Existence of Sequences of Polynomial Satisfying a Functional Equation Arising from Multiplication of Quantum Integers with a Given Support Base. (To appear in Journal of Number Theory)

6. Nguyen, L.: Extension of Supports of Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
7. Nguyen, L.: Solutions with Infinite Support Bases of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
8. Nguyen, L.: On the Polynomial and Maximal Solutions to a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
9. Nguyen, L.: On the Rational Function Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
10. Nguyen, L.: On the Classification Rational Function Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
11. Nguyen, L.: On the Extension of Support of Rational Function Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)
12. Nguyen, L.: Quantum Equivalence Relation of the set of Rational Function Solutions of a Functional Equation Arising from Multiplication of Quantum Integers (preprint)

Exponential Sums and Distinct Points on Arcs

Øystein J. Rødseth

Dedicated in honour of Mel Nathanson's 60 th birthday

Summary Suppose that some harmonic analysis arguments have been invoked to show that the indicator function of a set of residue classes modulo some integer has a large Fourier coefficient. To get information about the structure of the set of residue classes, we then need a certain type of complementary result. A solution to this problem was given by Gregory Freiman in 1961, when he proved a lemma which relates the value of an exponential sum with the distribution of summands in semicircles of the unit circle in the complex plane. Since then, Freiman's Lemma has been extended by several authors. Rather than residue classes, one has considered the situation for finitely many arbitrary points on the unit circle. So far, Lev is the only author who has taken into consideration that the summands may be bounded away from each other, as is the case with distinct residue classes. In this paper, we extend Lev's result by lifting a recent result of ours to the case of the points being bounded away from each other.

Keywords Arcs · Distribution · Exponential sums · Unit circle

Mathematics Subject Classifications (2010). 11J71, 11K36, 11T23

1 Introduction

In additive combinatorics, and in additive combinatorial number theory, in particular, situations of the following type are rather common. Let A be a set of N residue classes modulo an integer m. Suppose that some harmonic analysis arguments have

Ø.J. Rødseth
Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N-5008 Bergen, Norway
e-mail: rodseth@math.uib.no; http://folk.uib.no/nmaoy

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_22,

been invoked to show that the indicator function of A has a large Fourier coefficient; that is

$$\max_{0\neq x\in\mathbb{Z}/m\mathbb{Z}} |\widehat{1}_A(x)| \geq \alpha N$$

for some $\alpha \in (0,1]$. Following Green [1], or see Tao and Vu [13], we might say that A has "Fourier bias". Using this information, one wishes to conclude that A has "combinatorial bias"; perhaps integer representatives of the residue classes in some affine image of A concentrate on some interval. For this, we then need tight upper bounds for the absolute values of the exponential sums $\widehat{1}_A(x)$. Now, the basic idea is that the absolute value of an exponential sum is small if the terms are, in some sense, uniformly distributed. The case m prime was applied by Freiman [2,3] in the proof of his "2.4-theorem". A proof of this theorem is presented in both Freiman's classical monograph [4] and in Mel Nathanson's beautiful book [9].

Freiman [2,3], Postnikova [10], Moran and Pollington [8], Lev [5,6], and Rødseth [11, 12] considered the situation in which one has finitely many arbitrary points on the unit circle in the complex plane, rather than a subset of $\mathbb{Z}/m\mathbb{Z}$. In possible applications, the points on the unit circle will sometimes be bounded away from each other, like, for instance, if we are looking at a subset of $\mathbb{Z}/m\mathbb{Z}$. If we add the condition that the points should be bounded away from each other, we could hope for sharper results. Indeed, by adding this assumption, Lev sharpened Freiman's Lemma to (2) below. Lev's result [5, Theorem 2] seems, however, to be the only result in the literature addressing this issue. In this paper, we prove a result which extends both Lev's result about points being bounded away from each other, and our main result in [11, 12], where we did not take this property into consideration.

First, we shall, however, present a version of Freiman's Lemma. We include two different proofs in the hope of giving the reader an impression of two recent techniques used in the search and study of results related to Freiman's Lemma.

2 Three Theorems

In the following, n, N, κ, and k are non-negative or positive integers. We write $\mathbb{U}$ for the unit circle in the complex plane; that is,

$$\mathbb{U} := \{z \in \mathbb{C} : |z| = 1\}.$$

An empty sum is taken as zero.

We now state Freiman's Lemma [2,3].

Theorem 1 (Freiman's Lemma). *Suppose that the complex numbers $z_1, \ldots, z_N \in \mathbb{U}$ have the property that any open semi-circle of $\mathbb{U}$ contains at most n of them. Then,*

$$|z_1 + \cdots + z_N| \leq 2n - N. \tag{1}$$

The complex numbers $z_1, \dots, z_N$ are not necessarily all distinct. The assumptions of the lemma imply that $N \le 2n$, and the result is *sharp* in the range $n \le N \le 2n$. That is, for every n and N satisfying these inequalities, there exist sequences $z_1, \dots, z_N$ with $z_j \in \mathbb{U}$, such that the hypotheses are satisfied and $|z_1 + \cdots + z_N|$ meets the bound $2n - N$. In this sense, Freiman's Lemma is best possible. The result has, however, been extended by Moran and Pollington [8], Lev [5, 6], and by Rødseth [11, 12].

The next theorem is Lev's sharpening of Freiman's Lemma at the expense of requiring the points to be bounded away from each other. Lev [5, Theorem 2] proved the following theorem.

Theorem 2 (Lev). *Let $\delta \in (0, \pi]$ satisfy $n\delta \le \pi$. Suppose that the complex numbers $z_1, \dots, z_N \in \mathbb{U}$ have the following two properties:*

(a) Any open semi-circle of $\mathbb{U}$ contains at most n of them.
(ii) Any open arc of $\mathbb{U}$ of length δ contains at most one of them.

Then

$$|z_1 + \cdots + z_N| \le \frac{\sin((2n - N)\delta/2)}{\sin(\delta/2)}. \tag{2}$$

So, by introducing the condition (ii), Lev reduced Freiman's bound $2n - N$ in (1) to the bound in (2), a refinement that, according to Lev, was crucial in [7]. The result is sharp for $n \le N \le 2n$. By letting $\delta \to 0^+$ in Lev's result, we recover Freiman's Lemma.

In this paper, we shall prove the following theorem.

Theorem 3. *Let $\delta, \varphi \in (0, \pi]$ satisfy $n\delta \le \varphi$. Suppose that the complex numbers $z_1, \dots, z_N \in \mathbb{U}$ have the following two properties:*

(i) Any open arc of $\mathbb{U}$ of length φ contains at most n of them.
(ii) Any open arc of $\mathbb{U}$ of length δ contains at most one of them.

Let $N = \kappa n + r$, $1 \le r \le n$, and assume that $(\kappa + 1)\varphi \le 2\pi$. Then we have

$$|z_1 + \cdots + z_N| \le \frac{\sin(r\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin((\kappa + 1)\varphi/2)}{\sin(\varphi/2)} + \frac{\sin((n - r)\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin(\kappa\varphi/2)}{\sin(\varphi/2)}. \tag{3}$$

Notice that $\kappa = \lceil N/n \rceil - 1$. Also notice that we do need some condition like $(\kappa + 1)\varphi \le 2\pi$, to be certain that there is room enough on $\mathbb{U}$ to place N points such that they satisfy the other conditions set in the theorem. This condition can also be written as $N \le \lfloor 2\pi/\varphi \rfloor n$.

The restriction $n\delta \le \varphi$ is no problem. For if $n\delta \ge \varphi$, then (i) follows from (ii) and can be omitted. Now, $|z_1 + \cdots + z_N|$ attains its maximum on any N-term geometric progression with ratio $\exp(\mathrm{i}\delta)$; hence,

$$|z_1 + \cdots + z_N| \le \left| \sum_{j=0}^{N-1} \exp(\mathrm{i}j\delta) \right| = \frac{\sin(N\delta/2)}{\sin(\delta/2)}. \tag{4}$$

The upper bound (3) is attained on the union of two finite 2-dimensional geometric progressions on $\mathbb{U}$, one consisting of geometric progressions with ratio $\exp(\mathrm{i}\delta)$ and r terms each, centered around the points

$$\exp((-\kappa + 2j)\mathrm{i}\varphi/2), \quad j = 0, 1, \ldots, \kappa,$$

and the other consisting of geometric progressions with ratio $\exp(\mathrm{i}\delta)$ and $n-r$ terms each, centered around the points

$$\exp((-\kappa + 1 + 2j)\mathrm{i}\varphi/2), \quad j = 0, 1, \ldots, \kappa - 1.$$

This shows that Theorem 3 is sharp for $N \le \lfloor 2\pi/\varphi \rfloor n$.

Clearly, the bound (3) in Theorem 3 can be replaced by the weaker, but smooth and nice, bound

$$|z_1 + \cdots + z_N| \le \frac{\sin(n\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin(\varphi N/(2n))}{\sin(\varphi/2)};$$

cf. Lev [6].

3 Two Proofs of Freiman's Lemma

For real numbers $\alpha < \beta$, the set of $z \in \mathbb{U}$ satisfying $\alpha < \arg z \le \beta$ for some value of $\arg z$, is an *open-closed* arc. In Freiman's Lemma, one often assumes that any open-closed (or closed-open) semi-circle of $\mathbb{U}$ contains at most n of the points z_j, instead of taking open semi-circles; cf. [4,9]. It is, however, easy to see that the two variants of the hypotheses are equivalent; cf. Sect. 4.2.

Freiman's proof of Theorem 1 was simplified by Postnikova [10], and it is this proof we find in the books [4] and [9]. Here, we shall present two other proofs in an attempt to give the reader a pleasant introduction to two techniques recently employed in the quest for extensions of Freiman's Lemma. The two proofs are rather different. One proof can be characterized as topological–combinatorial or as a perturbation method, and is due to Lev (extracted from the proof of [5, Theorem 2]). The other proof uses properties of a certain Fourier coefficient, and is independently due to Lev and the present author.

3.1 First Proof

Assume that Freiman's Lemma is false. We consider the smallest N for which there exists an N-term sequence which satisfies the hypotheses, but violates (1). Then $N > 1$. By considering open semi-circles, the set of N-term sequences satisfying

the hypotheses forms a closed subset of the compact topological space $\mathbb{U}^N$, and is itself compact. By the continuity of the function $z_1, \ldots, z_N \mapsto |S|$, where $S := z_1 + \cdots + z_N$, we thus have that $|S|$ attains a maximum value on some N-term sequence $Z := z_1, \ldots, z_N$, which satisfies the hypotheses. Then $|S| > 2n - N$. A rotation of Z shows that we may assume that S is real and nonnegative.

Suppose that there is a $z \neq 1$ in Z. By symmetry, we may assume that $\operatorname{Arg} z < 0$, using the interval $[-\pi, \pi)$ for the principal argument. Replacing z by $z\exp(\mathrm{i}\varepsilon)$ for a small $\varepsilon > 0$, we get an increase in $|S|$. Thus the replacement results in violation of the hypotheses in Freiman's Lemma. The only possibility is that the replacement produces an open semi-circle with more than n points from Z; hence $-z$ also belongs to Z.

We now remove $\pm z$ from Z. This gives us a sequence Z' satisfying the hypotheses, and with parameters $N' = N - 2$ and $n' = n - 1$. Denoting the sum of the terms of Z' by S', we have by the minimality of N,

$$S = S' \leq 2n' - N' = 2n - N,$$

a contradiction. Thus, all terms of Z are equal to 1, and $N = S > 2n - N$. But a semi-circle containing 1 contains N terms from Z; hence $n \geq N$, and again we have a contradiction.

3.2 Second Proof

Assume that $z_1, \ldots, z_N$ satisfy the hypotheses. We shall prove (1), and can without loss of generality assume that S is real and nonnegative. Let $K(\theta)$ denote the number of values $j \in [1, N]$ such that $\theta - \pi \leq \arg z_j < \theta$ for some value of $\arg z_j$. Then we have

$$K(\theta) + K(\theta + \pi) = N, \tag{5}$$

so that $N - n \leq K(\theta) \leq n$.

Moreover,

$$\int_{-\pi}^{\pi} K(\theta) \sin\theta \, \mathrm{d}\theta = \sum_{j=1}^{N} \int_{\arg z_j}^{\arg z_j + \pi} \sin\theta \, \mathrm{d}\theta = 2S,$$

and we obtain

$$\begin{aligned} 2S &= \left(\int_{-\pi}^{0} + \int_{0}^{\pi} \right) K(\theta) \sin\theta \, \mathrm{d}\theta \\ &\leq (N - n) \int_{-\pi}^{0} \sin\theta \, \mathrm{d}\theta + n \int_{0}^{\pi} \sin\theta \, \mathrm{d}\theta \\ &= 4n - 2N. \end{aligned}$$

4 Proof of Theorem 3

We now turn to the proof of Theorem 3. In an attempt to make the proof more readable, we split the proof into several parts.

4.1 Notation

Set

$$\omega = \exp(\mathrm{i}\varphi/2) \quad \text{and} \quad \rho = \exp(\mathrm{i}\delta/2).$$

We often denote a sequence $z_1, \ldots, z_N \in \mathbb{U}^N$ by Z, and we write $S = z_1 + \cdots + z_N$. If the sequence $Z \in \mathbb{U}^N$ satisfies the assumptions of Theorem 3 for a certain value of n, we say that Z is an (N, n)-*admissible* sequence. We use the interval $[-\pi, \pi)$ for the principal argument $\operatorname{Arg} z$ of a nonzero complex number z.

4.2 Arcs

An arc (u, v), where $u, v \in \mathbb{U}$, consists of the points we pass in moving counterclockwise from u to v. Almost all arcs in this paper have lengths at most 2π. In spite of the similarity of notation, an arc cannot be confused with a real interval.

Consider the two statements:

(i) Any open arc of $\mathbb{U}$ of length φ contains at most n of the points z_j.
(i′) Any open-closed arc of $\mathbb{U}$ of length φ contains at most n of the points z_j.

Clearly, (i′) implies (i). On the other hand, if (i′) fails, then (i) fails. For if an open-closed arc $(u, u\exp(\mathrm{i}\varphi)]$, $u \in \mathbb{U}$, of $\mathbb{U}$ contains $n + 1$ of the points z_j, then, for a sufficiently small real $\varepsilon > 0$, the open arc $(u\exp(\mathrm{i}\varepsilon), u\exp(\mathrm{i}\varphi + \mathrm{i}\varepsilon))$ contains the same $n + 1$ points. Thus (i) and (i′) are equivalent. (And the reason is, of course, that we only consider finitely many points z_j.)

4.3 Assumptions

The two bounds (3) and (4) coincide for $n\delta = \varphi$. For the proof of Theorem 3, we may therefore assume that $n\delta < \varphi$. Moreover, we observe that the right-hand side of (3) is continuous as a function of φ on the real interval $(n\delta, 2\pi/(\kappa + 1)]$. Hence it suffices to prove the assertion of Theorem 3 in the case $\varphi < 2\pi/(\kappa + 1)$.

We shall prove Theorem 3 by contradiction. We therefore assume the theorem false. Choose the smallest nonnegative integer N for which there exists an n such that (3) fails for some (N, n)-admissible sequence. Then $N > 1$.

4.4 Geometric Progressions

A geometric progression Δ in $\mathbb{U}$ with ratio ρ^2 is called a δ-*progression*. If we write this progression as

$$u\rho^{r-1-2j}, \quad j = 0, \ldots, r-1, \tag{6}$$

for some $u \in \mathbb{U}$, then Δ is a progression of *length* r, *centered around* u. The point u may, or may not, belong to the progression. The point $u\rho^{r-1}$ is the *first* and $u\rho^{-(r-1)}$ is the *end* point (element, term) of the progression. If we multiply each term of Δ by $v \in \mathbb{U}$, we get a new δ-progression denoted Δv. If all terms of Δ belong to Z, we say that Δ is *in* Z. The δ-progression (6) is *maximal* in Z, if the progression is in Z, but neither $u\rho^{r+1}$ nor $u\rho^{-r-1}$ belongs to Z.

Notice that if Δ is a δ-progression in Z of length r, then $r \le n$. The reason is the inequality $n\delta < \varphi$, and that an open arc of $\mathbb{U}$ of length φ contains at most n points from Z.

4.5 Compactness/Continuity

The arcs in both (i) and (ii) are open. Therefore, by a standard compactness continuity argument, $|S|$ attains a maximum on the set of (N, n)-admissible sequences. Let the maximum be attained at the (N, n)-admissible sequence $Z := z_1, \ldots, z_N$. A rotation of Z shows that we may assume that the matching $S = z_1 + \cdots + z_N$ is real and nonnegative.

4.6 Dispersion

Recall that $N > 1$ and $N = \kappa n + r$, $1 \le r \le n$. Suppose that Z is a δ-progression. The length of this δ-progression is N.

If $\kappa = 0$, then

$$|z_1 + \cdots + z_N| \le \frac{\sin(r\delta/2)}{\sin(\delta/2)},$$

and (3) holds, contrary to hypothesis. Thus $N \ge n + 1$, and there is a closed arc of $\mathbb{U}$ of length $n\delta$ containing $n + 1$ of the points z_j. But we have $n\delta < \varphi$, so this contradicts condition (i) of Theorem 3. Therefore, Z is not a δ-progression.

4.7 Perturbation

We have just seen that Z is not a δ-progression. This implies that there exists at least one point $z \in Z$, $z \ne 1$, such that $\operatorname{Arg} z > 0$ and $z\rho^{-2} \notin Z$, or $\operatorname{Arg} z < 0$ and $z\rho^2 \notin Z$. Choose such a z, as close to -1 as possible. By symmetry, we can assume that $\operatorname{Arg} z < 0$. We split this case into the two cases determined by (7) and (8).

4.7.1 Case I

First, we assume that there is an integer λ in the interval $1 \le \lambda \le \kappa$ satisfying

$$-\frac{\lambda\varphi}{2} \le \operatorname{Arg} z < -\frac{(\lambda-1)\varphi}{2}. \tag{7}$$

The length of the shortest arc (z, z') for $z \ne z' \in Z$ is at least δ. Since $z\rho^2 \notin Z$, the arc (z, z') has length greater than δ. We start the perturbation procedure by performing a small counter-clockwise rotation of z along the unit circle; that is, we replace z by $z\exp(\mathrm{i}\varepsilon)$ for a small $\varepsilon > 0$. If ε is sufficiently small, then the length of the open arc $(z\exp(\mathrm{i}\varepsilon), z')$ is still at least δ, and the rotation of z does not disturb the truth of (ii). Since $\operatorname{Arg} z < 0$, there will, however, be an increase in $|S|$. Thus, the rotation violates (i′). Therefore, we have $z\omega^2 \in Z$, and the open-closed arc $(z, z\omega^2]$ contains exactly n terms from Z. If $z\omega^2\rho^2 \in Z$, then the arc $(z\rho^2, z\omega^2\rho^2]$ contains $n+1$ points of Z. Therefore, $z\omega^2\rho^2 \notin Z$.

Next, we perform a small rotation of both the points z and $z\omega^2$ simultaneously and counter-clockwise along the unit circle. We have

$$\operatorname{Arg}\left(z + z\omega^2\right) = \operatorname{Arg} z + \frac{\varphi}{2},$$

and if $\lambda \ge 2$, the rotation makes $|S|$ increase. Hence $z\omega^4 \in Z$. We also see that $z\omega^4\rho^2 \notin Z$, and that the arc $(z\omega^2, z\omega^4]$ contains exactly n terms from Z.

Using the right inequality in (7), we find for $1 \le j \le \lambda$, that

$$\operatorname{Arg}\left(\sum_{\ell=0}^{j-1} z\omega^{2\ell}\right) = \operatorname{Arg} z + \frac{(j-1)\varphi}{2} < -\frac{(\lambda-j)\varphi}{2} \le 0.$$

This shows that we may continue the perturbation process until we eventually obtain

$$z\omega^{2j} \in Z \quad \text{and} \quad z\omega^{2j}\rho^2 \notin Z \quad \text{for} \quad j = 0, 1, \ldots, \lambda.$$

In addition, for $j = 1, 2, \ldots, \lambda$, each arc $(z\omega^{2(j-1)}, z\omega^{2j}]$ contains exactly n points from Z.

By the left inequality in (7), we have

$$\operatorname{Arg}(z\omega^{2\lambda}) \ge -\operatorname{Arg} z.$$

Using the definition of z, it follows that if we start at the point $z\omega^{2\lambda}\rho^2$, which is in $\mathbb{U}$ but not in Z, and move counter-clockwise on $\mathbb{U}$, then the first point in Z we meet, is the end point of a maximal δ-progression in Z with z as its first element.

The open-closed arc $(z, z\omega^{2\lambda}]$ contains λn points from Z, and in addition we have $r' \geq 1$ points in the maximal δ-progression in Z with z as first point. Thus, we have

$$\lambda n + r' = N = \kappa n + r, \quad 1 \leq r', r \leq n,$$

so that $\lambda = \kappa$ and $r' = r$. This completes the first case (7).

4.7.2 Case II

Second, if z does not satisfy (7) for $\lambda = \kappa$, then

$$-\pi \leq \operatorname{Arg} z < -\frac{\kappa\varphi}{2}, \tag{8}$$

and we can do one more step in the perturbation process. By the inequality $(\kappa + 1)\varphi < 2\pi$, we have that the arc $(z, z\omega^{2(\kappa+1)}]$ contains exactly $(\kappa + 1)n$ points from Z. In addition, we have the point z itself. Thus we have found $(\kappa + 1)n + 1$ points in Z. Since Z contains at most $(\kappa + 1)n$ points, we have a contradiction. Thus, there are no points in Z satisfying (8).

4.8 Primary Points

For the $z \in Z$ defined at the beginning of Sect. 4.7, we now know that for $j = 1, 2, \ldots, \kappa$, each open-closed arc $(z\omega^{2(j-1)}, z\omega^{2j}]$ contains exactly n points from Z. In addition, there is the maximal δ-progression Δ of length r with z as first element. This accounts for all elements in Z.

Thus we have

$$z\omega^{2j} \in Z, \quad j = 0, 1, \ldots, \kappa.$$

Suppose that $z\rho^{-2} \in \Delta$. The open-closed arc $(z, z\omega^2]$ contains exactly n points from Z. If $z\omega^2\rho^{-2} \notin Z$, then the closed-open arc $[z\rho^{-2}, z\omega^2\rho^{-2})$ contains $n + 1$ points from Z, contrary to hypotheses. Thus $z\omega^2\rho^{-2} \in Z$. Continuing in the same manner, we get (with a slight abuse of notation) $\Delta\omega^2 \subseteq Z$. Repeating the argument, we ultimately obtain,

$$\Delta\omega^{2j} \subseteq Z \quad \text{for } j = 0, 1, \ldots, \kappa.$$

Notice that for $j \geq 1$, the δ-progression $\Delta\omega^{2j}$ is not necessarily maximal.

We set

$$Z_1 = \bigcup_{j=0}^{\kappa} \Delta\omega^{2j}.$$

Then $N_1 := Z_1 = \kappa r + r$. We write S_1 for the sum of all elements in Z_1, and we want to find an upper bound for $|S_1|$. (We may have $N_1 = N$, so we cannot use (3).) We can, however, easily determine the exact value of $|S_1|$,

$$|S_1| = \left| \sum_{\ell=0}^{r-1} \rho^{-2\ell} \sum_{j=0}^{\kappa} z\omega^{2j} \right| = \frac{\sin(r\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin((\kappa+1)\varphi/2)}{\sin(\varphi/2)}. \tag{9}$$

If $\kappa = 0$ or $r = n$, then $Z = Z_1$, and by (9), (3) holds, contrary to hypothesis. Thus we have $\kappa \geq 1$ and $1 \leq r < n$.

4.9 Secondary Points

We set

$$Z_2 = Z \setminus Z_1,$$

and put $N_2 = Z_2$. Then $N_2 = (\kappa - 1)(n - r) + (n - r)$. We have that Z_2 is $(N_2, n - r)$-admissible, and since $N_2 < N$, we can use (3). Writing S_2 for the sum of the terms in Z_2, we obtain

$$|S_2| \leq \frac{\sin((n-r)\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin(\kappa\varphi/2)}{\sin(\varphi/2)}. \tag{10}$$

We have

$$S \leq |S_1| + |S_2|,$$

and (9) and (10) show that (3) holds; again we have reached a contradiction. This concludes the proof of Theorem 3.

5 Closing Remarks

We can also state Theorem 3 in a slightly different way.

Theorem 4. *Let $\delta, \varphi \in (0, \pi]$ satisfy $n\delta \leq \varphi$. Assume that the complex numbers $z_1, \ldots, z_N \in \mathbb{U}$ have the two properties (i) and (ii) stated in Theorem 3. Then we have*

$$\begin{aligned} |z_1 + \cdots + z_N| \leq & \frac{\sin((N-(k-1)n)\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin(k\varphi/2)}{\sin(\varphi/2)} \\ & + \frac{\sin((kn-N)\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin((k-1)\varphi/2)}{\sin(\varphi/2)}, \end{aligned} \tag{11}$$

for any positive integer $k \leq 2\pi/\varphi$.

Denote the right-hand side of (11) by L_k. We easily see that

$$L_{k+1} - L_k = 2\frac{\sin((kn-N)\delta/2)}{\sin(\delta/2)} \cdot \frac{\sin(k\varphi/2)}{\sin(\varphi/2)} \left(\cos\frac{n\delta}{2} - \cos\frac{\varphi}{2}\right),$$

so that

$$L_1 \geq \ldots \geq L_{\lceil N/n \rceil} \leq L_{\lceil N/n \rceil + 1} \leq \ldots \leq L_{\lfloor 2\pi/\varphi \rfloor}.$$

Thus L_k attains its minimum at $k = \lceil N/n \rceil$ (and also at $k = N/n + 1$ if n divides N). Therefore, Theorems 3 and 4 are essentially one and the same.

Now, by letting $\delta \to 0^+$, we recover the main result of [11, 12].

The following direct extension of Theorem 2 is an immediate consequence of Theorem 4.

Corollary 1. *Let $k \geq 2$ be an integer, and let $\delta \in (0, \pi]$ satisfy $n\delta \leq 2\pi/k$. Suppose that the complex numbers $z_1, \ldots, z_N \in \mathbb{U}$ have the following two properties:*

(b) Any open arc of $\mathbb{U}$ of length $2\pi/k$ contains at most n of them.
(ii) Any open arc of $\mathbb{U}$ of length δ contains at most one of them.

Then

$$|z_1 + \cdots + z_N| \leq \frac{\sin((kn-N)\delta/2)}{\sin(\delta/2)}.$$

If k is even, then Corollary 1 is an easy consequence of Theorem 2. This does not seem to be the case for k odd. Corollary 1 is sharp for $(k-1)n \leq N \leq kn$. By letting $\delta \to 0^+$, we recover the result of Moran and Pollington [8].

For applications, one would perhaps, although not always necessary, turn the upper bound for $|z_1 + \cdots + z_N|$ into a lower bound for the number of terms z_j in at least one open arc of $\mathbb{U}$ of length φ.

In closing, let us consider such a lower bound for the residue class situation mentioned in the introduction. Then we have an N-set A of residue classes modulo an integer $m > 1$. Set $\delta = 2\pi/m$, and put $\varphi = 2\pi/k$ for some integer $k \geq 2$.

Let

$$\widehat{1}_A(1) = \sum_{a \in A} \exp(2\pi i a/m).$$

By Corollary 1, there exist integers u, v satisfying $u \leq v < u + m/k$ such that the image of the interval $[u, v]$ under the canonical homomorphism $\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ contains n_0 elements of A, where

$$n_0 \geq \left\lceil \frac{1}{k}\left(N + \frac{\operatorname{Arcsin}(|\widehat{1}_A(1)|\sin(\pi/m))}{\pi/m}\right)\right\rceil;$$

cf. [5, Corollary 2], [11, Sect. 6].

References

1. B. Green, Review MR2192089(2007a:11029) in MathSciNet
2. G. A. Freiman, Inverse problems of additive number theory. On the addition of sets of residues with respect to a prime modulus (Russian). Dokl. Akad. Nauk SSSR **141**, 571–573 (1961)
3. G. A. Freiman, Inverse problems of additive number theory. On the addition of sets of residues with respect to a prime modulus. Sov. Math.-Dokl. **2**, 1520–1522 (1961)
4. G. A. Freiman, *Foundations of a Structural Theory of Set Addition*. Translations of Mathematical Monographs, Vol. 37, American Mathematical Society, Providence, R. I., 1973
5. V. F. Lev, Distribution of points on arcs. Integers **5** (2),#A11, 6 pp. (electronic) (2005)
6. V. F. Lev, More on points and arcs. Integers **7** (2), #A24, 3 pp. (electronic) (2007)
7. V. F. Lev, Large sum-free sets in $\mathbb{Z}/p\mathbb{Z}$. Israel J. Math **154**, 221–234 (2006)
8. W. Moran, A. D. Pollington, On a result of Freiman. Preprint 1992
9. M. B. Nathanson, *Additive Number Theory: Inverse Problems and the Geometry of Sumsets*. Graduate Texts in Mathematics, Vol. 165, Springer, New York, 1996
10. L. P. Postnikova, Fluctuations in the distribution of fractional parts. Dokl. Akad. Nauk SSSR, **161**, 1282–1284 (1965)
11. Ø. J. Rødseth, Distribution of points on the circle. J. Number Theory **127**, (1), 127–135 (2007)
12. Ø. J. Rødseth, Addendum to "Distribution of points on the circle". J. Number Theory **128**, (6), 1889–1892 (2008)
13. T. Tao, V. Vu, *Additive Combinatorics*. Cambridge Studies in Advanced Mathematics, Vol. 105, Cambridge University Press, 2006

New Vacca-Type Rational Series for Euler's Constant γ and Its "Alternating" Analog $\ln \frac{4}{\pi}$

Jonathan Sondow

Summary We recall a pair of logarithmic series that reveals $\ln(4/\pi)$ to be an "alternating" analog of Euler's constant γ. Using the binary expansion of an integer, we derive linear, quadratic, and cubic analogs for $\ln(4/\pi)$ of Vacca's rational series for γ. Using a generalization of Vacca's series to integer bases $b \geq 2$, due in part to Ramanujan, we extend Addison's cubic, rational, base 2 series for γ to faster base b series. Open problems on further extensions of the results are discussed, and a history of the formulas is given.

Keywords Alternating Euler constant · Acceleration of series · Binary expansion · Euler's constant · Generalized Euler constant · Rational series · Vacca's series

Mathematics Subject Classifications (2010). 11Y60, 65B10

1 Introduction

If we denote

$$\gamma^+ := \gamma \quad \text{and} \quad \gamma^- := \ln \frac{4}{\pi},$$

where γ is *Euler's constant*, then the formulas

$$\gamma^{\pm} = \sum_{n=1}^{\infty} (\pm 1)^{n-1} \left(\frac{1}{n} - \ln \frac{n+1}{n} \right) \tag{1}$$

show that $\ln(4/\pi)$ is an *"alternating Euler constant"* [18].

J. Sondow
209 West 97th Street, New York, NY 10025, USA
e-mail: jsondow@alumni.princeton.edu

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_23,

The following *rational, linear series* (i.e., the nth term is a rational number with denominator a linear function of n) is known as *Vacca's series for Euler's constant* [22]:

$$\gamma = \sum_{n=1}^{\infty} (-1)^n \frac{\lfloor \log_2 n \rfloor}{n}. \tag{2}$$

Here $\lfloor . \rfloor$ is the floor function and $\log_b n := \ln n / \ln b$ is the logarithm of n to the base $b > 1$.

We give an analogous series for $\ln(4/\pi)$ and prove the two formulas simultaneously.

Theorem 1. *For $i = 0, 1$ and $n = 1, 2, 3, \ldots$, denote by $N_i(n)$ the number of i's in the binary expansion of n, and set $N_i(0) := 0$. Then $\gamma = \gamma^+$ and $\ln(4/\pi) = \gamma^-$ are given by the linear, rational series*

$$\gamma^{\pm} = \sum_{n=1}^{\infty} (-1)^n \frac{N_1(\lfloor n/2 \rfloor) \pm N_0(\lfloor n/2 \rfloor)}{n}, \tag{3}$$

where the one for γ^+ is Vacca's series for Euler's constant.

The two series begin

$$\begin{aligned} \gamma &= \frac{1}{2} - \frac{1}{3} + \frac{2}{4} - \frac{2}{5} + \frac{2}{6} - \frac{2}{7} + \frac{3}{8} - \frac{3}{9} + \frac{3}{10} - \frac{3}{11} + \frac{3}{12} - \frac{3}{13} + \frac{3}{14} - \frac{3}{15} + \cdots, \\ \ln \frac{4}{\pi} &= \frac{1}{2} - \frac{1}{3} + \frac{2}{6} - \frac{2}{7} - \frac{1}{8} + \frac{1}{9} + \frac{1}{10} - \frac{1}{11} + \frac{1}{12} - \frac{1}{13} + \frac{3}{14} - \frac{3}{15} - \frac{2}{16} + \frac{2}{17} + \cdots. \end{aligned} \tag{4}$$

If in each of them we group the terms in pairs, we obtain simpler formulas than (3).

Corollary 1. *We have the following quadratic, rational series for $\gamma = \gamma^+$ and $\ln(4/\pi) = \gamma^-$:*

$$\gamma^{\pm} = \sum_{n=1}^{\infty} \frac{N_1(n) \pm N_0(n)}{2n(2n+1)}. \tag{5}$$

In 1967 Addison [1] (see also Behrmann and van Lint [4, 23]) used an integral representation of the Riemann zeta function to obtain a cubic, rational series for γ. In our notation, *Addison's series for Euler's constant* can be written in either of the equivalent forms

$$\gamma = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{N_1(n) + N_0(n)}{2n(2n+1)(2n+2)} = \frac{1}{2} + \sum_{\text{even } n>0} \frac{\lfloor \log_2 n \rfloor}{n(n+1)(n+2)}.$$

We derive a similar series for the alternating Euler constant, by modifying a formula in [2].

Proposition 1. *We have the cubic, rational series*

$$\ln\frac{4}{\pi} = \frac{1}{4} - \sum_{\text{even } n>0} \frac{N_0(n) - N_1(n)}{n(n+1)(n+2)}.$$

Next we generalize Addison's base 2 series to faster base b series for γ.

Theorem 2. *For $b = 2, 3, 4, \ldots$, let $P_b(x) \in \mathbf{Z}[x]$ be the polynomial*

$$P_b(x) := (x+1)(x+2)\cdots(x+b-1)\sum_{k=1}^{b-1}\frac{k(b-k)}{x+k}$$

of degree $b-2$. Then Euler's constant is given by the base b, rational series

$$\gamma = \frac{1}{2} + \sum_{\substack{n>0\\ b|n}} \frac{\lfloor \log_b n\rfloor P_b(n)}{n(n+1)\cdots(n+b)}. \tag{6}$$

Moreover, the series converges faster as b increases. In fact, for fixed b, the nth term is

$$\frac{\lfloor \log_b bn\rfloor P_b(bn)}{bn(bn+1)\cdots(bn+b)} \sim \frac{1-b^{-2}}{\ln b}\frac{\ln n}{6n^3} \quad \text{as } n\to\infty\,. \tag{7}$$

Example 1. Since $P_2(x) = 1$, the base 2 series is Addison's. The base 3 series is the faster

$$\gamma = \frac{1}{2} + \sum_{\substack{n>0\\ 3|n}} \frac{\lfloor \log_3 n\rfloor (4n+6)}{n(n+1)(n+2)(n+3)}.$$

For instance, the 4th partial sums of the base 2 and base 3 series are equal to $0.5684\ldots$ and $0.5702\ldots$, which approximate $\gamma = 0.5772\ldots$ correctly to one and two decimal places, respectively.

To prove Proposition 1 and Theorem 2, we generalize an averaging technique which Krämer [11, Sect. 5.2.2] used in proving Addison's formula by accelerating Vacca's series. To derive the base b series of Theorem 2, we accelerate the *generalized Vacca series*

$$\gamma = \sum_{n=1}^{\infty} \sigma_n \frac{\lfloor \log_b n\rfloor}{n}, \tag{8}$$

where

$$\sigma_n = \sigma_{n,b} := \begin{cases} b-1 & \text{if } b \mid n,\\ -1 & \text{otherwise.}\end{cases}$$

(For other methods of accelerating (8), leading to base b, rational series for γ different from (6), see [21].)

The generalized Vacca series (8) appeared in 1968 as a problem in the Monthly [6, 9] proposed by Carlitz, with solution by Harborth. In 2000, Berndt and Bowman [5] derived (8) from an integral for Euler's constant due to Ramanujan [14, pp. 274–275].

In 1897, Nielsen [12] discovered a series closely related to (5) for γ^+. "Vacca's series" (2) was obtained independently by Franklin [7] in 1883, Jacobsthal [10] in 1906, Vacca [22] in 1909, Sandham [3, 16] in 1949, and Gerst [8] in 1969. Gerst also rediscovered Nielsen's series, and used it to give a new proof of Addison's formula. Papers on Vacca-type series for γ by these and other authors, including Glaisher (1910), Hardy (1912), Kluyver (1924), Brun (1938), S. Selberg (1940, 1967), Gosper (1972), Koecher (1989), and Bauer (1990), are discussed in [5] and [11, Sects. 5.2.2 and 10.1].

The genesis of Theorem 1 was as follows. Rivoal found a *nonrational* analog for $\ln(4/\pi)$ of Vacca's series for γ (see [15, Théorème 2]). We then tried to find a *rational* analog. Using relations (1) for γ^-, (9), and (10), we computed the first few terms of the series (4) for $\ln(4/\pi)$. A search for their numerators $1, -1, 2, -2, -1, 1, 1, -1, \ldots$ in The On-Line Encyclopedia of Integer Sequences [17] uncovered the related Sequence A037861. It suggested the formulas (3), and we proved them in the 2005 preprint [19] (an early version of the present paper). The proof given here is the same, except that to prove Lemma 1 we use a geometric interpretation of the series (1).

Recently, inspired by [19], Allouche, Shallit, and the author generalized Corollary 1 and Addison's formula to rational, base 2 series for certain classes of constants [2, Theorems 1 and 2]. (The proofs in [2] are quite different from those below. Results from [2] are used only in the proof of Proposition 1 and in the final section.)

In Sect. 2 we prove our main results. Open problems on extensions of them are discussed in Sect. 3.

2 Proofs

We first establish three lemmas.

Lemma 1. *For $n > 0$, the area A_n of the curvilinear triangular region bounded by the hyperbola $y = 1/x$ and the lines $y = 1/n$ and $x = n + 1$, is equal to*

$$A_n = \frac{1}{n} - \ln\frac{n+1}{n}\,. \tag{9}$$

The areas satisfy the relation

$$A_n = \frac{1}{2n(2n+1)} + A_{2n} + A_{2n+1} \tag{10}$$

and the inequalities

$$\frac{1}{2n(n+1)} < A_n < \frac{1}{n(n+1)}. \tag{11}$$

Proof. Calculation of an integral yields (9), and (10) follows by substitution. Since the region is contained in the rectangle bounded by the lines $y = 1/n$, $y = 1/(n+1)$, $x = n$, and $x = n+1$, and contains a triangle formed by bisecting the rectangle with a diagonal, (11) holds. □

As an application, substitute (9) into the series (1), then replace $A_1, A_2, \ldots$ with suitable bounds from (11). Summing the resulting series, we deduce the estimates $1/2 < \gamma < 1$ and $1/2 < \ln(\pi/\sqrt{2}) < 1$.

Lemma 2. *For $n \geq 0$, set*

$$\Delta^{\pm}(n) := N_1(n) \pm N_0(n).$$

Then the following relation holds when $n > 0$:

$$\Delta^{\pm}(\lfloor n/2 \rfloor) + (\pm 1)^{n-1} = \Delta^{\pm}(n).$$

Proof. This is easily verified by considering the cases n even and n odd. □

Lemma 3. *For $k > 0$, denote the kth partial sum of the series (1) for $\gamma^{\pm}$ by*

$$S_k^{\pm} := \sum_{n=1}^{k} (\pm 1)^{n-1} A_n.$$

Then

$$S_{2^k-1}^{\pm} = \sum_{n=1}^{2^{k-1}-1} \frac{\Delta^{\pm}(n)}{2n(2n+1)} + R_k^{\pm}, \tag{12}$$

where the remainder term is

$$R_k^{\pm} := \sum_{n=2^{k-1}}^{2^k-1} \Delta^{\pm}(n) A_n, \tag{13}$$

Proof. We induct on k. For $k = 1$, we have $S_1^{\pm} = A_1 = R_1^{\pm}$, as required (since the sum in (12) is empty). Now write

$$S_{2^{k+1}-1}^{\pm} = S_{2^k-1}^{\pm} + \sum_{n=2^k}^{2^{k+1}-1} (\pm 1)^{n-1} A_n,$$

and invoke the inductive hypotheses (12) and (13), substituting (10) into (13). Using the identity

$$\sum_{n=2^{k-1}}^{2^k-1} \Delta^{\pm}(n)\,(A_{2n} + A_{2n+1}) = \sum_{n=2^k}^{2^{k+1}-1} \Delta^{\pm}(\lfloor n/2 \rfloor)\,A_n$$

and Lemma 2, the inductive step follows. □

We now prove the results stated in Sect. 1.

Proof (Theorem 1 and Corollary 1). It is easy to see that the formulas (3) and (5) are equivalent. Indeed, for even $n = 2\nu$ we have $\lfloor n/2 \rfloor = \lfloor (n+1)/2 \rfloor = \nu$, and it follows that the even-odd pairs of terms in (3) correspond to the terms of (5). Thus, to prove the formulas, it suffices to show that

$$\gamma^{\pm} = \sum_{n=1}^{\infty} \frac{\Delta^{\pm}(n)}{2n(2n+1)}. \tag{14}$$

Since by its definition $S_k^{\pm} \to \gamma^{\pm}$ as $k \to \infty$, we can deduce (14) using (12) if $R_k^{\pm} \to 0$ as $k \to \infty$. But by (13), (11), and telescoping, we see that

$$\left|R_k^{\pm}\right| \le \sum_{n=2^{k-1}}^{2^k-1} \left|\Delta^{\pm}(n)\right| A_n < \sum_{n=2^{k-1}}^{2^k-1} k\left(\frac{1}{n} - \frac{1}{n+1}\right) = \frac{k}{2^k} \to 0.$$

Finally, since

$$N_1(\lfloor n/2 \rfloor) + N_0(\lfloor n/2 \rfloor) = \lfloor \log_2 n \rfloor \quad \text{for } n \ge 1,$$

the series (3) for γ^+ is indeed Vacca's series (2) for γ. □

Proof (Proposition 1). The series $\ln 2 = 1 - 1/2 + 1/3 - \cdots$ can be written in the two equivalent forms

$$1 - \ln 2 = \sum_{n=1}^{\infty}\left(\frac{1}{2n} - \frac{1}{2n+1}\right) = \frac{1}{2} + \sum_{n=1}^{\infty}\left(-\frac{1}{2n+1} + \frac{1}{2n+2}\right).$$

Averaging them gives the accelerated series

$$\frac{3}{4} - \ln 2 = \sum_{n=1}^{\infty} \frac{1}{2n(2n+1)(2n+2)}. \tag{15}$$

Now consider the formula [2, Example 2]

$$\frac{1}{2} - \ln\frac{\pi}{2} = \sum_{n=1}^{\infty} \frac{N_1(n) - N_0(n)}{2n(2n+1)(2n+2)}.$$

Substituting $N_1(n) = N_1(2n)$ and $N_0(n) = N_0(2n) - 1$ enables us to use the formula (15), and to replace $2n$ with even n. The result is the desired series for $\ln 2 - \ln(\pi/2) = \ln(4/\pi)$. □

Proof (Theorem 2). The idea is to apply the acceleration method in the preceding proof to the base b series (8) for γ.

If we group the terms of (8) in b-tuples with positive first member, we obtain

$$\gamma = \sum_{\substack{n>0 \\ b|n}} \lfloor \log_b n \rfloor \left(\frac{b-1}{n} - \frac{1}{n+1} - \frac{1}{n+2} - \cdots - \frac{1}{n+b-1} \right).$$

If instead we group the terms in b-tuples with positive last member, we deduce that

$$\gamma = 1 + \sum_{\substack{n>0 \\ b|n}} \lfloor \log_b n \rfloor \left(-\frac{1}{n+1} - \frac{1}{n+2} - \cdots - \frac{1}{n+b-1} + \frac{b-1}{n+b} \right),$$

where the first 1 is the sum of the series $\sum_{r=1}^{\infty}(b-1)/b^r = 1$. Averaging the two series for γ gives

$$\gamma = \frac{1}{2} + \frac{1}{2} \sum_{\substack{n>0 \\ b|n}} \lfloor \log_b n \rfloor \left(\frac{b-1}{n} - \frac{2}{n+1} - \frac{2}{n+2} - \cdots - \frac{2}{n+b-1} + \frac{b-1}{n+b} \right).$$

The expression in parentheses is equal to

$$\sum_{k=1}^{b-1} \left(\frac{1}{n} - \frac{2}{n+k} + \frac{1}{n+b} \right) = \frac{1}{n(n+b)} \sum_{k=1}^{b-1} \left(2k - b + \frac{2k(b-k)}{n+k} \right).$$

Since $\sum_{k=1}^{b-1}(2k-b) = 0$, we obtain the formula

$$\gamma = \frac{1}{2} + \sum_{\substack{n>0 \\ b|n}} \sum_{k=1}^{b-1} \frac{\lfloor \log_b n \rfloor k(b-k)}{n(n+k)(n+b)},$$

which is a rewriting of the desired series for γ.

Finally, we show that the series converges faster as b increases. Fixing b, we use the formula $\sum_{k=1}^{b-1} k(b-k) = b(b^2-1)/6$ to verify the asymptotic relation (7). In it, the factor $(1-b^{-2})/\ln b$ is a decreasing function of $b > 1$, and the proof of Theorem 2 is complete. □

3 Open Problems

We conclude by mentioning some possible extensions of our results to values of the following function.

Definition 1. The *generalized-Euler-constant function* $\gamma(z)$ [20] (see [13] for the function $f_1(z) := z \cdot \gamma(z)$) is defined for complex numbers z with $|z| \le 1$ by the convergent series

$$\gamma(z) := \sum_{n=1}^{\infty} z^{n-1} \left(\frac{1}{n} - \ln \frac{n+1}{n} \right).$$

Problem 1. Generalize Theorem 1 and Corollary 1 to base 2, rational series for values of $\gamma(z)$ at rational numbers $z \in [-1, 1]$.

In this direction, we have the following result.

Theorem 3. *If $|z| \le 1$, then the value $\gamma(z)$ of the generalized-Euler-constant function is also given by the series*

$$\gamma(z) = \sum_{n=1}^{\infty} \sum_{k=0}^{\lfloor \log_2 n \rfloor} \frac{z^{\lfloor n/2^k \rfloor - 1}}{2n(2n+1)}, \tag{16}$$

which has rational terms when z is a rational number.

Proof. In [2, Theorem 3] (which is easily extended to series with complex terms), take $r_n := z^{n-1}$ for $n > 0$. □

Example 2. Taking $z = 0$ gives the series

$$\gamma(0) = 1 - \ln 2 = \sum_{n=1}^{\infty} \frac{1}{2n(2n+1)}.$$

With $z = \pm 1$, we recover the formulas (5) for $\gamma(\pm 1) = \gamma^{\pm}$ (see [2, Example 3]).

In order to solve Problem 1, it would suffice to find a closed expression for the nth term of the series (16) (that is, for the sum on k), in terms of z and functions like $N_0(n)$, $N_1(n)$. For this, one might be able to use [2, Theorem 1].

Problem 2. Extend Theorem 2 to base $b \geq 2$, rational series for the alternating Euler constant $\ln(4/\pi)$ and for other values of $\gamma(z)$ at rational points.

In this connection, see the series in [2, Theorem 2]. Also, see [2, Sect. 4.2] for a base b generalization of the base 2 relation (10) among the areas A_n.

Postscript

Shortly after Sect. 1, 2, 3 were written, a preprint appeared which generalizes some of our results and which solves in part some of the problems raised here and in [2] – see Pilehrood, K.H., Pilehrood, T.H.: Vacca-type series for values of the generalized-Euler-constant function and its derivative (preprint, 2008); available at http://arxiv.org/abs/0808.0410v1.

Acknowledgements I am grateful to Stefan Krämer and Wadim Zudilin for valuable comments, and to Tanguy Rivoal for sending me a draft of [15].

References

1. Addison, A.W.: A series representation for Euler's constant. Am. Math. Mon. **74**, 823–824 (1967)
2. Allouche, J.-P., Shallit, J., Sondow, J.: Summation of series defined by counting blocks of digits, J. Number Theory **123**, 133–143 (2007)
3. Barrow, D.F.: Solution to Problem 4353, Am. Math. Mon. **58**, 117 (1951)
4. Behrmann, A.: Problem 5460, Am. Math. Mon. **74**, 206 (1967)
5. Berndt, B.C., Bowman, D.C.: Ramanujan's short unpublished manuscript on integrals and series related to Euler's constant. In: Thera, M. (ed.) Constructive, Experimental and Nonlinear Analysis, pp. 19–27. American Mathematical Society, Providence (2000)
6. Carlitz, L.: Advanced Problem 5601, Am. Math. Mon. **75**, 685 (1968)
7. Franklin, F.: On an expression for Euler's constant, J. Hopkins Circ. **II**, 143 (1883)
8. Gerst, L.: Some series for Euler's constant, Am. Math. Mon. **76**, 273–275 (1969)
9. Harborth, H.: Solution to Problem 5601, Am. Math. Mon. **76**, 568 (1969)
10. Jacobsthal, E.: Ueber die Eulersche Konstante, Math.-Naturwiss. Blätter **9**, 153–154 (1906)
11. Krämer, S.: Die Eulersche Konstante γ und verwandte Zahlen. Diplomarbeit, Mathematisches Institut der Georg-August-Universität Göttingen (2005)
12. Nielsen, N.: Een Raekke for Euler's Konstant, Nyt. Tidss. Math. **88**, 10–12 (1897)
13. Pilehrood, K.H., Pilehrood, T.H.: Arithmetical properties of some series with logarithmic coefficients, Math. Z. **255**, 117–131 (2007)
14. Ramanujan, S.: The Lost Notebook and Other Unpublished Papers (Introduction by G.E. Andrews). Springer, Berlin; Narosa Publishing House, New Delhi (1988)
15. Rivoal, T.: Polynômes de type Legendre et approximations de la constante d'Euler (2005, unpublished notes); available at http://www-fourier.ujf-grenoble.fr/~rivoal/
16. Sandham, H.F.: Advanced Problem 4353, Am. Math. Mon. **56**, 414 (1949)
17. Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences (2008); published at http://www.research.att.com/~njas/sequences/
18. Sondow, J.: Double integrals for Euler's constant and $\ln(4/\pi)$ and an analog of Hadjicostas's formula, Am. Math. Mon. **112**, 61–65 (2005)

19. Sondow, J.: New Vacca-type rational series for Euler's constant and its "alternating" analog $\ln(4/\pi)$ (2005, preprint); available at http://arXiv.org/abs/math/0508042v1
20. Sondow, J., Hadjicostas, P.: The generalized-Euler-constant function $\gamma(z)$ and a generalization of Somos's quadratic recurrence constant, J. Math. Anal. Appl. **332**, 292–314 (2007)
21. Sondow, J., Zudilin, W.: Euler's constant, q-logarithms, and formulas of Ramanujan and Gosper, Ramanujan J. **12**, 225–244 (2006); expanded version available at http://arXiv.org/abs/math/0304021
22. Vacca, G.: A new series for the Eulerian constant $\gamma = .577\ldots$, Quart. J. Pure Appl. Math. **41**, 363–364 (1910)
23. van Lint, J. H.: Solution to Problem 5460, Am. Math. Mon. **75**, 202 (1968)

Mixed Sums of Primes and Other Terms

Zhi-Wei Sun

In honor of Prof. M.B. Nathanson on the occasion of his 60th birthday

Summary In this paper, we study mixed sums of primes and linear recurrences. We show that if $m \equiv 2 \pmod 4$ and $m+1$ is a prime, then $(m^{2^n-1}-1)/(m-1) \neq m^n + p^a$ for any $n = 3, 4, \ldots$ and prime power p^a. We also prove that if $a > 1$ is an integer, $u_0 = 0$, $u_1 = 1$, and $u_{i+1} = au_i + u_{i-1}$ for $i = 1, 2, 3, \ldots$, then all the sums $u_m + au_n$ $(m, n = 1, 2, 3, \ldots)$ are distinct. One of our conjectures states that any integer $n > 4$ can be written as the sum of an odd prime and two positive Fibonacci numbers.

Keywords Fibonacci number · Linear recurrence · Mixed sum · Prime · Representation

Mathematics Subject Classifications (2000). Primary 11P32, Secondary 11A41, 11B37, 11B39, 11B75, 11Y99

1 Introduction

Let us first recall the famous Goldbach conjecture in additive number theory.

Conjecture 1.1 (Goldbach's Conjecture). Any even integer $n \geqslant 4$ can be written as the sum of two primes.

The number of primes not exceeding $n \geqslant 2$ is approximately $n/\log n$ by the prime number theorem. Hardy and Littlewood conjectured that the number of ways to write an even integer $n \geqslant 4$ as the sum of two primes is given asymptotically by

Z.-W. Sun
Department of Mathematics, Nanjing University, Nanjing 210093, People's Republic of China
and
State Key Laboratory of Novel Software Technology, Nanjing University,
Nanjing 210093, People's Republic of China
e-mail: zwsun@nju.edu.cn; http://math.nju.edu.cn/~zwsun

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_24,

$$\frac{cn}{\log^2 n}\prod_{p\mid n}\left(1+\frac{1}{p-2}\right),$$

where $c = 2\prod_p(1-(p-1)^{-2}) = 1.3203\cdots$ is a constant and p runs over odd primes. (Cf. [7, pp. 159-164].)

Goldbach's conjecture remains open, and the best result in this direction is Chen's theorem (cf. [1]): each large even integer can be written as the sum of a prime and a product of at most two primes.

Those integers $T_x = x(x+1)/2$ with $x \in \mathbb{N} = \{0, 1, 2\ldots\}$ are called triangular numbers. There are less than $\sqrt{2n}$ positive triangular numbers below an integer $n \geqslant 2$, so triangular numbers are more sparse than prime numbers. In 2008, the author made the following conjecture.

Conjecture 1.2 (Sun [18]).

(i) Each natural number $n \neq 216$ can be written in the form $p + T_x$ with $x \in \mathbb{Z}$, where p is zero or a prime.

(ii) Any odd integer greater than 3 can be written in the form $p + x(x+1)$, where p is a prime and x is a positive integer.

Douglas McNeil (University of London) (cf. [12]) has verified parts (i) and (ii) up to 10^{10} and 10^{12}, respectively. The author [22] would like to offer 1,000 US dollars for the first positive solutions to both (i) and (ii), and \$200 for the first explicit counterexample to (i) or (ii).

Powers of two are even much more sparse than triangular numbers. In a letter to Goldbach, Euler posed the problem whether any odd integer $n > 1$ can be expressed in the form $p+2^a$, where p is a prime and $a \in \mathbb{N}$. This question was reformulated by Polignac in 1849. By introducing covers of the integers by residue classes, Erdős [4] showed that there exists an infinite arithmetic progression of positive odd integers no term of which is of the form $p + 2^a$. (See also Nathanson [14, pp. 204–208].) On the basis of the work of Cohen and Selfridge [2], the author [17] proved that if

$$x \equiv 47867742232066880047611079 \pmod{M}$$

with

$$\begin{aligned} M &= 2\times 3\times 5\times 7\times 11\times 13\times 17\times 19\times 31\times 37\\ &\quad\times 41\times 61\times 73\times 97\times 109\times 151\times 241\times 257\times 331\\ &= 6648308496158851012401069159 0, \end{aligned}$$

then x is not of the form $\pm p^a \pm q^b$ where p, q are primes and $a, b \in \mathbb{N}$.

In 1971, Crocker [3] proved that there are infinitely many positive odd integers not of the form $p + 2^a + 2^b$ where p is a prime and $a, b \in \mathbb{Z}^+ = \{1, 2, 3, \ldots\}$. Here are the first few such numbers greater than 5 recently found by Charles Greathouse (USA):

6495105, 848629545, 1117175145, 2544265305, 3147056235, 3366991695.

Note that 1117175145 even cannot be written in the form $p + 2^a + 2^b$ with p a prime and $a, b \in \mathbb{N}$.

Erdős (cf. [5]) asked whether there is a positive integer k such that any odd number greater than 3 can be written the sum of an odd prime and at most k positive powers of two. Gallagher [6] proved that for any $\varepsilon > 0$ there is a positive integer $k = k(\varepsilon)$ such that those positive odd integers not representable as the sum of a prime and k powers of two form a subset of $\{1, 3, 5, \ldots\}$ with lower asymptotic density at least $1-\varepsilon$. In 1951, Linnik [10] showed that there exists a positive integer k such that each large even number can be written as the sum of two primes and k positive powers of two; Heath-Brown and Puchta [8] proved that we can take $k = 13$. (See also Pintz and Ruzsa [15].) We conjecture that any odd integer $n > 8$ can be expressed as the sum of an odd prime and three positive powers of two.

In March 2005, Georges Zeller-Meier [25] asked whether $2^{2^n-1} - 2^n - 1$ is composite for every $n = 3, 4, \ldots$. Clearly, an affirmative answer follows from part (i) of our following theorem in the case $m = 2$.

Theorem 1.3. *(i) Let $m \equiv 2 \pmod 4$ be an integer with $m + 1$ a prime. Then, for each $n = 3, 4, \ldots$, we have*

$$\frac{m^{2^n-1} - 1}{m - 1} \neq m^n + p^a,$$

where p is any prime and a is any nonnegative integer.

(i) Let m and n be integers greater than one. Then,

$$\frac{m^{2^n} - 1}{m - 1} \neq p + m^a + m^b,$$

where p is any prime, $a, b \in \mathbb{N}$ and $a \neq b$.

Remark 1.4. In the case $m = 2$, part (ii) of Theorem 1.3 was observed by A. Schinzel and Crocker independently in 1960s, and this plays an important role in Crocker's result about $p + 2^a + 2^b$. In 2001, the author and Le [23] proved that for $n = 4, 5, \ldots$ we cannot write $2^{2^n-1} - 1$ in the form $p^\alpha + 2^a + 2^b$, where p is a prime, $a, b, \alpha \in \mathbb{N}$ and $a \neq b$.

For any integer $m > 1$, the sequence $\{m^n\}_{n \geqslant 0}$ is a first-order linear recurrence with earlier terms dividing all later terms. To seek for good representations of integers, we'd better turn resort to second-order linear recurrences whose general term usually does not divide all later terms.

The famous Fibonacci sequence $\{F_n\}_{n \geqslant 0}$ is defined as follows:

$$F_0 = 0,\ F_1 = 1,\ \text{and } F_{n+1} = F_n + F_{n-1} \text{ for } n = 1, 2, 3, \ldots.$$

Here are few initial Fibonacci numbers:

$$\begin{aligned} F_0 = 0 < F_1 = F_2 = 1 < F_3 = 2 < F_4 = 3 < F_5 = 5 < F_6 \\ = 8 < F_7 = 13 < F_8 = 21 < \cdots. \end{aligned}$$

It is well known that

$$F_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right) \quad \text{for all } n \in \mathbb{N}.$$

Clearly $F_n < 2^{n-1}$ for $n = 2, 3, \ldots$, and

$$F_n \sim \frac{\varphi^n}{\sqrt{5}} \qquad (n \to +\infty),$$

where

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.618\cdots.$$

Note that $2 \mid F_n$ if and only if $3 \mid n$.

It is not known whether the positive integers not of the form $p + F_n$ with p a prime and $n \in \mathbb{N}$ form a subset of $\mathbb{Z}^+$ with positive lower asymptotic density. However, Wu and Sun [24] were able to construct a residue class containing no integers of the form $p^a + F_{3n}/2$ with p a prime and $a, n \in \mathbb{N}$. Note that $u_n = F_{3n}/2$ is just half of an even Fibonacci number; also $u_0 = 0$, $u_1 = 1$, and $u_{n+1} = 4u_n + u_{n-1}$ for $n = 1, 2, 3, \ldots$.

On December 23, 2008 the author [19] formulated the following conjecture.

Conjecture 1.5 (Conjecture on Sums of Primes and Fibonacci Numbers). Any integer $n > 4$ can be written as the sum of an odd prime and two positive Fibonacci numbers. We can require further that one of the two Fibonacci numbers is odd.

Remark 1.6. For a large integer n, there are about $\log n/\log\varphi$ Fibonacci numbers below n but there are about $n/\log n$ primes below n. So, Fibonacci numbers are much more sparse than prime numbers and hence the above conjecture looks more difficult than the Goldbach conjecture. McNeil (cf. [12, 13]) has verified Conjecture 1.5 up to 10^{14}. The author (cf. [22]) would like to offer 5,000 US dollars for the first positive solution published in a well-known mathematical journal and \$250 for the first explicit counterexample which can be rechecked by the author via computer. Note that Conjecture 1.5 implies that for any odd prime p we can find an odd prime $q < p$ such that $p - q$ can be written as the sum of two odd Fibonacci numbers. Perhaps it is safe to modify Conjecture 1.5 by substituting "two or three" for the word "two".

Recall that the Pell sequence $\{P_n\}_{n\geqslant 0}$ is defined as follows.

$$P_0 = 0,\ P_1 = 1,\ \text{and } P_{n+1} = 2P_n + P_{n-1} \ \text{ for } n = 1, 2, 3, \ldots.$$

It is well known that

$$P_n = \frac{1}{2\sqrt{2}}\left((1+\sqrt{2})^n - (1-\sqrt{2})^n\right) \quad \text{for all } n \in \mathbb{N}.$$

Clearly $P_n > 2^n$ for $n = 6, 7, \ldots$, and

$$P_n \sim \frac{(1+\sqrt{2})^n}{2\sqrt{2}} \qquad (n \to +\infty).$$

On January 10, 2009, the author [20] posed the following conjecture which is an analogue of Conjecture 1.5.

Conjecture 1.7 (Conjecture on Sums of Primes and Pell Numbers). Any integer $n > 5$ can be written as the sum of an odd prime, a Pell number and twice a Pell number. We can require further that the two Pell numbers are positive.

Remark 1.8. McNeil (cf. [22]) has verified Conjecture 1.7 up to 5×10^{13} and found no counterexample. The author (cf. [22]) would like to offer 1,000 US dollars for the first positive solution published in a well-known mathematical journal.

Soon after he learned Conjecture 1.7 from the author, Qing-Hu Hou (Nankai University) observed (without proof) that all the sums $P_s + 2P_t$ $(s, t = 1, 2, 3, \ldots)$ are distinct. Clearly Hou's observation follows from our following theorem.

Theorem 1.9. *Let $a > 1$ be an integer, and set*

$$u_0 = 0,\ u_1 = 1,\ \text{and } u_{i+1} = au_i + u_{i-1} \text{ for } i = 1, 2, 3, \ldots.$$

Then no integer x can be written as $u_m + au_n$ (with $m \in \mathbb{N}$ and $n \in \mathbb{Z}^+$) in at least two ways, except in the case $a = 2$ and $x = u_0 + au_2 = u_2 + au_1 = 4$.

Remark 1.10. Note that if $n \in \mathbb{Z}^+$ then $u_{n+1} + au_0 = au_n + u_{n-1}$.

Corollary 1.11. *Let $k, l, m, n \in \mathbb{Z}^+$. Then $P_k + 2P_l = P_m + 2P_n$ if and only if $k = m$ and $l = n$.*

Remark 1.12. In view of Corollary 1.11, we can assign an ordered pair $\langle m, n\rangle \in \mathbb{Z}^+ \times \mathbb{Z}^+$ the code $P_m + 2P_n$. Recall that a sequence $a_1 < a_2 < a_3 < \cdots$ of positive integers is called a Sidon sequence if all the sums of pairs, $a_i + a_j$, are all distinct. An unsolved problem of Erdős (cf. [7, p. 403]) asks for a polynomial $P(x) \in \mathbb{Z}[x]$ such that all the sums $P(m) + P(n)$ $(0 \leqslant m < n)$ are distinct.

Motivated by Conjecture 1.5 and its variants, Qing-Hu Hou and Jiang Zeng (University of Lyon-I) formulated the following conjecture during their visit to the author in January 2009.

Conjecture 1.13 (Hou and Zeng [9]). Any integer $n > 4$ can be written as the sum of an odd prime, a positive Fibonacci number and a Catalan number.

Remark 1.14. Catalan numbers are integers of the form

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n+1} \quad (n \in \mathbb{N}),$$

which play important roles in combinatorics (see, e.g., Stanley [16, Chapter 6]). They are also determined by $C_0 = 1$ and the recurrence

$$C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k} \quad (n = 0, 1, 2, \ldots).$$

By Stirling's formula, $C_n \sim 4^n/(n\sqrt{n\pi})$ as $n \to +\infty$. McNeil [13] has verified Conjecture 1.13 up to 3×10^{13} and found no counterexample. Hou and Zeng would like to offer 1,000 US dollars for the first positive solution published in a well-known mathematical journal and \$200 for the first explicit counterexample which can be rechecked by them via computer. Note that 3627586 cannot be written in the form $p + 2F_S + C_t$ with p a prime and $s, t \in \mathbb{N}$.

The Lucas sequence $\{L_n\}_{n\geqslant 0}$ is defined as follows.

$$L_0 = 2,\ L_1 = 1,\ \text{and } L_{n+1} = L_n + L_{n-1}\ (n = 1, 2, 3, \ldots).$$

It is known that

$$L_n = 2F_{n+1} - F_n = \left(\frac{1+\sqrt{5}}{2}\right)^n + \left(\frac{1-\sqrt{5}}{2}\right)^n$$

for every $n = 0, 1, 2, 3, \ldots$.

On January 16, 2009 the author made the following conjecture which is similar to Conjecture 1.13.

Conjecture 1.15. Each integer $n > 4$ can be written as the sum of an odd prime, a Lucas number and a Catalan number.

Remark 1.16. McNeil [13] has verified Conjecture 1.15 up to 10^{13} and found no counterexample. Note that 1389082 cannot be written in the form $p + 2L_s + C_t$ with p a prime and $s, t \in \mathbb{N}$.

We are going to prove Theorems 1.3 and 1.9 in Sect. 2. Section 3 is devoted to our discussion of Conjecture 1.5 and its variants.

2 Proofs of Theorems 1.3 and 1.9

Proof of Theorem 1.3. For $n = 2, 3, \ldots$ we clearly have

$$\begin{aligned}(m-1)\prod_{k=0}^{n-1}\left(m^{2^k}+1\right) &= \left(m^{2^0}-1\right)\left(m^{2^0}+1\right)\left(m^{2^1}+1\right)\cdots\left(m^{2^{n-1}}+1\right)\\ &= \left(m^{2^1}-1\right)\left(m^{2^1}+1\right)\cdots\left(m^{2^{n-1}}+1\right)\\ &= \cdots = \left(m^{2^{n-1}}-1\right)\left(m^{2^{n-1}}+1\right) = m^{2^n}-1.\end{aligned}$$

(i) Fix an integer $n \geqslant 3$. Write $n+1 = 2^k q$ with $k \in \mathbb{N}$, $q \in \mathbb{Z}^+$ and $2 \nmid q$. Since

$$2^n = (1+1)^n \geqslant 1 + n + \frac{n(n-1)}{2} > n+1,$$

we must have $0 \leqslant k \leqslant n-1$. Thus $m^{2^k}+1$ divides both $(m^{2^n}-1)/(m-1)$ and $m^{n+1}+1 = (m^{2^k})^q + 1$. Set

$$d_n = \frac{m^{2^n-1}-1}{m-1} - m^n.$$

Then

$$md_n = \frac{m^{2^n}-m}{m-1} - m^{n+1} = \frac{m^{2^n}-1}{m-1} - (m^{n+1}+1)$$

and hence $m^{2^k}+1$ divides d_n.

Suppose that d_n is a prime power. By the above, we can write $d_n = p^a$, where $a \in \mathbb{N}$ and p is a prime divisor of $m^{2^k}+1$. As m is even, p is an odd prime. Since

$$m^{p-1} \equiv 1 \pmod{p} \text{ and } m^{2^{k+1}} \equiv (-1)^2 = 1 \pmod{p},$$

we have

$$m^{\gcd(p-1,2^{k+1})} \equiv 1 \pmod{p}.$$

But

$$m^{2^k} \equiv -1 \not\equiv 1 \pmod{p},$$

so $p \equiv 1 \pmod{2^{k+1}}$. Note that

$$p^a = \frac{m^{2^n-1}-1}{m-1} - m^n = \sum_{k=0}^{2^n-2} m^k - m^n \equiv 1 + m + m^2 \pmod{m^3}.$$

If $k > 0$, then $p \equiv 1 \pmod{2^2}$ and hence

$$p^a \equiv 1 \not\equiv 1 + m \pmod{2^2},$$

which contradicts the congruence $p^a \equiv 1+m \pmod{m^2}$. So $k = 0$, $p \mid m^{2^0}+1$ and hence $p = m+1$. (Recall that $m+1$ is a prime.) It follows that p^a is congruent to 1 or $m+1$ modulo 8. Since $1 + m + m^2 \not\equiv 1, m+1 \pmod 8$, we get a contradiction. This proves part (i).

(ii) Let $a > b \geqslant 0$ be integers with $m^a + m^b < (m^{2^n}-1)/(m-1)$. Clearly $2^n > a > b$. Write $a - b = 2^k q$ with $k \in \mathbb{N}$, $q \in \mathbb{Z}^+$ and $2 \nmid q$. Then $0 \leqslant k < n$

and hence $d = m^{2^k} + 1$ divides both $(m^{2^n} - 1)/(m - 1)$ and $m^{a-b} + 1 = (m^{2^k})^q + 1$. Thus

$$\frac{m^{2^n} - 1}{m - 1} - m^a - m^b$$

is a multiple of d. Observe that

$$\begin{aligned}\frac{m^{2^n} - 1}{m - 1} &= \frac{m^{2^n-2} - 1}{m - 1}\left(m^{2^n-2} + 1\right)\left(m^{2^n-1} + 1\right)\\ &> \left(m^{2^n-2} + 1\right)\left(m^{2^n-1} + 1\right)\\ &\geqslant (m^b + 1)(m^{a-b} + 1) \geqslant m^a + m^b + d.\end{aligned}$$

So d is a proper divisor of $D = (m^{2^n} - 1)/(m - 1) - m^a - m^b$. This shows that D cannot be a prime. We are done. □

Proof of Theorem 1.9. Observe that

$$u_0 = 0 < u_1 = 1 < u_2 = a < u_3 < u_4 < \cdots.$$

By induction,

$$u_{2i} \equiv u_0 = 0 \pmod a \text{ and } u_{2i+1} \equiv u_1 = 1 \pmod a \quad \text{for } i = 0, 1, 2, \ldots.$$

We will make use of these simple properties.

Let $k, m \in \mathbb{N}$ and $l, n \in \mathbb{Z}^+$ with $k \leqslant m$. Below we discuss the equation $u_k + au_l = u_m + au_n$.

Case 1. $k = m$.

In this case,

$$u_k + au_l = u_m + au_n \Rightarrow u_l = u_n \Rightarrow l = n.$$

Case 2. $k = l < m$.

If $k = l < m - 1$ then

$$u_k + au_l < u_{m-2} + au_{m-1} = u_m < u_m + au_n.$$

When $k = l = m - 1$, as $u_m \not\equiv u_{m-1} \pmod a$ we have

$$u_k + au_l = (a + 1)u_{m-1} \neq u_m + au_n.$$

Case 3. $l < k < m$.

In this case,

$$u_k + au_l \leqslant u_k + au_{k-1} < au_k + u_{k-1} = u_{k+1} \leqslant u_m < u_m + au_n.$$

Case 4. $k < l < m$.

In this case,

$$u_k + au_l \leqslant au_l + u_{l-1} = u_{l+1} \leqslant u_m < u_m + au_n.$$

Case 5. $k < m \leqslant l$.

Suppose that $u_k + au_l = u_m + au_n$. Then

$$u_l > \frac{u_k + au_l - u_m}{a} = u_n \geqslant u_l - \frac{u_m}{a} \geqslant \frac{a-1}{a}u_l \geqslant (a-1)u_{l-1} \geqslant u_{l-1}.$$

It follows that

$$k = 0,\ m = l = 2,\ \text{and}\ u_n = u_{l-1} = u_1 = 1.$$

Thus $au_2 = u_2 + au_1$, i.e., $a^2 = a + a$ and hence $a = 2$.

Combining the above we have completed the proof. □

Remark 2.1. By modifying the proof of Theorem 1.9, we can determine all the solutions of the equation $F_k + F_e = F_m + F_n$ with $k, l, m, n \in \mathbb{N}$.

3 Discussion of Conjecture 1.5 and Its Variants

Concerning Conjecture 1.5, we mention that there are very few natural numbers not representable as the sum of a prime $p \equiv 5 \pmod 6$ and two Fibonacci numbers. Bjorn Poonen (MIT) informed me that by a heuristic argument there should be infinitely many positive integers not in the form $p + F_s + F_t$ if we require that the prime p lies in a fixed residue class with modulus greater than one. McNeil [11,13] made a computer search to find natural numbers not representable as the sum of a prime $p \equiv 5 \pmod 6$ an odd Fibonacci number and a positive Fibonacci number; he found that there are totally 729 such numbers in the interval $[0, 10^{14}]$, 277 of which (such as 857530546) even cannot be written as the sum of a prime $p = 5 \pmod 6$ and two Fibonacci numbers.

In 2008, the author (cf. [19, 20]) also made the following conjecture which is similar to Conjecture 1.5.

Conjecture 3.1. (i) Any integer $n > 4$ can be written as the sum of an odd prime, a positive Fibonacci number and the square of a positive Fibonacci number. We can require further that one of the two Fibonacci numbers is odd.

(ii) Each integer $n > 4$ can be written as the sum of an odd prime, a positive Fibonacci number and the cube of a positive Fibonacci number. We can require further that one of the two Fibonacci numbers is odd.

Remark 3.2. Note that 900068 cannot be written as the sum of a prime, a Fibonacci number and the fourth power of a Fibonacci number. Also,

$$F_n^3 \sim \frac{\varphi^{3n}}{(\sqrt{5})^3} = \frac{(4.236\cdots)^n}{5\sqrt{5}} \quad (n \to +\infty).$$

Let $k \in \{1,2,3\}$. For $n \in \mathbb{Z}^+$ let $r_k(n)$ denote the number of ways to write n as the sum of an odd prime, a positive Fibonacci number and the kth power of a positive Fibonacci number with one of the two Fibonacci numbers odd. That is,

$$r_k(n) = |\{\langle p,s,t\rangle : \ p + F_s + F_t^k = n, \\ p \text{ is an odd prime}, \ s,t \ge 2, \text{ and } 2 \nmid F_s \text{ or } 2 \nmid F_t\}|.$$

The author has investigated values of the quotient

$$s_k(n) = \frac{r_k(n)}{\log n}$$

via computer, and guessed that

$$c_k = \liminf_{n\to+\infty} s_k(n) > 0.$$

Numerical data suggest that $2 < c_1 < 3$. In fact, the author computed all values of $s_1(n)$ with $10^{50} \leqslant n \leqslant 10^{50} + 4\times 10^4$, and here are the two smallest values:

$$s_1(10^{50}+39030) = 2.22359\cdots \text{ and } s_1(10^{50}+5864) = 2.29037\cdots.$$

Here is another variant of Conjecture 1.5 made by the author (cf. [19, 21]).

Conjecture 3.3. (i) Any integer $n > 4$ can be written as the sum of an odd prime, an odd Lucas number and a positive Lucas number. For $k = 2,3$, we can write any integer $n > 4$ in the form $p + L_s + L_t^k$, where p is an odd prime, $s,t \ge 0$, and L_s or L_t is odd.

(ii) Each integer $n > 4$ can be written as the sum of an odd prime, a positive Fibonacci number and twice a positive Fibonacci number (or half of a positive Fibonacci number). We can also represent any integer $n > 4$ as the sum of an odd prime, twice a positive Fibonacci number, and the square of a positive Fibonacci number.

(iii) Any integer $n > 4$ can be written in the form $p + F_s + L_t$ with p an odd prime, $s > 0$, and F_s or L_t odd.

Remark 3.4. The author verified Conjectures 3.1 and 3.3 for $n \leqslant 3\times 10^7$. Qing-Hu Hou found that 17540144 cannot be written as the sum of a prime, a Lucas number and the fourth power of a Lucas number. McNeil (cf. [12]) has verified the first assertions in parts (i) and (ii) of Conjectures 3.1 and 3.3 up to 10^{12}. He (cf. [13]) has

also verified part (iii) of Conjecture 3.3 up to 10^{13}, and found that 36930553345551 cannot be written as the sum of a prime, a Fibonacci number and an even Lucas number.

What about the representations $n = p + P_s + kP_t$ with $k \in \{1, 3, 4\}$ related to Conjecture 1.7? Note that 2176 cannot be written as the sum of a prime and two Pell numbers. McNeil [13] found that 393185153350 cannot be written as the sum of a prime, a Pell number and three times a Pell number, and the smallest integer greater than 7 not representable as the sum of a prime, a Pell number and four times a Pell number is

$$872377759846 \approx 8.7 \times 10^{11}.$$

The companion Pell sequence $\{Q_n\}_{n\geqslant 0}$ is defined by

$$Q_0 = Q_1 = 2 \text{ and } Q_{n+1} = 2Q_n + Q_{n-1} \ (n = 1, 2, 3, \ldots).$$

McNeil [13] found that the smallest integer greater than 5 not representable as the sum of a prime, a Pell number and a companion Pell number is 169421772576.

McNeil's counterexamples seem to suggest that Conjecture 1.7 might also have large counterexamples. However, in the author's opinion, the large counterexamples to the representations $n = p + P_s + 3P_t$ and $n = p + P_s + 4P_t$ hint that they are very close to the "truth" (Conjecture 1.7). Corollary 1.11 is also a good evidence to support Conjecture 1.7. To expel suspicion, the author has investigated the behaviour of the representation function

$$r(n) = |\{\langle p, s, t\rangle : \ p + P_s + 2P_t = n \text{ with } p \text{ a prime and } s, t \geqslant 0\}|.$$

For $n \in [10^{50}, 10^{50} + 10081]$ most values of $s(n) = r(n)/\log n$ lies in the interval $(1, 2)$, the smallest value of $s(n)$ with n in the range is

$$s(10^{50} + 10045) = \frac{76}{\log(10^{50} + 10045)} \approx 0.66.$$

The author also computed the values of $s(n)$ with $n \in [10^{200}, 10^{200} + 100]$, the smallest value and the largest value are

$$s(10^{200} + 33) = \frac{443}{\log(10^{200} + 33)} \approx 0.96$$

and

$$s(10^{200} + 18) = \frac{824}{\log(10^{200} + 18)} \approx 1.79,$$

respectively. It seems that

$$c = \liminf_{n\to+\infty} s(n) \in (0.6, 1.2).$$

Acknowledgements Research supported by the National Natural Science Foundation of China (grant 10871087). The author wishes to thank Dr. Douglas McNeil who has checked almost all conjectures mentioned in this paper via his quite efficient and powerful computation.

References

1. J.-R. Chen, *On the representation of a large even integer as the sum of a prime and the product of at most two primes*, Sci. Sinica **16** (1973), 157–176
2. F. Cohen and J. L. Selfridge, *Not every number is the sum or difference of two prime powers*, Math. Comp. **29** (1975), 79–81
3. R. Crocker, *On a sum of a prime and two powers of two*, Pacific J. Math. **36** (1971), 103–107
4. P. Erdős, *On integers of the form* $2^k + p$ *and some related problems*, Summa Brasil. Math. **2** (1950), 113–123
5. P. Erdős, *Some of my favorite problems and results*, in: The Mathematics of Paul Erdős, I (R. L. Graham and J. Nešetřil, eds.), Algorithms and Combinatorics 13, Springer, Berlin, 1997, pp. 47–67
6. P. X. Gallagher, *Primes and powers of 2*, Invent. Math. **29** (1975), 125–142
7. R. K. Guy, *Unsolved Problems in Number Theory*, 3rd Edition, Springer, New York, 2004
8. D. R. Heath-Brown and J.-C. Puchta, *Integers represented as a sum of primes and powers of two*, Asian J. Math. **6** (2002), 535–565
9. Q. H. Hou and J. Zeng, Sequences A154404 in On-Line Encyclopedia of Integer Sequences, http://www.research.att.com/~njas/sequences/A154404
10. Yu. V. Linnik, *Prime numbers and powers of two*, Trudy Mat. Inst. Steklov. **38** (1951), 152–169
11. D. McNeil, *Sun's strong conjecture* (a message to Number Theory Mailing List in Dec. 2008), http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0812\&L=nmbrthry\&T=0\&P=3020
12. D. McNeil, *Various and sundry* (a message to Number Theory Mailing List in Jan. 2009), http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0901\&L=nmbrthry\&T=0\&P=840
13. D. McNeil, Personal communications in January 2009
14. M. B. Nathanson, *Additive Number Theory: The Classical Bases*, Grad. Texts in Math., Vol. 164, Springer, New York, 1996
15. J. Pintz and I. Z. Ruzsa, *On Linnik's appproximation to Goldbach's problem, I*, Acta Arith. **109** (2003), 169–194
16. R. P. Stanley, *Enumerative Combinatorics*, Vol. 2, Cambridge University Press, Cambridge, 1999
17. Z. W. Sun, *On integers not of the form* $\pm p^a \pm q^b$, Proc. Amer. Math. Soc. **128** (2000), 997–1002
18. Z. W. Sun, *On sums of primes and triangular numbers*, Journal of Combinatorics and Number Theory **1** (2009), 65–76
19. Z. W. Sun, Three messages to the Number Theory Mailing List,
http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0812\&L=nmbrthry\&T=0\&P=2140
http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0812\&L=nmbrthry\&T=0\&P=2704
http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0812\&L=nmbrthry\&T=0\&P=3124
20. Z. W. Sun, Sequences A154257, A154258, A154263, A154536 in On-Line Encyclopedia of Integer Sequences,
http://www.research.att.com/~njas/sequences/A154257
http://www.research.att.com/~njas/sequences/A154258
http://www.research.att.com/~njas/sequences/A154263
http://www.research.att.com/~njas/sequences/A154536

21. Z. W. Sun, Sequences A154285, A154417, A154290, A154421 in On-Line Encyclopedia of Integer Sequences,
http://www.research.att.com/~njas/sequences/A154285
http://www.research.att.com/~njas/sequences/A154290
http://www.research.att.com/~njas/sequences/A154417
22. Z. W. Sun, *Offer prizes for solutions to my main conjectures involving primes* (a message to the Number Theory Mailing List in January 2009), http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0901\&L=nmbrthry\&T=0\&P=1395
23. Z. W. Sun and M. H. Le, *Integers not of the form* $c(2^a + 2^b) + p^\alpha$, Acta Arith. **99** (2001), 183–190
24. K. J. Wu and Z. W. Sun, *Covers of the integers with odd moduli and their applications to the forms* $x^m - 2^n$ *and* $x^2 - F_{3n}/2$, Math. Comp., **78** (2009), 1853–1866
25. G. Zeller-Meier, *Not prime for each* $n \geqslant 3$ (a message to Number Theory Mailing List in March 2005), http://listserv.nodak.edu/cgi-bin/wa.exe?A2=ind0503\&L=nmbrthry\&T=0\&P=2173

Classes of Permutation Polynomials Based on Cyclotomy and an Additive Analogue

Michael E. Zieve

Dedicated to Mel Nathanson on the occasion of his sixtieth birthday

Summary I present a construction of permutation polynomials based on cyclotomy, an additive analogue of this construction, and a generalization of this additive analogue which appears to have no multiplicative analogue. These constructions generalize recent results of José Marcos.

Keywords Cyclotomy · Permutation polynomial

Mathematics Subject Classifications (2010). 11T06, 11T22

1 Introduction

Writing $\mathbb{F}_q$ for the field with q elements, we consider *permutation polynomials* over $\mathbb{F}_q$, namely polynomials $f \in \mathbb{F}_q[x]$ for which the map $\alpha \mapsto f(\alpha)$ induces a permutation of $\mathbb{F}_q$. These polynomials first arose in the work of Betti [6], Mathieu [28], and Hermite [19], as a tool for representing and studying permutations.

Since every mapping $\mathbb{F}_q \to \mathbb{F}_q$ is induced by a polynomial, the study of permutation polynomials focuses on polynomials with unusual properties beyond inducing a permutation. In particular, permutation polynomials of 'nice' shapes have been a topic of interest since the work of Hermite, in which he noted that there are many permutation polynomials of the form

$$f(x) := ax^i \left(x^{(q-1)/2} + 1\right) - bx^j \left(x^{(q-1)/2} - 1\right)$$

M.E. Zieve
University of Michigan, Department of Mathematics, 530 Church Street,
Ann Arbor, MI 48109-1043, USA
e-mail: zieve@umich.edu; www.math.lsa.umich.edu/~zieve

D. Chudnovsky and G. Chudnovsky (eds.), *Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson*,
DOI 10.1007/978-0-387-68361-4_25,

with q odd, $i, j > 0$, and $a, b \in \mathbb{F}_q^*$. The reason for this is that $f(\alpha) = 2a\alpha^i$ if $\alpha \in \mathbb{F}_q$ is a square, and $f(\alpha) = 2b\alpha^j$ otherwise; thus, for instance, f is a permutation polynomial if $2a$ and $2b$ are squares and $\gcd(ij, q-1) = 1$.

More generally, any polynomial of the form $f(x) := x^r h(x^{(q-1)/d})$ induces a mapping on $\mathbb{F}_q$ modulo dth powers, so testing whether f permutes $\mathbb{F}_q$ reduces to testing whether the induced mapping on cosets is bijective (assuming that f is injective on each coset, or equivalently that $\gcd(r, (q-1)/d) = 1$). The vast majority of known examples of 'nice' permutation polynomials have this 'cyclotomic' form for some $d < q-1$; see for instance [1–5, 7, 9–25, 29–34, 36–43]. Moreover, there is a much longer list of papers proving the nonexistence of permutation polynomials of certain shapes, and nearly all such papers again address these polynomials $f(x)$ having cyclotomic behavior.

In the recent paper [26], Marcos gives five constructions of permutation polynomials. His first two constructions are new classes of permutation polynomials having the above cyclotomic form. His third construction is a kind of additive analogue of the first, resulting in polynomials of the form $L(x) + h(T(x))$ where $T(x) := x^{q/p} + x^{q/p^2} + \cdots + x$ is the trace polynomial from $\mathbb{F}_q$ to its prime field $\mathbb{F}_p$, and $L(x) = \sum a_i x^{p^i}$ is any additive polynomial. The idea of the analogy is that $T(x)$ induces a homomorphism $\mathbb{F}_q \to \mathbb{F}_p$, just as $x^{(q-1)/d}$ induces a homomorphism from $\mathbb{F}_q^*$ to its subgroup of dth roots of unity. The fourth construction in [26] is a variant of the third for polynomials of the form $L(x)+h(T(x))(L(x)+c)$, and the fifth construction replaces $T(x)$ with other symmetric functions in $x^{q/p}$, $x^{q/p^2}, \ldots, x$.

In this paper, I present rather more general versions of the first four constructions from [26], together with simplified proofs. I can say nothing new about the fifth construction from [26], although that construction is quite interesting and I encourage the interested reader to look into it.

2 Permutation Polynomials from Cyclotomy

In this section, we prove the following result, where for $d \geq 1$, we write $h_d(x) := x^{d-1} + x^{d-2} + \cdots + x + 1$.

Theorem 1. *Fix a divisor $d > 2$ of $q-1$, integers $u \geq 1$ and $k \geq 0$, an element $b \in \mathbb{F}_q$, and a polynomial $g \in \mathbb{F}_q[x]$ divisible by h_d. Then*

$$f(x) := x^u \left(b x^{k(q-1)/d} + g(x^{(q-1)/d}) \right)$$

permutes $\mathbb{F}_q$ if and only if the following four conditions hold:

1. $\gcd(u, (q-1)/d) = 1$
2. $\gcd(d, u + k(q-1)/d) = 1$
3. $b \neq 0$
4. $1 + g(1)/b$ *is a dth power in* $\mathbb{F}_q^*$

The proof uses the following simple lemma.

Lemma 1. *Fix a divisor d of $q-1$, an integer $u > 0$, and a polynomial $h \in \mathbb{F}_q[x]$. Then $f(x) := x^u h(x^{(q-1)/d})$ permutes $\mathbb{F}_q$ if and only if the following two conditions hold:*

1. $\gcd(u, (q-1)/d) = 1$
2. $\widehat{f}(x) := x^u h(x)^{(q-1)/d}$ *permutes the set μ_d of dth roots of unity in $\mathbb{F}_q^*$.*

I discovered this lemma in 1997 when writing [35], and used it in seminars and private correspondence, but I did not publish it until recently [42, Lemma 2.1]. For other applications of this lemma, see [27, 42, 43].

Proof of Theorem 1. In light of the lemma, we just need to determine when $\widehat{f}(x)$ permutes μ_d, where

$$\widehat{f}(x) := x^u (bx^k + g(x))^{(q-1)/d}.$$

For $\zeta \in \mu_d \backslash \{1\}$ we have $g(\zeta) = 0$, so $\widehat{f}(\zeta) = b^{(q-1)/d} \zeta^{u+k(q-1)/d}$. Thus, $\widehat{f}$ is injective on $\mu_d \backslash \{1\}$ if and only if $b \neq 0$ and $\gcd(d, u + k(q-1)/d) = 1$. When these conditions hold, $\widehat{f}(\mu_d \backslash \{1\}) = \mu_d \backslash \{b^{(q-1)/d}\}$, so $\widehat{f}$ permutes μ_d if and only if $\widehat{f}(1) = b^{(q-1)/d}$. Since $\widehat{f}(1) = (b + g(1))^{(q-1)/d}$, the latter condition is equivalent to $(1 + g(1)/b)^{(q-1)/d} = 1$, as desired. □

The case $g = h_d$ of Theorem 1 is [26, Theorem 3] ([25, Theorem 2]), and the case that $g = h_5 - x^3 - x^4$ and $d = u = k - 2 = 5$ is [26, Proposition 5] ([25, Proposition 4]).

Remark 1. The key feature of the polynomials in Theorem 1 as a particular case of Lemma 1 is that the induced mapping $\widehat{f}$ on $\mu_d \backslash \{1\}$ is a monomial, and we know when monomials permute μ_d. For certain values of d, we know other permutations of μ_d: for instance, if $q = q_0^2$ and $d = q_0 - 1$ then $\mu_d = \mathbb{F}_{q_0}^*$, so we can obtain permutation polynomials over $\mathbb{F}_q$ by applying Lemma 1 to polynomials $f(x)$ for which the induced map $\widehat{f}$ on $\mathbb{F}_{q_0}^*$ is any prescribed permutation polynomial. This construction already yields interesting permutation polynomials of $\mathbb{F}_q$ coming from degree-3 permutation polynomials of $\mathbb{F}_{q_0}$; see [35] for details and related results.

3 Permutation Polynomials from Additive Cyclotomy

Lemma 1 addresses maps $\mathbb{F}_q \to \mathbb{F}_q$ which respect the partition of $\mathbb{F}_q^*$ into cosets modulo a certain subgroup. In this section, we give an analogous result in terms of cosets of the additive group of $\mathbb{F}_q$ modulo a subgroup. Let p be the characteristic of $\mathbb{F}_q$. An *additive* polynomial over $\mathbb{F}_q$ is a polynomial of the form $\sum_{i=0}^{k} a_i x^{p^i}$ with $a_i \in \mathbb{F}_q$. The key property of additive polynomials $A(x)$ is that they induce

homomorphisms on the additive group of $\mathbb{F}_q$, since $A(\alpha+\beta)=A(\alpha)+A(\beta)$ for $\alpha,\beta\in\mathbb{F}_q$. The additive analogue of Lemma 1 is as follows, where we write $\mathrm{im}B$ and $\ker B$ for the image and kernel of the mapping $B\colon\mathbb{F}_q\to\mathbb{F}_q$.

Proposition 1. *Pick additive $A,B\in\mathbb{F}_q[x]$ and an arbitrary $g\in\mathbb{F}_q[x]$. Then $f(x):=A(x)+g(B(x))$ permutes $\mathbb{F}_q$ if and only if $A(\ker B)+\widehat{f}(\mathrm{im}B)=\mathbb{F}_q$, where $\widehat{f}(x):=g(x)+A(\widehat{B}(x))$ and $\widehat{B}\in\mathbb{F}_q[x]$ is any polynomial for which $B(\widehat{B}(x))$ is the identity on $\mathrm{im}B$. In other words, f permutes $\mathbb{F}_q$ if and only if $\widehat{f}$ induces a bijection $\mathrm{im}B\to\mathbb{F}_q/A(\ker B)$, where $\mathbb{F}_q/A(\ker B)$ is the quotient of the additive group of $\mathbb{F}_q$ by the subgroup $A(\ker B)$.*

Proof. For $\beta\in\ker B$, we have $f(x+\beta)=A(x)+A(\beta)+g(B(x))=f(x)+A(\beta)$. Thus, for $\alpha\in\mathbb{F}_q$ we have $f(\alpha+\ker B)=f(\alpha)+A(\ker B)$. Since $\mathbb{F}_q=\ker B+\widehat{B}(\mathrm{im}B)$, it follows that $f(\mathbb{F}_q)=f(\widehat{B}(\mathrm{im}B))+A(\ker B)$. Since $f(\widehat{B}(\gamma))=A(\widehat{B}(\gamma))+g(B(\widehat{B}(\gamma)))=A(\widehat{B}(\gamma))+g(\gamma)$ for $\gamma\in\mathrm{im}B$, the result follows. □

Corollary 1. *If f permutes $\mathbb{F}_q$, then A is injective on $\ker B$ and $\widehat{f}$ is injective on $\mathrm{im}B$.*

Proof. If $A(\ker B)+\widehat{f}(\mathrm{im}B)=\mathbb{F}_q$, then

$$q\le \#A(\ker B)\cdot\#\widehat{f}(\mathrm{im}B)\le\#(\ker B)\cdot\#(\mathrm{im}B)=q,$$

where the last equality holds because B defines a homomorphism on the additive group of $\mathbb{F}_q$. The result follows. □

Corollary 2. *Suppose $A(B(\alpha))=B(A(\alpha))$ for all $\alpha\in\mathbb{F}_q$. Then f permutes $\mathbb{F}_q$ if and only if A permutes $\ker B$ and $A(x)+B(g(x))$ permutes $\mathrm{im}B$.*

Proof. Since A and B commute, and $A(0)=0$, it follows that $A(\ker B)\subseteq\ker B$. Thus, by the previous corollary, if f permutes $\mathbb{F}_q$ then A permutes $\ker B$. Henceforth assume that A permutes $\ker B$. By the proposition, f permutes $\mathbb{F}_q$ if and only if $\ker B+\widehat{f}(\mathrm{im}B)=\mathbb{F}_q$; since the left side is the preimage under B of $B(\widehat{f}(\mathrm{im}B))$, this condition may be restated as $B(\widehat{f}(\mathrm{im}B))=\mathrm{im}B$. For $\gamma\in\mathrm{im}B$, we have $B(\widehat{f}(\gamma))=B(g(\gamma))+B(A(\widehat{B}(\gamma)))=B(g(\gamma))+A(B(\widehat{B}(\gamma)))=B(g(\gamma))+A(\gamma)$, so $B(\widehat{f}(x))$ permutes $\mathrm{im}B$ if and only if $B(g(x))+A(x)$ permutes $\mathrm{im}B$. □

One way to get explicit examples satisfying the conditions of this result is as follows: if $B=x^{q/p}+x^{q/p^2}+\cdots+x^p+x$ and $A\in\mathbb{F}_p[x]$, then $A(B(x))=B(A(x))$, so f permutes $\mathbb{F}_q$ if and only if A permutes $\ker B$ and $A(x)+B(g(x))$ permutes $\mathrm{im}B=\mathbb{F}_p$. In case g is a constant (in $\mathbb{F}_q$) times a polynomial over $\mathbb{F}_p$, this becomes (a slight generalization of) [26, Theorem 1] ([25, Theorem 6]). The following case of [26, Corollary 8] ([25, Corollary 8]) exhibits this.

Example 1. In case $q = p^2$ and $B = x^p + x$ and $A = x$, the previous corollary says $f(x) := x + g(x^p + x)$ permutes $\mathbb{F}_{p^2}$ if and only if $x + g(x)^p + g(x)$ permutes $\mathbb{F}_p$, which trivially holds when $g = \gamma h(x)$ with $h \in \mathbb{F}_p[x]$ and $\gamma^{p-1} = -1$. For instance, taking $h(x) = x^2$, it follows that $x + \gamma(x^p + x)^2$ permutes $\mathbb{F}_{p^2}$. By using other choices of h, we can make many permutation polynomials over $\mathbb{F}_{p^2}$ whose degree is a small multiple of p. This is of interest because heuristics suggest that 'at random' there would be no permutation polynomials over $\mathbb{F}_q$ of degree less than $q/(2\log q)$. The bulk of the known low-degree permutation polynomials are *exceptional*, in the sense that they permute $\mathbb{F}_{q^k}$ for infinitely many k; a great deal is known about these exceptional polynomials, for instance see [18]. It is known that any permutation polynomial of degree at most $q^{1/4}$ is exceptional. However, the examples described above have degree on the order of $q^{1/2}$ and are generally not exceptional.

Our final result generalizes the above example in a different direction than Proposition 1.

Theorem 2. *Pick any $g \in \mathbb{F}_q[x]$, any additive $A \in \mathbb{F}_p[x]$, and any $h \in \mathbb{F}_p[x]$. For $B := x^{q/p} + x^{q/p^2} + \cdots + x^p + x$, the polynomial $f(x) := g(B(x)) + h(B(x))A(x)$ permutes $\mathbb{F}_q$ if and only if A permutes $\ker B$ and $B(g(x)) + h(x)A(x)$ permutes $\mathbb{F}_p$ and h has no roots in $\mathbb{F}_p$.*

Proof. For $\beta \in \ker B$ we have $f(x + \beta) = f(x) + h(B(x))A(\beta)$. Thus, if f permutes $\mathbb{F}_q$ then A is injective on $\ker B$ and h has no roots in $\mathbb{F}_p$. Since $A(B(x)) = B(A(x))$ and $A(0) = 0$, also $A(\ker B) \subseteq \ker B$, so if f permutes $\mathbb{F}_q$ then A permutes $\ker B$. Henceforth assume A permutes $\ker B$ and h has no roots in $\mathbb{F}_p$. Since $\mathrm{im} B = \mathbb{F}_p$ and $h(\mathbb{F}_p) \subseteq \mathbb{F}_p \backslash \{0\}$, we have $h(B(\alpha)) \in \mathbb{F}_p \backslash \{0\}$ for $\alpha \in \mathbb{F}_q$. Thus, for $\alpha \in \mathbb{F}_q$ we have $f(\alpha + \ker B) = f(\alpha) + \ker B$, so f permutes $\mathbb{F}_q$ if and only if $B(f(\mathbb{F}_q)) = \mathrm{im} B$. Now for $\alpha \in \mathbb{F}_q$ we have $B(f(\alpha)) = B(g(B(\alpha))) + B(h(B(\alpha))A(\alpha))$, and since $h(B(\alpha)) \in \mathbb{F}_p$ this becomes $B(f(\alpha)) = B(g(B(\alpha))) + h(B(\alpha))B(A(\alpha)) = B(g(B(\alpha)) + h(B(\alpha))A(B(\alpha))$, so $B(f(\mathbb{F}_q))$ is the image of $\mathrm{im} B$ under $B(g(x)) + h(x)A(x)$. The result follows. □

In case $g = \gamma h + \delta$ with $\gamma, \delta \in \mathbb{F}_q$, the above result becomes a generalization of [26, Theorem 10] ([25, Theorem 10]). In view of the analogy between Lemma 1 and Proposition 1, it is natural to seek a 'multiplicative' analogue of Theorem 2. However, I have been unable to find such a result: the obstacle is that the polynomial f in Theorem 2 is the sum of products of polynomials, which apparently should correspond to a product of powers of polynomials, but the latter is already included in Lemma 1.

Acknowledgements I thank José Marcos for sending me preliminary versions of his paper [26], and for encouraging me to develop consequences of his ideas while his paper was still under review.

References

1. S. Ahmad: Split dilations of finite cyclic groups with applications to finite fields. Duke Math. J. **37**, 547–554 (1970)
2. A. Akbary, S. Alaric and Q. Wang: On some classes of permutation polynomials. Int. J. Number Theory **4**, 121–133 (2008)
3. A. Akbary and Q. Wang: On some permutation polynomials over finite fields. Int. J. Math. Math. Sci. **16**, 2631–2640 (2005)
4. A. Akbary and Q. Wang: A generalized Lucas sequence and permutation binomials. Proc. Am. Math. Soc. **134**, 15–22 (2006)
5. A. Akbary and Q. Wang: On polynomials of the form $x^r f(x^{(q-1)/l})$. Int. J. Math. Math. Sci. (2007) art. ID 23408
6. E. Betti: Sopra la risolubilitá per radicali delle equazioni algebriche irriduttibili di grado primo. Ann. Sci. Mat. Fis. **2**, 5–19 (1851) [= Op. Mat. **1**, 17–27 (1903)]
7. F. Brioschi: Des substitutions de la forme $\theta(r) \equiv \varepsilon(r^{n-2} + ar^{(n-3)/2})$ pour un nombre n premier de lettres. Math. Ann. **2**, 467–470 (1870) [= Op. Mat. **5**, 193–197 (1909)]
8. L. Carlitz: Some theorems on permutation polynomials. Bull. Am. Math. Soc. **68**, 120–122 (1962)
9. L. Carlitz: Permutations in finite fields. Acta Sci. Math. (Szeged) **24**, 196–203 (1963)
10. L. Carlitz and C. Wells: The number of solutions of a special system of equations in a finite field. Acta Arith. **12**, 77–84 (1966)
11. S. D. Cohen and R. W. Matthews: A class of exceptional polynomials. Trans. Am. Math. Soc. **345**, 897–909 (1994)
12. S. D. Cohen and R. W. Matthews: Exceptional polynomials over finite fields. Finite Fields Appl. **1**, 261–277 (1995)
13. L. E. Dickson: The analytic representation of substitutions on a power of a prime number of letters with a discussion of the linear group. Ann. of Math. **11**, 65–120 (1896)
14. L. E. Dickson: Linear Groups with an Exposition of the Galois Field Theory, Teubner, Leipzig (1901) [Reprinted by Dover, New York (1958)]
15. A. B. Evans: Orthomorphism Graphs of Groups, Springer, Heidelberg (1992)
16. A. B. Evans: Cyclotomy and orthomorphisms: a survey. Congr. Numer. **101**, 97–107 (1994)
17. J. P. Fillmore: A note on split dilations defined by higher residues. Proc. Am. Math. Soc. **18**, 171–174 (1967)
18. R. M. Guralnick and M. E. Zieve: Polynomials with PSL(2) monodromy. Ann. Math., to appear, arXiv:0707.1835
19. Ch. Hermite: Sur les fonctions de sept lettres. C. R. Acad. Sci. Paris. **57**, 750–757 (1863) [= Ouvres **2**, 280–288 (1908)]
20. N. S. James and R. Lidl: Permutation polynomials on matrices. Linear Algebra Appl. **96**, 181–190 (1987)
21. S. Y. Kim and J. B. Lee: Permutation polynomials of the type $x^{1+(q-1)/m} + ax$. Commun. Korean Math. Soc. **10**, 823–829 (1995)
22. Y. Laigle-Chapuy: Permutation polynomials and applications to coding theory. Finite Fields Appl. **13**, 58–70 (2007)
23. J. B. Lee and Y. H. Park: Some permuting trinomials over finite fields. Acta Math. Sci. **17**, 250–254 (1997)
24. H. W. Lenstra, Jr. and M. Zieve: A family of exceptional polynomials in characteristic three. In: Finite Fields and Applications, Cambridge University Press, Cambridge 209–218 (1996)
25. J. E. Marcos: Specific permutation polynomials over finite fields. arXiv:0810.2738v1 (preliminary version of [26])
26. J. E. Marcos: Specific permutation polynomials over finite fields. Finite Fields Appl., to appear
27. A. M. Masuda and M. E. Zieve: Permutation binomials over finite fields. Trans. Am. Math. Soc. **361**, 4169–4180 (2009), arXiv:0707.1108
28. E. Mathieu: Mémoire sur l'étude des fonctions de plusieurs quantités sur la manière de les former, et sur les substitutions qui les laissent invariables. J. Math. Pures Appl. **6**, 241–323 (1861)

29. G. Mullen and H. Niederreiter: The structure of a group of permutation polynomials. J. Aust. Math. Soc. (Ser. A) **38**, 164–170 (1985)
30. H. Niederreiter and K. H. Robinson: Complete mappings of finite fields. J. Aust. Math. Soc. (Ser. A) **33**, 197–212 (1982)
31. H. Niederreiter and A. Winterhof: Cyclotomic $\mathscr{R}$-orthomorphisms of finite fields. Discrete Math. **295**, 161–171 (2005)
32. Y. H. Park and J. B. Lee: Permutation polynomials with exponents in an arithmetic progression. Bull. Aust. Math. Soc. **57**, 243–252 (1998)
33. Y. H. Park and J. B. Lee: Permutation polynomials and group permutation polynomials. Bull. Aust. Math. Soc. **63**, 67–74 (2001)
34. L. J. Rogers: On the analytical representation of heptagrams. Proc. London Math. Soc. **22**, 37–52 (1890)
35. T. J. Tucker and M. E. Zieve: Permutation polynomials, curves without points, and Latin squares. preprint (2000)
36. D. Wan: Permutation polynomials over finite fields. Acta Math. Sin. **3**, 1–5 (1987)
37. D. Wan: Permutation binomials over finite fields. Acta Math. Sin. **10**, 30–35 (1994)
38. D. Wan and R. Lidl: Permutation polynomials of the form $x^r f(x^{(q-1)/d})$ and their group structure. Monatsh. Math. **112**, 149–163 (1991)
39. L. Wang: On permutation polynomials. Finite Fields Appl. **8**, 311–322 (2002)
40. C. Wells: Groups of permutation polynomials. Monatsh. Math. **71**, 248–262 (1967)
41. C. Wells: A generalization of the regular representation of finite abelian groups. Monatsh. Math. **72**, 152–156 (1968)
42. M. E. Zieve: Some families of permutation polynomials over finite fields. Int. J. Number Theory **4**, 851–857 (2008), arXiv:0707.1111
43. M. E. Zieve: On some permutation polynomials over $\mathbb{F}_q$ of the form $x^r h(x^{(q-1)/d})$. Proc. Am. Math. Soc. **137**, 2209–2216 (2009), arXiv:0707.1110