Springer Proceedings in Mathematics & Statistics

Melvyn B. Nathanson *Editor*

Combinatorial and Additive Number Theory II

CANT, New York, NY, USA, 2015 and 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 220

Springer Proceedings in Mathematics & Statistics

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

More information about this series at http://www.springer.com/series/10533

Melvyn B. Nathanson
Editor

Combinatorial and Additive Number Theory II

CANT, New York, NY, USA, 2015 and 2016

Springer

Editor
Melvyn B. Nathanson
Department of Mathematics
Lehman College (CUNY)
Bronx, NY
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-68030-9 ISBN 978-3-319-68032-3 (eBook)
https://doi.org/10.1007/978-3-319-68032-3

Library of Congress Control Number: 2017956728

Mathematics Subject Classification (2010): 05A17, 11A05, 11A41, 11A51, 11B05, 11B13, 11B75, 11N05, 11P21, 11P81, 11P84, 11Y05

Printed on acid-free paper

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

Preface

The CUNY Graduate Center Workshops on Combinatorial and Additive Number Theory (CANT) have been organized every year, beginning in 2003, by the New York Number Theory Seminar. The seminar was started in 1981 by David and Gregory Chudnovsky, Harvey Cohn, and Mel Nathanson, and for 36 years has been meeting every Thursday afternoon during the academic year, and also in the summer.

The four-day CANT conferences are held in May at the CUNY Graduate Center in Manhattan, usually from Tuesday to Friday of the week immediately preceding Memorial Day. They have become a fixed point in the number theory calendar.

This collection derives from talks at the CANT 2015 and CANT 2016 workshops. There are 20 papers on important topics in number theory and related parts of mathematics. These topics include sumsets, partitions, convex polytopes and discrete geometry, Ramsey theory, primality testing, and cryptography.

I am grateful to Springer and its mathematics editor, Marc Strauss, for publishing the proceedings of these meetings. A previous volume is [1].

Bronx, NY, USA
Melvyn B. Nathanson

Reference

1. M.B. Nathanson, editor, Combinatorial and additive number theory–CANT 2011 and 2012. Springer Proc. Math. Stat. vol. 101, Springer, New York, 2014

Contents

On a Conjecture of Fox and Kleitman on the Degree of Regularity of a Certain Linear Equation

Sukumar Das Adhikari and Shalom Eliahou

Abstract Fox and Kleitman proved in 2006 that for any positive integer b, the $2n$-variable equation $x_1 + \cdots + x_n - x_{n+1} - \cdots - x_{2n} = b$ is not $2n$-regular. Moreover, they conjectured the existence of an integer $b_n \geq 1$ such that for $b = b_n$, this equation is $(2n-1)$-regular. In this note, we settle the first nontrivial case of the conjecture, namely for $n = 2$, and we propose a slight refinement of it.

Keywords Partition regularity · Diophantine equation · Finite coloring Monochromatic solution

1 Introduction

Here, $\mathbb{Z}$ denotes the set of integers and $\mathbb{N}_+$ the set of positive integers. For given integers $\alpha_1, \ldots, \alpha_k$ and c, consider the linear Diophantine equation L:

$$\sum_{i=1}^{k} \alpha_i x_i = c.$$

Following Rado [5], given $n \in \mathbb{N}_+$, equation L is said to be *n-regular* if, for every n-coloring of $\mathbb{N}_+$, there exists a *monochromatic* solution $x \in \mathbb{N}_+^k$ to L.

S. Das Adhikari
Harish-Chandra Research Institute, Chhathnag Road, Jhunsi, Allahabad 211 019, India
e-mail: adhikari@hri.res.in

S. Eliahou (✉)
EA 2597 - LMPA - Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, Université du Littoral Côte d'Opale (ULCO), CS, 62228 Calais, France
e-mail: eliahou@lmpa.univ-littoral.fr

S. Eliahou
CNRS FR 2956, Paris, France

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_1

The *degree of regularity* of L is the largest integer $n \geq 0$, if any, such that L is n-regular. This (possibly infinite) number is denoted by $dor(L)$. If $dor(L) = \infty$, then L is said to be *regular*.

A conjecture of Rado [5] states that there is a function $r \colon \mathbb{N}_+ \to \mathbb{N}_+$ such that given any $n \in \mathbb{N}_+$ and any equation $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$ with integer coefficients, if this equation is not regular over $\mathbb{N}_+$, then it already fails to be $r(n)$-regular. Even though there is a more general version, we state it here for a single homogeneous equation, as it has been proved by Rado that if the conjecture is true for a single equation, then it is true for a system of finitely many linear equations [5], and as Fox and Kleitman have shown that if the conjecture is true for a linear homogeneous equation, then it is true for any linear equation [3]. This conjecture is known as *Rado's Boundedness Conjecture*. The first nontrivial case of the conjecture has been proved by Fox and Kleitman [3]; more precisely, they established the bound $r(3) \leq 24$. In the same paper, the authors made the following conjecture for a very specific linear Diophantine equation [3].

Conjecture 1 *Let $n \geq 1$. There exists an integer $b_n \geq 1$ such that the degree of regularity of the $2n$-variable equation*

$$x_1 + \cdots + x_n - x_{n+1} - \cdots - x_{2n} \;=\; b_n$$

is exactly $2n - 1$.

If true, that would be best possible, since they proved in the same paper that for any $b_n \in \mathbb{N}_+$, the above equation is not $2n$-regular.

In this note, we settle the first nontrivial case of the conjecture, namely the case $n = 2$. Indeed, we shall show that if b_2 is any positive multiple of 6, then the corresponding equation has degree of regularity exactly 3.

More generally, we shall determine the degree of regularity over $\mathbb{N}_+$ of the equation

$$x_1 + x_2 - x_3 - x_4 \;=\; b$$

for all $b \in \mathbb{N}_+$. See Theorem 1 for the exact statement.

A related conjecture of Rado [5], stating that for every positive integer n, there exists a linear homogeneous equation with degree of regularity equal to n, was proved by Alexeev and Tsimerman [1]. Before that paper, Fox and Radoičić [2] had shown that for $n \geq 2$, the equation

$$x_1 + 2x_2 + \cdots + 2^{n-2}x_{n-1} - 2^{n-1}x_n = 0 \tag{1}$$

is not n-regular and had conjectured that it is $(n-1)$-regular; Golowich [4] proved their conjecture, thus providing another proof of the above-mentioned conjecture of Rado.

2 Main Result

Here is the main result of this note, which solves the case $n = 2$ of the conjecture of Fox and Kleitman.

Theorem 1 *For any $b \in \mathbb{N}_+$, let L_b be the equation*

$$x_1 + x_2 - x_3 - x_4 = b.$$

The degree of regularity $dor(L_b)$ over $\mathbb{N}_+$ only depends on the class of b mod 6, as follows:

$$\mathrm{dor}(L_b) = \begin{cases} 1 & \text{if } b \equiv 1, 3, 5 \bmod 6, \\ 2 & \text{if } b \equiv 2, 4 \bmod 6, \\ 3 & \text{if } b \equiv 0 \bmod 6. \end{cases}$$

Proof There are several steps.
Step 0. For any $b \in \mathbb{N}_+$, we have

$$1 \le dor(L_b) \le 3.$$

Indeed, equation L_b is obviously 1-regular since it is solvable in $\mathbb{N}_+$. Moreover, as mentioned above, it is not 4-regular [3].

Step 1. Assume first that b is odd. Consider the 2-coloring mod_2 of $\mathbb{N}_+$ given by the class mod 2. Let $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ be a mod_2-monochromatic vector in $\mathbb{N}_+^4$. Then the λ_is all have the same parity, whence

$$\lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 \equiv 0 \bmod 2.$$

Since $b \not\equiv 0 \bmod 2$, it follows that L_b admits no mod_2-monochromatic solution in $\mathbb{N}_+^4$. Therefore, $\mathrm{dor}(L_b) < 2$, implying $\mathrm{dor}(L_b) = 1$ by Step 0 above. This covers the cases $b \equiv 1, 3, 5 \bmod 6$.

Step 2. Assume now that b is even. Let us show then that L_b is 2-regular. Indeed, letting $h = b/2$ with $h \in \mathbb{N}_+$, the following three vectors in $\mathbb{N}_+^4$ are solutions to L_b:

$$\begin{aligned} &(b + 1, 1, 1, 1), \\ &(h + 1, h + 1, 1, 1), \\ &(b + 1, b + 1, h + 1, h + 1). \end{aligned}$$

For any given 2-coloring of $\mathbb{N}_+$, at least two elements in the set $\{b + 1, h + 1, 1\}$ must have the same color. Therefore, at least one of the above three solutions must be monochromatic. This shows that $\mathrm{dor}(L_b) \ge 2$, as asserted.

Step 3. Assume that $b \not\equiv 0 \bmod 3$. Let us show then that $\mathrm{dor}(L_b) \le 2$ in that case, i.e., that L_b is not 3-regular. Consider the 3-coloring mod_3 of $\mathbb{N}_+$ given by the class mod 3. Let $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ be a mod_3-monochromatic vector in $\mathbb{N}_+^4$. Then the λ_is all have the same class mod 3, whence

$$\lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 \equiv 0 \bmod 3.$$

Since $b \not\equiv 0 \bmod 3$, it follows that L_b admits no mod_3-monochromatic solution, whence $\mathrm{dor}(L_b) \leq 2$ as claimed. In particular, when b is even, this covers the cases $b \equiv 2, 4 \bmod 6$.

Step 4. In the remaining case $b \equiv 0 \bmod 6$, we claim that the equation L_b is 3-regular. To that end, let us show here that it suffices to treat the case $b = 6$.

Indeed, assume that L_6 is 3-regular, and let $b = 6k$ with $k \geq 2$. Let c be a 3-coloring of $\mathbb{N}_+$. Let c' be the new 3-coloring of $\mathbb{N}_+$ defined by the formula

$$c'(n) = c(nk)$$

for all $n \in \mathbb{N}_+$. Since L_6 is 3-regular, there is a vector $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \mathbb{N}_+^4$ satisfying L_6, i.e., such that

$$\lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 = 6,$$

and which is monochromatic under c'. Therefore, the vector $(\lambda_1 k, \lambda_2 k, \lambda_3 k, \lambda_4 k)$ satisfies L_{6k}, i.e.,

$$\lambda_1 k + \lambda_2 k - \lambda_3 k - \lambda_4 k = 6k,$$

and is monochromatic under c by construction. This shows that if L_6 is 3-regular, then L_{6k} also is 3-regular for all $k \geq 2$.

Step 5. We now complete the proof of the theorem by establishing the 3-regularity of L_6.

Let $c\colon \mathbb{N}_+ \to \{0, 1, 2\}$ be an arbitrary 3-coloring of $\mathbb{N}_+$. We need to show that at least one solution $x \in \mathbb{N}_+^4$ of equation L_6, i.e., of

$$x_1 + x_2 - x_3 - x_4 = 6,$$

is monochromatic under c.

Here are five families of special solutions to L_6 with only two or three distinct entries, parametrized by $a \in \mathbb{N}_+$:

$$\begin{aligned} &\{a + 6, a, a, a\},\\ &\{a + 5, a + 1, a, a\},\\ &\{a + 4, a + 2, a, a\},\\ &\{a + 3, a + 3, a, a\},\\ &\{a + 8, a, a + 1, a + 1\}. \end{aligned}$$

Consider now the family of underlying sets of these special solutions:

$$\mathcal{E} = \big\{\{a, a+3\},\ \{a, a+6\},\ \{a, a+2, a+4\},\ \{a, a+1, a+5\},\ \{a, a+1, a+8\}\big\},$$

where a ranges through $\mathbb{N}_+$.

If any element in $\mathcal{E}$ happens to be a monochromatic set, we are done. So, from now on, we may and will make the following assumption:

(H) *All elements in $\mathcal{E}$ are multichromatic sets under c,*

where *multichromatic* means nonmonochromatic here. We now proceed to show that this hypothesis leads to a contradiction, thereby completing the proof of the theorem.

First observe that

$$\{c(a), c(a+3), c(a+6)\} \ = \ \{0, 1, 2\} \tag{2}$$

for all $a \in \mathbb{N}_+$. Indeed, by (H), the colors of $a, a+3, a+6$ are pairwise distinct since $a+6 = (a+3)+3$. This implies the following.

Claim 1 *For all $a \in \mathbb{N}_+$, we have*

$$c(a) \ = \ c(a+9). \tag{3}$$

Indeed, by (2) we have

$$\{c(a), c(a+3), c(a+6)\} \ = \ \{0, 1, 2\} \ = \ \{c(a+3), c(a+6), c(a+9)\}.$$

Since the sets on the left and on the right are equal and have

$$I = \{c(a+3), c(a+6)\}$$

in common, they remain equal when removing I. This implies $c(a) = c(a+9)$, as claimed.

Consequently, our 3-coloring c induces a well-defined 3-coloring on $\mathbb{Z}/9\mathbb{Z}$ that we still denote by c,

$$c \colon \mathbb{Z}/9\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}.$$

For simplicity, let us denote the elements of $\mathbb{Z}/9\mathbb{Z}$ by 0, 1, ..., 8 and their respective colors under c by $c_0, c_1, \dots, c_8$. Moreover, let us depict the distribution of these colors in the following table C (Table 1):

Claim 2 *For all $i \in \mathbb{Z}/9\mathbb{Z}$, we have*

Table 1 Color table C

c_0	c_1	c_2
c_3	c_4	c_5
c_6	c_7	c_8

Table 2 Color table C revisited

0	c_1	c_2
1	c_4	c_5
2	c_7	c_8

$$\left|\{c_i, c_{i+2}, c_{i+4}\}\right| \geq 2, \tag{4}$$
$$\left|\{c_i, c_{i+1}, c_{i+5}\}\right| \geq 2, \tag{5}$$
$$\left|\{c_i, c_{i+1}, c_{i+2}\}\right| \geq 2. \tag{6}$$

Indeed, this follows from the fact that the sets $\{a, a+2, a+4\}$, $\{a, a+1, a+5\}$, and $\{a, a+1, a+8\}$ belong to $\mathcal{E}$ for all $a \in \mathbb{N}_+$ and hence by (H) are assumed to be multichromatic under c.

Now, by (2), (3) and up to symmetry, we may assume that the first column (c_0, c_3, c_6) of C is equal to $(0, 1, 2)$, as depicted in Table 2.

Moreover, it follows from (2) again that the second and third columns of C are both permutations of its first column. Now, there are nine possible pairs holding the remaining two 0s in C, namely

$$(c_1, c_2),\ (c_1, c_5),\ (c_1, c_8);$$
$$(c_4, c_2),\ (c_4, c_5),\ (c_4, c_8);$$
$$(c_7, c_2),\ (c_7, c_5),\ (c_7, c_8).$$

But each one in turn is excluded by an appropriate argument, recalling that $c_0 = 0$:

$$\begin{array}{lll}
\left|\{c_0, c_1, c_2\}\right| \geq 2 \text{ by } (6), & \left|\{c_0, c_1, c_5\}\right| \geq 2 \text{ by } (5), & \left|\{c_8, c_0, c_1\}\right| \geq 2 \text{ by } (6); \\
\left|\{c_0, c_2, c_4\}\right| \geq 2 \text{ by } (4), & \left|\{c_4, c_5, c_0\}\right| \geq 2 \text{ by } (5), & \left|\{c_8, c_0, c_4\}\right| \geq 2 \text{ by } (5); \\
\left|\{c_7, c_0, c_2\}\right| \geq 2 \text{ by } (4), & \left|\{c_5, c_7, c_0\}\right| \geq 2 \text{ by } (4), & \left|\{c_7, c_8, c_0\}\right| \geq 2 \text{ by } (6).
\end{array}$$

This contradiction shows that (H) is absurd and concludes the proof of the theorem. □

Slight changes in the above proof actually give a somewhat stronger result.

Proposition 1 *For every integer interval $A = [r, r+17] \subseteq \mathbb{N}_+$ of cardinality 18, and for every 3-coloring of A, the equation*

$$x_1 + x_2 - x_3 - x_4 \ = \ 6$$

admits a monochromatic solution $x \in A^4$.

Proof (Sketch) Indeed, one observes that the whole proof of Theorem 1 goes through by considering colorings on $[r, r+17]$ only.

In Step 5, one argues with an arbitrary 3-coloring $c' \colon [r, r+17] \to \{0, 1, 2\}$, and the family of underlying sets of special solutions:

$$\mathcal{E}' = \big\{\{a, a+3\},\ \{a, a+6\},\ \{a, a+2, a+4\},\ \{a, a+1, a+5\},\ \{a, a+1, a+8\}\big\},$$

where a ranges through $[r, r+8]$. This restricted range for a is the only difference between $\mathcal{E}'$ and the family $\mathcal{E}$ considered in the proof of Theorem 1. Thus here, any set of special solutions is contained in the interval $[r, r+17]$.

One then considers the map

$$c' \colon (\mathbb{Z} \cap [r, r+17])/9\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}$$

induced by the 3-coloring, and the rest of the argument is the same. □

We now consider the equation

$$x_1 + \cdots + x_n - x_{n+1} - \cdots - x_{2n} = 6$$

for $n \geq 3$ by means of the following remark.

Lemma 1 *Let $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_l \in \mathbb{Z}$. Assume that the k-variable equation $\sum_{i=1}^{k} \alpha_i x_i = 0$ is r-regular and that $\sum_{j=1}^{l} \beta_j = 0$. Then, the $(k+l)$-variable equation $\sum_{i=1}^{k} \alpha_i x_i + \sum_{j=1}^{l} \beta_j x_{k+j} = 0$ is also r-regular.*

Proof Let c be any r-coloring of $\mathbb{N}_+$. By assumption, there is a monochromatic solution $(u_1, \ldots, u_k) \in \mathbb{N}_+^k$ to the first equation, namely satisfying $\sum_{i=1}^{k} \alpha_i u_i = 0$. Since $(\sum_{j=1}^{l} \beta_j) u_k = 0$, it follows that $(u_1, \ldots, u_k, u_k, \ldots, u_k) \in \mathbb{N}_+^{k+l}$ is a monochromatic solution to the extended equation. □

This yields the following extension of Theorem 1.

Corollary 1 *For every integer $n \geq 2$, the $2n$-variable equation*

$$x_1 + \cdots + x_n - x_{n+1} - \cdots - x_{2n} = 6$$

is 3-regular.

Proof The case $n = 2$ is settled in Theorem 1. For $n \geq 3$, note that the given $2n$-variable equation is extended from the corresponding $2(n-1)$-variable one by adding the zero coefficient sum term $x_n - x_{2n}$. Therefore, Lemma 1 applies, and a repeated application of it from the case $n = 2$ yields the claimed statement. □

3 A Refined Conjecture

We conclude this note with a slight refinement of the conjecture of Fox and Kleitman. Consider again the $2n$-variable equation

$$x_1 + \cdots + x_n - x_{n+1} - \cdots - x_{2n} = b \tag{7}$$

with $b \in \mathbb{N}_+$. As recalled in the Introduction, Fox and Kleitman proved that this equation is never $2n$-regular [3].

Conjecture 2 *The degree of regularity of Eq.* (7) *only depends on the class of b mod* $(2n-1)!$. *Moreover, Eq.* (7) *is* $(2n-1)$*-regular exactly when b is a multiple of* $(2n-1)!$.

Note that Theorem 1 settles the case $n = 2$ of this refined conjecture. As for $n = 3$, for instance, the conjecture states that the equation

$$x_1 + x_2 + x_3 - x_4 - x_5 - x_6 \;=\; 120k$$

should be 5-regular for all $k \geq 1$. As in Step 4 of the proof of Theorem 1, it would suffice to show it for $k = 1$.

Let now $N(n)$ be the lowest common multiple of all integers from 1 to $2n-1$. For instance, $N(3) = 60$. If the above right-hand side $120k$ is replaced by any $b \not\equiv 0 \bmod 60$, the resulting equation fails to be 5-regular; this follows from the following more general statement, a tiny step toward Conjecture 2.

Proposition 2 *If $b \not\equiv 0$ mod $N(n)$, then Eq.* (7) *is not* $(2n-1)$*-regular.*

Proof By assumption on b, there exists $1 \leq k \leq 2n-1$ such that $b \not\equiv 0 \bmod k$. Consider then the k-coloring mod_k of $\mathbb{N}_+$ given by the class mod k. Let $(u_1, \ldots, u_{2n}) \in \mathbb{N}_+^{2n}$ be any monochromatic vector under mod_k, i.e., satisfying $u_i \equiv a \bmod k$ for some $a \in \mathbb{N}$ and for all $1 \leq i \leq 2n$. Then

$$u_1 + \cdots + u_n - u_{n+1} - \cdots - u_{2n} \;\equiv\; 0 \bmod k.$$

Since $b \not\equiv 0 \bmod k$, it follows that Eq. (7) admits no monochromatic solution for this specific k-coloring. Therefore, Eq. (7) is not k-regular, whence it is not $(2n-1)$-regular either since $2n-1 \geq k$. □

In view of the above result, one might wonder whether Conjecture 2 might hold with $(2n-1)!$ replaced by its factor $N(n)$. The answer is no. As it happens, the equation $x_1 + x_2 + x_3 - x_4 - x_5 - x_6 \;=\; 60$ is not 5-regular. In fact, it is not even 4-regular, as will be shown in a subsequent paper.

References

1. B. Alexeev, J. Tsimerman, Equations resolving a conjecture of Rado on partition regularity. J. Comb. Theory Ser. A **117**, 1008–1010 (2010)
2. J. Fox, D.J. Kleitman, On rado's boundedness conjecture. J. Comb. Theory Ser. A **113**, 84–100 (2006)
3. J. Fox, R. Radoičič, The axiom of choice and the degree of regularity of equations over the reals. Preprint (2005)
4. N. Golowich, Resolving a conjecture on degree of regularity of linear homogeneous equations. Electron. J. Comb. **21**(3), paper 3.28 (2014)
5. R. Rado, Studien zur Kombinatorik. Math. Z. **36**, 424–480 (1933)

Open Problems About Sumsets in Finite Abelian Groups: Minimum Sizes and Critical Numbers

Béla Bajnok

Abstract For a positive integer h and a subset A of a given finite abelian group, we let hA, $h\hat{}A$, and $h_{\pm}A$ denote the h-fold sumset, restricted sumset, and signed sumset of A, respectively. Here we review some of what is known and not yet known about the minimum sizes of these three types of sumsets, as well as their corresponding critical numbers. In particular, we discuss several new open direct and inverse problems.

Keywords Additive combinatorics · Finite abelian group · Sumset
Critical number

1 Introduction and Notations

Throughout this paper, G denotes a finite abelian group of order $n \geq 2$, written in additive notation. If G is cyclic, we identify it with $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. We say that G has type $(n_1, \ldots, n_r)$ if

$$G \cong \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_r}$$

for integers $2 \leq n_1 | n_2 | \cdots | n_r$; here r is the rank, and n_r is the exponent of G. For an m-subset $A = \{a_1, \ldots, a_m\}$ of G and for a nonnegative integer h, we consider three types of sumsets:

- the *h-fold sumset*:

$$hA = \left\{ \Sigma_{i=1}^m \lambda_i a_i \mid \lambda_1, \cdots, \lambda_m \in \mathbb{N}_0,\ \Sigma_{i=1}^m \lambda_i = h \right\},$$

- the *h-fold restricted sumset*:

$$h\hat{}A = \left\{ \Sigma_{i=1}^m \lambda_i a_i \mid \lambda_1, \cdots, \lambda_m \in \{0, 1\},\ \Sigma_{i=1}^m \lambda_i = h \right\},$$

- the *h-fold signed sumset*:

B. Bajnok (✉)
Department of Mathematics, Gettysburg College, Gettysburg, PA 17325-1486, USA
e-mail: bbajnok@gettysburg.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_2

$$h_{\pm}A = \left\{ \Sigma_{i=1}^{m} \lambda_i a_i \mid \lambda_1, \cdots, \lambda_m \in \mathbb{Z},\ \Sigma_{i=1}^{m} |\lambda_i| = h \right\}.$$

We denote the set formed by the inverses of the elements of A by $-A$; we say that A is *symmetric* if $A = -A$ and that A is *asymmetric* if A and $-A$ are disjoint. For an element $b \in \mathbb{Z}$, we write $b \cdot A$ for the *b-fold dilation* $\{b \cdot a_1, \ldots, b \cdot a_m\}$ of A. The subgroup of G generated by A is denoted by $\langle A \rangle$.

It is a central question in additive combinatorics to evaluate minimum sumset sizes, in particular, for given G, h, and $1 \leq m \leq n$ the quantities

$$\rho(G, m, h) = \min\{|hA| \mid A \subseteq G, |A| = m\},$$

$$\rho\hat{}(G, m, h) = \min\{|h\hat{}A| \mid A \subseteq G, |A| = m\},$$

$$\rho_{\pm}(G, m, h) = \min\{|h_{\pm}A| \mid A \subseteq G, |A| = m\}.$$

Trivially, each value is 1 whenever $h = 0$, and each value equals m whenever $h = 1$. (To see that $\rho_{\pm}(G, m, 1) = m$, note that every group has a symmetric subset of any size $m \leq n$.) Below we assume that $h \geq 2$; in the case of restricted sums, we may and will also assume that $h \leq m - 2$.

The study of $\rho(G, m, h)$ goes back two hundred years to the work of Cauchy [14] who determined it for groups of prime order and $h = 2$ and is now known for all parameters—see Sect. 2. However, only partial results have been found for $\rho\hat{}(G, m, h)$ and $\rho_{\pm}(G, m, h)$—we discuss these in Sects. 3 and 4.

We also consider minimum sumset sizes without restrictions on the number of terms added:

$$\rho(G, m, \mathbb{N}_0) = \min\{| \cup_{h=0}^{\infty} hA| \mid A \subseteq G, |A| = m\},$$

$$\rho\hat{}(G, m, \mathbb{N}_0) = \min\{| \cup_{h=0}^{\infty} h\hat{}A| \mid A \subseteq G, |A| = m\},$$

$$\rho_{\pm}(G, m, \mathbb{N}_0) = \min\{| \cup_{h=0}^{\infty} h_{\pm}A| \mid A \subseteq G, |A| = m\}.$$

Since $\cup_{h=0}^{\infty} hA$ and $\cup_{h=0}^{\infty} h_{\pm}A$ both equal $\langle A \rangle$, we have

$$\rho(G, m, \mathbb{N}_0) = \rho_{\pm}(G, m, \mathbb{N}_0) = \min\{d \in D(n) \mid d \geq m\},$$

where $D(n)$ is the set of positive divisors of n. The set $\cup_{h=0}^{\infty} h\hat{}A$, also denoted by ΣA, is less understood; we discuss $\rho\hat{}(G, m, \mathbb{N}_0)$ in Sect. 5.

Related to each function, we study the corresponding *critical number*: the minimum value of m, if it exists, for which the corresponding sumset of any m-subset of G is G itself. The study of critical numbers originated with the 1964 paper [21] of Erdős and Heilbronn; in Sect. 6, we review what is known and not yet known about them.

Furthermore, we also examine the so-called *inverse problems* corresponding to some of these quantities, that is, we look for subsets of the group that achieve the extremal values of our functions.

The open problems mentioned here are just some of the many intriguing questions about sumsets.

2 Minimum Size of h-fold Sumsets

Given G, m, and h, which m-subsets of G have the smallest h-fold sumsets? Two ideas come to mind: Place the elements into a coset of some subgroup, or have the elements form an arithmetic progression. We may also combine these two ideas; for example, in the cyclic group $\mathbb{Z}_n$, we take an arithmetic progression of cosets, as follows.

For any divisor d of n, we take the subgroup $H = \cup_{j=0}^{d-1}\{j \cdot n/d\}$, then set

$$A_d(n,m) = \cup_{i=0}^{c-1}(i+H) \bigcup \cup_{j=0}^{k-1}\{c + j \cdot n/d\},$$

where $m = cd + k$ and $1 \leq k \leq d$. An easy computation shows that

$$|hA_d(n,m)| = \min\{n,\ (hc+1)d,\ hm - h + 1\};$$

letting

$$f_d(m,h) = (hc+1)d = (h\lceil m/d \rceil - h + 1)d,$$

and noting that $f_n(m,h) = n$ and $f_1(m,h) = hm - h + 1$, allows us to write

$$|hA_d(n,m)| = \min\{f_n(m,h),\ f_d(m,h),\ f_1(m,h)\}.$$

Therefore, with

$$u(n,m,h) = \min\{f_d(m,h) \mid d \in D(n)\}$$

we get $\rho(\mathbb{Z}_n, m, h) \leq u(n,m,h)$. It turns out that a similar construction works in any group and that we cannot do better:

Theorem 1 (Plagne [38]) *For every G, m, and h, we have $\rho(G,m,h) = u(n,m,h)$.*

(Here $u(n,m,h)$ is a relative of the Hopf–Stiefel function used also in topology and bilinear algebra; see, for example, [20], [30], [37], and [40].)

With $\rho(G,m,h)$ thus determined, let us turn to the inverse problem of classifying all m-subsets A of G for which hA has minimum size $\rho(G,m,h)$. The general question seems complicated. For example, while one can show that for a 6-subset A of $\mathbb{Z}_{15}$ to have a twofold sumset of size $\rho(\mathbb{Z}_{15}, 6, 2) = 9$, A must be the union of two cosets of the order 3 subgroup of $\mathbb{Z}_{15}$; there are three different possibilities for

$\rho(\mathbb{Z}_{15}, 7, 2) = 13$: A can be the union of two cosets of the order 3 subgroup plus one additional element, or a coset of the order 5 subgroup together with two more elements, or an arithmetic progression of length 7.

We are able to say more for m values that are not more than the smallest prime divisor p of n. Note that, as a special case of Theorem 1, when $m \leq p$, we get

$$\rho(G, m, h) = \min\{p, hm - h + 1\}. \tag{2.1}$$

The case when p is greater than $hm - h + 1$ easily follows from [31]:

Theorem 2 (Kemperman [31]) *Let p be the smallest prime divisor of n, and assume that $h \geq 2$ and $p > hm - h + 1$. Then for an m-subset A of G, we have $|hA| = \rho(G, m, h) = hm - h + 1$ if, and only if, A is an arithmetic progression in G.*

For the case when p is less than $hm - h + 1$, we propose:

Conjecture 3 *Let p be the smallest prime divisor of n, and assume that $m \leq p < hm - h + 1$. Then for an m-subset A of G, we have $|hA| = \rho(G, m, h) = p$ if, and only if, A is contained in a coset of some subgroup H of G with $|H| = p$.*

This leaves the case when $p = hm - h + 1$, where arithmetic progressions of length m and m-subsets in a coset of a subgroup of order p are two of several possibilities. It may be an interesting problem to classify all such subsets.

3 Minimum Size of *h*-fold Restricted Sumsets

While the value of $\rho\hat{}(G, m, h)$ is not even known for cyclic groups in general, as it turns out, we get an extremely close approximation for it by considering the sets $A_d(n, m) \subseteq \mathbb{Z}_n$ of Sect. 2 above. A somewhat tedious computation shows that we get

$$|h\hat{}A_d(n, m)| = \begin{cases} \min\{n,\ (hc+1)d,\ hm - h^2 + 1\} \text{ if } h \leq \min\{k, d-1\}; \\ \\ \min\{n,\ hm - h^2 + 1 + \delta_d\} \qquad \text{otherwise,} \end{cases}$$

where δ_d is an explicitly computed *correction term* (see [4] for details). Letting

$$u\hat{}(n, m, h) = \min\{|h\hat{}A_d(n, m)| \mid d \in D(n)\},$$

we thus get $\rho\hat{}(\mathbb{Z}_n, m, h) \leq u\hat{}(n, m, h)$. Since $|h\hat{}A_d(n, m)|$ equals $\min\{n, hm - h^2 + 1\}$ for both $d = 1$ and $d = n$, we always have

$$\rho\hat{}(\mathbb{Z}_n, m, h) \leq \min\{n, hm - h^2 + 1\}.$$

As is well known, equality holds for prime n:

Theorem 4 (Dias Da Silva, Hamidoune [16]; Alon et al. [1, 2]) For a prime p, we have

$$\rho\hat{}(\mathbb{Z}_p, m, h) = \min\{p, hm - h^2 + 1\}.$$

The lower bound $u\hat{}(n, m, h)$ is surprisingly accurate for cyclic groups of composite order as well: For all (n, m, h) with $n \leq 40$, we find that equality holds in over 99.9% of cases, and when it does not, then $\rho\hat{}(\mathbb{Z}_n, m, h)$ and $u\hat{}(n, m, h)$ differ only by 1. All the exceptions that are known come from the construction that we explain next.

Recall that the m elements in $A_d(n, m)$ are within $c + 1 = \lceil m/d \rceil$ cosets of the order d subgroup H of $\mathbb{Z}_n$, and at most one of these cosets is not contained entirely in $A_d(n, m)$. We now consider the variation when the m elements are still within $c + 1$ cosets of H, but exactly two of the cosets do not lie entirely in our set. In order to do so, we write

$$m = k_1 + (c - 1)d + k_2$$

with positive integers k_1 and k_2; we assume that $k_1 < d$, $k_2 < d$, but $k_1 + k_2 > d$. We then set

$$B_d(n, m) = \cup_{j=0}^{k_1-1}\{j \cdot n/d\} \bigcup \cup_{i=1}^{c-1}(i \cdot g + H) \bigcup \cup_{j=0}^{k_2-1}\{c \cdot g + (j_0 + j) \cdot n/d\},$$

where $0 \leq j_0 \leq d - 1$ and $g \in \mathbb{Z}_n$. As it turns out, $|h\hat{}B_d(n, m)|$ is less than $|h\hat{}A_d(n, m)|$ in just three specific cases: When $h = 2$, n is divisible by $2m - 2$, and $m - 1$ is not a power of 2; when $h = 3$, $m = 6$, and n is divisible by 10; and when h is odd, n is divisible by $hm - h^2$, and $m + 2$ is divisible by $h + 2$ (see [4]). Moreover, every known instance when $\rho\hat{}(\mathbb{Z}_n, mh)$ is less than $u\hat{}(n, m, h)$ arises as one of these three cases. Letting

$$w\hat{}(n, m, h) = \min\{|h\hat{}B_d(n, m)| \mid d \in D(n)\},$$

we see that $\rho\hat{}(\mathbb{Z}_n, m, h)$ is at most $\min\{u\hat{}(n, m, h), w\hat{}(n, m, h)\}$, but we also believe that equality holds:

Conjecture 5 *For all n, m, and h, we have*

$$\rho\hat{}(\mathbb{Z}_n, m, h) = \min\{u\hat{}(n, m, h), w\hat{}(n, m, h)\}.$$

Let us highlight the case $h = 2$. First, note that Conjecture 5 then becomes:

$$\rho\hat{}(\mathbb{Z}_n, m, 2) = \begin{cases} \min\{\rho(\mathbb{Z}_n, m, 2), 2m - 4\} \text{ if } 2|n \text{ and } 2|m, \\ \qquad\qquad\qquad\qquad \text{or } (2m - 2)|n \text{ and } \log_2(m - 1) \notin \mathbb{N}; \\ \min\{\rho(\mathbb{Z}_n, m, 2), 2m - 3\} \text{ otherwise.} \end{cases}$$

Some general inequalities are known: Plagne [39] proved that the upper bound

$$\rho\hat{}(G, m, 2) \leq \min\{\rho(G, m, 2), 2m - 2\}$$

holds for all groups, and Eliahou and Kervaire [19] proved that the lower bound

$$\rho\hat{}(G, m, 2) \geq \min\{\rho(G, m, 2), 2m - 3\}$$

holds for all elementary abelian p-groups for odd p. Furthermore, Lev [32] conjectured the lower bound

$$\rho\hat{}(G, m, 2) \geq \min\{\rho(G, m, 2), 2m - 3 - |\mathrm{Ord}(G, 2)|\},$$

where $\mathrm{Ord}(G, 2)$ is the set of elements of G that have order 2, and Plagne [39] conjectured that $\rho\hat{}(G, m, 2)$ and $\rho(G, m, 2)$ can differ by at most 2. (We should add that no such statement is possible for higher h values: As was proven in [4], when $h \geq 3$, for any $C \in \mathbb{N}$, one can find a group G and a positive integer m so that $\rho\hat{}(G, m, h)$ and $\rho(G, m, h)$ differ by C or more.)

As in Sect. 2, we are able to say more when $m \leq p$ with p being the smallest prime divisor of n. We believe that the following analogue of (2.1) holds:

Conjecture 6 *If p is the smallest prime divisor of n and $h < m \leq p$, then*

$$\rho\hat{}(G, m, h) = \min\{p, hm - h^2 + 1\}.$$

Note that Conjecture 6 is a generalization of Theorem 4.

Turning to inverse problems: our analogues for Theorem 2 and Conjecture 3 are:

Conjecture 7 *Let p be the smallest prime divisor of n, and assume that $2 \leq h \leq m - 2$ and $p > hm - h^2 + 1$. Then for an m-subset A of G, we have $|h\hat{}A| = hm - h^2 + 1$ if, and only if, $h = 2$, $m = 4$, and $A = \{a, a + g_1, a + g_2, a + g_1 + g_2\}$ for some $a, g_1, g_2 \in G$, or A is an arithmetic progression in G.*

Conjecture 8 *Let p be the smallest prime divisor of n, and assume that $m \leq p < hm - h^2 + 1$. Then for an m-subset A of G, we have $|h\hat{}A| = p$ if, and only if, A is contained in a coset of some subgroup H of G with $|H| = p$.*

Károlyi proved Conjectures 6 and 7 for $h = 2$ [28, 29].

4 Minimum Size of h-fold Signed Sumsets

Studying the function $\rho_{\pm}(G, m, h)$ provides us with several surprises. First, we realize that, unlike it is the case for $\rho(G, m, h)$, the value of $\rho_{\pm}(G, m, h)$ depends on the structure of G and not just on the order n of G. Second, while the size of the signed sumset of a subset is usually much greater than the size of its sumset, the value of $\rho_{\pm}(G, m, h)$ equals $\rho(G, m, h)$ surprisingly often; in fact, there is only one case with $n \leq 24$ where the two are not equal: $\rho_{\pm}(\mathbb{Z}_3^2, 4, 2) = 8$ while $\rho(\mathbb{Z}_3^2, 4, 2) = 7$. Furthermore, one might think that symmetric sets provide the smallest minimum size, but sometimes asymmetric sets or even *near-symmetric sets*—sets that become symmetric by the removal of one element—are better; we are able to prove, though, that one of these three types always provides the minimum size.

For our treatment below, we use the functions $f_d(m, h)$ and $u(n, m, h)$ defined in Sect. 2. For cyclic groups, we have the following result:

Theorem 9 (Bajnok and Matzke [9]) *For cyclic groups G, m, and h, we have* $\rho_{\pm}(G, m, h) = \rho(G, m, h)$.

The proof of Theorem 9 follows from the fact that for each $d \in D(n)$ one can find a symmetric subset R of G of size at least (but not necessarily equal to) m for which $|hR| \leq f_d(n, m)$.

More generally, for a group of type $(n_1, \ldots, n_r)$ one can prove that

$$\rho_{\pm}(G, m, h) \leq \min\{\Pi_{i=1}^{r} \rho_{\pm}(\mathbb{Z}_{n_i}, m_i, h) \mid m_i \leq n_i, \Pi_{i=1}^{r} m_i \geq m\},$$

so by Theorems 1 and 9,

$$\rho_{\pm}(G, m, h) \leq \min\{\Pi_{i=1}^{r} u(n_i, m_i, h) \mid m_i \leq n_i, \Pi_{i=1}^{r} m_i \geq m\}.$$

Furthermore, in [9] we proved that

$$\min\{\Pi_{i=1}^{r} u(n_i, m_i, h) \mid m_i \leq n_i, \Pi_{i=1}^{r} m_i \geq m\} = \min\{f_d(m, h) \mid d \in D(G, m)\},$$

where $D(G, m)$ consists of all $d \in D(n)$ that can be written as $d = \Pi_{i=1}^{r} d_i$ with $d_i \in D(n_i)$ and $dn_r \geq d_r m$. (We may observe that for cyclic groups $D(G, m) = D(n)$.) Letting

$$u_{\pm}(G, m, h) = \min\{f_d(m, h) \mid d \in D(G, m)\}$$

thus results in the upper bound $\rho_{\pm}(G, m, h) \leq u_{\pm}(G, m, h)$. Of course, we also have $\rho_{\pm}(G, m, h) \geq u(n, m, h)$, so to get lower and upper bounds for $\rho_{\pm}(G, m, h)$, one can minimize the values of $f_d(m, h)$ for all $d \in D(n)$ and for all $d \in D(G, m)$, respectively. In fact, with one specific exception, that we are about to explain, we believe that $\rho_{\pm}(G, m, h) = u_{\pm}(G, m, h)$ holds for all G, m, and h.

We can observe that if A is asymmetric, then $0 \notin 2_{\pm}A$. Consequently, if d is an odd divisor of n and $d \geq 2m + 1$, then we can choose an m-subset of G whose twofold signed sumset has size less than d, and thus $\rho_{\pm}(G, m, 2) \leq d - 1$. We believe that this is the only possibility for $\rho_{\pm}(G, m, h)$ to be less than $u_{\pm}(G, m, h)$:

Conjecture 10 (Bajnok and Matzke [9]) *For all G, m, and* $h \geq 3$, *we have* $\rho_{\pm}(G, m, h) = u_{\pm}(G, m, h)$.

Furthermore, with $D_o(n)$ denoting the set of odd divisors of n that are greater than $2m$, we have

$$\rho_{\pm}(G,m,2)=\begin{cases} u_{\pm}(G,m,2) & \text{if } D_o(n)=\emptyset, \\ \\ \min\{u_{\pm}(G,m,2),d_m-1\} & \text{if } d_m=\min D_o(n). \end{cases}$$

We can say more about elementary abelian groups. Clearly, $\rho_{\pm}(\mathbb{Z}_2^r,m,h)=\rho(\mathbb{Z}_2^r,m,h)$, so consider $\mathbb{Z}_p^r$ where p is an odd prime. When $p \leq h$, one can prove that $\rho_{\pm}(\mathbb{Z}_p^r,m,h)=\rho(\mathbb{Z}_p^r,m,h)$ [10]. The case when h is less than p is more delicate; we need the following notations. First, set k equal to the largest integer for which $p^k+\delta \leq hm-h+1$, where $\delta=0$ if $p-1$ is divisible by h and $\delta=1$ otherwise. Second, set q equal to the largest integer for which $(hq+1)p^k+\delta \leq hm-h+1$. With these notations, we have the following result:

Theorem 11 (Bajnok and Matzke [10]) *Suppose that either $p \leq h$ or that $h < p$ and $m \leq (q+1)p^k$ with k and q defined as above. Then $\rho_{\pm}(\mathbb{Z}_p^r,m,h)=\rho(\mathbb{Z}_p^r,m,h)$.*

We believe that $\rho_{\pm}(\mathbb{Z}_p^r,m,h)$ is greater than $\rho(\mathbb{Z}_p^r,m,h)$ in the remaining case:

Conjecture 12 (Bajnok and Matzke [10]) *If $h < p$ and $m > (q+1)p^k$ with k and q defined as above, then $\rho_{\pm}(\mathbb{Z}_p^r,m,h) > \rho(\mathbb{Z}_p^r,m,h)$.*

Using Vosper's Theorem [41, 42] and (Lev's improvement [34] of) Kemperman's results on so-called *critical pairs* [31], in [10] we were able to prove Conjecture 12 for the case when $r=2$ and $h=2$; therefore, we have a complete account for all m for which $\rho_{\pm}(\mathbb{Z}_p^2,m,2)=\rho(\mathbb{Z}_p^2,m,2)$. In particular, we found that there are exactly $(p-1)^2/4$ values of m where equality does not hold. We have not been able to find any groups where this proportion is higher than $1/4$ and believe that there are none:

Conjecture 13 *For any abelian group of order n, $\rho_{\pm}(G,m,2)$ and $\rho(G,m,2)$ disagree for fewer than $n/4$ values of m.*

Let us turn now to the inverse problem of classifying all m-subsets A of G for which $|h_{\pm}A|=\rho_{\pm}(G,m,h)$. Letting $\mathrm{Sym}(G,m)$, $\mathrm{Nsym}(G,m)$, and $\mathrm{Asym}(G,m)$ denote the collection of m-subsets of G that are, respectively, symmetric, near-symmetric (that is, become symmetric after removing one element), and asymmetric, in [9] we proved that

$$\rho_{\pm}(G,m,h)=\min\{|h_{\pm}A| \mid A \in \mathrm{Sym}(G,m)\cup \mathrm{Nsym}(G,m)\cup \mathrm{Asym}(G,m)\}.$$

(None of the three types are superfluous.) This does not completely solve the inverse problem: We may have other subsets with $|h_{\pm}A|=\rho_{\pm}(G,m,h)$. Furthermore, it would be interesting to know exactly when each of the three types of sets just described yields a signed sumset of minimum size.

5 Minimum Size of Restricted Sumsets with an Arbitrary Number of Terms

In this section, we attempt to find the minimum size $\rho\hat{}(G, m, \mathbb{N}_0)$ of $\Sigma A = \cup_{h=0}^{\infty} h\hat{}A$ among all m-subsets of G. We restrict our attention to cyclic groups.

As before, we choose a divisor d of n and consider an arithmetic progression of cosets of the subgroup H of order d in $\mathbb{Z}_n$. We again write $m = cd + k$ with $1 \leq k \leq d$ and construct a set $C_d(n, m)$ that lies in exactly $c + 1 = \lceil m/d \rceil$ cosets of H, as follows.

Assume first that c is even. In this case, we let $C_d(n, m)$ consist of the collection of c cosets

$$\{i + H \mid -c/2 \leq i \leq c/2 - 1\},$$

together with k elements of the coset $c/2 + H$. (It makes no difference which k elements we choose.) It is easy to see then that

$$\Sigma C_d(n, m) = \{i + H \mid -(c^2 + 2c)/8 \cdot d \leq i \leq (c^2 - 2c)/8 \cdot d + c/2 \cdot k\},$$

and thus

$$\begin{aligned} |\Sigma C_d(n, m)| &= \min\left\{n,\ (c^2/4 \cdot d + c/2 \cdot k + 1) \cdot d\right\} \\ &= \min\left\{n,\ (c/2 \cdot m - c^2/4 \cdot d + 1) \cdot d\right\}. \end{aligned}$$

Similarly, when c is odd, we set $C_d(n, m)$ equal to the collection

$$\{i + H \mid -(c-1)/2 \leq i \leq (c-1)/2\},$$

together with k elements of the coset $(c + 1)/2 + H$; this time we find that

$$|\Sigma C_d(n, m)| = \min\left\{n,\ \big((c+1)/2 \cdot m - (c+1)^2/4 \cdot d + 1\big) \cdot d\right\}.$$

Therefore, letting

$$\begin{aligned} F_d(m) &= \big(\lceil c/2 \rceil \cdot m - \lceil c/2 \rceil^2 \cdot d + 1\big) \cdot d \\ &= \big(\lceil (m/d - 1)/2 \rceil \cdot m - \lceil (m/d - 1)/2 \rceil^2 \cdot d + 1\big) \cdot d, \end{aligned}$$

and noting that $F_n(m) = n$, we may write $|\Sigma C_d(n, m)| = \min\{F_n(m), F_d(m)\}$. Setting

$$u(n, m, \mathbb{N}_0) = \min\{F_d(m) \mid d \in D(n)\},$$

we get:

Theorem 14 *For all positive integers n and $m \leq n$, we have $\rho\hat{}(\mathbb{Z}_n, m, \mathbb{N}_0) \leq u(n, m, \mathbb{N}_0)$.*

After some numerical experimentation, we believe that equality holds:

Conjecture 15 *For all positive integers n and $m \leq n$, we have $\rho\hat{}(\mathbb{Z}_n, m, \mathbb{N}_0) = u(n, m, \mathbb{N}_0)$.*

Note that, since $F_1(m) = \lfloor m^2/4 \rfloor^2 + 1$, Theorem 14 implies that

$$\rho\hat{}(\mathbb{Z}_n, m, \mathbb{N}_0) \leq \min \left\{n, \ \lfloor m^2/4 \rfloor + 1\right\} \tag{5.1}$$

holds for all n and $m \leq n$.

Next, we examine groups of prime order. Let p be a positive prime. Trivially, for any subset A and any positive integer h, the h-fold restricted sumset of A is contained in ΣA and, therefore, $\rho\hat{}(\mathbb{Z}_p, m, \mathbb{N}_0)$ cannot be less than $\rho\hat{}(\mathbb{Z}_p, m, \lfloor m/2 \rfloor)$. By Theorem 4,

$$\rho\hat{}(\mathbb{Z}_p, m, \lfloor m/2 \rfloor) = \min \left\{p, \ \lfloor m/2 \rfloor \cdot m - \lfloor m/2 \rfloor^2 + 1\right\} = \min \left\{p, \ \lfloor m^2/4 \rfloor + 1\right\};$$

together with our upper bound (5.1), we arrive at:

Theorem 16 *Conjecture 15 holds for groups of prime order; in particular,*

$$\rho\hat{}(\mathbb{Z}_p, m, \mathbb{N}_0) = \min \left\{p, \ \lfloor m^2/4 \rfloor + 1\right\}$$

for all primes p and $m \leq p$.

There have been some studies of several variations of $\rho\hat{}(G, m, \mathbb{N}_0)$ provided by various restrictions on the subsets A of G. We mention only one pair of such results. Recall that $\text{Asym}(G, m)$ denotes the collection of asymmetrical m-subsets of G; also set $\Sigma^* A = \cup_{h=1}^{\infty} h\hat{}A$. Furthermore, let

$$\rho_{\text{A}}\hat{}(G, m, \mathbb{N}_0) = \min\{|\Sigma A| \mid A \in \text{Asym}(G, m)\},$$

$$\rho_{\text{A}}\hat{}(G, m, \mathbb{N}) = \min\{|\Sigma^* A| \mid A \in \text{Asym}(G, m)\}.$$

With these notations:

Theorem 17 (Balandraud [11–13]) *For every odd prime p and every $m \leq (p-1)/2$, we have*

$$\rho_{\text{A}}\hat{}(\mathbb{Z}_p, m, \mathbb{N}_0) = \min\{p, (m^2 + m)/2 + 1\},$$

$$\rho_{\text{A}}\hat{}(\mathbb{Z}_p, m, \mathbb{N}) = \min\{p, (m^2 + m)/2\}.$$

The fact that the values are upper bounds is provided by the set $\{1, 2, \ldots, m\}$.

6 Critical Numbers

Given G and h, we define the *h-critical number* $\chi(G, h)$ as the least integer m for which $hA = G$ holds for all m-subsets A of G; we define $\chi\hat{}(G, h)$ and $\chi_{\pm}(G, h)$ analogously. We also define the *critical number* $\chi(G, \mathbb{N}_0)$ as the smallest value of m for which $\cup_{h=0}^{\infty} hA = G$ holds for all m-subsets A of G; we define $\chi\hat{}(G, \mathbb{N}_0)$ and $\chi_{\pm}(G, \mathbb{N}_0)$ analogously.

The study of critical numbers originated with the 1964 paper [21] of Erdős and Heilbronn: They studied the variation (in groups of prime order) where only m-subsets of $G \setminus \{0\}$ were considered. (As we now know, the restriction to subsets that do not contain 0 does not change the critical numbers when the number of terms is a fixed value of h but reduces them by 1 when the number of terms is arbitrary; for example, the least integer m for which $hA = G$ holds for all m-subsets A of $G \setminus \{0\}$ equals $\chi(G, h)$, but the least integer m for which $\Sigma A = G$ holds for all m-subsets A of $G \setminus \{0\}$ equals $\chi\hat{}(G, \mathbb{N}_0) - 1$; see [5–8].)

Two of these six quantities are obvious: Since $\cup_{h=0}^{\infty} hA$ and $\cup_{h=0}^{\infty} h_{\pm}A$ are both equal to $\langle A \rangle$, we have

$$\chi(G, \mathbb{N}_0) = \chi_{\pm}(G, \mathbb{N}_0) = n/p + 1,$$

where p is the smallest prime divisor of n. Furthermore, $\chi(G, h)$ and $\chi\hat{}(G, \mathbb{N}_0)$ have now been determined, but the remaining two quantities are not known in general. Let us review what we know.

To state the result for $\chi(G, h)$, we need to introduce the—perhaps already familiar—function

$$v_g(n, h) = \max \{(\lfloor (d - 1 - \gcd(d, g)) / h \rfloor + 1) \cdot n/d \ : \ d \in D(n)\}$$

($n, g, h \in \mathbb{N}$). We should note that this function has appeared elsewhere in additive combinatorics already. For example, according to the classical result of Diananda and Yap [15], the maximum size of a sum-free set (that is, a set A that is disjoint from $2A$) in the cyclic group $\mathbb{Z}_n$ is given by

$$v_1(n, 3) = \begin{cases} (1 + 1/p) \cdot n/3 & \text{if } n \text{ has prime divisors } p \equiv 2 \bmod 3 \\ & \text{and } p \text{ is the smallest,} \\ \lfloor n/3 \rfloor & \text{otherwise.} \end{cases}$$

Similarly, we proved in [3] that the maximum size of a (3, 1)-sum-free set in $\mathbb{Z}_n$ (where A is disjoint from $3A$) equals $v_2(n, 4)$. More generally, $v_{k-l}(n, k + l)$ provides a lower bound for the maximum size of (k, l)-sum-free sets in $\mathbb{Z}_n$ (where $kA \cap lA = \emptyset$ for positive integers $k > l$) (see [3]); equality holds in the case when $k - l$ and n are relatively prime (see the paper [27] of Hamidoune and Plagne). We can now state our result for $\chi(G, h)$:

Theorem 18 (Bajnok [5]) *For all G and h, we have $\chi(G,h) = v_1(n,h)+1$.*

Let us now see what we can say about $\chi\hat{}(G,h)$. First, we can prove that $\chi\hat{}(G,h)$ is well defined, except when $h \in \{2, n-2\}$ and G is isomorphic to an elementary abelian 2-group. Furthermore, for all G with exponent at least 3, we have

$$\chi\hat{}(G,2) = (n + |\mathrm{Ord}(G,2)| + 1)/2 + 1$$

and, as a consequence, when $h \geq (n + |\mathrm{Ord}(G,2)| - 1)/2$, we have $\chi\hat{}(G,h) = h+2$ [5]. Regarding other values of h, few exact results are known; in particular, for $3 \leq h \leq \lfloor n/2 \rfloor - 1$, we only know the value of $\chi\hat{}(\mathbb{Z}_n, h)$ when n is prime or even.

Indeed, for prime values of p, Theorem 4 allows us to derive that

$$\chi\hat{}(\mathbb{Z}_p, h) = \lfloor (p-2)/h \rfloor + h + 1.$$

The case of even n and $h = 3$ was established by Gallardo, Grekos, et al. in [24]; we generalized this in [5] (see also [7]) to prove that that for any h and even $n \geq 12$, we have

$$\chi\hat{}(\mathbb{Z}_n, h) = \begin{cases} n/2+1 \text{ if } 3 \leq h \leq n/2-2; \\ n/2+2 \text{ if } h = n/2-1. \end{cases}$$

Let us take a closer look at the case of $h = 3$. In [5], we proved that if $n \geq 16$ and has prime divisors congruent to 2 mod 3 and p is the smallest such divisor, then

$$\chi\hat{}(\mathbb{Z}_n, 3) \geq \begin{cases} (1+1/p)\cdot n/3 + 3 \text{ if } n = p, \\ (1+1/p)\cdot n/3 + 2 \text{ if } n = 3p, \\ (1+1/p)\cdot n/3 + 1 \text{ otherwise;} \end{cases}$$

and if n has no prime divisors congruent to 2 mod 3, then

$$\chi\hat{}(\mathbb{Z}_n, 3) \geq \begin{cases} \lfloor n/3 \rfloor + 4 \text{ if n is divisible by 9,} \\ \lfloor n/3 \rfloor + 3 \text{ otherwise.} \end{cases}$$

We also believe that, actually, equality holds above for all n—this is certainly the case if n is even or prime; we have verified this (by computer) for all $n \leq 50$. Our conjecture is a generalization of the one made by Gallardo, Grekos, et al. in [24] that was proved (for large n) by Lev in [33].

The study of $\chi\hat{}(G, \mathbb{N}_0)$ posed considerable amount of challenges, but after several decades of attempts, due to the combined results of Diderrich and Mann [18], Diderrich [17], Mann and Wou [35], Dias Da Silva and Hamidoune [16], Gao and Hamidoune [25], Griggs [26], and Freeze et al. [22, 23], we have the value for every group:

Theorem 19 (The combined results of authors above)
Suppose that $n \geq 10$, and let p be the smallest prime divisor of n. Then[1]

$$\chi\hat{}(G, \mathbb{N}_0) = \begin{cases} \lfloor 2\sqrt{n-2} \rfloor + 1 & \textit{if G is cyclic of order n=p or n=pq where} \\ & q \textit{ is prime and } 3 \leq p \leq q \leq p + \lfloor 2\sqrt{p-2} \rfloor + 1, \\ & \textit{(see footnote 1)} \\ n/p + p - 1 & \textit{otherwise.} \end{cases}$$

In closing, we state an intriguing question for the inverse problem regarding $\chi\hat{}(\mathbb{Z}_p, \mathbb{N}_0)$, that is, the attempt to classify all subsets A of size $\chi\hat{}(\mathbb{Z}_p, \mathbb{N}_0) - 1 = \lfloor 2\sqrt{p-2} \rfloor$ in the cyclic group $\mathbb{Z}_p$ of odd prime order p for which $\Sigma A \neq \mathbb{Z}_p$. First, some notations and an observation. Following standard techniques, we identify the elements of $\mathbb{Z}_p$ with integers between $-(p-1)/2$ and $(p-1)/2$ (inclusive); therefore, we can write $A = A_1 \cup A_2$ where A_1 consists of the nonnegative elements of A, and A_2 consists of its negative elements. We define the *norm* of $A \subseteq \mathbb{Z}_p$, denoted by $||A||$, as the sum of the absolute values of its elements, thus

$$||A|| = \Sigma_{a \in A_1} a - \Sigma_{a \in A_2} a.$$

We note that if $||A|| \leq p - 2$, then

$$1 + \Sigma_{a \in A_1} a \notin \Sigma A;$$

in particular, $\Sigma A \neq \mathbb{Z}_p$. Consequently, if $||A|| \leq p - 2$, then $\Sigma(b \cdot A) \neq \mathbb{Z}_p$ for any $b \in \mathbb{Z}_p$. We believe that this simple condition answers our inverse problem for all large enough primes; namely: There is a positive integer p_0 so that if $p > p_0$ is prime and $A \subseteq \mathbb{Z}_p$ has size $\chi\hat{}(\mathbb{Z}_p, \mathbb{N}_0) - 1 = \lfloor 2\sqrt{p-2} \rfloor$, then $\Sigma A \neq \mathbb{Z}_p$ if, and only if, there is a nonzero element $b \in \mathbb{Z}_p$ for which $||b \cdot A|| \leq p - 2$. (We verified that all primes under 40, with the exception of $p = 17$, satisfy this condition.) We mention the following related result:

Theorem 20 (Nguyen et al. [36]) *Let p be an odd prime, and let $A \subseteq \mathbb{Z}_p$ have size $|A| \geq 1.99\sqrt{p}$. If $\Sigma A \neq \mathbb{Z}_p$, then there is a nonzero element $b \in \mathbb{Z}_p$ for which $||b \cdot A|| \leq p + O(\sqrt{p})$.*

References

1. N. Alon, M.B. Nathanson, I. Ruzsa, Adding distinct congruence classes modulo a prime. Am. Math. Monthly **102**, 250–255 (1995)
2. N. Alon, M.B. Nathanson, I. Ruzsa, The polynomial method and restricted sums of congruence classes. J. Number Theory **56**, 404–417 (1996)
3. B. Bajnok, On the maximum size of a (k, l)-sum-free subset of an abelian group. Int. J. Number Theory **5**(6), 953–971 (2009)

[1] Note that $\lfloor 2\sqrt{n-2} \rfloor + 1 = n/p + p$ in this case.

4. B. Bajnok, On the minimum size of restricted sumsets in cyclic groups. Acta Math. Hungar. **148**(1), 228–256 (2016)
5. B. Bajnok, The h-critical number of finite abelian groups. Unif. Distrib. Theory **10**(2), 93–115 (2015)
6. B. Bajnok, More on the h-critical numbers of finite abelian groups. Ann. Univ. Sci. Budapest. Eötvös Sect. Math. **59**, 113–122 (2016)
7. B. Bajnok, Corrigendum to "The h-critical number of finite abelian groups". Unif. Distrib. Theory (to appear)
8. B. Bajnok, S. Edwards, On two questions about restricted sumsets in finite abelian groups. Australas. J. Comb. **68**(2), 229–244 (2017)
9. B. Bajnok, R. Matzke, The minimum size of signed sumsets. *Electron. J. Comb.***22**(2), paper 2.50, 17 (2015)
10. B. Bajnok, R. Matzke, On the minimum size of signed sumsets in elementary abelian groups. J. Number Theory **159**, 384–401 (2016)
11. É. Balandraud, An addition theorem and maximal zero-sum free sets in $\mathbb{Z}/p\mathbb{Z}$. Israel J. Math. **188**, 405–429 (2012)
12. É. Balandraud, Erratum to: "An addition theorem and maximal zero-sum free sets in $\mathbb{Z}/p\mathbb{Z}$". Israel J. Math. **192**(2), 1009–1010 (2012)
13. É. Balandraud, Addition theorems in $\mathbb{F}_p$ via the polynomial method, arXiv:1702.06419v1 (math.CO)
14. A.-L. Cauchy, Recherches sur les nombres. J. École Polytech. **9**, 99–123 (1813)
15. P.H. Diananda, H.P. Yap, Maximal sum-free sets of elements of finite groups. Proc. Jpn. Acad. **45**, 1–5 (1969)
16. J.A. Dias Da Silva, Y.O. Hamidoune, Cyclic space for Grassmann derivatives and additive theory. Bull. London Math. Soc. **26**, 140–146 (1994)
17. G.T. Diderrich, An addition theorem for abelian groups of order pq. J. Number Theory **7**, 33–48 (1975)
18. G.T. Diderrich, H.B. Mann, Combinatorial problems in finite abelian groups. in *A Survey of Combinatorial Theory*, ed. by J.N. Srivastava et al. (North-Holland 1973)
19. S. Eliahou, M. Kervaire, Sumsets in vector spaces over finite fields. J. Number Theory **71**, 12–39 (1998)
20. S. Eliahou, M. Kervaire, Old and new formulas for the Hopf-Stiefel and related functions. Expo. Math. **23**(2), 127–145 (2005)
21. P. Erdős, H. Heilbronn, On the addition of residue classes mod p. Acta Arith. **9**, 149–159 (1964)
22. M. Freeze, W. Gao, A. Geroldinger, The critical number of finite abelian groups. J. Number Theory **129**, 2766–2777 (2009)
23. M. Freeze, W. Gao, A. Geroldinger, Coorigendum to "The critical number of finite abelian groups". J. Number Theory **152**, 205–207 (2015)
24. L. Gallardo, G. Grekos et al., Restricted addition in $\mathbb{Z}\_/n\mathbb{Z}$ and an application to the Erdős–Ginzburg–Ziv problem. J. London Math. Soc. **65**(2), 513–523 (2002)
25. W. Gao, Y.O. Hamidoune, On additive bases. Acta Arith. **88**(3), 233–237 (1999)
26. J.R. Griggs, Spanning subset sums for finite abelian groups. Discret. Math. **229**, 89–99 (2001)
27. Y.O. Hamidoune, A. Plagne, A new critical pair theorem applied to sum-free sets in Abelian groups. Comment. Math. Helv. **79**, 1–25 (2003)
28. Gy. Károlyi, On restricted set addition in abelian groups. Ann. Univ. Sci. Budapest, Eötvös Sect. Math. **46**, 47–54 (2003)
29. Gy. Károlyi, The Erdős–Heilbronn problem in abelian groups. Israel J. of Math. **139**, 349–359 (2004)
30. Gy. Károlyi, A note on the Hopf–Stiefel function. Eur. J. Combin. **27**, 1135–1137 (2006)
31. J.H.B. Kemperman, On small sumsets in an abelian group. Acta Math. **103**, 63–88 (1960)
32. V.F. Lev, Restricted set addition in groups I: the classical setting. J. London Math. Soc.**62**(2), 27–40 (2000)

33. V.F. Lev, Three-fold Restricted Set Addition in Groups. Europ. J. Combinatorics **23**, 613–617 (2002)
34. V.F. Lev, Critical pairs in abelian groups and Kemperman's structure theorem. Int. J. Number Theory **3**, 379–396 (2006)
35. H.B. Mann, Y.F. Wou, Addition theorem for the elementary abelian group of type (p, p). Monatshefte für Math. **102**, 273–308 (1986)
36. N.H. Nguyen, E. Szemerédi, V.H. Vu, Subset sums modulo a prime. Acta Arith. **131**(4), 303–316 (2008)
37. A. Plagne, Additive number theory sheds extra light on the Hopf–Stiefel $\circ$ function. Enseign. Math., II Sér. **49** (1–2), 109–116 (2003)
38. A. Plagne, Optimally small sumsets in groups, I. The supersmall sumset property, the $\mu_G^{(k)}$ and the $\nu_G^{(k)}$ functions. Unif. Distrib. Theory **1**(1), 27–44 (2006)
39. A. Plagne, Optimally small sumsets in groups, II. The hypersmall sumset property and restricted addition. Unif. Distrib. Theory **1**(1), 111–124 (2006)
40. D. Shapiro, Products of sums of squares. Expo. Math. **2**, 235–261 (1984)
41. A.G. Vosper, The critical pairs of subsets of a group of prime order. J. Lond. Math. Soc. **31**, 200–205 (1956)
42. A.G. Vosper, Addendum to "The critical pairs of subsets of a group of prime order". J. Lond. Math. Soc. **31**, 280–282 (1956)

Benford Behavior of Generalized Zeckendorf Decompositions

Andrew Best, Patrick Dynes, Xixi Edelsbrunner, Brian McDonald, Steven J. Miller, Kimsy Tor, Caroline Turnage-Butterbaugh and Madeleine Weinstein

Abstract We prove connections between Zeckendorf decompositions and Benford's law. Recall that if we define the Fibonacci numbers by $F_1 = 1$, $F_2 = 2$, and $F_{n+1} = F_n + F_{n-1}$, every positive integer can be written uniquely as a sum of nonadjacent elements of this sequence; this is called the Zeckendorf decomposition, and similar unique decompositions exist for sequences arising from recurrence relations of the form $G_{n+1} = c_1 G_n + \cdots + c_L G_{n+1-L}$ with c_i positive and some other restrictions. Additionally, a set $S \subset \mathbb{Z}$ is said to satisfy Benford's law base 10 if the density of the elements in S with leading digit d is $\log_{10}(1 + \frac{1}{d})$; in other words, smaller leading digits are more likely to occur. We prove that as $n \to \infty$ for a randomly selected integer m in $[0, G_{n+1})$ the distribution of the leading digits of the summands in its

A. Best
Department of Mathematics, The Ohio State University, Columbus, OH, USA
e-mail: andrewbest312@gmail.com

P. Dynes
Department of Mathematical Sciences, Clemson University, Clemson, SC, USA
e-mail: pdynes@clemson.edu

X. Edelsbrunner · S. J. Miller (✉)
Department of Mathematics and Statistics, Williams College, Williamstown, MA, USA
e-mail: sjm1@williams.edu; Steven.Miller.MC.96@aya.yale.edu

X. Edelsbrunner
e-mail: xe1@williams.edu

B. McDonald
Department of Mathematics, University of Chicago, Chicago, IL, USA
e-mail: bdmcdonald@uchicago.edu

K. Tor
Department of Mathematics, Pierre and Marie Curie University–Paris 6, Paris, France
e-mail: kimsy.tor@gmail.com

C. Turnage-Butterbaugh
Department of Mathematics, Duke University, Durham, NC, USA
e-mail: cturnagebutterbaugh@gmail.com

M. Weinstein
Department of Mathematics, University of California Berkeley, Berkeley, CA, USA
e-mail: maddie@math.berkeley.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_3

generalized Zeckendorf decomposition converges to Benford's law almost surely. Our results hold more generally: One obtains similar theorems to those regarding the distribution of leading digits when considering how often values in sets with density are attained in the summands in the decompositions.

Keywords Zeckendorf decompositions · Fibonacci numbers
Positive linear recurrence relations · Benford's law

MSC 2010: 11B39 · 11B05 · 60F05 (primary)11K06 · 65Q30 · 62E20 (secondary)

1 Introduction

Zeckendorf's theorem states that every positive integer m can be written uniquely as a sum of nonconsecutive Fibonacci numbers, where the Fibonacci numbers are defined by $F_{n+1} = F_n + F_{n-1}$ with $F_1 = 1$ and $F_2 = 2$ (we must re-index the Fibonaccis, as if we included 0 or had two 1s we clearly could not have uniqueness). Such a sum is called the Zeckendorf decomposition of m, and each number in the sum is called a summand. Zeckendorf decompositions have been generalized to many other sequences, specifically those arising from positive linear recurrences. More generally, we can consider a positive linear recurrence sequence given by

$$G_{n+1} \ = \ c_1 G_n + \cdots + c_L G_{n+1-L}, \tag{1}$$

with c_i nonnegative, L, c_1, and c_L positive, as well as rules to specify the first L terms of the sequence and a generalization of the nonadjacency constraint to what is a legal decomposition. Unique decompositions exist both here and for other sequences; see [1, 9–16, 19–21, 24, 25, 28, 29, 31] for a sample of the vast literature on this topic.

Our purpose is to connect generalized Zeckendorf decompositions and Benford's law. In fact, what we show is more general, and the connection with Benford's law follows as a corollary. Still, Benford's law was the motivation for our investigation, so we discuss its history. First discovered by Simon Newcomb [26] in the 1880s, it was rediscovered by Benford [3] approximately fifty years later, who noticed that the distributions of the leading digits of numbers in many data sets were not uniform. In fact, there was a strong bias toward lower values. For example, the leading digit 1 appeared about 30% of the time and the leading digit 9 under 5% of the time. Data sets with such leading digit distributions are said to follow Benford's law. More precisely, the probability of a first digit base B of d is $\log_B(1 + 1/d)$, or more generally the probability that the significand[1] is at most s is $\log_B(s)$. Benford's law appears in astoundingly many data sets, from physical constants to census information to financial and behavioral data and has a variety of applications (two of the most

[1] If $x > 0$ and $B > 1$ we may uniquely write x as $S_B(x) \cdot B^{k_B(x)}$, where $S_B(x) \in [1, B)$ is the significand of x and $k_B(x)$ is an integer.

interesting being its use to detect accounting or voting fraud). This digit bias is in fact quite natural once one realizes that a data set will follow Benford's law if its logarithms modulo 1 are equidistributed.[2] See [4, 17, 18, 23, 27] for more on the theory of Benford's law, as well as the edited volume [22] for a compilation of articles on its theory and applications.

Before exploring whether or not the summands in Zeckendorf decompositions obey Benford's law, it's natural to ask the question about the sequence of Fibonacci numbers themselves. The answer is yes and follows almost immediately from Binet's formula,

$$F_n \;=\; \frac{5+\sqrt{5}}{10}\left(\frac{1+\sqrt{5}}{2}\right)^n+\frac{5-\sqrt{5}}{10}\left(\frac{1-\sqrt{5}}{2}\right)^n \tag{2}$$

(note this is slightly different than the standard expression for Binet's formula as we have re-indexed our sequence so that the Fibonaccis begin 1, 2, 3, 5). The proof is completed by showing the logarithms modulo 1 are equidistributed, which is immediate from the irrationality of $\log_{10}(\frac{1+\sqrt{5}}{2})$ and Kronecker's theorem (if α is irrational, then $n\alpha$ is equidistributed modulo 1) and simple book-keeping to bound the error of the secondary piece; see [12, 23, 30] for details.

Instead of studying Benfordness of summands in Zeckendorf decompositions, we could instead look at other properties of the summands, such as how often we have an even number or how often they are a square modulo B for some fixed B. So long as our sequence has a positive density, our arguments will be applicable.[3] We quickly review this notion. Given a set of positive integers $\mathscr{G}=\{G_n\}_{n=1}^{\infty}$ and a subset $S\subset\mathscr{G}$, we let $q(S,n)$ be the fraction of elements of $\mathscr{G}$ with index at most n that are in S:

$$q(S,n)\;:=\;\frac{\#\{G_i\in S\;:\;1\le i\le n\}}{n}. \tag{3}$$

When $\lim_{n\to\infty} q(S,n)$ exists, we define the **asymptotic density** $q(S)$ as

$$q(S)\;:=\;\lim_{n\to\infty} q(S,n), \tag{4}$$

and for brevity often say the sequence S has **density** $q(S)$.

In an earlier work, we proved that if a set S has a positive density $q(S)$ in the Fibonaccis, then so too do its summands in the Zeckendorf decompositions, and in particular Zeckendorf decompositions using Fibonacci numbers follow Benford's law [5]. Our main result below is generalizing these results to the case of a **positive linear recurrence sequence**, which is a sequence of positive integers $\{G_n\}_{n=1}^{\infty}$ and

[2] Given a data set $\{x_n\}$, let $y_n=\log_{10}x_n \bmod 1$. If $\{y_n\}$ is equidistributed modulo 1 then in the limit the percentage of the time it is in $[\alpha,\beta]\subset[0,1]$ is just $\beta-\alpha$. For example, to restrict to significands of d take $\alpha=\log_{10}d$ and $\beta=\log_{10}(d+1)$.

[3] For example, in the limit one-third of the Fibonacci numbers are even. To see this we look at the sequence modulo 2 and find it is $1,0,1,1,0,1,1,0,1,\ldots$; it is thus periodic with period 3 and one-third of the numbers are even.

a set of nonnegative coefficients $c_1, \ldots, c_L$ with $L, c_1, c_L > 0$,

$$G_{n+1} \;=\; c_1 G_n + c_1 G_{n-1} + \cdots + c_L G_{n+1-L}, \tag{5}$$

and prescribed positive initial terms $G_1, G_2, \ldots, G_L$.

Theorem 1.1 *Fix a positive linear recurrence sequence* $\{G_n\}$. *Let* $S \subseteq \{G_n\}_{n=1}^{\infty}$ *be a set with positive density* $q(S)$*, and fix an* $\varepsilon > 0$. *As* $n \to \infty$*, for an integer* m *selected uniformly at random from* $[0, G_{n+1})$ *the proportion of the summands in* m*'s Zeckendorf decomposition which belongs to* S *is within* ε *of* $q(S)$ *with probability* $1 + o(1)$.

The Benfordness of the summands follows immediately from Theorem 1.1. Let S be the set of numbers in $\{G_n\}_{n=1}^{\infty}$ that start with a given digit. Since G_n is a positive linear recurrence sequence, the density of S in $\{G_n\}_{n=1}^{\infty}$ will follow Benford's law in base B, provided that $\log_B \lambda$ is irrational, where λ is the characteristic polynomial of $\{G_n\}^{\infty}$. If we have a Zeckendorf decomposition with summands from $\{G_n\}_{n=1}^{\infty}$, the proportion of those summands which are in S will also follow Benford's law. We can state this more precisely as follows.

Corollary 1.1 *Fix a positive linear recurrence sequence* $\{G_n\}$. *Let* $S_d \subseteq \{G_n\}_{n=1}^{\infty}$ *be a set of numbers with a given first digit* d. *Then* S *has Benford density (base* B*)* $q(S_d) = log_B(1 + \frac{1}{d})$. *Fix an* $\varepsilon > 0$. *As* $n \to \infty$*, for an integer* m *selected uniformly at random from* $[0, G_{n+1})$ *the proportion of the summands in* m*'s Zeckendorf decomposition which belong to* S_d *is within* ε *of* $q(S_d)$ *with probability* $1 + o(1)$.

We define some concepts needed to prove Theorem 1.1 in Sect. 2, in particular the notion of a super-legal decomposition. We derive some needed properties of these decompositions, and then prove our main result in Sect. 3 by showing related random variables (the number of summands, and the number of summands in our set with positive density in our recurrence sequence) are strongly concentrated.

2 Legal and Super-Legal Decompositions

For the rest of the paper any positive linear recurrence sequence $\{G_n\}_{n=1}^{\infty}$ ***satisfies*** (5) ***with*** $c_i \geq 0$ ***and*** $L, c_1, c_L \geq 1$.

Let $\{G_n\}_{n=1}^{\infty}$ be a positive linear recurrence sequence. Its characteristic polynomial is

$$f(\lambda) \;=\; c_0 + c_1\lambda + \cdots + c_{L-1}\lambda^{L-1} + c_L\lambda^L, \tag{6}$$

with roots $\lambda_1, \ldots, \lambda_L$. By the Generalized Binet Formula, (for a proof see, for example, Appendix A of [2]) we have λ_1 is the unique positive root, $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_L|$, and there exists an $A > 0$ such that

$$G_n \;=\; A\lambda_1^n + O(n^{L-2}\lambda_2^n). \tag{7}$$

We introduce a few important terms needed to state our results.

Definition 2.1 Let $\{G_n\}$ be a positive linear recurrence sequence. Given nonnegative integers $a_1, \dots, a_n$, the sum $\sum_{i=1}^{n} a_i G_{n+1-i}$ is a **legal** Zeckendorf decomposition if one of the following conditions holds.

1. We have $n < L$ and $a_i = c_i$ for $1 \le i \le n$.
2. There exists an $s \in \{1, \dots, L\}$ such that

$$a_1 = c_1, \quad a_2 = c_2, \quad \cdots \quad , \quad a_{s-1} = c_{s-1}, \quad \text{and} \quad a_s < c_s, \tag{8}$$

$a_{s+1}, \dots, a_{s+\ell} = 0$ for some $\ell \ge 0$, and $\{b_i\}_{i=1}^{n-s-\ell}$ with $b_i = a_{s+\ell+i}$ is either legal or empty.

Definition 2.2 Let $\{G_n\}$ be a positive linear recurrence sequence. Given nonnegative integers $a_1, \dots, a_n$, the sum $\sum_{i=1}^{n} a_i G_{n+1-i}$ is a **super-legal** Zeckendorf decomposition if there exists an $s \in \{1, \dots, L\}$ such that

$$a_1 = c_1, \quad a_2 = c_2, \quad \cdots \quad , \quad a_{s-1} = c_{s-1}, \quad \text{and} \quad a_s < c_s, \tag{9}$$

$a_{s+1}, \dots, a_{s+\ell} = 0$ for some $\ell \ge 0$, and $\{b_i\}_{i=1}^{n-s-\ell}$ with $b_i = a_{s+\ell+i}$ is either super-legal or empty.

In other words, a decomposition is super-legal if it satisfies condition (2) of Definition 2.1.

Definition 2.3 Let $\{G_n\}$ be a positive linear recurrence sequence and assume that the sum $\sum_{i=1}^{n} a_i G_{n+1-i}$ is a legal Zeckendorf decomposition. We call each string described by one of the conditions of Definition 2.1 (not counting the additional 0s) a **block** and call the number of terms in each block its **length**.

We note that every super-legal Zeckendorf decomposition is legal and that a concatenation of super-legal Zeckendorf decompositions makes a super-legal Zeckendorf decomposition.

Example 2.1 The recurrence $G_{n+1} = G_n + 2G_{n-1} + 3G_{n-2}$ with $G_1 = 1,\ G_2 = 2,\ G_3 = 5$ produces the sequence $1, 2, 5, 12, 28, 67, 159, 377, \dots$. The decomposition of 858 is

$$858 \;=\; 377 + 2(159) + 2(67) + 28 + 1 \;=\; G_8 + 2G_7 + 2G_6 + G_5 + G_1. \tag{10}$$

This example gives coefficients $(1, 2, 2, 1, 0, 0, 0, 1)$, so the blocks of 858 are $(1, 2, 2)$, $(1, 0)$, and (1), with lengths 3, 2, and 1, respectively. Note that even though the definition of a block excludes the additional 0s (i.e., the $a_{s+1} = a_{s+2} = \cdots = a_{s+\ell} = 0$ from the Definition 2.1), it is still permissible for a block to end with a

0. The decomposition for 858 is legal but not super-legal, since the final block (1) satisfies condition (1) but not condition (2) from Definition 2.1.

Example 2.2 An example of a super-legal decomposition using the recurrence from Example 2.1 is

$$860 \;=\; 377 + 2(159) + 2(67) + 28 + 2 + 1 \;=\; G_8 + 2G_7 + 2G_6 + G_5 + G_2 + G_1, \tag{11}$$

which gives coefficients (1, 2, 2, 1, 0, 0, 1, 1). In this case, the final block is (1, 1), which satisfies the condition of Definition 2.2.

Given two legal decompositions, we do not necessarily obtain a new legal sequence by concatenating the coefficients. However, if we require that the leading block be super-legal, we do obtain a new legal decomposition by concatenation. With the help of a few lemmas which help us count the number of super-legal decompositions, we can circumvent this obstruction.

Lemma 2.1 *Let $\{G_n\}$ be a positive linear recurrence sequence with relation given by (5), and let H_n be the number of super-legal decompositions using only $G_1, G_2, \ldots,$ G_n. We have*

$$H_{n+1} \;=\; c_1 H_n + c_2 H_{n-1} + \cdots + c_L H_{n+1-L}. \tag{12}$$

Proof Note that $H_{n+1} - H_n$ is the number of super-legal decompositions with largest element G_{n+1}. We count how many such decompositions there are by summing over the possible lengths of the leading block. Say the leading block is of length j with $1 < j \le L$. Then the leading block is $(c_1, c_2, \ldots, c_{j-1}, a_j)$, where a_j is chosen from $\{0, 1, \ldots, c_j - 1\}$. Therefore, there are c_j ways of choosing this leading block. Because we require G_{n+1} to be included in the decomposition, if $j = 1$ there are $c_1 - 1$ ways of choosing this leading block, since the leading coefficient must be nonzero. For any choice of leading block of length j, there are H_{n+1-j} ways of choosing the remaining coefficients. Therefore, we find that

$$H_{n+1} - H_n \;=\; \sum_{j=1}^{L} c_j H_{n+1-j} - H_n, \tag{13}$$

completing the proof.

Lemma 2.2 *Let $\{G_n\}$ be a positive linear recurrence sequence, and let H_n be the number of super-legal decompositions using only $G_1, G_2, \ldots, G_n$. Then* $\lim_{n\to\infty} H_n/G_n$ *exists and is positive.*

Proof Since H_n is generated by the same recursion as G_n, it has the same characteristic polynomial, which then has the same roots. Therefore, for some $B \ge 0$, we have

$$H_n \;=\; B\lambda_1^n + O(n^{L-2}\lambda_2^n). \tag{14}$$

Thus, $\lim_{n\to\infty} H_n/G_n = B/A$ and it suffices to show that $B > 0$. Note that we always have $H_j > 0$, so we must have

$$\alpha \;:=\; \min_{1\le j\le L} \frac{H_j}{G_j} \;>\; 0. \tag{15}$$

It follows by induction on n that $H_n \ge \alpha G_n$ for all n. Thus, we conclude that $B > 0$, as desired.

3 Density Theorem

To prove the main result as stated in Theorem 1.1, we compute expected values and variances of the relevant random variables. An essential part of the ensuing analysis is the following estimate on the probability that $a_j = k$ for a fixed k, and showing that it has little dependence on j. We prove the theorem via casework based on the structure of the blocks in the decomposition of m. Namely, in the case that a_j is in the rth position of a block of length ℓ, the two subcases are $r = \ell$ (that is, a_j is the last element in the block) or $r < \ell$ (that is, a_j is not the last element in the block). This is why the notion of a super-legal decomposition is useful; if we want to know whether the legal decomposition $(a_1, a_2, \dots, a_n)$ has a block that terminates at a_r, this is equivalent to whether $(a_1, a_2, \dots, a_r)$ forms a super-legal decomposition. So, we first prove some useful lemmas and then collect our results to prove Theorem 1.1.

Lemma 3.1 *Let $\{G_n\}$ be a positive linear recurrence sequence, and choose an integer m uniformly at random from $[0, G_{n+1})$, with legal decomposition*

$$m \;=\; \sum_{j=1}^{n} a_j G_{n+1-j}. \tag{16}$$

Note that this defines random variables $A_1, \dots, A_n$ taking on values $a_1, \dots, a_n$.

Let $p_{j,k}(n) := \text{Prob}\left(A_j = k\right)$. Then, for $\log n < j < n - \log n$, we have

$$p_{j,k}(n) \;=\; p_k(n)(1 + o(1)), \tag{17}$$

where $p_k(n)$ is computable and does not depend on j.

Proof We divide the argument into cases based on the length of the block containing a_j, as well as the position a_j takes in this block. Suppose that a_j is in the rth place in a block of length ℓ. In order to have $a_j = k$, we must either have $r < \ell$ and $k = c_r$, or $r = \ell$ and $k < c_r$.

In the former case, there are c_ℓ ways to choose the terms in the block containing a_j, due to the c_ℓ choices there are for the final term, and everything else is fixed. There are H_{j-r} ways to choose the coefficients for the terms greater than those in the block containing a_j, and $G_{n-j-\ell+r+1}$ ways to choose the smaller terms.

We now consider the latter case, where $r = \ell$ and $k < c_r$. There is now only one possibility for the coefficients in the block containing a_j, but the rest of the argument remains the same as in the previous case. Therefore, by Lemma 2.2 we find that

$$\begin{aligned} N_{j,k,\ell,r}(n) &:= \#\{m \in \mathbb{Z} \cap [0, G_{n+1}) : a_j = k,\ a_j \text{ in } rth \text{ position in block of length } \ell\} \\ &= \begin{cases} c_\ell G_{n-j-\ell+r+1} H_{j-r} & \text{if } r < \ell,\ k = c_r, \\ G_{n-j-\ell+r+1} H_{j-r} & \text{if } r = \ell,\ k < c_r, \\ 0 & \text{otherwise} \end{cases} \\ &= N_{k,\ell,r}(n)(1 + o(1)), \end{aligned} \tag{18}$$

where

$$N_{k,\ell,r}(n) \ := \ \begin{cases} c_\ell AB\lambda_1^{n-\ell+1} & \text{if } r < \ell,\ k = c_r, \\ AB\lambda_1^{n-\ell+1} & \text{if } r = \ell,\ k < c_r, \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

and $N_{k,\ell,r}(n)$ does not depend on j; these formulas follow from applications of the Generalized Binet Formula to the sequences for the G_n's and H_n's. We conclude the proof by noting that

$$p_{j,k}(n) \ = \ \frac{1}{G_{n+1}} \sum_{\ell=1}^{L} \sum_{r=1}^{\ell} N_{j,k,\ell,r}(n) \ = \ \left(\frac{1}{G_{n+1}} \sum_{\ell=1}^{L} \sum_{r=1}^{\ell} N_{k,\ell,r}(n) \right) \cdot (1 + o(1)), \tag{20}$$

where we used (18) to replace $N_{j,k,\ell,r}(n)$. The claim now follows by defining

$$p_k(n) \ := \ \frac{1}{G_{n+1}} \sum_{\ell=1}^{L} \sum_{r=1}^{\ell} N_{k,\ell,r}(n) \tag{21}$$

and noting that its size is independent of j. More is true, as the Generalized Binet Formula again gives us that G_{n+1} is a constant times λ_1^{n+1} (up to lower order terms), and similarly the sum for $p_k(n)$ is a multiple of λ_1^{n+1} plus lower order terms.

We also use the following result, which follows immediately from Theorems 1.2 and 1.3 in [24] (see also [25] for a survey on the subject).

Lemma 3.2 *Let $\{G_n\}$ be a positive linear recurrence sequence, with $s(m)$ the number of summands in the decomposition of m. That is, for $m = \sum_{j=1}^{n} a_j G_{n+1-j}$, let $s(m) := \sum_{j=1}^{n} a_j$. Let $X_n(m)$ be the random variable defined by $X_n(m) = s(m)$, where m is chosen uniformly at random from* $[0, G_{n+1})$. *Then*

$$\mathbb{E}[X_n(m)] \;=\; nC + o(n) \quad \textit{and} \quad \mathrm{Var}[X_n(m)] \;=\; o(n^2). \tag{22}$$

We define another random variable similarly.

Lemma 3.3 *Let $\{G_n\}$ be a positive linear recurrence sequence, and let $S \subseteq \{G_n\}$ be a set with positive density $q(S)$ in $\{G_n\}$. For m chosen uniformly at random in $[0, G_{n+1})$, let*

$$Y_n(m) \;:=\; \sum_{j \in T_n} a_j, \tag{23}$$

where $T_n = \{j \leq n | G_{n+1-j} \in S\}$.

Then, for some constant $C > 0$, we have

$$\mathbb{E}[Y_n(m)] \;=\; dnC + o(n), \quad \mathrm{Var}[Y_n(m)] \;=\; o(n^2). \tag{24}$$

Proof We first compute the expected value. We have

$$\begin{aligned}
\mathbb{E}[Y_n(m)] = \;& \mathbb{E}\left[\sum_{j \in T_n} a_j\right] \;=\; \sum_{j \in T_n} \mathbb{E}[a_j] \;=\; \sum_{j \in T_n} \sum_{k=1}^{\infty} k p_{j,k}(n) \\
=\;& O(\log n) + \sum_{j \in T_n} \sum_{k=1}^{\infty} k p_k(n)(1 + o(1)). \\
=\;& O(\log n) + dn(1 + o(1)) \sum_{k=1}^{\infty} k p_k(n) \\
=\;& O(\log n) + d(1 + o(1)) \sum_{j=1}^{n} \sum_{k=1}^{\infty} k p_k(n) \\
=\;& O(\log n) + d(1 + o(1)) \sum_{j=1}^{n} \sum_{k=1}^{\infty} k p_{j,k}(n) \\
=\;& O(\log n) + d(1 + o(1)) \sum_{j=1}^{n} \mathbb{E}[a_j] \\
=\;& O(\log n) + \mathbb{E}[X_n(m)] d(1 + o(1)) \\
=\;& dnC + o(n).
\end{aligned} \tag{25}$$

Note that the above sums are actually finite, since $p_{j,k} = p_k = 0$ for sufficiently large k. The $\log n$ term appears since Lemma 3.1 only allows us to say $p_{j,k} = p_k(1 + o(1))$ when $\log n < j < n - \log n$.

We now must consider the variance. First note that if $i + \log n < j$, then letting

$$q_{i,r}(n) \;:=\; \text{Prob (the block containing } a_i \text{ ends at } a_{i+r} | a_i = k) \,, \tag{26}$$

we have

$$\begin{aligned}\text{Prob}\left(a_j=\ell|a_i=k\right) &= \sum_{r=1}^{L-1} q_{i,r}(n)p_{j-i-r,\ell}(n)\\ &= (1+o(1))p_\ell(n)\sum_{r=1}^{L-1} q_{i,r}(n)\\ &= p_\ell(n)(1+o(1)). \end{aligned}\tag{27}$$

Thus, we compute

$$\begin{aligned}\mathbb{E}[Y_n(m)^2] &= \mathbb{E}\left[\sum_{i,j\in T_n} a_ia_j\right] = \sum_{i,j\in T_n}\mathbb{E}[a_ia_j]\\ &= \sum_{i,j\in T_n}\sum_{k,\ell=1}^{\infty} k\ell p_{i,k}(n)\text{Prob}\left(a_j=\ell|a_i=k\right)\\ &= O(n\log n)+2\sum_{\substack{i,j\in T_n\\ 2\log n<i+\log n<j<n-\log n}}\sum_{k,\ell=1}^{\infty} k\ell p_{i,k}(n)\text{Prob}\left(a_j=\ell|a_i=k\right)\\ &\leq O(n\log n)+2\sum_{\substack{i,j\in T_n\\ 2\log n<i+\log n<j<n-\log n}}\sum_{k,\ell=1}^{\infty} k\ell p_k(n)p_\ell(n)(1+o(1))\\ &= O(n\log n)+(1+o(1))d^2n^2\sum_{k,\ell=1}^{\infty}k\ell p_k(n)p_\ell(n)\\ &= O(n\log n)+(1+o(1))d^2n^2\left(\sum_{k=1}^{\infty}kp_k(n)\right)^2\\ &= O(n\log n)+d^2n^2C^2(1+o(1)) = d^2n^2C^2+o(n^2). \end{aligned}\tag{28}$$

Therefore,

$$\text{Var}[Y_n(m)] = \mathbb{E}[Y_n(m)^2]-\mathbb{E}[Y_n(m)]^2 = o(n^2),\tag{29}$$

completing the proof.

We are now ready to prove our main result. The idea of the proof is that the results above strongly concentrate $Y_n(m)$ and $X_n(m)$.

Proof (Proof of Theorem 1.1). Note that the proportion of the summands in m's Zeckendorf decomposition which belong to S is $\frac{Y_n(m)}{X_n(m)}$, where $X_n(m)$, $Y_n(m)$ are defined as in the previous lemmas. Therefore, it suffices to show that for any $\varepsilon>0$, with probability $1+o(1)$ we have

$$\left|\frac{Y_n(m)}{X_n(m)} - d\right| \; < \; \varepsilon. \tag{30}$$

By Chebyshev's inequality, letting $g(n) = n^{1/2}\mathrm{Var}[X_n(m)]^{-1/4}$, we obtain

$$\mathrm{Prob}\left(|X_n(m) - \mathbb{E}[X_n(m)]| > \frac{\mathbb{E}[X_n(m)]}{g(n)}\right) \; \leq \; \frac{\mathrm{Var}[X_n(m)]g(n)^2}{\mathbb{E}[X_n(m)]^2} \; = \; o(1). \tag{31}$$

Letting

$$e_1(n) \; := \; \frac{1}{nC}\left(\frac{\mathbb{E}[X_n(m)]}{g(n)} + |\mathbb{E}[X_n(m)] - nC|\right), \tag{32}$$

we have with probability $1 + o(1)$ that

$$nC(1 - e_1(n)) \; \leq \; X_n(m) \; \leq \; Cn(1 + e_1(n)). \tag{33}$$

Note that $e_1(n) = o(1)$. A similar argument for $Y_n(m)$ shows that there exists some $e_2(n) = o(1)$ such that with probability $1 + o(1)$ we have

$$dnC(1 - e_2(n)) \; \leq \; Y_n(m) \; \leq \; dnC(1 + e_2(n)). \tag{34}$$

Therefore, we have that

$$\frac{Y_n(m)}{X_n(m)} \; \leq \; \frac{dnC(1 + e_2(n))}{nC(1 - e_1(n))} \; < \; d + \varepsilon, \tag{35}$$

with probability $1 + o(1)$, and we can similarly obtain

$$\frac{Y_n(m)}{X_n(m)} \; > \; d - \varepsilon. \tag{36}$$

Thus, we conclude that with probability $1 + o(1)$

$$\left|\frac{Y_n(m)}{X_n(m)} - d\right| \; < \; \varepsilon, \tag{37}$$

completing the proof.

4 Conclusion and Future Work

We were able to handle the behavior of Zeckendorf decompositions in fairly general settings by cleverly separating any decomposition into manageable blocks. The key step was the notion of a super-legal decomposition, which simplified the combinatorial analysis of the generalized Zeckendorf decompositions significantly. This allowed us to prove not just Benford behavior for the leading digits, but also similar results for other sequences with positive density.

We obtained results for a large class of linear recurrences by considering only the main term of Binet's formula for each linear recurrence. In future work, we plan on revisiting these problems for other sequences. Obvious candidates include far-difference representations [1, 11], f-decompositions [10], and recurrences with leading term zero (some of which do not have unique decompositions) [7, 8].

Acknowledgements This research was conducted as part of the 2014 SMALL REU program at Williams College and was supported by NSF grants DMS1265673, DMS1561945, DMS1347804, Williams College, and the Clare Boothe Luce Program of the Henry Luce Foundation. It is a pleasure to thank them for their support, and the participants at SMALL and at the 16th International Conference on Fibonacci Numbers and their Applications for helpful discussions.

References

1. H. Alpert, Differences of multiple fibonacci numbers, Int. Electron. J. Combinat. Num. Theor. **9**, 745–749 (2009)
2. O. Beckwith, A. Bower, L. Gaudet, R. Insoft, S. Li, S.J. Miller, P. Tosteson, The average gap distribution for generalized Zeckendorf decompositions. Fibonacci Quart. **51**, 13–27 (2013)
3. F. Benford, The law of anomalous numbers, Proc. Am. Philos. Soc. **78**, 551–572 (1938), https://www.jstor.org/stable/984802
4. A. Berger, T. Hill, *An Introduction to Benford's Law*, (Princeton University Press, 2015)
5. A. Best, P. Dynes, X. Edelsbrunner, B. McDonald, S.J. Miller, K. Tor, C. Turnage-Butterbaugh, M. Weinstein, Benford behavior of Zeckendorf decompositions. Fibonacci Quart. **52**(5), 35–46 (2014)
6. J. Brown, R. Duncan, Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences, Fibonacci Quart. **8**, 482–486 (1970), https://www.mathstat.dal.ca/FQ/Scanned/8-5/brown.pdf
7. M. Catral, P. Ford, P. Harris, S.J. Miller, D. Nelson, Generalizing Zeckendorf's theorem: the Kentucky sequence. Fibonacci Quart. **52**(5), 68–90 (2014)
8. M. Catral, P. Ford, P. Harris, S.J. Miller, D. Nelson, Legal decompositions arising from non-positive linear recurrences. Fibonacci Quart. **54**(4), 3448–3465 (2016)
9. D.E. Daykin, Representation of natural numbers as sums of generalized fibonacci numbers. J. London Math. Soc. **35**, 143–160 (1960)
10. P. Demontigny, T. Do, A. Kulkarni, S.J. Miller, D. Moon, U. Varma, Generalizing Zeckendorf's theorem to f-decompositions. J. Num. Theor. **141**, 136–158 (2014)
11. P. Demontigny, T. Do, A. Kulkarni, S.J. Miller, U. Varma, A generalization of fibonacci far-difference representations and Gaussian behavior. Fibonacci Quart. **52**(3), 247–273 (2014)
12. M. Drmota, J. Gajdosik, The distribution of the sum-of-digits function. J. Théor. Nombrés Bordeaux **10**(1), 17–32 (1998)

13. J.M. Dumont, A. Thomas, Gaussian asymptotic properties of the sum-of-digits function. J. Num. Theor. **62**(1), 19–38 (1997)
14. P. Filipponi, P.J. Grabner, I. Nemes, A. Pethö, R.F. Tichy, Corrigendum to: generalized Zeckendorf expansions. Appl. Math. Lett. **7**(6), 25–26 (1994)
15. P.J. Grabner, R.F. Tichy, Contributions to digit expansions with respect to linear recurrences. J. Num. Theor. **36**(2), 160–169 (1990)
16. P.J. Grabner, R.F. Tichy, I. Nemes, A. Pethö, Generalized Zeckendorf expansions. Appl. Math. Lett. **7**(2), 25–28 (1994)
17. T. Hill, The first-digit phenomenon. Am. Scient. **86**, 358–363 (1996)
18. T. Hill, A statistical derivation of the significant-digit law. Statist. Sci. **10**, 354–363 (1996)
19. T. J. Keller, Generalizations of Zeckendorf's theorem, Fibonacci Quart. **101**(special issue on representations), 95–102 (1972)
20. M. Kologlu, G. Kopp, S.J. Miller, Y. Wang, On the number of summands in Zeckendorf decompositions. Fibonacci Quart. **49**(2), 116–130 (2011)
21. T. Lengyel, A counting based proof of the generalized Zeckendorf's theorem. Fibonacci Quart. **44**(4), 324–325 (2006)
22. S. J. Miller (ed.), *Benford's Law: Theory and Applications*, (Princeton University Press, 2015)
23. S.J. Miller, R. Takloo-Bighash, *An Invitation to Modern Number Theory* (Princeton University Press, Princeton, NJ, 2006)
24. S.J. Miller, Y. Wang, From fibonacci numbers to central limit type theorems, J. Combinator. Theor. Series A **119**
25. S.J. Miller, Y. Wang, Gaussian Behavior in Generalized Zeckendorf Decompositions, in *Combinatorial and Additive Number Theory, CANT 2011 and 2012* ed. by Melvyn B. Nathanson, Springer Proceedings in Mathematics & Statistics (2014), pp. 159–173, https://arXiv.org/pdf/1107.2718v1
26. S. Newcomb, Note on the frequency of use of the different digits in natural numbers. Amer. J. Math. **4**, 39–40 (1881)
27. R.A. Raimi, The first digit problem. Amer. Math. Monthly **83**(7), 521–538 (1976)
28. W. Steiner, *Parry expansions of polynomial sequences*, Integers 2 (2002), Paper A14
29. W. Steiner, The joint distribution of Greedy and lazy fibonacci expansions. Fibonacci Quart. **43**, 60–69 (2005)
30. L. Washington, Benford's law for fibonacci and Lucas numbers, Fibonacci Quart. **19**(2), 175–177 (1981), http://www.fq.math.ca/Scanned/19-2/washington.pdf
31. E. Zeckendorf, Représentation des nombres naturels par une somme des nombres de Fibonacci ou de nombres de Lucas. Bulletin de la Société Royale des Sciences de Liége **41**, 179–182 (1972)

Ramsey Theory Problems over the Integers: Avoiding Generalized Progressions

Andrew Best, Karen Huan, Nathan McNew, Steven J. Miller, Jasmine Powell, Kimsy Tor and Madeleine Weinstein

Abstract Two well-studied Ramsey-theoretic problems consider subsets of the natural numbers which either contain no three elements in arithmetic progression or in geometric progression. We study generalizations of this problem by varying the kinds of progressions to be avoided and the metrics used to evaluate the density of the resulting subsets. One can view a 3-term arithmetic progression as a sequence $x,\ f_n(x),\ f_n(f_n(x))$, where $f_n(x) = x + n$, n a nonzero integer. Thus, avoiding 3-term arithmetic progressions are equivalent to containing no three elements of the form $x,\ f_n(x),\ f_n(f_n(x))$ with $f_n \in \mathscr{F}_{\mathrm{t}}$, the set of integer translations. One can similarly construct related progressions using different families of functions. We investigate several such families, including geometric progressions ($f_n(x) = nx$ with $n > 1$

This work supported in part by NSF Grants DMS1265673, DMS1561945, DMS1347804 and Williams College.

A. Best
Department of Mathematics, The Ohio State University, Columbus, OH, USA
e-mail: andrewbest312@gmail.com

K. Huan · S. J. Miller (✉)
Department of Mathematics and Statistics, Williams College, Williamstown, MA, USA
e-mail: sjm1@williams.edu; Steven.Miller.MC.96@aya.yale.edu

K. Huan
e-mail: klh1@williams.edu

N. McNew
Department of Mathematics, Towson University, Towson, MD, USA
e-mail: nmcnew@towson.edu

J. Powell
Department of Mathematics, Michigan University, Ann Arbor, MI, USA
e-mail: jtpowell@umich.edu

K. Tor
Department of Mathematics, Pierre and Marie Curie University–Paris 6, Paris, France
e-mail: kimsy.tor@gmail.com

M. Weinstein
University of California Berkeley, Berkeley, CA, USA
e-mail: maddie@math.berkeley.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_4

a natural number) and exponential progressions ($f_n(x) = x^n$). Progression-free sets are often constructed "greedily," including every number so long as it is not in progression with any of the previous elements. Rankin characterized the greedy geometric-progression-free set in terms of the greedy arithmetic set. We characterize the greedy exponential set and prove that it has asymptotic density 1 and then discuss how the optimality of the greedy set depends on the family of functions used to define progressions. Traditionally, the size of a progression-free set is measured using the (upper) asymptotic density; however, we consider several different notions of density, including the uniform and exponential densities.

Keywords Ramsey theory · Progressions

1 Background

A classic Ramsey-theoretic problem is to consider how large a set of integers can be without containing three terms in the set that are in arithmetic progression. In other words, no three integers in the set are of the form $a, a+n, a+2n$. An analogous problem involves looking at sets avoiding 3-term geometric progressions of the form a, na, n^2a. This question was first introduced by Rankin in 1961. More recently, Nathanson and O'Bryant [13] and the third-named author [12] have made further progress toward characterizing such sets and finding bounds on their maximal densities.

Progression-free sets are often constructed "greedily": Starting with an initial included integer, every successive number is included so long as doing so does not create a progression involving any of the previously included elements. We consider two possible generalizations of the greedy arithmetic and geometric-progression-free sets. Let $A_3^* = \{0, 1, 3, 4, 9, 10, \dots\}$ be the greedy set of nonnegative integers free of arithmetic progressions. Note that A_3^* consists of exactly those nonnegative integers with no digit 2in their ternary expansions. Let $G_3^* = \{1, 2, 3, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 17, 19, 21, 22, 23, \dots\}$ be the greedy set of positive integers free of geometric progressions. In 1961, Rankin [15] characterized this set as the set of those integers where all of the exponents in their prime factorization are contained in A_3^*. We will use this characterization below to compute the size of Rankin's set with respect to various densities.

One can view arithmetic and geometric progressions as part of a larger class of functional progressions consisting of three terms of the form $x, f_n(x), f_n(f_n(x))$. From this perspective, a natural generalization of arithmetic and geometric progressions would be to let $f_n(x) = x^n$ and so consider exponential-progression-free sets. We characterize the greedy set in this case, which we call E_3^*. We show that its uniform density is 1 (Theorem 3), and the exponential density of the set of integers excluded from the greedy set E_3^* is 1/4 (Proposition 2).

Additionally, we consider the relationship between G_3^* and A_3^*, namely that the geometric-progression-free set is constructed by taking those numbers with prime exponents in the arithmetic-progression-free set. This leads us to consider iterating this idea so that in each step the permissible set of exponents comes from the prior iteration. We show that the asymptotic densities of the sets are produced in each iteration of this construction approach 1 (Theorem 4) but that each has a lower uniform density of 0 (Theorem 5).

2 Comparing Asymptotic and Uniform Densities

2.1 Definitions

The density most frequently encountered is the asymptotic density, $d(A)$. When the asymptotic density does not exist, the upper asymptotic density $\overline{d}(A)$, and the lower asymptotic density, $\underline{d}(A)$ can be used instead. Their definitions are as follows.

Definition 1 The asymptotic density of a set $A \subseteq \mathbb{N}$, if it exists, is defined to be

$$d(A) = \lim_{N\to\infty} \frac{|A \cap \{1, \ldots, N\}|}{N}. \tag{1}$$

The upper asymptotic density of a set $A \subseteq \mathbb{N}$ is defined to be

$$\overline{d}(A) = \limsup_{N\to\infty} \frac{|A \cap \{1, \ldots, N\}|}{N}, \tag{2}$$

and the lower asymptotic density of a set $A \subseteq \mathbb{N}$ is defined to be

$$\underline{d}(A) = \liminf_{N\to\infty} \frac{|A \cap \{1, \ldots, N\}|}{N}. \tag{3}$$

Using Rankin's characterization of G_3^* in Sect. 1, its asymptotic density can be computed as follows:

$$\begin{aligned} d(G_3^*) &= \prod_p \left(\frac{p-1}{p}\right) \sum_{i\in A_3^*} \frac{1}{p^i} = \prod_p \left(1 - \frac{1}{p}\right) \prod_{i=0}^{\infty} \left(1 + \frac{1}{p^{3^i}}\right) \\ &= \prod_p \left(1 - \frac{1}{p^2}\right) \prod_{i=1}^{\infty} \left(1 + \frac{1}{p^{3^i}}\right) \\ &= \prod_p \left(1 - \frac{1}{p^2}\right) \prod_{i=1}^{\infty} \frac{1 - \frac{1}{p^{2\cdot 3^i}}}{1 - \frac{1}{p^{3^i}}} \\ &= \frac{1}{\zeta(2)} \prod_{i=1}^{\infty} \frac{\zeta(3^i)}{\zeta(2\cdot 3^i)} \approx 0.71974. \end{aligned} \tag{4}$$

Though the asymptotic density is usually the preferred notion of size of a progression-free set when it exists, other types of density can be computed that reveal different information about the size of a set and the spacing of its elements or that are more sensitive in the case of very small or large sets. Another way to measure the "size" of a set is the *uniform density*, also known as Banach density, first described in [2].

Definition 2 The upper uniform density of a set $A \subseteq \mathbb{N}$, if it exists, is defined to be

$$\overline{u}(A) = \lim_{s\to\infty} \max_{n\geq 0} \sum_{a\in A, n<a\leq n+s} \frac{1}{s}, \tag{5}$$

and the lower uniform density of a set $A \subseteq \mathbb{N}$, if it exists, is defined to be

$$\underline{u}(A) = \lim_{s\to\infty} \min_{n\geq 0} \sum_{a\in A, n<a\leq n+s} \frac{1}{s}. \tag{6}$$

The uniform density exists if the upper and lower uniform densities are the same, in which case $u(A) = \overline{u}(A) = \underline{u}(A)$. Intuitively, the uniform density measures how sparse or dense a set can be locally. Notice that uniform density is more sensitive than asymptotic density, specifically to local densities in any interval past the initial interval. This is particularly helpful to us because our sets tend to have increasing gaps between elements. For more information and background on the uniform density see [3, 6, 8]. For any set A of natural numbers, we have (see [8])

$$0 \leq \underline{u}(A) \leq \underline{d}(A) \leq \overline{d}(A) \leq \overline{u}(A) \leq 1. \tag{7}$$

Furthermore, notice that if both the uniform and the asymptotic densities exist, then they are equal. These values can differ substantially, however. It is shown in [11] that for any $0 \leq \alpha \leq \beta \leq \gamma \leq \delta \leq 1$ there exists a set of integers, A, with $\underline{u}(A) = \alpha$, $\underline{d}(A) = \beta$, $\overline{d}(A) = \gamma$, and $\overline{u}(A) = \delta$.

2.2 Sets Free of Geometric Progressions

In [12], a set S is constructed to have high upper asymptotic density as follows. For any fixed N, let

$$\mathbb{S}_N = \left(\frac{N}{48}, \frac{N}{45}\right] \bigcup \left(\frac{N}{40}, \frac{N}{36}\right] \bigcup \left(\frac{N}{32}, \frac{N}{27}\right] \bigcup \left(\frac{N}{24}, \frac{N}{12}\right] \bigcup \left(\frac{N}{9}, \frac{N}{8}\right] \bigcup \left(\frac{N}{4}, N\right]. \tag{8}$$

Now, fix $N_1 = N$, let

$$N_i = \frac{48^2 N_{i-1}^2}{N_1}, \tag{9}$$

and let S be the union of all such S_{N_i}. This set is free of geometric progressions with integral ratios and has upper asymptotic density approximately 0.815509. However, its lower asymptotic density, and therefore its lower uniform density, is 0, and it is readily seen that its upper uniform density is 1, because S contains arbitrarily long stretches of included numbers.

An open problem, stated by Beiglböck et al. [1], asks whether it is possible to find a set of integers free of geometric progressions such that the number of consecutive excluded terms is bounded. (Such a set is sometimes called syndetic). Using a Chinese remainder theorem-type argument, one find that Rankin's greedy set does not have this property. To see this, let p_n denote the n-th prime number and consider the congruences

$$\begin{aligned} a &\equiv p_1^2 \pmod{p_1^3} \\ a+1 &\equiv p_2^2 \pmod{p_2^3} \\ &\vdots \\ a+n-1 &\equiv p_n^2 \pmod{p_n^3}. \end{aligned} \tag{10}$$

By the Chinese remainder theorem, there exists an integer a that satisfies these congruences, so that the n consecutive integers $a, \ldots, a+n-1$ are all excluded from Rankin's greedy set.

The problem above is equivalent to asking whether there exists a set of integers with positive lower uniform density which avoid geometric progressions, which leads us to consider the uniform density of similar sets. This problem has also been considered recently by [10].

2.3 Upper Uniform Density of Rankin's Set

We know the asymptotic density of Rankin's set, G_3^*, as well as its lower uniform density by the argument above. We now consider the upper uniform density of Rankin's set starting with a simple upper bound.

Theorem 1 *An upper bound on the upper uniform density of G_3^* is* $7/8$.

Proof Note that all integers that are exactly divisible by 2^2 are excluded from Rankin's set. That is, all integers that are congruent to 4 mod 8 are excluded. We know that $\overline{u}(\{x : x \not\equiv 4 \bmod 8\}) = 7/8$, and therefore we have that $\overline{u}(G_3^*) \leq 7/8$.

By extending this kind of argument to primes other than 2 and powers greater than 2 which must also be excluded, we are able to determine the exact upper uniform density of this set. Enumerate the primes by $\{p_j\}_{j=1}^{\infty}$ and recall that for any prime p, if any n in our set is exactly divisible by p^k for some k in A_3^* then n is excluded from Rankin's set.

Theorem 2 *The upper uniform density of* G_3^* *equals its asymptotic density:* $\overline{u}(G_3^*) = d(G_3^*)$.

Proof By (7), we know that $d(G_3^*) \leq \overline{u}(G_3^*)$. Thus, to prove our result, it is sufficient to show that $\overline{u}(G_3^*) \leq d(G_3^*)$.

Let

$$T_i \;:=\; \{k \;:\; p_j^b \mid k \Rightarrow p_j^{b+1} \mid k \text{ holds for all } \; j \leq i \; \text{ and } b \leq i, b \notin A_3^*\} \qquad (11)$$

be the set of integers not exactly divisible by any of the first i primes raised to a power (at most i) that is not in A_3^*.

Then, as a first step, notice that the proportion of integers not exactly divisible by p_j^2 in any interval of length p_j^3 is $\left(1 - \frac{1}{p_j^2} + \frac{1}{p_j^3}\right)$. Generalizing this, the proportion of integers not exactly divisible by p_j^b for any $b \leq i$, that is not in A_3^* in any interval of length p_j^{i+1} is

$$1 - \sum_{\substack{0 \leq b \leq i \\ b \notin A_3^*}} \left(\frac{1}{p_j^b} - \frac{1}{p_j^{b+1}} \right). \qquad (12)$$

Thus, by the Chinese remainder theorem, the proportion of integers contained in T_i in any interval of length $\prod_{j=1}^{i} p_j^{i+1}$ is

$$\prod_{j=1}^{i} \left(1 - \sum_{\substack{0 \leq b \leq i \\ b \notin A_3^*}} \left(\frac{1}{p_j^b} - \frac{1}{p_j^{b+1}} \right) \right), \qquad (13)$$

so (13) gives the uniform density of T_i, and thus the upper uniform density as well.

Because $G_3^* \subset T_i$ for each i, we have $\overline{u}(G_3^*) \leq \overline{u}(T_i)$ for each i. Using the expression (13) for $\overline{u}(T_i)$, and letting i go to infinity,

$$\begin{aligned}\overline{u}(G_3^*) \ \le\ \overline{\lim}_{i\to\infty} u(T_i) &= \lim_{i\to\infty}\prod_{j=1}^{i}\left(1-\sum_{\substack{0\le b\le i\\ b\notin A_3^*}}\left(\frac{1}{p_j^b}-\frac{1}{p_j^{b+1}}\right)\right)\\ &= \prod_{j=1}^{\infty}\left(1-\left(1-\frac{1}{p_j}\right)\sum_{b\in\mathbb{N}\setminus A_3^*}\frac{1}{p_j^b}\right)\\ &= \prod_{j=1}^{\infty}\left(1-\frac{1}{p_j}\right)\left[\sum_{b=0}^{\infty}\frac{1}{p_j^b}-\sum_{b\in\mathbb{N}\setminus A_3^*}\frac{1}{p_j^b}\right]\\ &= \prod_{j=1}^{\infty}\left(1-\frac{1}{p_j}\right)\sum_{a\in A_3^*}\frac{1}{p_j^a}=d(G_3^*).\end{aligned} \tag{14}$$

3 Greedy Set Avoiding Exponential Progressions

We can view both a 3-term arithmetic progression and a 3-term geometric progression as a sequence $x,\ f_n(x),\ f_n(f_n(x))$, where $f_n(x)=x+n$ or $f_n(x)=nx$, respectively. We can similarly construct other sequences in terms of different families of functions. We consider first exponential progressions with $f(x)=x^n$.

3.1 Characterization and Density

Let $E_3^*=\{1,2,3,\ldots,14,15,17,\ldots,79,80,82,\ldots\}$ be the greedy set of integers free of exponential progressions, that is, E_3^* avoids progressions of the form x, x^n, x^{n^2}, where x, n are natural numbers greater than 1.

Proposition 1 *An integer $k=p_1^{a_1}p_2^{a_2}\cdots p_n^{a_n}$ is included in E_3^* if and only if $g=\gcd(a_1,a_2,\ldots,a_n)$ is included in G_3^*.*

Proof We proceed by induction on k. Clearly, $1\in E_3^*$. Assume that for all integers less than k, our inductive hypothesis holds, and that $k=p_1^{a_1}p_2^{a_2}\cdots p_n^{a_n}$, with $g=\gcd(a_1,a_2,\ldots,a_n)$.

Suppose first that $g\notin G_3^*$. Since g is excluded from G_3^*, it must be the last term of a geometric progression. Thus, there exists a natural number $r>1$ such that $g/r^2, g/r, g$ is a geometric progression with g/r^2 and g/r both in G_3^*. Setting $b_i=a_i/r$, it is clear that $\gcd(b_i)=g/r$, and by the inductive hypothesis, the number $k_1=p_1^{b_1}p_2^{b_2}\cdots p_n^{b_n}$ is in E_3^*. Similarly, if we set $c_i=a_i/r^2$, it is clear that $\gcd(c_i)=g/r^2$, and by the inductive hypothesis, it follows again that the number $k_0=p_1^{c_1}p_2^{c_2}\cdots p_n^{c_n}$ is in E_3^*. Then, since $k_0^r=k_1$ and $k_1^r=k$, it follows that k_0, k_1, k is an exponential progression, so that k is not in E_3^*.

Now suppose that $k \notin E_3^*$. Since k is excluded from E_3^*, it must be the last term of an exponential progression; thus, there exists a natural number $m > 1$ such that $\sqrt[m^2]{k},\ \sqrt[m]{k},\ k$ is an exponential progression with the first two terms in E_3^*. In particular, since

$$\sqrt[m^2]{k} \;=\; p_1^{\frac{a_1}{m^2}} p_2^{\frac{a_2}{m^2}} \cdots p_n^{\frac{a_n}{m^2}} \;\in E_3^*,$$

we have by the inductive hypothesis that the number

$$g/m^2 \;=\; \gcd\left(\frac{a_1}{m^2}, \frac{a_2}{m^2}, \ldots, \frac{a_n}{m^2}\right)$$

is in G_3^*. Similarly, $g/m \in G_3^*$. Then, since g/m^2, g/m, g is a geometric progression, it follows that g is not in G_3^*.

Throughout the subsequent sections, we will refer to the set of squareful numbers.

Definition 3 An integer n is squareful if, for any prime p dividing n, p^2 also divides n.

Unlike the cases of arithmetic progressions and geometric progressions, where the greedy sets are not necessarily optimal, we find that it is not really possible to do better than E_3^* while avoiding exponential progressions. We see first that E_3^* already has uniform (and asymptotic) density 1.

Theorem 3 *We have* $u(E_3^*) = d(E_3^*) = 1$.

Proof With Eq. (7) in mind, we prove that the uniform density of E_3^* is 1 by showing that the lower uniform density is 1. Equivalently, we show that the upper uniform density of $\mathbb{N} \setminus E_3^*$ is 0. Since $\mathbb{N} \setminus E_3^*$ is a subset of the squareful numbers, it is sufficient to show that the upper uniform density of the squareful numbers is 0, which we do by considering yet another superset, namely, the set of numbers not exactly divisible by the first power of any small primes.

Let

$$R_i \;:=\; \{k \;:\; p_j \mid k \Rightarrow p_j^2 \mid k \text{ holds for all } \; j \le i\} \tag{1}$$

be the set of integers not exactly divisible by any of the first i primes to the first power. Notice that the proportion of integers not exactly divisible by p_j in any interval of length p_j^2 is $\left(1 - \frac{1}{p_j} + \frac{1}{p_j^2}\right)$. Thus, by the Chinese remainder theorem, the proportion of integers contained in R_i in any interval of length $\prod_{j=1}^{i} p_j^2$ is

$$\prod_{j=1}^{i}\left(1 - \frac{1}{p_j} + \frac{1}{p_j^2}\right), \tag{2}$$

so (2) gives the uniform density of R_i, and thus the upper uniform density as well.

Because $\mathbb{N} \setminus E_3^* \subset R_i$ for each i, we have $\overline{u}(\mathbb{N} \setminus E_3^*) \le \overline{u}(R_i)$ for each i. Using the expression (2) for $\overline{u}(R_i)$, and letting i go to infinity,

$$\overline{u}(\mathbb{N} \setminus E_3^*) \;\le\; \lim_{i\to\infty} u(R_i) = \prod_{j=1}^{\infty}\left(1 - \frac{1}{p_j} + \frac{1}{p_j^2}\right) \;=\; 0 \tag{3}$$

Thus, we must have that $\underline{u}(E_3^*) = 1 - \overline{u}(\mathbb{N} \setminus E_3^*) = 1$, and so both the uniform and asymptotic densities of E_3^* are 1.

Because E_3^* has density 1, we focus now on the excluded set of integers, $\mathbb{N} \setminus E_3^*$, which has density 0, and ask whether it is possible to do better, creating a set which avoids exponential progressions while excluding fewer integers. Using the exponential density, which can be used to further analyze sets with density zero, we will see that E_3^* is essentially best possible.

Definition 4 The upper exponential density of a set $A \subseteq \mathbb{N}$ is defined to be

$$\overline{e}(A) \;:=\; \limsup_{n\to\infty} \frac{1}{\log(n)} \log\left(\sum_{a\in A, a\le n} 1\right). \tag{4}$$

The lower exponential density $\underline{e}$ and the exponential density e are defined similarly in the usual way.

Note that the exponential density is defined such that the kth-powers have density $1/k$ and that any set with positive lower asymptotic density will have exponential density 1.

Proposition 2 *The exponential density of the set of integers excluded from the greedy exponential-progression-free set is* $e(\mathbb{N} \setminus E_3^*) = 1/4$.

Proof We first consider exponential progressions, x, x^n, x^{n^2}, in the case when $n = 2$, the smallest nontrivial case. We will see that numbers excluded from this sort of progression form the bulk of the numbers that are excluded.

In the interval $[1, N]$, a first approximation for the number of integers that are excluded from E_3^* is $N^{1/4}$. If $m \le N^{1/4}$, then $m^4 \le N$, and there is a progression of the form m, m^2, m^4. However, not every fourth power is thus excluded. Specifically, if m or m^2 is already excluded from E_3^* then m^4 will not be. For example, $4^4 = 2^8$ would not be excluded even though it is a fourth power, since 2^4 is already excluded.

Because this situation only occurs when the initial term, m, of an exponential progression is already part of a smaller progression, and thus a number, we account for this sort of integer with an error term counting all the squareful numbers less than $N^{1/4}$. The count of the squareful numbers up to M is $O\left(M^{1/2}\right)$, see for example [7], so the number of squareful numbers up to $N^{1/4}$ is $O\left(N^{1/8}\right)$. Thus, simply looking at the exponential progressions where $n = 2$, we exclude $\sqrt[4]{N} + O(\sqrt[8]{N})$ elements from the interval $[1, N]$.

Moreover, for each $n > 2$, we see that the number of excluded terms due to progressions x, x^n, x^{n^2} is $O\left(N^{1/n^2}\right)$ which is smaller than the error term in the expression above.

Finally, we use this to compute the exponential density of $\mathbb{N} \setminus E_3^*$,

$$\begin{aligned} e(\mathbb{N} \setminus E_3^*) &= \lim_{N\to\infty} \frac{\log(\sqrt[4]{N} + O(\sqrt[8]{N}))}{\log N} \\ &= \lim_{N\to\infty} \frac{\log(\sqrt[4]{N}(1 + O(N^{-1/8})))}{\log N} = \frac{1}{4}. \end{aligned} \tag{5}$$

Note that, any set that avoids exponential progressions will have to exclude on the order of $\sqrt[4]{N}$ terms to account for fourth powers, producing a set of excluded integers with exponential density at least 1/4. So we see that in this sense, E_3^* is the optimal set containing no exponential progressions.

4 Excluded Exponent Sets

Another possible way to generalize the sets A_3^* and G_3^* is to iterate the method used to construct G_3^* by taking those numbers whose prime exponents are contained in A_3^*. We can construct a third set of numbers where the admissible prime exponents are the integers in G_3^*. By repeating this pattern, we construct a family of sets.

4.1 Characterization and Density

We obtain the nth set by taking all of the numbers whose primes are raised only to the powers in the $(n-1)$th set. Let S_n be the nth set so constructed, where $S_1 = A_3^*$ and $S_2 = G_3^*$.

4.2 Density of Iterated Construction

We consider the asymptotic densities of sets with this construction, and then we consider the lower uniform densities. First, we define a generalization of the square-free numbers and prove two results useful for our discussion.

Definition 5 Let $x > 0$ be an integer with factorization $p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}$. We say x is *k-free*, if $a_i < k$ for each $1 \le i \le n$.

Lemma 1 *For each* $k \ge 2$, *let* Q_k *be the set of k-free numbers. Then* $\lim_{k\to\infty} d(Q_k) = 1$.

Proof From, for example, Pappalardi [17], we know that

$$S^k(x) \ := \ \#\{n \leq x \mid n \text{ is } k\text{-free}\} \ = \ \frac{x}{\zeta(k)} + O(x^{\frac{1}{k}}), \tag{1}$$

where ζ is the Riemann zeta function. Thus, we have

$$\lim_{k\to\infty} d(Q_k) \ = \ \lim_{k\to\infty} \frac{1}{\zeta(k)} \ = \ 1. \tag{2}$$

Lemma 2 *Let B_m be the set of positive integers whose prime factorizations have at least one prime raised to the mth power. Then, $d(B_m) > 0$ for each $m \geq 2$.*

Proof To compute the density, we rewrite B_m as $Q_{m+1} \setminus Q_m$. Then, since $Q_m \subset Q_{m+1}$, we have

$$d(B_m) \ = \ d(Q_{m+1}) - d(Q_m) \ = \ \frac{1}{\zeta(m+1)} - \frac{1}{\zeta(m)} \ > \ 0, \tag{3}$$

for each $m \geq 2$, as desired.

Theorem 4 *The asymptotic density of S_n approaches* 1 *as n goes to infinity, but there is no n for which the density of S_n equals* 1.

Proof By the definition of S_n for $n > 1$, we have

$$d(S_n) = \prod_p \left(\frac{p-1}{p}\right) \sum_{i\in S_{n-1}} \frac{1}{p^i}. \tag{4}$$

S_n contains as a subset the k-free numbers for some k. As $n \to \infty$, $k \to \infty$ as well. By Lemma 1, we know that as $k \to \infty$, the density of the k-free numbers approaches 1. Thus, we get that $d(S_n) \to \infty$ as $n \to \infty$.

However, in each set, there exists some m such that no numbers whose factorizations have a prime raised to the mth power are included. The set of numbers with at least one prime raised to the mth power has positive density by Lemma 2. Thus, no set in this family has density 1.

Nevertheless, the sets S_n increase in size very quickly. For example, in the fourth iteration of this family, the first element that is excluded is $2^{2^{2^2}} = 65536$. The densities of S_n for the first few values of n are given in Table 1.

Despite the high densities of these sets, they all still miss arbitrarily long sequences of consecutive integers.

Theorem 5 *For each n, S_n has lower uniform density 0.*

Proof Fix $n > 1$, and consider the set S_n. Using the Chinese remainder theorem as in 10, we can construct an arbitrarily long sequence of consecutive numbers all of which are excluded from S_n.

Table 1 Densities of the sets S_n

n	$d(S_n)$
1	0
2	0.719745
3	0.957964
4	0.999992

Let m be a number excluded from S_{n-1}. Then any number with a prime raised to the mth power in its prime factorization is excluded from S_n. We construct a list of l numbers each of which is exactly divisible by some prime raised to the mth power. Take the first l primes, $p_1, \ldots, p_l$ and consider the system of congruences

$$\begin{aligned} a &\equiv p_1^m \pmod{p_1^{m+1}} \\ a+1 &\equiv p_2^m \pmod{p_2^{m+1}} \\ &\vdots \\ a+l-1 &\equiv p_l^m \pmod{p_l^{m+1}}. \end{aligned} \tag{5}$$

By the Chinese remainder theorem, there exists an a that solves this system of congruences, and so the integers $a, a+1, \ldots a+l-1$ are all excluded from S_n.

Note that an argument analogous to that of the proof of Theorem 2 would show that the upper uniform density of S_n is equal to its asymptotic density.

5 Conclusion and Future Work

In addition to calculating the upper uniform density of Rankin's set, we have characterized the greedy set of integers avoiding 3-term exponential progressions and analyzed it using the asymptotic, uniform, and exponential densities. We have also generalized the construction of the set G_3^* and analyzed the densities of the resulting sets.

We end with some additional topics we hope to pursue in later work.

Question 1 *What other functions $f_n(x)$ could we use to define interesting progression-free sets? How does the resulting progression-free set depend on the function?*

Question 2 *Can the sets S_n be characterized as being free of some form of progression or pattern?*

Question 3 *What other notions of density reveal meaningful information about the size of a progression-free set? The multiplicative density, defined by Davenport and*

Erdős [5], might be particularly interesting to consider. How does the use of different measures of density affect the structure of an optimal progression-free set?

Question 4 *One might consider a family of sets where the set after E_3^* extends from E_3^* analogously to how E_3^* extends from G_3^*, that is, an integer $k = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}$ is included in the nth set if and only if $g = \gcd(a_1, a_2, \ldots, a_n)$ is included in the $(n-1)$th set. Can the sets after the first three in this family be characterized as avoiding some meaningful kind of progression?*

Question 5 *What about exponential-progression-free sets over Gaussian integers? Or other number fields? In particular, what can be said about the densities of the sets of ideals which avoid exponential progressions?*

Acknowledgements This research was conducted as part of the 2014 SMALL REU program at Williams College and was supported by NSF grants DMS1265673, DMS1561945, DMS1347804, Williams College, and the Clare Boothe Luce Program of the Henry Luce Foundation. It is a pleasure to thank them for their support.

References

1. V. Bergelson, M. Beiglböck, N. Hindman, D. Strauss, Multiplicative structures in additively large sets, J. Comb. Theor. Ser A **113**, 1219–1242, (2006) http://www.sciencedirect.com/science/article/pii/S0097316505002141
2. T.C. Brown, A.R. Freedman, Arithmetic progressions in lacunary sets. Rocky Mountain J. Math. **17**, 587–596 (1987)
3. T.C. Brown, A.R. Freedman, The Uniform Density of sets of integers and Fermat's Last theorem, C. R. Math. Rep. Acad. Sci. Canada **12**, 1–6 (1990) http://people.math.sfu.ca/~vjungic/tbrown/tom-37.pdf
4. A. Best, K. Huan, N. McNew, S.J. Miller, J. Powell, K. Tor, M. Weinstein, Geometric-progression-free sets over quadratic number fields (2014), arXiv:1412.0999
5. H. Davenport, P. Erdős, On sequences of positive integers. J. Indian Math. Soc. **15**, 19–24 (1951)
6. Z. Gáliková, B. László, T. Salát, Remarks on uniform density of sets of integers, Acta Acad. Paed. Agriensis, Sectio Mathematicae **29**, 3–13 (2002) http://www.kurims.kyoto-u.ac.jp/EMIS/journals/AMI/2002/acta2002-galikova-laszlo-salat.pdf
7. S.W. Golomb, Powerfulnumbers. Amer. Math. Monthly **77**, 848–855 (1970)
8. G. Grekos, On various definitions of density (survey), Tatra Mt. Math. Publ. **31**, 17–27 (2005) http://www.mis.sav.sk/journals/uploads/0131150501GREK02.ps
9. G. Grekos, V. Toma, J. Tomanová, A note on uniform or Banach density, Annales Mathématiques Blaise Pascal. **17**, 153–163 (2010) http://ambp.cedram.org/cedram-bin/article/AMBP_2010__17_1_153_0.pdf
10. X. He, Geometric progression-free sequences with small gaps. J. Num. Theor. **151**, 197–210 (2015)
11. L. Mišík, Sets of positive integers with prescribed values of densities. Math. Slovaca **52**, 289–296 (2002)
12. N. McNew, On sets of integers which contain no three terms in geometric progression, Math. Comp., (Electronically published on May 14, 2015), https://dx.doi.org/10.1090/mcom/2979 (to appear in print)
13. M.B. Nathanson, K. O'Bryant, A problem of Rankin on sets without geometric progressions. Acta Arithmet. **170**, 327–342 (2015)

14. M.B. Nathanson, K. O' Bryant, On sequences without geometric progressions, Integers **13** (2013), Paper No. A73
15. R. A. Rankin, Sets of integers containing not more than a given number of terms in arithmetical progression, Proc. Roy. Soc. Edinburgh Sect. A **65**(1960/61), 332–344 (1960/61)
16. J. Riddell, Sets of integers containing no n terms in geometric progression. Glasgow Math. J. **10**, 137–146 (1969)
17. F. Pappalardi, A Survey on $k-$ Freeness, Number theory, Ramanujan Math. Soc. Lect. Notes Ser., Ramanujan Math. Soc., **1** Mysore, 71–88 (2005)

Recurrence Identities of b-ary Partitions

Dakota Blair

Abstract Solving the b-ary partition problem, counting the number $p_b(n)$ of partitions of n into powers of b, is a pursuit which dates back to Euler. The function $p_b(n)$ satisfies a recurrence, and this note examines a family of identities which can be deduced by iterating the recurrence in a suitable way. These identities can then be used to calculate $p_b(n)$ for large values of n. Further, these identities correspond to generating function identities involving a sequence of polynomials which have suggestive connections to Eulerian polynomials.

Keywords Integer partitions · Partition functions · Recurrence · Congruences
Generating functions · Eulerian polynomials

1 History

A *partition* of a nonnegative integer n is an expression consisting of a sum of positive integers whose value is n. A *b-ary partition* of n is a partition of n where each term in the sum is a power of a base b. Denote the number of partitions of n as $p(n)$ and the number of b-ary partitions[1] as $p_b(n)$. For example, the partitions of 4 are $4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1$, and therefore $p(4) = 5$ and $p_2(4) = 4$. The problem of calculating $p_b(n)$ dates to Euler [1] who first studied $p_2(n)$ in his celebrated 1748 paper *De partitione numerorum*. In 1918, Tanturri [2] examined the $p_2(n)$ problem, stating its recurrence and proving several identities. In that same year, Hardy and Ramanujan [3] published their asymptotic formula for the general partition function $p(n)$. To achieve this, they pioneered the circle method, noting that for the generating function for $p(n)$:

D. Blair (✉)
Graduate Center of the City University of New York, New York, NY, USA
e-mail: dblair@gradcenter.cuny.edu

[1] See Table 1 for values of $p_b(bn)$ for small values of b and n. The expression $p_b(bn)$ is chosen because by Theorem 3.1 the value of $p_b(n)$ is constant on runs of b.

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_5

Table 1 Values of $p_b(n)$ for $2 \leq b \leq 9$ and $1 \leq n \leq 32$

n	$p_2(2n)$	$p_3(3n)$	$p_4(4n)$	$p_5(5n)$	$p_6(6n)$	$p_7(7n)$	$p_8(8n)$	$p_9(9n)$
1	2	2	2	2	2	2	2	2
2	4	3	3	3	3	3	3	3
3	6	5	4	4	4	4	4	4
4	10	7	6	5	5	5	5	5
5	14	9	8	7	6	6	6	6
6	20	12	10	9	8	7	7	7
7	26	15	12	11	10	9	8	8
8	36	18	15	13	12	11	10	9
9	46	23	18	15	14	13	12	11
10	60	28	21	18	16	15	14	13
11	74	33	24	21	18	17	16	15
12	94	40	28	24	21	19	18	17
13	114	47	32	27	24	21	20	19
14	140	54	36	30	27	24	22	21
15	166	63	40	34	30	27	24	23
16	202	72	46	38	33	30	27	25
17	238	81	52	42	36	33	30	27
18	284	93	58	46	40	36	33	30
19	330	105	64	50	44	39	36	33
20	390	117	72	55	48	42	39	36
21	450	132	80	60	52	46	42	39
22	524	147	88	65	56	50	45	42
23	598	162	96	70	60	54	48	45
24	692	180	106	75	65	58	52	48
25	786	198	116	82	70	62	56	51
26	900	216	126	89	75	66	60	54
27	1014	239	136	96	80	70	64	58
28	1154	262	148	103	85	75	68	62
29	1294	285	160	110	90	80	72	66
30	1460	313	172	119	96	85	76	70
31	1626	341	184	128	102	90	80	74
32	1828	369	199	137	108	95	85	78

> Every point of the circle is an essential singularity of the function, and no part of the contour of integration can be deformed in such a manner as to make its contribution obviously negligible. Every element of the contour requires special study; there is no obvious method of writing down a "dominant term."

In a 1940 paper, Mahler [4] established an oft-cited estimate which implies that $p_b(n)$ has intermediate growth, namely

$$\log p_b(n) \sim \frac{(\log n)^2}{2\log b}.$$

Later, in 1948, de Bruijn [5] made use of residue calculations to refine Mahler's work on the asymptotics of $p_b(n)$. Subsequently in 1966, Knuth [6] refined the asymptotic estimates of $p_2(n)$. Churchhouse [7] in 1969 proved theorems regarding congruences of $p_2(n)$ by iterating the recurrence. He also posited a conjecture related to these congruences. Then, Rødseth [8] in 1970 proved Churchhouse's conjecture as well as congruences in the cases where b is a prime. Many later authors adapted Rødseth's method, about which he says:

> The method we use below in proving the above results goes back to Ramanujan, and has been exploited since then by many writers, notably Watson. We use the technique of Atkin and O'Brien.

Building on Rødseth's work, Andrews [9] used generating function algebra to generalize Churchhouse's conjecture to all bases. This year also saw an independent proof of Churchhouse's conjecture by Gupta [10]. Then, in 1972, Gupta [11] proved Churchhouse's result in a simpler way by making use of Kemmer's identity. In 1975, Hirschhorn and Loxton [12] proved several congruences for $p_2(n)$ for n along certain arithmetic progressions. Dirdal [13, 14] also proved congruences for $p_b(n)$ realizing these as limits of congruences of $p_{b,d}(n)$, the number of partitions of n into powers of b repeating each power at most d times. Gupta and Pleasants [15] used Kemmer's identity and matrix methods in 1979 to prove properties of $p_b(n)$ based on the base b expansion of n. Then, in 1990, Reznick [16] proved properties of $p_{2,d}(n)$ while relating them to $p_2(n)$. His terrific bibliography in that paper is an excellent resource on the history of this subject. In a 2011 paper, Rødseth and Sellers [17] gave the problem a fresh look and proved congruences for $p_b(n)$ along the lines of Hirschhorn and Loxton.

2 Notation

Denote the set of integers by $\mathbb{Z}$ and the nonnegative integers by $\mathbb{N}$. Let $p_b(n)$ be the number of b-ary partitions of n, that is, the number of partitions of n whose parts are powers of b with no restriction on how often each power is used. Let $B_b(m,q)$ be the generating function of $p_b(b^m n)$, that is,

$$B_b(m,q) = \sum_{n\in\mathbb{Z}} p_b(b^m n)q^n.$$

Consider a sequence $(a_i)_{i\in I}$. Denote the length of the sequence as $|I|$, and note that a sequence can be interpreted as a $1 \times |I|$ matrix. Given a matrix M, denote its transposition as M^T.

The subsequent notations follow those of Graham et al. [18]. If S is any statement, then let $[[S]]$ denote the Iverson bracket:

$$[[S]] = \begin{cases} 1 \text{ if } S \text{ is true;} \\ 0 \text{ if } S \text{ is false.} \end{cases}$$

Denote the nth falling power of x as $x^{\underline{n}} = x(x-1)(x-2)\cdots(x-n+1)$. Let $\left[{n \atop k}\right]$ and $\left\{{n \atop k}\right\}$ be Stirling numbers of the first and second kind, respectively. In particular, define

$$\begin{bmatrix} n \\ 0 \end{bmatrix} = \begin{Bmatrix} n \\ 0 \end{Bmatrix} = [[n = 0]] \qquad \text{and} \qquad \begin{bmatrix} n \\ k \end{bmatrix} = \begin{Bmatrix} n \\ k \end{Bmatrix} = [[k = 0]]$$

and

$$\begin{aligned} \begin{Bmatrix} n \\ k \end{Bmatrix} &= k\begin{Bmatrix} n-1 \\ k \end{Bmatrix} + \begin{Bmatrix} n-1 \\ k-1 \end{Bmatrix} \\ \begin{bmatrix} n \\ k \end{bmatrix} &= (n-1)\begin{Bmatrix} n-1 \\ k \end{Bmatrix} + \begin{Bmatrix} n-1 \\ k-1 \end{Bmatrix}. \end{aligned}$$

Further, let $\left\langle{n \atop k}\right\rangle$ denote the Eulerian numbers, that is,

$$\left\langle{n \atop k}\right\rangle = \sum_{j=0}^{k} (-1)^j \binom{n+1}{j} (k+1-j)^n.$$

3 The Recurrence

This section concerns itself with proving basic identities for $p_b(n)$.

Theorem 3.1 *The b-ary partition function satisfies the following recurrence:*

$$p_b(n) = 0 \textit{ for } n < 0$$

$$p_b(n) = 1 \textit{ for } 0 \le n < b$$

$$p_b(bn+i) = p_b(bn) \textit{ for } 0 \le i < b \tag{RI}$$

$$p_b(bn) = p_b(bn-1) + p_b(n) \tag{RII}$$

Proof Let $0 \le i < b$. Consider a b-ary partition of $bn + i$. Such a partition must contain at least i copies of 1. Let f be a map which removes i ones from a b-ary partition of $bn + i$, and similarly let g be a map which adds i ones to a b-ary partition of bn. These operations are inverses since removing i ones from a b-ary partition of $bn + i$, and then adding i ones to the result produces the initial b-ary partition, that is, fg is the identity map and therefore f is a bijection. Thus, $p_b(bn) = p_b(bn + i)$ which proves RI.

To see RII, partition the set of b-ary partitions of bn into two sets: those involving ones and those not. For those involving ones, removing a single one will result in a b-ary partition of $bn - 1$, and vice versa. Note that there are $p_b(bn - 1)$ such partitions. For those not, each part is a positive power of b, from which b may be factored out, and the resulting sum will be a b-ary partition of n. Similarly for any b-ary partition of n, multiplying each part by b will result in a b-ary partition of bn. As before, this defines a bijection from b-ary partitions of bn without ones to the b-ary partitions of n. Therefore, the number of b-ary partitions of bn without ones is $p_b(n)$. Consequently, $p_b(bn) = p_b(bn - 1) + p_b(n)$.

The following corollary is the primary way the recurrence for $p_b(n)$ will be used in what is to follow.

Corollary 3.2 *The b-ary partition counting function $p_b(n)$ satisfies the following identity:*

$$p_b(bn) = p_b(bn - b) + p_b(n). \tag{RIII}$$

Proof Combining RI and RII reveals

$$\begin{aligned} p_b(bn) &= p_b(bn - 1) + p_b(n) \text{ by } \textbf{RII} \\ &= p_b(b(n - 1) + b - 1) + p_b(n) \text{ by } \textbf{RI} \\ &= p_b(b(n - 1)) + p_b(n) \end{aligned}$$

and hence the corollary.

4 Generalizations of Tanturri and Churchhouse

The following lemma is a generalization of an identity which goes back to Tanturri.

Lemma 4.1 *The b-ary partition counting function $p_b(n)$ satisfies the following identity:*

$$p_b(bn) = \sum_{k=0}^{n} p_b(n - k)$$

Proof By RIII, $p_b(n) = p_b(bn) - p_b(b(n - 1))$, so

$$\sum_{k=0}^{n} p_b(b(n - k)) - p_b(b(n - 1 - k)) = \sum_{k=0}^{n} p_b(n - k)$$

where the left-hand side is a telescoping sum, leaving

$$p_b(bn) - p_b(-b) = \sum_{k=0}^{n} p_b(n-k)$$

hence

$$p_b(bn) = \sum_{k=0}^{n} p_b(n-k)$$

as desired.

Churchhouse extended this for $b = 2$ to calculate $p_2(2^m n)$. This may be further extended to all b.

Theorem 4.2 *There exist positive integers $C_{b,m}(k)$ such that*

$$p_b(b^m n) = \sum_{k=0}^{n} C_{b,m}(k) p_b(n-k). \qquad (\mathrm{IH}(m))$$

Proof The proof proceeds by induction on m, with the case $m = 1$ being provided by Lemma 4.1. The assertion $\mathrm{IH}(m+1)$ can be shown by assuming $\mathrm{IH}(m)$, applying this to $p_b(b^m(bn))$, separating the first term, reindexing the remaining terms by setting $k = bj - i$, and using RI:

$$\begin{aligned}
p_b(b^{m+1}n) &= p_b(b^m(bn)) \\
&= \sum_{k=0}^{bn} C_{b,m}(k) p_b(bn-k) \\
&= C_{b,m}(0) p_b(bn) + \sum_{k=1}^{bn} C_{b,m}(k) p_b(bn-k) \\
&= C_{b,m}(0) p_b(bn) + \sum_{j=1}^{n} \sum_{i=0}^{b-1} C_{b,m}(bj-i) p_b(bn-bj+i) \\
&= C_{b,m}(0) p_b(bn) + \sum_{j=1}^{n} \sum_{i=0}^{b-1} C_{b,m}(bj-i) p_b(bn-bj).
\end{aligned}$$

Now, applying Lemma 4.1 and reindexing by setting $h = n - j - \ell$ reveals

$$\begin{aligned} p_b(bn - bj) &= p_b(b(n-j)) \\ &= \sum_{\ell=0}^{n-j} p_b(n - j - \ell) \\ &= \sum_{h=0}^{n-j} p_b(h) \end{aligned}$$

and therefore this yields

$$\begin{aligned} p_b(b^{m+1}n) &= C_{b,m}(0)p_b(bn) + \sum_{j=1}^{n}\sum_{i=0}^{b-1} C_{b,m}(bj-i)p_b(bn-bj) \\ &= C_{b,m}(0)p_b(bn) + \sum_{j=1}^{n}\sum_{i=0}^{b-1} C_{b,m}(bj-i)\sum_{h=0}^{n-j} p_b(h). \end{aligned}$$

This sum may be reordered by factoring out the sum indexed by h, extending the range of the sum indexed by h, making the substitution $s = n - h$, interchanging the sums indexed by j and s, limiting the range of the sum indexed by j, and recalling that $k = bj - i$, that is,

$$\begin{aligned} \sum_{j=1}^{n}\sum_{i=0}^{b-1} C_{b,m}(bj-i)\sum_{h=0}^{n-j} p_b(h) &= \sum_{j=1}^{n}\sum_{h=0}^{n-j} p_b(h)\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{j=1}^{n}\sum_{h=0}^{n}[[h \le n-j]]p_b(h)\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{j=1}^{n}\sum_{s=0}^{n}[[n-s \le n-j]]p_b(n-s)\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{j=1}^{n}\sum_{s=0}^{n}[[j \le s]]p_b(n-s)\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{s=0}^{n} p_b(n-s)\sum_{j=1}^{n}[[j \le s]]\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{s=0}^{n} p_b(n-s)\sum_{j=1}^{s}\sum_{i=0}^{b-1} C_{b,m}(bj-i) \\ &= \sum_{s=0}^{n} p_b(n-s)\sum_{k=1}^{sb} C_{b,m}(k). \end{aligned}$$

Finally, the first term may be combined with this sum using Lemma 4.1:

$$\begin{aligned}
p_b(b^{m+1}n) &= C_{b,m}(0)p_b(bn) + \sum_{s=0}^{n} p_b(n-s) \sum_{k=1}^{sb} C_{b,m}(k) \\
&= C_{b,m}(0) \sum_{s=0}^{n} p_b(n-s) + \sum_{s=0}^{n} p_b(n-s) \sum_{k=1}^{sb} C_{b,m}(k) \\
&= \sum_{s=0}^{n} p_b(n-s) \sum_{k=0}^{sb} C_{b,m}(k) \\
&= \sum_{s=0}^{n} \sum_{k=0}^{bs} C_{b,m}(k) p_b(n-s)
\end{aligned}$$

Thus,

$$p_b(b^{m+1}n) = \sum_{s=0}^{n} C_{b,m+1}(s) p_b(n-s)$$

where

$$C_{b,m+1}(s) = \sum_{k=0}^{bs} C_{b,m}(k)$$

proving IH$(m+1)$ and hence the theorem.

The coefficients $C_{b,m}(k)$ are, in fact, more than simply coefficients, and they are indeed polynomials of degree $m-1$.

Theorem 4.3 *The values $C_{b,m}(k)$ are polynomials of degree at most $m-1$ evaluated at k.*

Proof Note that $C_{b,1} = 1$, a degree 0 polynomial in k. By the inductive hypothesis $C_{b,m}(k) = \sum_{i=0}^{m-1} \alpha_{m,i} k^i$, therefore[2]

$$\begin{aligned}
C_{b,m+1}(k) &= \sum_{j=0}^{bk} C_{b,m}(j) \\
&= \sum_{j=0}^{bk} \sum_{i=0}^{m-1} \alpha_{m-1,i}\, j^i \\
&= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{j=0}^{bk} j^i \\
&= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{j=0}^{bk} \sum_{l=0}^{i} \left\{ {i \atop l} \right\} j^{\underline{l}}
\end{aligned}$$

[2] Note that the Stirling numbers $\left[{n \atop k}\right]$ and $\left\{{n \atop k}\right\}$ are defined on page 56.

by an identity in [18, p. 264] which gives powers as a sum of falling powers. Then, by interchanging the order of summation and using the power rule for falling powers (*Ibid.*, p. 50 (2.50)):

$$\begin{aligned} C_{b,m+1}(k) &= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \left\{ {i \atop l} \right\} \sum_{j=0}^{bk} j^{\underline{l}} \\ &= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \left\{ {i \atop l} \right\} \frac{(bk)^{\underline{l+1}}}{l+1}. \end{aligned}$$

Now, using an identity writing falling powers as a sum of powers (*Ibid.*, p. 264), noting that $l + 1 \leq m$ and interchanging the order of summation reveals:

$$\begin{aligned} C_{b,m+1}(k) &= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \frac{1}{l+1} \left\{ {i \atop l} \right\} \sum_{j=0}^{l+1} \left[{l+1 \atop j} \right] (-1)^{l+1-j} (bk)^j \\ &= \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \frac{1}{l+1} \left\{ {i \atop l} \right\} \sum_{j=0}^{l+1} k^j b^j \left[{l+1 \atop j} \right] (-1)^{l+1-j} \\ &= \sum_{j=0}^{m} k^j b^j \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \frac{1}{l+1} \left\{ {i \atop l} \right\} \left[{l+1 \atop j} \right] (-1)^{l+1-j} \end{aligned}$$

Thus, $C_{b,m+1}(k)$ is a polynomial in k of degree at most m with coefficients

$$\alpha_{m+1,j} = b^j \sum_{i=0}^{m-1} \alpha_{m,i} \sum_{l=0}^{i} \frac{1}{l+1} \left\{ {i \atop l} \right\} \left[{l+1 \atop j} \right] (-1)^{l+1-j}$$

concluding the proof.

Recall that the generating function for $p_b(b^m n)$ is $B_b(m, q) = \sum_{n \in \mathbb{Z}} p_b(b^m n) q^n$.

Lemma 4.4 *The generating function for $p_b(bn)$ satisfies the identity:*

$$(1 - q) B_b(1, q) = B_b(0, q)$$

Proof By RIII, $p_b(bn) = p_b(b(n-1)) + p_b(n)$ and therefore $p_b(n) = p_b(bn) - p_b(b(n-1))$, so multiplying by q^n on both sides and summing over all integers n

$$\begin{aligned}\sum_{n\in\mathbb{Z}} p_b(n)q^n &= \sum_{n\in\mathbb{Z}} p_b(bn)q^n - \sum_{n\in\mathbb{Z}} p_b(b(n-1))q^n \\ &= \sum_{n\in\mathbb{Z}} p_b(bn)q^n - q\sum_{n\in\mathbb{Z}} p_b(bn)q^n \\ &= (1-q)\sum_{n\in\mathbb{Z}} p_b(bn)q^n \\ B_b(0,q) &= (1-q)B_b(1,q)\end{aligned}$$

establishing the claim.

5 A Family of Generating Function Identities

This section contains a proof of the main theorem which reveals a family of generating function identities. These identities correspond to a sequence of polynomials which have suggestive connections to Eulerian polynomials.

First, this lemma shows the recurrence may be iterated to express any value of $p_b(n)$ as the sum of multiples of $p_b(b^m)$ for suitable m.

Lemma 5.1 *For all* $n, m \geq 1$, *and* $1 \leq k < b^m$,

$$p_b(b^m n + ub) = p_b(b^m n) + \sum_{k=1}^{u} p_b(b^{m-1}n + k).$$

Proof Let $k = ub + v$ with $0 \leq v < b$. It may be assumed that $v = 0$ because if $v > 0$, then by RI

$$p_b(b^m n + k) = p_b(b^m n + ub + v) = p_b(b^m n + ub).$$

Therefore, applying RIII once, twice, and finally a total of u times iteratively to the leading term, it may be seen that

$$\begin{aligned}p_b(b^m n + ub) &= p_b(b^m n + (u-1)b) + p_b(b^{m-1}n + u) \\ &= p_b(b^m n + (u-2)b) + p_b(b^{m-1}n + u-1) + p_b(b^{m-1}n + u) \\ &= p_b(b^m n + (u-2)b) + \sum_{j=0}^{1} p_b(b^{m-1}n + u - j) \\ p_b(b^m n + ub) &= p_b(b^m n) + \sum_{j=0}^{u-1} p_b(b^{m-1}n + u - j).\end{aligned}$$

Letting $\ell = u - j$ then reveals

$$p_b(b^m n + ub) = p_b(b^m n) + \sum_{\ell=1}^{u} p_b(b^{m-1}n + \ell)$$

concluding the proof.

Lemma 5.2 *For all n and $m \geq 2$,*

$$p_b(b^m n) = p_b(b^m(n-1)) + p_b(b^{m-1}n) + (b-1)p_b(b^{m-1}(n-1)) \\ + [[m > 2]]b \sum_{u=1}^{b^{m-2}-1} p_b(b^{m-1}(n-1) + ub)$$

Proof First apply RIII to $p_b(b^m n)$ to obtain

$$\begin{aligned} p_b(b^m n) &= p_b(b^m n - b) + p_b(b^{m-1}n) \\ &= p_b(b^m(n-1) + b^m - b) + p_b(b^{m-1}n). \end{aligned}$$

Then, Lemma 5.1 may be applied to the first term resulting in

$$p_b(b^m(n-1) + (b^{m-1}-1)b) = p_b(b^m(n-1)) + \sum_{k=1}^{b^{m-1}-1} p_b(b^{m-1}(n-1) + k)$$

which can then be substituted into the previous expression. Then, note that the first $(b-1)$ terms in the sum are identical by RI. When $m = 2$ these are the only terms, but if $m > 2$ there are more terms in the sum which is indicated by the factor $[[m > 2]]$ below.

$$\begin{aligned} p_b(b^m n) &= p_b(b^m(n-1) + b^m - b) + p_b(b^{m-1}n) \\ &= p_b(b^m(n-1)) + p_b(b^{m-1}n) + \sum_{k=1}^{b^{m-1}-1} p_b(b^{m-1}(n-1) + k) \\ &= p_b(b^m(n-1)) + p_b(b^{m-1}n) + (b-1)p_b(b^{m-1}n) \\ &\quad + [[m > 2]] \sum_{k=b}^{b^{m-1}-1} p_b(b^{m-1}(n-1) + k) \end{aligned}$$

When $m > 2$, the summation stratifies by RI:

$$\sum_{k=b}^{b^{m-1}-1} p_b(b^{m-1}(n-1)+k) = \sum_{u=1}^{b^{m-2}-1}\sum_{v=0}^{b-1} p_b(b^{m-1}(n-1)+ub+v)$$
$$= b\sum_{u=1}^{b^{m-2}-1} p_b(b^{m-1}(n-1)+ub)$$

Therefore, in the general case, the expression becomes

$$p_b(b^m n) = p_b(b^m(n-1)) + p_b(b^{m-1}n) + (b-1)p_b(b^{m-1}(n-1))$$
$$+ [[m > 2]]b\sum_{u=1}^{b^{m-2}-1} p_b(b^{m-1}(n-1)+ub)$$

as claimed.

Corollary 5.3 *The generating function for $p_b(b^2n)$ satisfies the identity:*

$$(1-q)^2 B_b(2,q) = (1+(b-1)q)B_b(0,q)$$

Proof When $m = 2$, Lemma 5.2 becomes

$$p_b(b^2n) = p_b(b^2(n-1)) + p_b(bn) + (b-1)p_b(b(n-1))$$

that is,

$$p_b(b^2n) - p_b(b^2(n-1)) = p_b(bn) + (b-1)p_b(b(n-1)).$$

By passing to generating functions, the result is achieved.

$$\sum_{n\in\mathbb{Z}} p_b(b^2n)q^n - \sum_{n\in\mathbb{Z}} p_b(b^2(n-1))q^n = \sum_{n\in\mathbb{Z}} p_b(bn)q^n + \sum_{n\in\mathbb{Z}}(b-1)p_b(b(n-1))q^n$$
$$B_b(2,q) - qB_b(2,q) = B_b(1,q) + (b-1)qB_b(1,q)$$
$$(1-q)B_b(2,q) = (1+(b-1)q)B_b(1,q)$$
$$(1-q)B_b(2,q) = (1+(b-1)q)(1-q)^{-1}B_b(0,q)$$

Therefore,

$$(1+(b-1)q)B_b(0,q) = (1-q)^2 B_b(2,q)$$

as stated.

Lemma 5.4 *For all $n, m \geq 1$, and $1 \leq k < b^m$, there exist polynomials $g_{m,k,j}(x) = g_j(x)$ of degree j with integer coefficients for $0 \leq j \leq m-1$ such that*

$$p_b(b^m n + k) = p_b(b^m n) + \sum_{j=1}^{m-1} g_j(b)p_b(b^j n). \qquad (\mathrm{IH}(m))$$

Proof The proof proceeds by induction on m. For $m = 1$, the induction hypothesis says $p_b(n + k) = p_b(n)$ for $1 \le k < b$ which is true by RI. Assume that IH(m') is true for all $m' < m$. From Lemma 5.1,

$$p_b(b^m n + ub) = p_b(b^m n) + \sum_{k=1}^{u} p_b(b^{m-1}n + k).$$

Then, by the induction hypothesis at $m - 1$,

$$p_b(b^{m-1} + k) = p_b(b^{m-1}n) + \sum_{l=1}^{m-2} g_{m,k,l}(b)p_b(b^l n).$$

Therefore,

$$\begin{aligned} p_b(b^m n + ub) &= p_b(b^m n) + \sum_{k=1}^{u}\left(p_b(b^{m-1}n) + \sum_{l=1}^{m-2} g_{m,k,l}(b)p_b(b^l n)\right) \\ &= p_b(b^m n) + up_b(b^{m-1}n) + \sum_{k=1}^{u}\sum_{l=1}^{m-2} g_{m,k,l}(b)p_b(b^l n). \end{aligned}$$

Finally, switching the order of summation reveals

$$p_b(b^m n + ub) = p_b(b^m n) + up_b(b^{m-1}n) + \sum_{l=1}^{m-2}\left(\sum_{k=1}^{u} g_{m,k,l}(b)\right)p_b(b^l n).$$

Let $w = b^{m-1} - u$, $g_{m-1}(x) = x^{m-1} - w$ and $g_j(x) = \left(\sum_{k=0}^{u-1} g_{k,j}(x)\right)$ for $1 \le j \le m - 2$. Then, $u = b^{m-1} - w$ and $g_{m-1}(b) = b^{m-1} - w = u$, and therefore

$$p_b(b^m n + ub + v) = p_b(b^m n) + \sum_{j=1}^{m-1} g_j(b)p_b(b^j n)$$

as stated.

With this preparation, the main theorem may be proven. This allows the generating function $B_b(m, q)$ to be written in terms of $B_b(0, q)$.

Theorem 5.5 *For all m, there exists a polynomial $f_m(x, q)$ of degree $m - 1$ in q and degree $\binom{m}{2}$ in x such that*

$$f_m(b, q)B_b(0, q) = (1 - q)^m B_b(m, q).$$

Proof The proof proceeds by induction on m. The base case $m = 0$ is trivial, that is, $f_0(x, q) = 1$. Assume that the theorem holds for all $m' < m$. Applying RIII to

$p_b(b^m n)$ yields the following:

$$\begin{aligned} p_b(b^m n) &= p_b(b^m n - b) + p_b(b^{m-1}n) \\ &= p_b(b^m(n-1) + b^m - b) + p_b(b^{m-1}n). \end{aligned}$$

By Lemma 5.4,

$$p_b(b^m(n-1) + b^m - b) = p_b(b^m(n-1)) + \sum_{j=1}^{m-1} g_j(b) p_b(b^j(n-1))$$

and therefore

$$p_b(b^m n) = p_b(b^m(n-1)) + \left(\sum_{j=1}^{m-1} g_j(b) p_b(b^j(n-1)) \right) + p_b(b^{m-1}n)$$

Then, multiplying by q^n on both sides and summing:

$$\begin{aligned} B_b(m,q) &= \sum_{n\in\mathbb{Z}} p_b(b^m n) q^n \\ &= \sum_{n\in\mathbb{Z}} p_b(b^m(n-1)) q^n + \sum_{n\in\mathbb{Z}} p_b(b^{m-1}n) q^n \\ &\quad + \sum_{n\in\mathbb{Z}} \sum_{j=1}^{m-1} g_j(b) p_b(b^j(n-1)) q^n \end{aligned}$$

and by reindexing in sums involving $n-1$ and combining the b^{m-1} terms,

$$\begin{aligned} B_b(m,q) &= q \sum_{n\in\mathbb{Z}} p_b(b^m n) q^n + \sum_{n\in\mathbb{Z}} p_b(b^{m-1}n) q^n + q \sum_{n\in\mathbb{Z}} \sum_{j=1}^{m-1} g_j(b) p_b(b^j n) q^n \\ &= q B_b(m,q) + (1 + g_{m-1}(b) q) \sum_{n\in\mathbb{Z}} p_b(b^{m-1}n) q^n \\ &\quad + q \sum_{j=1}^{m-2} g_j(b) \sum_{n\in\mathbb{Z}} p_b(b^j n) q^n \\ &= q B_b(m,q) + (1 + g_{m-1}(b) q) B_b(m-1,q) + q \sum_{j=1}^{m-2} g_j(b) B_b(j,q) \end{aligned}$$

Therefore,

$$(1-q)B_b(m,q) = (1+g_{m-1}(b)q)B_b(m-1,q) + q\sum_{j=1}^{m-2} g_j(b)B_b(j,q).$$

The induction hypothesis provides

$$(1-q)^j B_b(j,q) = B_b(0,q)f_j(q).$$

that is,

$$B_b(j,q) = \frac{(1-q)^{m-j-1}f_j(q)}{(1-q)^{m-1}}B_b(0,q).$$

Hence, substituting this into the previous sum and multiplying by $(1-q)^{m-1}$ reveals

$$\begin{aligned}(1-q)^m B_b(m,q) &= (1+g_{m-1}(b)q)f_{m-1}(q)B_b(0,q)\\ &\quad + q\sum_{j=1}^{m-2} g_j(b)(1-q)^{m-j-1}f_j(q)B_b(0,q)\\ &= \Bigg((1+g_{m-1}(b)q)f_{m-1}(q)\\ &\quad + q\sum_{j=1}^{m-2} g_j(b)(1-q)^{m-j-1}f_j(q)\Bigg)B_b(0,q).\end{aligned}$$

Consequently,

$$f_m(x,q) = \left((1+g_{m-1}(x)q)f_{m-1}(x,q) + q\sum_{j=1}^{m-2} g_j(x)(1-q)^{m-j-1}f_j(x,q)\right)$$

which is a polynomial of degree $m-1$ in q and degree $\binom{m}{2}$ in x, and therefore

$$f_m(b,q)B_b(0,q) = (1-q)^m B_b(m,q)$$

which proves the theorem.

6 The Polynomial Data

The polynomials in Theorem 5.5 provide a bridge between large values of $p_b(n)$ and its generating function identities. Lacking further theorems, evaluating these large values quickly exceeds the computational power of pencil and paper, but computers

Table 2 Polynomials $f_m(b, q)$ for $2 \leq m \leq 4$ and $2 \leq b \leq \binom{m}{2} + 2$

b	$f_2(b, q)$
2	$1{+}q$
3	$1{+}2q$

b	$f_3(b, q)$
2	$1{+}6q + q^2$
3	$1{+}19q + 7q^2$
4	$1{+}42q + 21q^2$
5	$1{+}78q + 46q^2$

b	$f_4(b, q)$
2	$1{+}31q + 31q^2 + q^3$
3	$1{+}234q + 447q^2 + 47q^3$
4	$1{+}1081q + 2635q^2 + 379q^3$
5	$1{+}3702q + 10218q^2 + 1704q^3$
6	$1{+}10335q + 30735q^2 + 5585q^3$
7	$1{+}24896q + 77801q^2 + 14951q^3$
8	$1{+}53669q + 173747q^2 + 34727q^3$

Table 3 Polynomials $f_5(b, q)$ for $2 \leq b \leq 12$

b	$f_5(b, q)$
2	$1{+}196q + 630q^2 + 196q^3 + q^4$
3	$1{+}5822q + 33504q^2 + 19040q^3 + 682q^4$
4	$1{+}79320q + 561714q^2 + 387600q^3 + 19941q^4$
5	$1{+}642451q + 5055891q^2 + 3835861q^3 + 231421q^4$
6	$1{+}3649340q + 30621390q^2 + 24573740q^3 + 1621705q^4$
7	$1{+}16077981q + 140871555q^2 + 117324441q^3 + 8201271q^4$
8	$1{+}58573732q + 529473294q^2 + 452753140q^3 + 32941657q^4$
9	$1{+}184174970q + 1704597594q^2 + 1486613030q^3 + 111398806q^4$
10	$1{+}515009556q + 4855552326q^2 + 4299866676q^3 + 329571441q^4$
11	$1{+}1308822280q + 12524820930q^2 + 11227696630q^3 + 876084760q^4$
12	$1{+}3072329216q + 29763241530q^2 + 26948358536q^3 + 2133434941q^4$

are ideally suited to calculating these large values. Each $f_m(b, q)$ provides an identity which provides a new way to calculate values of the form $p_b(b^m n)$. Theorem 4.2 provides an alternate way of computing these numbers. The tools used for this work were primarily Python and Sage with double-checking provided by Mathematica. The City University of New York High Performance Computing Center at the College of Staten Island helpfully provided hardware for long-running computations, but with the optimizations provided by Theorems 4.2 and 5.5 retail consumer hardware is capable of calculating $f_m(b, q)$ for high values of m. Tables 2 and 3 contain identities of the form

$$f_m(q) B_b(0, q) = (1 - q)^m \sum_{n \in \mathbb{Z}} p_b(b^m n) q^n$$

for various specific m and b.

Since $f_m(b, q)$ is a polynomial of degree $\binom{m}{2}$ in b, then so is each coefficient in q. Therefore, for a given m, by calculating $f_m(b, q)$ for $\binom{m}{2} + 1$ values of b, it is possible to determine a polynomial in b for each coefficient of q. This data

Table 4 Polynomials $f_m(b, q)$ for $1 \le m \le 4$.

m	$f_m(b, q)$
1	1
2	$bq - q + 1$
3	$\frac{1}{2}b^3q^2 + \frac{1}{2}b^3q - \frac{1}{2}b^2q^2 + \frac{1}{2}b^2q - bq^2 + bq + q^2 - 2q + 1$
4	$\frac{1}{6}b^6q^3 + \frac{2}{3}b^6q^2 - \frac{1}{4}b^5q^3 + \frac{1}{6}b^6q - \frac{1}{6}b^4q^3 + \frac{1}{4}b^5q - \frac{1}{6}b^4q^2 - \frac{1}{4}b^3q^3 + \frac{1}{3}b^4q - \frac{1}{2}b^3q^2$ $+\frac{1}{2}b^2q^3 + \frac{3}{4}b^3q - b^2q^2 + bq^3 + \frac{1}{2}b^2q - 2bq^2 - q^3 + bq + 3q^2 - 3q + 1$

Table 5 Coefficients of $f_5(b, q)$

	q^0	q^1	q^2	q^3	q^4
b^0	1	-4	6	-4	1
b^1	0	1	-3	3	-1
b^2	0	$\frac{1}{2}$	$-\frac{3}{2}$	$\frac{3}{2}$	$-\frac{1}{2}$
b^3	0	$\frac{3}{4}$	$-\frac{5}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
b^4	0	$\frac{11}{24}$	$-\frac{7}{8}$	$\frac{3}{8}$	$\frac{1}{24}$
b^5	0	$\frac{11}{24}$	$-\frac{5}{8}$	$-\frac{1}{8}$	$\frac{7}{24}$
b^6	0	$\frac{3}{8}$	$\frac{5}{24}$	$-\frac{13}{24}$	$-\frac{1}{24}$
b^7	0	$\frac{5}{24}$	$\frac{1}{8}$	$-\frac{3}{8}$	$\frac{1}{24}$
b^8	0	$\frac{1}{8}$	$\frac{5}{24}$	$-\frac{7}{24}$	$-\frac{1}{24}$
b^9	0	$\frac{1}{12}$	$\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{12}$
b^{10}	0	$\frac{1}{24}$	$\frac{11}{24}$	$\frac{11}{24}$	$\frac{1}{24}$

determines $f_m(b, q)$ for a given m and all b. An alternate way of calculating this polynomial is to continue to iterate the recurrence. This method is demonstrated for $m = 3$ in Theorem 10.1. For the case $m = 4$, this approach works, but the argument is significantly longer than the $m = 3$ case.

Table 4 shows $f_m(b, q)$ for $1 \le m \le 4$, written out brutally as polynomials. This representation does not, at first glance, appear particularly illuminating, but it may be the case something may be learned from it. Along these lines, tables of coefficients for the monomials in $f_m(b, q)$ for $5 \le m \le 8$ are presented in Tables 5, 6, 7, and 8. These tables may also be thought of as matrices M_m so that

$$f_m(b, q) = Q_m M_m B_m$$

where $Q_m = ((q^i)_{i=0}^{m-1})^T$ and $B_m = (b^i)_{i=0}^{\binom{m}{2}}$. Perhaps, this representation will suggest a combinatorial interpretation of these coefficients.

The polynomials $f_m(b, q)$ may be seen from an alternate viewpoint as polynomials in b where each coefficient of b is a polynomial in q. This viewpoint (see Tables 9, 10, 11, and 12) proves its usefulness in revealing certain repeating structures. These poly-

Table 6 Coefficients of $f_6(b, q)$

	q^0	q^1	q^2	q^3	q^4	q^5
b^0	1	-5	10	-10	5	-1
b^1	0	1	-4	6	-4	1
b^2	0	$\frac{1}{2}$	-2	3	-2	$\frac{1}{2}$
b^3	0	$\frac{3}{4}$	-2	$\frac{3}{2}$	0	$-\frac{1}{4}$
b^4	0	$\frac{11}{24}$	$-\frac{4}{3}$	$\frac{5}{4}$	$-\frac{1}{3}$	$-\frac{1}{24}$
b^5	0	$\frac{25}{48}$	$-\frac{4}{3}$	$\frac{7}{8}$	$\frac{1}{6}$	$-\frac{11}{48}$
b^6	0	$\frac{1}{2}$	$-\frac{13}{24}$	$-\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{24}$
b^7	0	$\frac{13}{36}$	$-\frac{35}{72}$	$-\frac{5}{24}$	$\frac{31}{72}$	$-\frac{7}{72}$
b^8	0	$\frac{7}{24}$	$-\frac{1}{8}$	$-\frac{5}{8}$	$\frac{11}{24}$	0
b^9	0	$\frac{11}{48}$	$\frac{1}{12}$	$-\frac{19}{24}$	$\frac{5}{12}$	$\frac{1}{16}$
b^{10}	0	$\frac{1}{6}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{6}$	0
b^{11}	0	$\frac{23}{240}$	$\frac{17}{60}$	$-\frac{41}{120}$	$-\frac{1}{20}$	$\frac{1}{80}$
b^{12}	0	$\frac{1}{16}$	$\frac{7}{24}$	$-\frac{1}{4}$	$-\frac{1}{8}$	$\frac{1}{48}$
b^{13}	0	$\frac{5}{144}$	$\frac{17}{72}$	$-\frac{1}{12}$	$-\frac{13}{72}$	$-\frac{1}{144}$
b^{14}	0	$\frac{1}{48}$	$\frac{5}{24}$	0	$-\frac{5}{24}$	$-\frac{1}{48}$
b^{15}	0	$\frac{1}{120}$	$\frac{13}{60}$	$\frac{11}{20}$	$\frac{13}{60}$	$\frac{1}{120}$

nomials have been calculated for values of m up to 23, and unfortunately these pages are unable to contain them. Or, with apologies to Fermat, "Hanc *paginis* exiguitas non caperet." Fortunately, this data is available for download at the following URL: http://dakota.tensen.net/2015/rp/

The form of these polynomials suggests a conjecture containing an unexpected appearance of Eulerian numbers:

Conjecture 6.1 *The polynomial $f_m(b, q)$ has the form*

$$f_m(b, q) = \sum_{i=0}^{\binom{m}{2}} (1 - q)^{m-y(i)} g_{m,i}(q) b^i$$

where $y(n) = \left\lfloor \frac{\sqrt{8n+1}}{2} \right\rfloor$ and $g_{m,i}(q)$ are polynomials. Further, with $\left\langle {n \atop k} \right\rangle$ denoting the Eulerian numbers[3]*:*

$$g_{m,\binom{m}{2}}(q) = \frac{q}{(m-1)!} \sum_{i=0}^{m-2} \left\langle {m-1 \atop i} \right\rangle q^i$$

[3] Note that the Eulerian numbers $\left\langle {n \atop k} \right\rangle$ are defined on page 56.

Table 7 Coefficients of $f_7(b, q)$

	q^0	q^1	q^2	q^3	q^4	q^5	q^6
b^0	1	-6	15	-20	15	-6	1
b^1	0	1	-5	10	-10	5	-1
b^2	0	$\frac{1}{2}$	$-\frac{5}{2}$	5	-5	$\frac{5}{2}$	$-\frac{1}{2}$
b^3	0	$\frac{3}{4}$	$-\frac{11}{4}$	$\frac{7}{2}$	$-\frac{3}{2}$	$-\frac{1}{4}$	$\frac{1}{4}$
b^4	0	$\frac{11}{24}$	$-\frac{43}{24}$	$\frac{31}{12}$	$-\frac{19}{12}$	$\frac{7}{24}$	$\frac{1}{24}$
b^5	0	$\frac{25}{48}$	$-\frac{89}{48}$	$\frac{53}{24}$	$-\frac{17}{24}$	$-\frac{19}{48}$	$\frac{11}{48}$
b^6	0	$\frac{17}{32}$	$-\frac{115}{96}$	$\frac{23}{48}$	$\frac{7}{16}$	$-\frac{17}{96}$	$-\frac{7}{96}$
b^7	0	$\frac{125}{288}$	$-\frac{331}{288}$	$\frac{109}{144}$	$\frac{41}{144}$	$-\frac{119}{288}$	$\frac{25}{288}$
b^8	0	$\frac{19}{48}$	$-\frac{13}{16}$	$\frac{1}{24}$	$\frac{19}{24}$	$-\frac{7}{16}$	$\frac{1}{48}$
b^9	0	$\frac{17}{48}$	$-\frac{71}{144}$	$-\frac{11}{18}$	$\frac{5}{4}$	$-\frac{67}{144}$	$-\frac{5}{144}$
b^{10}	0	$\frac{7}{24}$	$\frac{1}{16}$	$-\frac{23}{24}$	$\frac{7}{12}$	0	$\frac{1}{48}$
b^{11}	0	$\frac{313}{1440}$	$\frac{1}{32}$	$-\frac{127}{144}$	$\frac{113}{144}$	$-\frac{13}{96}$	$-\frac{23}{1440}$
b^{12}	0	$\frac{31}{180}$	$\frac{35}{144}$	$-\frac{71}{72}$	$\frac{19}{36}$	$\frac{5}{72}$	$-\frac{19}{720}$
b^{13}	0	$\frac{61}{480}$	$\frac{151}{480}$	$-\frac{69}{80}$	$\frac{21}{80}$	$\frac{27}{160}$	$-\frac{1}{96}$
b^{14}	0	$\frac{67}{720}$	$\frac{7}{18}$	$-\frac{53}{72}$	$-\frac{1}{18}$	$\frac{43}{144}$	$\frac{1}{90}$
b^{15}	0	$\frac{91}{1440}$	$\frac{43}{96}$	$-\frac{19}{144}$	$-\frac{49}{144}$	$-\frac{1}{32}$	$-\frac{11}{1440}$
b^{16}	0	$\frac{7}{180}$	$\frac{11}{36}$	$-\frac{11}{72}$	$-\frac{19}{72}$	$\frac{5}{72}$	$\frac{1}{360}$
b^{17}	0	$\frac{7}{288}$	$\frac{343}{1440}$	$-\frac{7}{240}$	$-\frac{181}{720}$	$\frac{23}{1440}$	$\frac{1}{480}$
b^{18}	0	$\frac{7}{480}$	$\frac{19}{96}$	$\frac{1}{16}$	$-\frac{13}{48}$	$-\frac{1}{96}$	$\frac{1}{160}$
b^{19}	0	$\frac{11}{1440}$	$\frac{13}{96}$	$\frac{19}{144}$	$-\frac{29}{144}$	$-\frac{7}{96}$	$-\frac{1}{1440}$
b^{20}	0	$\frac{1}{240}$	$\frac{5}{48}$	$\frac{1}{6}$	$-\frac{1}{6}$	$-\frac{5}{48}$	$-\frac{1}{240}$
b^{21}	0	$\frac{1}{720}$	$\frac{19}{240}$	$\frac{151}{360}$	$\frac{151}{360}$	$\frac{19}{240}$	$\frac{1}{720}$

Another conjecture also suggests itself:

Conjecture 6.2 *Let $h_i(q)$ be defined by $qh_i(q) = g_{i+1,i}(q)$. Then,*

$$\begin{aligned} f_m(b, q) = {} & (1-q)^{m-1} \\ & + q\sum_{i=0}^{m-1}(1-q)^{m-y(i)}h_i(q)b^i \\ & + \sum_{i=m}^{\binom{m}{2}-1}(1-q)^{m-y(i)}g_{m,i}(q)b^i \\ & + \frac{qb^{\binom{m}{2}}}{(m-1)!}\sum_{i=0}^{m-2}\left\langle {m-1 \atop i} \right\rangle q^i. \end{aligned}$$

Table 8 Coefficients of $f_8(b,q)$

	q^0	q^1	q^2	q^3	q^4	q^5	q^6	q^7
b^0	1	-7	21	-35	35	-21	7	-1
b^1	0	1	-6	15	-20	15	-6	1
b^2	0	$\frac{1}{2}$	-3	$\frac{15}{2}$	-10	$\frac{15}{2}$	-3	$\frac{1}{2}$
b^3	0	$\frac{3}{4}$	$-\frac{7}{2}$	$\frac{25}{4}$	-5	$\frac{5}{4}$	$\frac{1}{2}$	$-\frac{1}{4}$
b^4	0	$\frac{11}{24}$	$-\frac{9}{4}$	$\frac{35}{8}$	$-\frac{25}{6}$	$\frac{15}{8}$	$-\frac{1}{4}$	$-\frac{1}{24}$
b^5	0	$\frac{25}{48}$	$-\frac{19}{8}$	$\frac{65}{16}$	$-\frac{35}{12}$	$\frac{5}{16}$	$\frac{5}{8}$	$-\frac{11}{48}$
b^6	0	$\frac{17}{32}$	$-\frac{83}{48}$	$\frac{161}{96}$	$-\frac{1}{24}$	$-\frac{59}{96}$	$\frac{5}{48}$	$\frac{7}{96}$
b^7	0	$\frac{259}{576}$	$-\frac{161}{96}$	$\frac{137}{64}$	$-\frac{113}{144}$	$-\frac{89}{192}$	$\frac{13}{32}$	$-\frac{41}{576}$
b^8	0	$\frac{7}{16}$	$-\frac{137}{96}$	$\frac{127}{96}$	$\frac{11}{48}$	$-\frac{11}{12}$	$\frac{35}{96}$	$-\frac{1}{96}$
b^9	0	$\frac{27}{64}$	$-\frac{85}{72}$	$\frac{307}{576}$	$\frac{89}{72}$	$-\frac{823}{576}$	$\frac{7}{18}$	$\frac{17}{576}$
b^{10}	0	$\frac{329}{864}$	$-\frac{43}{72}$	$-\frac{131}{288}$	$\frac{127}{108}$	$-\frac{155}{288}$	$\frac{5}{72}$	$-\frac{31}{864}$
b^{11}	0	$\frac{911}{2880}$	$-\frac{389}{720}$	$-\frac{289}{576}$	$\frac{19}{12}$	$-\frac{623}{576}$	$\frac{161}{720}$	$\frac{1}{2880}$
b^{12}	0	$\frac{1189}{4320}$	$-\frac{323}{1440}$	$-\frac{13}{12}$	$\frac{787}{432}$	$-\frac{245}{288}$	$\frac{19}{480}$	$\frac{47}{2160}$
b^{13}	0	$\frac{329}{1440}$	$-\frac{1}{240}$	$-\frac{125}{96}$	$\frac{59}{36}$	$-\frac{47}{96}$	$-\frac{19}{240}$	$\frac{11}{1440}$
b^{14}	0	$\frac{325}{1728}$	$\frac{293}{1440}$	$-\frac{859}{576}$	$\frac{595}{432}$	$-\frac{23}{576}$	$-\frac{67}{288}$	$-\frac{43}{8640}$
b^{15}	0	$\frac{427}{2880}$	$\frac{281}{720}$	$-\frac{641}{576}$	$\frac{1}{2}$	$\frac{29}{576}$	$\frac{7}{720}$	$\frac{41}{2880}$
b^{16}	0	$\frac{61}{540}$	$\frac{521}{1440}$	$-\frac{1607}{1440}$	$\frac{1177}{2160}$	$\frac{179}{720}$	$-\frac{223}{1440}$	$\frac{11}{4320}$
b^{17}	0	$\frac{31}{360}$	$\frac{289}{720}$	$-\frac{127}{144}$	$\frac{1}{8}$	$\frac{13}{36}$	$-\frac{67}{720}$	$\frac{1}{720}$
b^{18}	0	$\frac{23}{360}$	$\frac{587}{1440}$	$-\frac{199}{288}$	$-\frac{25}{144}$	$\frac{67}{144}$	$-\frac{97}{1440}$	$-\frac{7}{1440}$
b^{19}	0	$\frac{131}{2880}$	$\frac{109}{288}$	$-\frac{247}{576}$	$-\frac{7}{16}$	$\frac{241}{576}$	$\frac{37}{1440}$	$-\frac{1}{576}$
b^{20}	0	$\frac{277}{8640}$	$\frac{11}{32}$	$-\frac{679}{2880}$	$-\frac{1283}{2160}$	$\frac{347}{960}$	$\frac{131}{1440}$	$\frac{13}{8640}$
b^{21}	0	$\frac{61}{2880}$	$\frac{5}{16}$	$\frac{13}{64}$	$-\frac{4}{9}$	$-\frac{5}{64}$	$-\frac{1}{80}$	$-\frac{1}{576}$
b^{22}	0	$\frac{269}{20160}$	$\frac{107}{504}$	$\frac{401}{4032}$	$-\frac{187}{504}$	$\frac{107}{4032}$	$\frac{13}{630}$	$-\frac{5}{4032}$
b^{23}	0	$\frac{1}{120}$	$\frac{1}{6}$	$\frac{1}{6}$	$-\frac{1}{3}$	$-\frac{1}{24}$	$\frac{1}{30}$	0
b^{24}	0	$\frac{43}{8640}$	$\frac{43}{360}$	$\frac{521}{2880}$	$-\frac{61}{270}$	$-\frac{269}{2880}$	$\frac{1}{72}$	$\frac{1}{8640}$
b^{25}	0	$\frac{1}{360}$	$\frac{4}{45}$	$\frac{29}{144}$	$-\frac{1}{6}$	$-\frac{5}{36}$	$\frac{1}{90}$	$\frac{1}{720}$
b^{26}	0	$\frac{1}{720}$	$\frac{1}{18}$	$\frac{13}{72}$	$-\frac{1}{18}$	$-\frac{23}{144}$	$-\frac{1}{45}$	0
b^{27}	0	$\frac{1}{1440}$	$\frac{7}{180}$	$\frac{49}{288}$	0	$-\frac{49}{288}$	$-\frac{7}{180}$	$-\frac{1}{1440}$
b^{28}	0	$\frac{1}{5040}$	$\frac{1}{42}$	$\frac{397}{1680}$	$\frac{151}{315}$	$\frac{397}{1680}$	$\frac{1}{42}$	$\frac{1}{5040}$

Table 9 Polynomial $f_4(b, q)$ as a polynomial in b

	$f_4(b,q) =$		$(1-q)^3$		
$+$		q	$(1-q)^2$		b
$+$	$(2)^{-1}$	q	$(1-q)^2$		b^2
$+$	$(4)^{-1}$	q	$(1-q)$	$(3+q)$	b^3
$+$	$(6)^{-1}$	q	$(1-q)$	$(2+q)$	b^4
$+$	$(4)^{-1}$	q	$(1-q)$	$(1+q)$	b^5
$+$	$(6)^{-1}$	q	$\cdot$	$(1+4q+q^2)$	b^6

Table 10 Polynomial $f_5(b, q)$ as a polynomial in b

	$f_5(b,q) =$		$(1-q)^4$		
$+$		q	$(1-q)^3$		b
$+$	$(2)^{-1}$	q	$(1-q)^3$		b^2
$+$	$(4)^{-1}$	q	$(1-q)^2$	$(3+q)$	b^3
$+$	$(24)^{-1}$	q	$(1-q)^2$	$(11+q)$	b^4
$+$	$(24)^{-1}$	q	$(1-q)^2$	$(11+7q)$	b^5
$+$	$(24)^{-1}$	q	$(1-q)$	$(9+14q+q^2)$	b^6
$+$	$(24)^{-1}$	q	$(1-q)$	$(5+8q-q^2)$	b^7
$+$	$(24)^{-1}$	q	$(1-q)$	$(3+8q+q^2)$	b^8
$+$	$(12)^{-1}$	q	$(1-q)$	$(1+4q+q^2)$	b^9
$+$	$(24)^{-1}$	q	$\cdot$	$(1+11q+11q^2+q^3)$	b^{10}

Assuming these conjectures indicates that the polynomials $f_b(m, q)$ are completely determined by the polynomials $(g_{m,i})_{i=m}^{\binom{m}{2}-1}$. The data so far obeys this conjecture, so these polynomials are given for $8 \le m \le 10$ (in Tables 13, 14, 15, 16, 17, 18) from which, by using the form above, one may construct $f_m(b, q)$.

7 A New Congruence

The conjectured form of $f_m(b, q)$ suggests several conjectures, including some regarding congruences of $p_b(n)$. For instance, $f_m(b, q) \equiv (1-q)^{m-1} \pmod{b}$ seems likely, and therefore Theorem 5.5 suggests

$$(1-q)^m B_b(m, q) \equiv (1-q)^{m-1} B_b(0, q) \pmod{b}$$

and hence

$$(1-q) B_b(m, q) \equiv B_b(0, q) \pmod{b}$$

Table 11 Polynomial $f_6(b, q)$ as a polynomial in b

$$
\begin{array}{rrlll}
 & f_6(b,q) = & (1-q)^5 & & \\
+ & q & (1-q)^4 & b & \\
+ & (2)^{-1}\,q & (1-q)^4 & b^2 & \\
+ & (4)^{-1}\,q & (1-q)^3 & (3+q)\,b^3 & \\
+ & (24)^{-1}\,q & (1-q)^3 & (11+q)\,b^4 & \\
+ & (48)^{-1}\,q & (1-q)^3 & (25+11q)\,b^5 & \\
+ & (24)^{-1}\,q & (1-q)^2 & (12+11q+q^2)\,b^6 & \\
+ & (72)^{-1}\,q & (1-q)^2 & (26+17q-7q^2)\,b^7 & \\
+ & (24)^{-1}\,q & (1-q)^2 & (7+11q)\,b^8 & \\
+ & (48)^{-1}\,q & (1-q)^2 & (11+26q+3q^2)\,b^9 & \\
+ & (6)^{-1}\,q & (1-q) & (1+4q+q^2)\,b^{10} & \\
+ & (240)^{-1}\,q & (1-q) & (23+91q+9q^2-3q^3)\,b^{11} & \\
+ & (48)^{-1}\,q & (1-q) & (3+17q+5q^2-q^3)\,b^{12} & \\
+ & (144)^{-1}\,q & (1-q) & (5+39q+27q^2+q^3)\,b^{13} & \\
+ & (48)^{-1}\,q & (1-q) & (1+11q+11q^2+q^3)\,b^{14} & \\
+ & (120)^{-1}\,q & \cdot & (1+26q+66q^2+26q^3+q^4)\,b^{15} &
\end{array}
$$

Fortunately, this can be proven independently of the conjectured form of $f_m(b, q)$ and this statement appears below as Theorem 7.2. Reducing Lemma 5.2 modulo b reveals the following corollary:

Corollary 7.1 *The partition counting function $p_b(n)$ satisfies the congruence:*

$$p_b(b^m n) \equiv p_b(b^m(n-1)) + p_b(b^{m-1}n) + (b-1)p_b(b^{m-1}(n-1)) \pmod{b}$$

This corollary can then be used to prove the following theorem.

Theorem 7.2 *The partition counting function $p_b(n)$ satisfies the congruence:*

$$p_b(b^m n) - p_b(b^m(n-1)) \equiv p_b(n) \pmod{b}$$

Proof Beginning with the statement Corollary 7.1,

$$p_b(b^m n) \equiv p_b(b^m(n-1)) + p_b(b^{m-1}n) + (b-1)p_b(b^{m-1}(n-1)) \pmod{b},$$

and applying Corollary 7.1 to the middle term $p_b(b^{m-1}n)$ reveals:

$$
\begin{aligned}
p_b(b^m n) &\equiv p_b(b^m(n-1)) + p_b(b^{m-1}(n-1)) + p_b(b^{m-2}n) + (b-1)p_b(b^{m-2}(n-1)) \\
&\qquad + (b-1)p_b(b^{m-1}(n-1)) \pmod{b} \\
&\equiv p_b(b^m(n-1)) + p_b(b^{m-2}n) + (b-1)p_b(b^{m-2}(n-1)) \pmod{b}
\end{aligned}
$$

Table 12 Polynomial $f_7(b,q)$ as a polynomial in b

$$
\begin{array}{rrlll}
f_7(b,q) = & & (1-q)^6 & & \\
+ & q & (1-q)^5 & b & \\
+ & (2)^{-1}\, q & (1-q)^5 & b^2 & \\
+ & (4)^{-1}\, q & (1-q)^4 & (3+q)\, b^3 & \\
+ & (24)^{-1}\, q & (1-q)^4 & (11+q)\, b^4 & \\
+ & (48)^{-1}\, q & (1-q)^4 & (25+11q)\, b^5 & \\
+ & (96)^{-1}\, q & (1-q)^3 & (51+38q+7q^2)\, b^6 & \\
+ & (288)^{-1}\, q & (1-q)^3 & (125+44q-25q^2)\, b^7 & \\
+ & (48)^{-1}\, q & (1-q)^3 & (19+18q-q^2)\, b^8 & \\
+ & (144)^{-1}\, q & (1-q)^3 & (51+82q+5q^2)\, b^9 & \\
+ & (48)^{-1}\, q & (1-q)^2 & (14+31q+2q^2+q^3)\, b^{10} & \\
+ & (1440)^{-1}\, q & (1-q)^2 & (313+671q-241q^2-23q^3)\, b^{11} & \\
+ & (720)^{-1}\, q & (1-q)^2 & (124+423q+12q^2-19q^3)\, b^{12} & \\
+ & (480)^{-1}\, q & (1-q)^2 & (61+273q+71q^2-5q^3)\, b^{13} & \\
+ & (720)^{-1}\, q & (1-q)^2 & (67+414q+231q^2+8q^3)\, b^{14} & \\
+ & (1440)^{-1}\, q & (1-q) & (91+736q+546q^2+56q^3+11q^4)\, b^{15} & \\
+ & (360)^{-1}\, q & (1-q) & (14+124q+69q^2-26q^3-q^4)\, b^{16} & \\
+ & (1440)^{-1}\, q & (1-q) & (35+378q+336q^2-26q^3-3q^4)\, b^{17} & \\
+ & (480)^{-1}\, q & (1-q) & (7+102q+132q^2+2q^3-3q^4)\, b^{18} & \\
+ & (1440)^{-1}\, q & (1-q) & (11+206q+396q^2+106q^3+q^4)\, b^{19} & \\
+ & (240)^{-1}\, q & (1-q) & (1+26q+66q^2+26q^3+q^4)\, b^{20} & \\
+ & (720)^{-1}\, q & \cdot & (1+57q+302q^2+302q^3+57q^4+q^5)\, b^{21} &
\end{array}
$$

Subsequently applying the Corollary to the middle term $m-3$ more times produces

$$p_b(b^m n) \equiv p_b(b^m(n-1)) + p_b(bn) + (b-1)p_b(b(n-1)) \pmod{b}$$

and finally applying RIII to $p_b(bn)$ yields

$$\begin{aligned}
p_b(b^m n) &\equiv p_b(b^m(n-1)) + p_b(b(n-1)) + p_b(n) + (b-1)p_b(b(n-1)) \pmod{b} \\
&\equiv p_b(b^m(n-1)) + p_b(n) \pmod{b}
\end{aligned}$$

that is,

$$p_b(b^m n) - p_b(b^m(n-1)) \equiv p_b(n) \pmod{b}$$

as stated.

Table 13 Polynomials $g_{8,i}$ for $8 \leq i \leq 27$

m	i	$g_{m,i}$(q)
8	8	$(96)^{-1}\ (42 + 31q - q^2)$
8	9	$(576)^{-1}\ (243 + 292q + 17q^2)$
8	10	$(864)^{-1}\ (329 + 471q + 33q^2 + 31q^3)$
8	11	$(2880)^{-1}\ (911 + 1177q - 647q^2 - q^3)$
8	12	$(4320)^{-1}\ (1189 + 2598q - 453q^2 - 94q^3)$
8	13	$(1440)^{-1}\ (329 + 981q + 81q^2 - 11q^3)$
8	14	$(8640)^{-1}\ (1625 + 6633q + 2139q^2 + 43q^3)$
8	15	$(2880)^{-1}\ (427 + 1978q + 324q^2 + 110q^3 + 41q^4)$
8	16	$(4320)^{-1}\ (488 + 2539q - 231q^2 - 647q^3 + 11q^4)$
8	17	$(720)^{-1}\ (62 + 413q + 129q^2 - 65q^3 + q^4)$
8	18	$(1440)^{-1}\ (92 + 771q + 455q^2 - 111q^3 - 7q^4)$
8	19	$(2880)^{-1}\ (131 + 1352q + 1338q^2 + 64q^3 - 5q^4)$
8	20	$(8640)^{-1}\ (277 + 3524q + 4734q^2 + 812q^3 + 13q^4)$
8	21	$(2880)^{-1}\ (61 + 961q + 1546q^2 + 266q^3 + 41q^4 + 5q^5)$
8	22	$(20160)^{-1}\ (269 + 4549q + 6554q^2 - 926q^3 - 391q^4 + 25q^5)$
8	23	$(120)^{-1}\ (1 + 21q + 41q^2 + q^3 - 4q^4)$
8	24	$(8640)^{-1}\ (43 + 1075q + 2638q^2 + 686q^3 - 121q^4 - q^5)$
8	25	$(720)^{-1}\ (2 + 66q + 211q^2 + 91q^3 - 9q^4 - q^5)$
8	26	$(720)^{-1}\ (1 + 41q + 171q^2 + 131q^3 + 16q^4)$
8	27	$(1440)^{-1}\ (1 + 57q + 302q^2 + 302q^3 + 57q^4 + q^5)$

8 Sellers' Question

In a Spring 2014 talk at the New York Number Theory Seminar, Sellers presented the following identities:

$$\sum_{n\in\mathbb{Z}} p_3(81n + 42)q^n = \frac{27(8q^2 + 17q + 2)}{(1-q)^4} B_3(0, q)$$

$$\sum_{n\in\mathbb{Z}} p_3(81n + 78)q^n = \frac{27(2q^2 + 17q + 8)}{(1-q)^4} B_3(0, q)$$

and asked, "Why do the polynomial factors in the numerator come in such natural pairs as 'reciprocal polynomials'?" Given that $8q^2 + 17q + 2$ appears, why should its reciprocal polynomial, the polynomial with its coefficients reversed, that is, $2q^2 + 17q + 8$, appear?

Table 14 Polynomials $g_{9,i}$ for $9 \leq i \leq 24$

m	i	$g_{m,i}(q)$
9	9	$(1152)^{-1}\ (513 + 548q + 43q^2)$
9	10	$(3456)^{-1}\ (1463 + 1587q + 285q^2 + 121q^3)$
9	11	$(17280)^{-1}\ (6521 + 5607q - 3597q^2 + 109q^3)$
9	12	$(8640)^{-1}\ (3018 + 4631q - 1066q^2 - 103q^3)$
9	13	$(1920)^{-1}\ (597 + 1263q - 17q^2 - 3q^3)$
9	14	$(8640)^{-1}\ (2369 + 6981q + 1551q^2 + 79q^3)$
9	15	$(8640)^{-1}\ (2024 + 6103q - 513q^2 + 901q^3 + 125q^4)$
9	16	$(17280)^{-1}\ (3379 + 11660q - 4458q^2 - 2020q^3 + 79q^4)$
9	17	$(17280)^{-1}\ (2827 + 12542q - 732q^2 - 1750q^3 + 73q^4)$
9	18	$(17280)^{-1}\ (2325 + 13126q + 2988q^2 - 1854q^3 - 25q^4)$
9	19	$(17280)^{-1}\ (1867 + 12770q + 6684q^2 - 466q^3 + 25q^4)$
9	20	$(8640)^{-1}\ (743 + 6185q + 4971q^2 + 319q^3 + 22q^4)$
9	21	$(5760)^{-1}\ (383 + 3533q + 1974q^2 - 470q^3 + 331q^4 + 9q^5)$
9	22	$(60480)^{-1}\ (3054 + 31679q + 15070q^2 - 18864q^3 - 788q^4 + 89q^5)$
9	23	$(120960)^{-1}\ (4589 + 57611q + 54038q^2 - 21950q^3 - 3683q^4 + 115q^5)$
9	24	$(120960)^{-1}\ (3371 + 49719q + 63010q^2 - 11342q^3 - 4077q^4 + 119q^5)$

Table 15 Polynomials $g_{9,i}$ for $25 \leq i \leq 35$

m	i	$g_{m,i}(q)$
9	25	$(120960)^{-1}\ (2417 + 42599q + 74486q^2 + 5818q^3 - 4295q^4 - 65q^5)$
9	26	$(120960)^{-1}\ (1689 + 34745q + 75466q^2 + 20322q^3 - 1187q^4 + 5q^5)$
9	27	$(120960)^{-1}\ (1157 + 28445q + 79022q^2 + 39334q^3 + 3181q^4 + 61q^5)$
9	28	$(120960)^{-1}\ (761 + 21712q + 64153q^2 + 27480q^3 + 4667q^4 + 2168q^5 + 19q^6)$
9	29	$(120960)^{-1}\ (481 + 15384q + 47073q^2 + 8608q^3 - 10677q^4 - 408q^5 + 19q^6)$
9	30	$(120960)^{-1}\ (301 + 11232q + 40293q^2 + 15328q^3 - 6417q^4 - 288q^5 + 31q^6)$
9	31	$(10080)^{-1}\ (15 + 680q + 3011q^2 + 1856q^3 - 419q^4 - 104q^5 + q^6)$
9	32	$(120960)^{-1}\ (103 + 5528q + 28775q^2 + 25632q^3 + 1069q^4 - 632q^5 + 5q^6)$
9	33	$(20160)^{-1}\ (9 + 632q + 4097q^2 + 4832q^3 + 667q^4 - 152q^5 - 5q^6)$
9	34	$(60480)^{-1}\ (13 + 1112q + 8861q^2 + 14496q^3 + 5431q^4 + 328q^5 - q^6)$
9	35	$(10080)^{-1}\ (1 + 120q + 1191q^2 + 2416q^3 + 1191q^4 + 120q^5 + q^6)$

Table 16 Polynomials $g_{10,i}$ for $10 \leq i \leq 24$

m	i	$g_{m,i}(q)$
10	10	$(3456)^{-1}\ (1508 + 1479q + 366q^2 + 103q^3)$
10	11	$(17280)^{-1}\ (6971 + 4662q - 3057q^2 + 64q^3)$
10	12	$(17280)^{-1}\ (6736 + 8167q - 1772q^2 - 171q^3)$
10	13	$(25920)^{-1}\ (9437 + 15513q - 207q^2 + 97q^3)$
10	14	$(17280)^{-1}\ (5828 + 13307q + 2582q^2 + 243q^3)$
10	15	$(103680)^{-1}\ (31513 + 66086q - 10716q^2 + 15722q^3 + 1075q^4)$
10	16	$(34560)^{-1}\ (9287 + 22450q - 12300q^2 - 2258q^3 + 101q^4)$
10	17	$(17280)^{-1}\ (4103 + 12979q - 3099q^2 - 1091q^3 + 68q^4)$
10	18	$(34560)^{-1}\ (7159 + 28984q + 210q^2 - 3208q^3 - 25q^4)$
10	19	$(51840)^{-1}\ (9217 + 45515q + 12507q^2 - 1507q^3 + 148q^4)$
10	20	$(17280)^{-1}\ (2619 + 15673q + 7953q^2 + 315q^3 + 80q^4)$
10	21	$(32400)^{-1}\ (4111 + 25660q + 1585q^2 - 1870q^3 + 2915q^4 - q^5)$
10	22	$(120960)^{-1}\ (12709 + 90224q - 1458q^2 - 43888q^3 + 2869q^4 + 24q^5)$
10	23	$(362880)^{-1}\ (31337 + 267179q + 92282q^2 - 116054q^3 - 2579q^4 - 5q^5)$
10	24	$(120960)^{-1}\ (8470 + 84599q + 53646q^2 - 29488q^3 - 1388q^4 + 81q^5)$

Table 17 Polynomials $g_{10,i}$ for $25 \leq i \leq 34$

m	i	$g_{m,i}(q)$
10	25	$(241920)^{-1}\ (13535 + 158685q + 157158q^2 - 30134q^3 - 6837q^4 - 87q^5)$
10	26	$(120960)^{-1}\ (5334 + 71929q + 93434q^2 + 1956q^3 - 1360q^4 + 67q^5)$
10	27	$(362880)^{-1}\ (12460 + 195325q + 332566q^2 + 75308q^3 + 3926q^4 + 335q^5)$
10	28	$(241920)^{-1}\ (6355 + 108566q + 147413q^2 - 43284q^3 + 14869q^4 + 8062q^5 - 61q^6)$
10	29	$(362880)^{-1}\ (7178 + 138609q + 210603q^2 - 121954q^3 - 55980q^4 + 3057q^5 - 73q^6)$
10	30	$(120960)^{-1}\ (1781 + 39601q + 78646q^2 - 15078q^3 - 15247q^4 + 1013q^5 + 4q^6)$
10	31	$(725760)^{-1}\ (7807 + 200214q + 493533q^2 + 11980q^3 - 106623q^4 - 2178q^5 + 67q^6)$
10	32	$(80640)^{-1}\ (623 + 18268q + 54323q^2 + 15184q^3 - 7435q^4 - 332q^5 + 9q^6)$
10	33	$(362880)^{-1}\ (1970 + 66843q + 237309q^2 + 114242q^3 - 23184q^4 - 4029q^5 - 31q^6)$
10	34	$(60480)^{-1}\ (226 + 8760q + 36747q^2 + 27718q^3 + 2274q^4 - 126q^5 + q^6)$

Why should one expect that these sorts of identities exist in the first place? Some combinatorial insight is desired, but failing that Lemma 5.4 and Theorem 5.5 guarantee that *some* relationship exists, although they fall short of explaining why such reciprocal polynomials appear.

By Lemma 5.4, applying RI and RIII to an expression like $p_b(b^m n + k)$ will produce identities between its generating function and $B_b(0, q)$. Consider the results when Lemma 5.4 applied to Sellers' example:

Table 18 Polynomials $g_{10,i}$ for $35 \leq i \leq 44$

m	i	$g_{m,i}(q)$
10	35	$(725760)^{-1}\,(1831 + 82308q + 403011q^2 + 395512q^3 + 80853q^4 + 4068q^5 + 97q^6)$
10	36	$(241920)^{-1}$ $(399 + 20449q + 106983q^2 + 95433q^3 + 10453q^4 + 6939q^5 + 1269q^6 - 5q^7)$
10	37	$(725760)^{-1}$ $(763 + 43703q + 236763q^2 + 167071q^3 - 72007q^4 - 15363q^5 + 1969q^6 - 19q^7)$
10	38	$(120960)^{-1}\,(79 + 5281q + 33561q^2 + 33127q^3 - 6983q^4 - 4617q^5 + 31q^6 + q^7)$
10	39	$(362880)^{-1}\,(143 + 11137q + 80253q^2 + 99011q^3 + q^4 - 9297q^5 + 179q^6 + 13q^7)$
10	40	$(241920)^{-1}\,(55 + 5137q + 43875q^2 + 69397q^3 + 10597q^4 - 7533q^5$ $-575q^6 + 7q^7)$
10	41	$(1209600)^{-1}\,(151 + 16753q + 163431q^2 + 316465q^3 + 117805q^4 - 8181q^5$ $-1643q^6 + 19q^7)$
10	42	$(80640)^{-1}\,(5 + 723q + 8577q^2 + 20519q^3 + 10719q^4 + 9q^5 - 229q^6 - 3q^7)$
10	43	$(241920)^{-1}\,(7 + 1217q + 17163q^2 + 51757q^3 + 41957q^4 + 8595q^5 + 265q^6 - q^7)$
10	44	$(80640)^{-1}\,(1 + 247q + 4293q^2 + 15619q^3 + 15619q^4 + 4293q^5 + 247q^6 + q^7)$

$$\begin{aligned} p_3(81n+42) &= p_3(81n) + 14p_3(27n) + 30p_3(9n) + 9p_3(3n) \\ p_3(81n+78) &= p_3(81n) + 26p_3(27n) + 108p_3(9n) + 81p_3(3n) \end{aligned}$$

Then, upon passing to generating functions

$$\begin{aligned} \sum_{n\in\mathbb{Z}} p_3(81n+42)q^n &= B_3(4,q) \;+\; 14B_3(3,q) + 30B_3(2,q) + 9B_3(1,q) \\ &= \left(\frac{f_4(3,q)}{(1-q)^4} + 14\frac{f_3(3,q)}{(1-q)^3} + 30\frac{f_2(3,q)}{(1-q)^2} + 9\frac{f_1(3,q)}{1-q}\right) B_3(0,q) \\ &= \frac{27(8q^2+17q+2)}{(1-q)^4} B_3(0,q) \end{aligned}$$

and

$$\begin{aligned} \sum_{n\in\mathbb{Z}} p_3(81n+78)q^n &= B_3(4,q) + 26B_3(3,q) + 108B_3(2,q) + 81B_3(1,q) \\ &= \left(\frac{f_4(3,q)}{(1-q)^4} + 26\frac{f_3(3,q)}{(1-q)^3} + 108\frac{f_2(3,q)}{(1-q)^2} + 81\frac{f_1(3,q)}{1-q}\right) B_3(0,q) \\ &= \frac{27(2q^2+17q+8)}{(1-q)^4} B_3(0,q). \end{aligned}$$

A full understanding of identities like these seems to require a thorough understanding of the polynomials $f_m(b,q)$ as well as the polynomials $g_{m,k,j}(b)$ from Lemma 5.4.

9 Some Computations

Doing this work without computing $p_b(n)$ for large values of n would be a waste, so here are the values for a few choice n and their prime factorization.

$$\begin{aligned} p_2(2^{10}) &= 2320518948 \\ &= 2^2 \cdot 3 \cdot 11 \cdot 197 \cdot 89237 \\ &\text{See also [7]} \end{aligned}$$

$$\begin{aligned} p_2(2^{30}) &= 1525223521662612652482573042270879062244863772153 30\backslash \\ &\quad 737509179365599818522093065697433856805421794 70233380 \\ &= 2^2 \cdot 5 \cdot 19 \cdot 31 \cdot 79 \cdot 1217 \cdot 46553987 \cdot 719224073 \\ &\quad \cdot 88243965275199121 \cdot 1201364132790744647 \\ &\quad \cdot 379393391071160025350141826238305857058 0931 \end{aligned}$$

$$\begin{aligned} p_3(3^{27}) &= 350364423551707258416807382080740574025054741900008\backslash \\ &\quad 668600126882878615683202075701898785282388145497481\backslash \\ &\quad 0418192030384012393566952227798779899585 2 \\ &= 2^2 \cdot 875911058879268146042018455201851435062636854750 02\backslash \\ &\quad 167150031720719653920800518925474696320597036374 37\backslash \\ &\quad 026045480075960030983917380569496949748963 \end{aligned}$$

10 Proving the Case $m = 3$

This section gives an iterative construction of the polynomial $f_3(b, q)$. The methods used here can be used to prove the $m = 4$ case, but the argument becomes significantly longer. It is likely that this method can be used to construct $f_m(b, q)$ for any fixed m, but the length of the argument becomes unwieldy.

Theorem 10.1 *The generating function for $p_b(b^3 n)$ satisfies the identity:*

$$f_3(b, q)B_b(q) = (1 - q)^3 B_b(3, q)$$

where

$$f_3(b, q) = (1 - q)^2 + q(1 - q)b - \tfrac{1}{2}q(1 - q)b^2 + \tfrac{1}{2}q(q + 1)b^3$$

Proof Begin as before, by iterating the recurrence via Lemma 5.2:

$$\begin{aligned} p_b(b^3n) &= p_b(b^3(n-1)) + p_b(b^2n) + (b-1)p_b(b^2(n-1)) \\ &\quad + b\sum_{u=1}^{b-1} p_b(b^2(n-1)+ub) \end{aligned}$$

The sum in the final term can be simplified further via Lemma 5.1:

$$\begin{aligned} \sum_{u=1}^{b-1} p_b(b^2(n-1)+ub) &= \sum_{u=1}^{b-1} p_b(b^2(n-1)) + \sum_{k=1}^{u} p_b(b(n-1)+k) \\ &= (b-1)p_b(b^2(n-1)) + \sum_{u=1}^{b-1}\sum_{k=1}^{u} p_b(b(n-1)+k) \end{aligned}$$

Now, $1 \le k \le b-1$ so by RI

$$\begin{aligned} \sum_{u=1}^{b-1}\sum_{k=1}^{u} p_b(b(n-1)+k) &= \sum_{u=1}^{b-1}\sum_{k=1}^{u} p_b(b(n-1)) \\ &= p_b(b(n-1))\sum_{u=1}^{b-1}\sum_{k=1}^{u} 1 \\ &= p_b(b(n-1))\sum_{u=1}^{b-1} u \\ &= \binom{b}{2} p_b(b(n-1)) \end{aligned}$$

Finally, the original expression becomes

$$\begin{aligned} p_b(b^3n) &= p_b(b^3(n-1)) + p_b(b^2n) + (b-1)p_b(b^2(n-1)) \\ &\quad + b(b-1)p_b(b^2(n-1)) + b\binom{b}{2} p_b(b(n-1)) \end{aligned}$$

Passing to generating functions by multiplying this identity by q^n and summing over all n yields

$$\begin{aligned}\sum_{n\in\mathbb{Z}} p_b(b^3n)q^n &= \sum_{n\in\mathbb{Z}} p_b(b^3(n-1)) + \sum_{n\in\mathbb{Z}} p_b(b^2n) + \sum_{n\in\mathbb{Z}}(b-1)p_b(b^2(n-1)) \\ &\quad + \sum_{n\in\mathbb{Z}} b(b-1)p_b(b^2(n-1)) + \sum_{n\in\mathbb{Z}} b\binom{b}{2}p_b(b(n-1)) \\ B_b(3,q) &= qB_b(3,q) + B_b(2,q) + (b-1)qB_b(2,q) \\ &\quad + b(b-1)qB_b(2,q) + b\binom{b}{2}qB_b(1,q)\end{aligned}$$

After moving terms of $B_b(3,q)$ to the right-hand side, the above equation becomes

$$\begin{aligned}(1-q)B_b(3,q) &= (1+(b-1)q + b(b-1)q)\,B_b(2,q) + b\binom{b}{2}qB_b(1,q) \\ &= (1+(b^2-1)q)B_b(2,q) + b\binom{b}{2}qB_b(1,q)\end{aligned}$$

Substituting in the results for $B_b(2,q)$ and $B_b(1,q)$ in Lemma 4.4, Corollary 5.3, and multiplying by $(1-q)^2$ yields

$$\begin{aligned}(1-q)^3B_b(3,q) &= (1+(b^2-1)q)\,((1+(b-1)q)B_b(0,q)) + b\binom{b}{2}q(1-q)B_b(0,q) \\ &= \left((1+(b^2-1)q)\left((1+(b-1)q) + b\binom{b}{2}q(1-q)\right)\right)B_b(0,q) \\ &= \left(1+\tfrac{1}{2}(b-1)\left((b^2+2b+4)q + (b^2-2)q^2\right)\right)B_b(0,q)\end{aligned}$$

Therefore,

$$(1-q)^3B_b(3,q) = \left((1-q)^2 + q(1-q)b - \tfrac{1}{2}q(1-q)b^2 + \tfrac{1}{2}q(q+1)b^3\right)B_b(0,q)$$

as desired.

Acknowledgements This research was supported, in part, under National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113 and the City University of New York High Performance Computing Center at the College of Staten Island.

References

1. L. Euler, *De partitione numerorum*, Novi comment. Acad. Sci. Imp. Petropol. **3**, 125–169 (1753). http://math.dartmouth.edu/~euler/docs/originals/E191.pdf
2. A. Tanturri, *Sul numero delle partizioni d'un numero in potenze di 2*, Att. della Sci. di Torino **54**, 97–110 (1918). http://biodiversitylibrary.org/page/12142624
3. G.H. Hardy, S. Ramanujan, Asymptotic formulaæn combinatory analysis. Proc. Lond. Math. Soc. **s2-17**(1), 75–115 (1918). http://plms.oxfordjournals.org/content/s2-17/1/75.short

4. K. Mahler, On a special functional equation. J. London Math. Soc. **15**, 115–123 (1940). MR 0002921 (2,133e)
5. N.G. de Bruijn, On Mahler's partition problem. Nederl. Akad. Wetensch., Proc. **51** (1948), 659–669 = Indagationes Math. 10, 210–220 (1948). MR 0025502 (10,16d)
6. D.F. Knuth, *An almost linear recurrence*, issue 2, pp. 117–128. http://www.fq.math.ca/Scanned/4-2/knuth.pdf. MR "33 #7317"
7. R.F. Churchhouse, Congruence properties of the binary partition function. Proc. Cambridge Philos. Soc. **66**, 371–376 (1969). MR 0248102 (40 #1356)
8. Ø. Rødseth, Some arithmetical properties of m-ary partitions. Proc. Cambridge Philos. Soc. **68**, 447–453 (1970). MR 0260695 (41 #5319)
9. G.E. Andrews, Congruence properties of the m-ary partition function. J. Number Theory **3**, 104–110 (1971). MR 0268144 (42 #3043)
10. H. Gupta, Proof of the Churchhouse conjecture concerning binary partitions. Proc. Cambridge Philos. Soc. **70**, 53–56 (1971). MR 0295924 (45 #4986)
11. H. Gupta, A simple proof of the Churchhouse conjecture concerning binary partitions. Indian J. Pure Appl. Math. **3**(5), 791–794 (1972). MR 0330038 (48 #8377)
12. M.D. Hirschhorn, J.H. Loxton, Congruence properties of the binary partition function. Math. Proc. Cambridge Philos. Soc. **78**(3), 437–442 (1975). MR 0382157 (52 #3045)
13. G. Dirdal, Congruences for m-ary partitions. Math. Scand. **37**(1), 76–82 (1975). MR 0389752 (52 #10583)
14. G. Dirdal, On restricted m-ary partitions. Math. Scand. **37**(1), 51–60 (1975). MR 0389751 (52 #10582)
15. H. Gupta, P.A.B. Pleasants, Partitions into powers of m. Indian J. Pure Appl. Math. **10**(6), 655–694 (1979). MR 534195 (80f:10014)
16. B. Reznick, *Some binary partition functions*, pp. 451–477. MR 1084197 (91k:11092)
17. Ø.J. Rødseth, J.A. Sellers, On m-ary partition function congruences: a fresh look at a past problem. J. Number Theory **87**(2), 270–281 (2001). MR 1824148 (2001m:11177). https://doi.org/10.1006/jnth.2000.2594
18. R.L. Graham, D.E. Knuth, O. Patashnik, *Concrete mathematics*, 2nd edn. (Addison-Wesley Publishing Company, Reading, MA, 1994), A Foundation for Computer Science. MR 1397498 (97d:68003)

Cryptographic Hash Functions and Some Applications to Information Security

Lisa Bromberg

Abstract We explore hashing with matrices over $\mathrm{SL}_2(\mathbb{F}_p)$, outlining known results of Tillich and Zémor. We then summarize the bounds on the girth of the Cayley graph of the subgroup of $\mathrm{SL}_2(\mathbb{F}_p)$ for specific generators A, B, work done by the author, Shpilrain, and Vdovina. We demonstrate that even without optimization, these hashes have comparable performance to hashes in the SHA family.

Keywords Information security · Cryptography · Group theory

1 Introduction

The aim of cryptography is to protect information from being stolen or modified by an adversary. In modern cryptography, specific security goals are achieved with the design of algorithms and also using the known computational hardness of certain mathematical problems.

There are currently two main classes of cryptographic primitives: *public-key (asymmetric)* and *symmetric-key*. Symmetric-key algorithms are older and in fact can be traced back to at least the time of Julius Caesar. In symmetric-key ciphers, knowledge of the encryption key is usually equivalent (or equal) to knowledge of the decryption key. Because of this, participating parties need to agree on a shared secret key before communicating through an open channel.

Public-key cryptography is a relatively young area of mathematics, but it has been a very active area of research since its inception in 1976, with a seminal paper of Diffie and Hellman [4]. In public-key algorithms, there are two separate keys: a public-key that is published and a private-key which each user keeps secret. Knowledge of the public-key does not imply knowledge of the private-key with any efficient computation. In fact, the public-key is generated from the private-key using a *one-way* function, with a *trapdoor*, which is a function that is easy (i.e., polynomial time with respect to the complexity of an input) to compute, but hard (no

L. Bromberg (✉)
United States Military Academy, West Point, NY, USA
e-mail: lisa.bromberg@usma.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_6

visible (probabilistic) polynomial-time algorithm on "most" inputs) to invert the image of a random input without special information; the special information is the above-mentioned "trapdoor." A well-known example of public-key encryption is the RSA cryptosystem, whose one-way function is the product of two large primes p, q. If p and q are known, then it is easy to compute their product, but it is hard to factor a large number into its prime factors.

Since public-key cryptosystems are more computationally costly than symmetric algorithms, some modern cryptosystems rely on an asymmetric cipher to produce a session key and then proceed with symmetric encryption for the remainder of the session.

1.1 Hash Functions

A very important cryptographic primitive is the *hash function*. Cryptographic hash functions have many applications to information security, including digital signatures and methods of authentication. They can also be used as ordinary hash functions, to index data in a hash table, fingerprinting (a procedure which maps an arbitrary large data item to a shorter bit string, or fingerprint which uniquely identifies the original data for all practical purposes), to detect duplicate data, and as checksums to detect (accidental) data corruption. In fact, in the context of information security, cryptographic hash values are often referred to as fingerprints, checksums, or just hash values.

We will first define the hash function and explain some properties we require a hash function to possess. Then, we introduce *Cayley hash functions*, which are a family of hash functions based on nonabelian groups. We then explore hashing with matrices and outline results of the authors, Shpilrain and Vdovina [1] using matrices over $\mathrm{SL}(2, \mathbb{F}_p)$ of a particular form which generate the group.

Definition 1 Let $n \in \mathbb{N}$ and let $H\colon \{0,1\}^* \to \{0,1\}^n$ such that $m \mapsto h = H(m)$. We require a hash function to satisfy the following:

(1) *Preimage resistance*: Given output y, it is hard to find input x such that $H(x) = y$;
(2) *Second preimage resistance*: Given input x_1, it is hard to find another input $x_2 \neq x_1$ such that $H(x_1) = H(x_2)$;
(3) *Collision resistance*: It is hard to find inputs $x_1 \neq x_2$ such that $H(x_1) = H(x_2)$.

Note that since hash functions are not injective, this "uniqueness" that we desire is purely computational. From a practical perspective, this means that no big cluster of computers can find the input based only on the output of a hash function.

There exist old hash function constructions whose collision resistance follows from the hardness of number-theoretic or group-theoretic problems. However, these hash functions can only be used in applications which require only collision resistance and are often too slow for practical purposes. Standardized hash functions, such as the SHA family, follow the *block cipher design*: Their use is not restricted to collision resistance, but their collision resistance is heuristic and not established by any precise mathematical problem. In fact, recent attacks against the SHA-1 algorithm have led to a competition for a new Standard Hash Algorithm [11].

Another direction, more relevant to our exposition, is the *expander hash function*, dating back to 1991 when Zémor proposed building a hash function based on the special linear group. This first attempt was quickly broken, but Tillich and Zémor quickly proposed a second function which was resistant to the attack on the first; see [16]. However, this newer hash function is also vulnerable to attack; see [17]. The Tillich–Zémor hash function is a type of expander hash called a *Cayley hash function* and is different from functions in the SHA family in that it is not a block hash function, but rather each bit is hashed individually. We discuss this particular hash function in further detail in Sect. 2.

The expander hash design is fundamentally different from classical hash designs in that it allows for relating important properties of hash functions such as collision resistance, preimage resistance (see Definition 1), and their output distribution to the graph-theoretical notions of cycle, girth, and expanding constants. When the graphs used are *Cayley graphs*, the design additionally provides efficient parallel computation and group-theoretical interpretations of the hash properties.

The expander hash design, though not so new anymore, is still little understood by the cryptographic community. The Tillich–Zémor hash function is often considered broken because of existing trapdoor attacks and attacks against specific parameters. In fact, relations between hash, graph, and group-theoretic properties have been sketched but no precise statements on these problems exist. Since the mathematical problems which underly the security of expander hashes do not belong to classical problems, it appears as though they have not been investigated. Hence, their actual hardness is unknown. Efficiency aspects have also only been sketched.

Cayley hash functions are based on the idea of using a pair of (semi)group elements, A and B, to hash the "0" and "1" bit, respectively, and then to hash an arbitrary bit string by using multiplication of elements in the (semi)group. We focus on hashing with 2×2 matrices over $\mathbb{F}_p$. Since there are many known pairs of 2×2 matrices over $\mathbb{Z}$ which generate a free monoid, this yields numerous pairs of matrices over $\mathbb{F}_p$ (for p sufficiently large) that are candidates for collision-resistant hashing. However, this trick can backfire and allow for a lifting of matrix elements to $\mathbb{Z}$ to find a collision. This "lifting attack" was used by Tillich and Zémor [16] in the case where two matrices A and B generate (as a monoid) all of $\mathrm{SL}(2, \mathbb{Z}_+)$. With other, "similar" pairs of matrices from $\mathrm{SL}(2, \mathbb{Z})$, the situation is different, and while the same "lifting attack" can (in some cases) produce collision in the *group* generated by A and B, it says nothing about the *monoid* generated by A and B; see [1]. Since only positive powers are used for hashing, this is all that is needed, and so, for these pairs of matrices, there are no known attacks at this time that would affect the security of the corresponding hash functions.

Additionally, we recall lower bounds on the length of collisions for hash functions corresponding to some particular pairs of matrices from $\mathrm{SL}(2, \mathbb{F}_p)$; again, see [1].

1.2 Cayley Hash Functions

Classical hash functions mix pieces of the message repeatedly so the result appears sufficiently random [13]. For this reason, they may be unappealing outside the area of cryptography. On the other hand, a particular type of expander hash function, the Cayley hash function, has a more straightforward design.

Given a group G and a subset $S = \{s_1, \ldots, s_k\}$ of G, their *Cayley graph* $\mathfrak{G}$ is a k-regular graph that has a vertex v_g associated with each element of G and an edge between vertices v_{g_1} and v_{g_2} if there exists $s_i \in S$ such that $g_2 = g_1 s_i$.

To build a hash function from the Cayley graph, let $\sigma\colon \{1, \ldots, k\} \to S$ be an ordering, fix an initial value g_0, and write the message m as a string $m_1 m_2 \cdots m_N$, where $m_i \in \{1, \ldots, k\}$. Then, the hash value is $H(m) := g_0 \sigma(m_1) \cdots \sigma(m_N)$. This is represented on the Cayley graph as a (nonbacktracking) walk; the endpoint of the walk is the hash value.

Two texts yielding the same hash value correspond to two paths with the same start and endpoints. We would like those two paths to differ necessarily by a "minimum amount." Such a vague notion can be guaranteed if there are no short cycles in the Cayley graph. More precisely, we want the Cayley graph to have a large *girth*:

Definition 2 The *directed girth* of a Cayley graph $\mathfrak{G}$ is the largest integer ∂ such that, given any two vertices u and v, any pair of distinct paths which joins u to v will be such that one of those paths has length (i.e., number of edges) ∂ or more.

The idea is that the girth of the Cayley graph is a relevant parameter to hashing. More precisely, if a Cayley graph has a large girth ∂, then the corresponding hash function will have the property that small modifications of the text will modify the hash value [16].

One of the main advantages of Cayley hash functions over classical hash functions is their ability to be parallelized. Namely, if messages x and y are concatenated, then the hashed value of xy is $H(xy) = H(x)H(y)$. Associativity of the group means we can break down a large message into more manageable pieces, hash each piece, and then recover the final result from the partial products.

Finally, a desirable feature of any hash function is the equidistribution of the hashed values. This property can be guaranteed if the associated Cayley graph satisfies the following property.

Proposition 1 [17, Proposition 2.3] *If the Cayley graph of a group G is such that the greatest common divisor of its cycle lengths equals* 1*, then for the corresponding hash function, the distribution of hashed values of texts of length n tends to equidistribution when n tends to infinity.*

This proposition is proved using classical graph-theoretic techniques, by studying successive powers A^n of the adjacency matrix of the graph. Equidistribution can be achieved with graphs that have a high expansion coefficient; see [18].

The collision, second preimage, and preimage resistance of classical hash functions easily translate to group-theoretic problems.

Definition 3 Let G be a group and let $S = \{s_1, \ldots, s_k\} \subset G$ be a generating set. Let $L \in \mathbb{Z}$ be "small."

(1) *Balance problem* Find an "efficient" algorithm that returns two words $m = m_1 \cdots m_\ell$ and $m' = m'_1 \cdots m'_{\ell'}$ with $\ell, \ell' < L, m_i, m'_i \in \{1, \ldots, k\}$ and $\prod s_{m_i} = \prod s_{m'_i}$.

(2) *Representation problem* Find an "efficient" algorithm that returns a word $m_1 \cdots m_\ell$ with $\ell < L, m_i \in \{1, \ldots, k\}$ and $\prod s_{m_i} = 1$.

(3) *Factorization problem* Find an "efficient" algorithm that, given any element $g \in G$, returns a word $m_1 \cdots m_\ell$ with $\ell < L, m_i \in \{1, \ldots, k\}$ and $\prod s_{m_i} = g$.

Note that since the group is finite, the length restriction is required, since for every $w \in G$, $w^{|G|} = 1$. Note also that Lubotzky described the factorization problem as a noncommutative analog of the discrete logarithm problem [8]. In fact, if we omit trivial solutions, then the representation and factorization problems are equivalent to the discrete logarithm problem in abelian groups.

In general, the factorization problem is at least as hard as the representation problem, which is itself at least as hard as the balance problem.

It is apparent that a Cayley hash function is collision resistant if and only if the balance problem is hard, second preimage resistant if and only if the representation problem is hard, and preimage resistant if and only if the factorization problem is hard.

Among all Cayley hash proposals, the Tillich–Zémor hash function is the only remaining current candidate. In general, the security of Cayley hashes depends on the hardness in general of the factorization problem, which remains a big open problem.

The efficiency of Cayley hashes depends on specific parameters: The Tillich–Zémor hash function is the most efficient expander hash, but it is still 10–20 times slower than the standard classical hash SHA. Computation in Cayley hashes can be easily parallelized, which could be a major benefit in applications. We outline a hash function based on the Tillich–Zémor hash function [1] which is resistant to known methods of attack and which is efficient in computation.

1.3 Possible Attacks

The mathematical structure of Cayley hash functions leaves them vulnerable to attacks which exploit this structure.

An important category of attack is the *subgroup attack*. A probabilistic attack was devised by Camion [2], based on the search for text whose hashcode falls into a subgroup.

A second important category of attack is the *lifting attack*. Let us illustrate how a lifting attack works with an example. Let $G = \mathrm{SL}(2, \mathbb{F}_p)$. There is a natural map, the *reduction modulo p map*, from $\mathrm{SL}(2, \mathbb{Z})$ to $\mathrm{SL}(2, \mathbb{F}_p)$. A lifting attack for $\mathrm{SL}(2, \mathbb{F}_p)$ will "lift" the generators of $\mathrm{SL}(2, \mathbb{Z})$ and then try to "lift" the element to be factored on the subgroup of $\mathrm{SL}(2, \mathbb{Z})$ generated by the lifts of the generators. Generally, if a factorization exists, it is easier to find over $\mathbb{Z}$ rather than over $\mathbb{F}_p$, since properly chosen generators of an infinite group will give us unique factorization. Once a factorization over $\mathbb{Z}$ has been obtained, reducing modulo p provides a factorization over $\mathbb{F}_p$. The most difficult part of the lifting attack is the lifting itself. For a specific example of how the lifting attack works in the case of the Tillich–Zémor hash function, see [16].

2 Hashing with Matrices

Hashing with matrices refers to the idea of using a pair of matrices, A and B (over a finite ring) to hash the "0" bit and the "1" bit, respectively. Then, an arbitrary bit string is hashed by using multiplication of matrices. So, the bit string 1001101 is hashed to the matrix BA^2B^2AB.

One way to help ensure the requirements of Definition 1 is to use a pair of elements, A and B, of a semigroup S such that the Cayley graph of the semigroup generated by A and B is an expander graph. The most popular implementation of this idea is the *Tillich–Zémor hash function* [17].

The use of the special linear group $\mathrm{SL}(2, \mathbb{F}_p)$ of 2×2 matrices with determinant 1 over a finite field $\mathbb{F}_p$ is a promising choice for devising hash functions. To begin with, we can choose simple matrices as generators, which yield a fast hash: Multiplication by such a matrix amounts to a few additions in $\mathbb{F}_p$. Cayley graphs over $\mathrm{SL}(2, \mathbb{F}_p)$ also have good expanding properties; see Sarnak [15], Lafferty and Rockmore [7], and Margulis [9, 10].

3 Hashing with $G = \mathrm{SL}(2, \mathbb{F}_p)$

Another idea is to use A and B over $\mathbb{Z}$ which generate a free monoid and then reduce the entries modulo a large prime p to get matrices over $\mathbb{F}_p$. Here, we have a lower bound on the length of bit string where a collision may occur, since there cannot be an equality of positive products of A and B unless at least one of the entries in at least one of the products is at least p. The bound is on the order of $\log p$.

We investigate the Cayley graphs of $\mathrm{SL}(2, \mathbb{F}_p)$ generated by

$$A(n) = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}, \quad B(n) = \begin{pmatrix} 1 & 0 \\ n & 1 \end{pmatrix},$$

where $n = 2, 3$, respectively, and p is a large prime. Particularly, we show their application to hashing.

The main difference is the difference between the *group* generated by $A(n)$ and $B(n)$ and the *monoid* generated by $A(n)$ and $B(n)$.

3.1 The Base Case

A pair of matrices over $\mathbb{Z}$ which generate a free monoid is

$$A(1) = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad B(1) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Note that these matrices as generators of the group $\mathrm{SL}(2, \mathbb{F}_p)$ have a Cayley graph which forms an expander, so they are good candidates for the basis of a hash function. Note also that these matrices are invertible and thus actually generate the group $\mathrm{SL}(2, \mathbb{Z})$. This group is not free, but the monoid generated by $A(1)$ and $B(1)$ is free. Since only positive powers are used in hashing, this is all we need.

However, since $A(1)$ and $B(1)$ generate all of $\mathrm{SL}(2, \mathbb{Z})$, we can use a lifting attack on the corresponding hash function: A collision is found using the Euclidean algorithm on the entries of a matrix; see [16]. In short, it is readily seen that a short factorization of the identity over $\mathrm{SL}(2, \mathbb{F}_p)$ produces collisions. To find such a factorization, the strategy is to reduce the problem to factoring in an infinite group: in this case, the group $\mathrm{SL}(2, \mathbb{Z})$. Find a matrix U in $\mathrm{SL}(2, \mathbb{Z})$ which reduces modulo p to the identity and which can be expressed as a product of $A(1)$s and $B(1)$s. In this case, that means that we only require U to have nonnegative coefficients. Then, we use the Euclidean algorithm, which is an efficient way to obtain the factorization of U.

For this attack to be effective, there must be a way of finding such a matrix U. Tillich and Zémor [16] describe a probabilistic algorithm which does this. It is based on the fact that the set of matrices of $\mathrm{SL}(2, \mathbb{Z})$ with nonnegative coefficients is “dense.”

To protect against such attacks, one should choose a set of generators that generate a sufficiently sparse submonoid of the infinite group associated with $\mathrm{SL}(2, \mathbb{F}_p)$. Tillich and Zémor proposed using the matrices

$$A = \begin{pmatrix} \alpha & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \alpha & \alpha + 1 \\ 1 & 1 \end{pmatrix},$$

where computations are made in the quotient field $\mathbb{F}_{2^n} = \mathbb{F}_2/\langle p(x) \rangle$, where $p(x)$ has degree n and α is a root of p. See [17] for details on the implementation of this hash.

Tillich and Zémor use matrices A, B from the group $\mathrm{SL}(2, R)$, where R is a commutative ring defined by $R = \mathbb{F}_2[x]/(p(x))$. They took $p(x)$ to be the irreducible polynomial $x^{131} + x^7 + x^6 + x^5 + x^4 + x + 1$ over $\mathbb{F}_2[x]$. Thus, R is isomorphic to $\mathbb{F}_{2^n}$, where n is the degree of the irreducible polynomial $p(x)$. Then, the matrices A and B are

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

This hash function was published in 1994 [17], but there have been several recent attacks. In 2010, Petit and Quisquater [12] describe a preimage attack; in 2011, Grassl, Ilic, Magliveras and Steinwandt [6] describe a collision attack.

3.2 Hashing with $A(2)$ and $B(2)$

In this section, we outline circuits in the Cayley graph of $\mathrm{SL}(2, \mathbb{F}_p)$ with generating set $A(2)$, $B(2)$, as presented in [1]. Note that these matrices also correspond to a Cayley graph which forms an expander graph.

We begin by noting that the lifting attack on the hash function depending on $A(1)$ and $B(1)$ described above is the only published attack on that hash function. This particular attack does not work with $A(2)$, $B(2)$. In particular, this gives evidence of the security of using these matrices for hashing over $\mathbb{F}_p$ for a large prime p.

First, we need to justify why these matrices are better candidates than $A(1)$ and $B(1)$. Recall that when considered as matrices over $\mathbb{Z}$, $A(1)$ and $B(1)$ generate (as a monoid) the entire monoid of 2×2 matrices over $\mathbb{Z}$ with positive entries, $\mathrm{SL}(2, \mathbb{Z}_{+})$.

However, this is not the case with $A(2)$ and $B(2)$.

Theorem 1 Sanov [14]

(1) The group generated by

$$A(2) = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad B(2) = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix},$$

is a free group.

(2) The subgroup of $\mathrm{SL}(2, \mathbb{Z})$ *generated by* $A(2)$ *and* $B(2)$ *consists of all invertible matrices of the form*

$$\begin{pmatrix} 1 + 4m_1 & 2m_2 \\ 2m_3 & 1 + 4m_4 \end{pmatrix}, \tag{*}$$

where the m_i are integers.

This does not say much about the *monoid* generated by these matrices, though. In fact, a generic matrix of the form above would not belong to this monoid. This is true for two reasons: First, by another result of Sanov [14], the matrices $A(2)$ and $B(2)$ generate a free group. Second, the number of matrices in the above form which are representable by positive words is negligible. In fact, the number of distinct elements represented by all freely reducible words in $A(2)$ and $B(2)$ of length $n \geq 2$ is $4 \cdot 3^{n-1}$, while the number of distinct elements represented by positive words of length $n \geq 2$ is 2^n.

Tillich and Zémor's lifting attack can still give an efficient algorithm which finds relations of length $O(\log p)$ in the group generated by $A(2)$ and $B(2)$ considered as matrices over $\mathbb{F}_p$. Note that it does not affect the security of the hash function based on $A(2)$ and $B(2)$ since only positive powers of $A(2)$ and $B(2)$ are used, and the group relations produced by the algorithm will involve both negative and positive powers with overwhelming probability.

Theorem 2 (Bromberg, Shpilrain and Vdovina [1, Theorem 1]) *There is an efficient heuristic algorithm that finds particular relations of the form $w(A(2), B(2)) = 1$, where w is a group word of length $O(\log p)$, and the matrices $A(2)$ and $B(2)$ are considered over $\mathbb{F}_p$.*

3.3 Girth of the Cayley Graph Generated by $A(k)$ and $B(k)$

For hashing, we use only positive powers, so we need only to consider products of positive powers of $A(k)$ and $B(k)$. We note that entries in a matrix of a length n product of positive powers of $A(k)$ and $B(k)$ grow faster (as functions of n) in the alternating product of $A(k)$ and $B(k)$. This is formalized below.

Proposition 2 ([1, Proposition 1]) *Let $w_n(a, b)$ be an arbitrary positive word of even length n, and let $W_n = w_n(A(k), B(k))$ with $k \geq 2$. Let $C_n = (A(k) \cdot B(k))^{n/2}$. Then:*

(1) The sum of entries in any row of C_n is at least as large as the sum of entries in any row of W_n.
(2) The largest entry of C_n is at least as large as the sum of entries of W_n.

Lemma 1 ([1, Lemma 1]) *Let (x, y) be a pair of positive integers, $x \neq y$, and let $k \geq 2$. One can apply transformations of the following two kinds: Transformation R takes (x, y) to $(x, y + kx)$; transformation L takes (x, y) to $(x + ky, y)$. Among all sequences of these transformations of the same length, the sequence where R and L alternate results in:*

(1) The largest sum of elements in the final pair;
(2) The largest maximum element in the final pair.

Thus, we consider powers of the matrix

$$C(k) := A(k)B(k) \tag{1}$$

to get to entries larger than p “as quickly as possible.”

3.3.1 Powers of $C(2)$

As seen in the work of the authors, Shpilrain and Vdovina [1], there are no collisions of the form

$$u(A(2), B(2)) = v(A(2), B(2))$$

if positive words u and v are of length less than $\log_{\sqrt{3+\sqrt{8}}} p$. In particular, the girth of the Cayley graph of the semigroup generated by $A(2)$ and $B(2)$ (considered as matrices over $\mathbb{F}_p$) is at least $\log_{\sqrt{3+\sqrt{8}}} p$.

The base of the logarithm here is $\sqrt{3+\sqrt{8}} \approx 2.4$. Thus, for example, if p is on the order of 2^{256}, then there are no collisions of the form $u(A(2), B(2)) = v(A(2), B(2))$ if positive words u and v are of length less than 203.

3.3.2 Powers of $C(3)$

If we consider the matrices $A(3)$ and $B(3)$ as generators of $\mathrm{SL}(2, \mathbb{F}_p)$, there are no collisions of the form

$$u(A(3), B(3)) = v(A(3), B(3))$$

if positive words u and v are of length less than $2 \log_{\frac{11+\sqrt{117}}{2}} p = \log_{\sqrt{\frac{11+\sqrt{117}}{2}}} p$. In particular, the girth of the Cayley graph of the semigroup generated by $A(3)$ and $B(3)$ (considered as matrices over $\mathbb{F}_p$) is at least $\log_{\sqrt{\frac{11+\sqrt{117}}{2}}} p$.

The base of the logarithm here is $\sqrt{\frac{11+\sqrt{117}}{2}} \approx 3.3$. For example, if p is on the order of 2^{256}, then there are no collisions of the form $u(A(2), B(2)) = v(A(2), B(2))$ if positive words u and v are of length less than 149.

3.4 Conclusions

First, the lifting attack by Tillich and Zémor [16] which produces explicit relations of length $O(\log p)$ in the monoid generated by $A(1)$ and $B(1)$ can be used in conjunction with Sanov's result [14] and some results from [5] to efficiently produce relations of length $O(\log p)$ in the group generated by $A(2)$ and $B(2)$. Generically, the relations produced by this method will involve both positive and negative powers of $A(2)$ and $B(2)$. Therefore, this method does not produce collision for the corresponding hash function, since the hash function only uses positive powers of $A(2)$ and $B(2)$.

Since there is no known analog of Sanov's result for $A(3)$ and $B(3)$, at this time there is no known efficient algorithm for even producing relations of length $O(\log p)$ in the group generated by $A(3)$ and $B(3)$, let alone in the monoid. We note that by the pigeonhole principle, such relations do in fact exist.

We have computed an explicit lower bound of $\log_b p$ for the length of relations in the monoid generated by $A(2)$ and $B(2)$, where $b \approx 2.4$. For the monoid generated by $A(3)$ and $B(3)$, we have a similar lower bound with base $b \approx 3.3$.

We conclude that at this time, there are no known attacks on hash functions corresponding to the pair $A(2)$ and $B(2)$ nor on the pair $A(3)$ and $B(3)$. Therefore, there is no visible threat to their security.

3.5 Problems for Future Research

We list here some problems for future research on these Cayley hash functions.

1. Find a description, similar to Sanov's, for matrices in the *monoid* generated by $A(2)$ and $B(2)$ over $\mathbb{Z}$.

2. Find an analog of "Sanov's form" for the subgroup of $\mathrm{SL}(2, \mathbb{Z})$ generated by $A(3), B(3)$.

3. Determine which words in the matrices $A(1), B(2)$ will have the fastest growth of their entries, i.e., find analogs to Proposition 2 and Lemma 1.

This problem is of interest because if we can show the alternating product again has fastest growth, then a similar calculation as was done for $A(2), B(2)$ and for $A(3), B(3)$ would show a lower bound with a smaller base. This means that the base of the logarithm is $\sqrt{2+\sqrt{3}}$, which is about 1.93. So this would mean that for p on the order of 2^{256}, there will be no collisions of the form $u(A(1), B(2)) = v(A(1), B(2))$ if positive words u and v are of length less than $269 = \log_{\sqrt{2+\sqrt{3}}}(p)$. This is a stronger bound than for either the $A(2), B(2)$ case or the $A(3), B(3)$ case.

4 Computations and Efficiency

In this section, we include results of some experiments done to test the efficiency of the hashes proposed in [1]. We hash with 2×2 matrices over $\mathbb{F}_p$ for a large prime p.

We conducted several tests, performed on a computer with an Intel Core i7 quad-core 4.0 GHz processor and 16 GB of RAM, running Linux Mint version 17.1 with Python version 3.4.1 and NumPy version 1.9.1.

Working with 2×2 matrices over a large field $\mathbb{F}_p$ for large prime p, we note that multiplication of the matrices themselves is quite fast (can be done in 7 multiplications), but reduction modulo p takes more work. To test the efficiency with multiplication in $\mathrm{SL}_2(\mathbb{F}_p)$, we conducted two experiments, both with $p = 2^{127} - 1$. In the first, we chose a random number between 1 and 1,000,000, found a matrix M as a word in $A(2)$ and $B(2)$ of that length, and then computed that it took approximately 80 ms to compute $M^{10,000}$. In the second experiment, we determined that it took approximately 30 milliseconds to compute a matrix as a word of length 10,000 in $A(2)$ and $B(2)$ over $\mathbb{F}_{2^{127}-1}$.

For comparison, see [3] for performance results of various cryptographic functions. In particular, SHA-512 hashes approximately 99 MiB/second (MiB stands for mebibyte, and 1 MiB $= 2^{20}$ bytes) and so this is roughly 10^8 bytes per second. Our proposed hash (the second experiment) also hashes approximately 10^8 bytes per second. Moreover, SHA-512 has been optimized; our hash performs at this speed without any optimization. For instance, our computations involve performing the reduction modulo p at each step.

Also, our computation can be parallelized, whereas SHA-512 (and others in the SHA family) cannot. This is because our bit strings can be broken up into smaller parts, hashed, and then "put back together": For instance, if H denotes the hash function, and the message $M = ABC$, then $H(M) = H(ABC) = H(A)H(B)H(C)$. This is not true with SHA hashes.

References

1. L. Bromberg, V. Shpilrain, A. Vdovina, Navigating the cayley graph of $\mathrm{SL}_2(\mathbb{F}_p)$ and applications to hashing. Semigroup Forum **94**, 314–324 (2017)
2. P. Camion, Can a fast signature scheme without secret key be secure?, in *Applied Algebra, Algorithmics and Error-Correcting Codes (Toulouse, 1984)*, volume 228 of Lecture Notes in Computer Science (Springer, Berlin, 1986), pp. 215–241
3. W. Dai, Crypto++ 5.6.0 benchmarks
4. W. Diffie and M. E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, IT-22(6):644–654, 1976
5. D.B.A. Epstein, J.W. Cannon, D.F. Holt, S.V.F. Levy, M.S. Paterson, W.P. Thurston, *Word Processing in Groups* (Jones and Bartlett Publishers, Boston, MA, 1992)
6. M. Grassl, I. Ilić, S. Magliveras, R. Steinwandt, Cryptanalysis of the Tillich-Zémor hash function. J. Cryptol. **24**(1), 148–156 (2011)
7. J. Lafferty, D. Rockmore, Numerical investigation of the spectrum for certain families of Cayley graphs, in *Expanding Graphs (Princeton, NJ, 1992)*, volume 10 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science (American Mathematical Society, Providence, RI, 1993), pp. 63–73
8. A. Lubotzky, *Discrete Groups, Expanding Graphs and Invariant Measures*, volume 125 of Progress in Mathematics (Birkhäuser Verlag, Basel, 1994). With an appendix by Jonathan D. Rogawski
9. G.A. Margulis, Explicit constructions of graphs without short cycles and low density codes. Combinatorica **2**(1), 71–78 (1982)
10. G.A. Margulis, Explicit group-theoretic constructions of combinatorial schemes and their applications in the construction of expanders and concentrators. Problemy Peredachi Informatsii **24**(1), 51–60 (1988)
11. C. Petit, Cryptographic hash functions from expander graphs. Ph.D. thesis (University College London, 2009)
12. C. Petit, J.-J. Quisquater, Preimages for the Tillich-Zémor hash function, in *Selected Areas in Cryptography*, volume 6544 of Lecture Notes in Computer Science (Springer, 2011), pp. 282–301
13. C. Petit, J.-J. Quisquater, Rubik's for cryptographers. Not. Am. Math. Soc. **60**(6), 733–740 (2013)
14. I.N. Sanov, A property of a representation of a free group. Doklady Akad. Nauk SSSR (N. S.) **57**, 657–659 (1947)
15. P. Sarnak, *Some Applications of Modular Forms*. Cambridge Tracts in Mathematics, vol. 99 (Cambridge University Press, Cambridge, 1990)

16. J.-P. Tillich, G. Zémor, Group-theoretic hash functions, in *Algebraic Coding (Paris, 1993)*, volume 781 of Lecture Notes in Computer Science (Springer, Berlin, 1994), pp. 90–110
17. J.-P. Tillich, G. Zémor, Hashing with SL_2, in *Advances in Cryptology—CRYPTO '94*, volume 839 of Lecture Notes in Computer Science (Springer, Berlin, 1994), pp. 40–49
18. G. Zémor, Hash functions and Cayley graphs. Des. Codes Cryptogr. **4**(4), 381–394 (1994)

Numerical Sets, Core Partitions, and Integer Points in Polytopes

Hannah Constantin, Ben Houston-Edwards and Nathan Kaplan

Abstract We study a correspondence between numerical sets and integer partitions that leads to a bijection between simultaneous core partitions and the integer points of a certain polytope. We use this correspondence to prove combinatorial results about core partitions. For small values of a, we give formulas for the number of (a, b)-core partitions corresponding to numerical semigroups. We also study the number of partitions with a given hook set.

Keywords Numerical semigroups · Numerical sets · Core partitions
Simultaneous core partitions · Hook sets of partitions

1 Introduction

A large number of recent papers have studied statistical questions about sizes of simultaneous core partitions [1, 2, 5, 11, 14, 19, 25, 26, 28–31]. One of the larger successes in this area is Johnson's proof of Armstrong's conjecture, which we state as Theorem 3 below [19]. Broadly, these problems address questions of the following type: Given a finite set of partitions, for example, the set of simultaneous (a, b)-core partitions, what can we say about statistical properties of their sizes? We use a correspondence between numerical sets and partitions to study these types of questions for partitions coming from families of numerical semigroups and for partitions with a fixed hook set.

H. Constantin
Department of Mathematics, University of Toronto, Toronto, ON M5S 2E4, Canada
e-mail: hconstan@math.utoronto.ca

B. Houston-Edwards
Department of Mathematics, Yale University, New Haven, CT 06511, USA
e-mail: benjamin.houston-edwards@yale.edu

N. Kaplan (✉)
Department of Mathematics, University of California, Irvine, CA 92697, USA
e-mail: nckaplan@math.uci.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_7

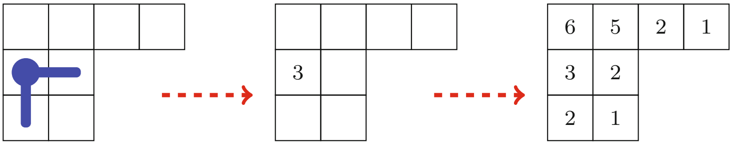


Fig. 1 Young diagram and hook lengths of the partition (4, 2, 2). This partition is both a 4-core and a 7-core

We first briefly introduce some notation necessary to explain our main results. A *partition* λ of n is a sequence of positive integers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 1$ whose sum is n. We refer to the λ_i as the *parts* of the partition λ. We represent a partition by its *Young diagram*, a series of left aligned rows of boxes in which there λ_i boxes in row i. For any box of the Young diagram, its *hook length* is the number of boxes directly to the right of it, plus the number of boxes directly below it, plus one for the box itself. We denote by $H(\lambda)$ and $\mathcal{H}(\lambda)$ the *hook set* and *hook multiset* of λ—the set and multiset of hook lengths, respectively (Fig. 1)

Hook lengths play an important role in the representation theory of the symmetric group. For example, the Frame-Robinson-Thrall hook-length formula [15] expresses the dimension of the irreducible representation π_λ of S_n corresponding to a partition λ of n:

$$\dim \pi_\lambda = \frac{n!}{\prod_{h \in \mathcal{H}(\lambda)} h}.$$

A partition λ of n with no hook lengths divisible by a is called an *a-core partition* or more simply, an *a-core*. When a is prime the corresponding irreducible representations have maximal a-adic valuation and play a role in the modular representation theory of S_n [17].

There has been an explosion of recent papers studying enumerative questions about special classes of a-core partitions. The set of a-cores is clearly infinite but the number of partitions that are both a-cores and b-cores, *simultaneous* (a, b)*-cores*, is finite. Similarly, an $(a_1, a_2, \ldots, a_k)$-core partition is an a_i-core for all $i \in [1, k]$. Anderson gives a nice formula for the number of simultaneous (a, b)-core partitions by establishing a bijection with a certain set of Dyck paths.

Theorem 1 [3, Theorem 1] *For coprime a and b, the number of simultaneous* (a, b)*-core partitions is* $\frac{1}{a+b}\binom{a+b}{a}$.

It is natural to ask about the sizes of the partitions making up this finite set. A formula for the size of the largest simultaneous (a, b)-core partition was first given by Olsson and Stanton.

Theorem 2 [24, Theorem 4.1] *For relatively prime positive integers a and b, the largest* (a, b)*-core has size* $(a^2 - 1)(b^2 - 1)/24$. *Moreover, there is a unique* (a, b)*-core of this size.*

Different proofs have been given by Tripathi [27] and Johnson [19].

In 2011, Armstrong conjectured that the average size of an (a, b)-core partition has a simple relation to the maximum size [5]. This conjecture was proven in the special case where $b = a + 1$ by Stanley and Zanello [25] and more generally when $b \equiv 1 \pmod{a}$ by Aggarwal [1]. The *conjugate* of a partition $\lambda = (\lambda_1, \dots, \lambda_k)$ is the partition $\widetilde{\lambda} = (\lambda'_1, \dots, \lambda'_\ell)$ where λ'_j is the number of parts of $\lambda \geq j$. This partition comes from exchanging the rows and columns of the Young diagram of λ. A partition is *self-conjugate* if it is equal to its conjugate. Armstrong's conjecture was proven for self-conjugate (a, b)-cores by Chen, Huang, and Wang [11]. After these partial results, the full conjecture was proven by Johnson [19]. Another proof was given by Wang [28].

Theorem 3 [19, Theorem 1.7] *For relatively prime positive integers a and b, the average size of an (a, b)-core partition is $(a + b + 1)(a - 1)(b - 1)/24$.*

Johnson's work [19] is of special interest to us because he proves Theorem 3 by studying a bijection of a-core partitions with the lattice

$$A_{a-1} = \left\{ (x_1, \dots, x_a) \in \mathbb{Z}^a : \sum_{i=1}^{a} x_i = 0 \right\}$$

under which the simultaneous (a, b)-cores correspond to the integer points of a rational simplex. This bijection is given by the "signed abacus construction." Under this bijection, the size of a partition is given by a certain quadratic function [19]. Johnson's works also gives the ability to compute higher moments of the distribution of the sizes of simultaneous (a, b)-cores, a problem also addressed in [14]. Thiel and Williams also consider these higher moments and extend this approach to affine Weyl groups [26].

Our approach is similar to Johnson's in that we study a correspondence between simultaneous (a, b)-core partitions and the integer points of a rational polytope; however, we do not use the abacus construction. Instead, we study a bijection φ between partitions and *numerical sets*, subsets of $\mathbb{N} = \{0, 1, 2, \dots\}$ that contain 0 and have finite complement. A numerical set that is closed under addition is called a *numerical semigroup*. The bijection is given by considering the *profile* of a partition λ, the sequence of southmost and eastmost edges of its Young diagram. These steps are labeled by elements of $\mathbb{N}$ starting with the lower left corner of the Young diagram and moving to the upper right where the vertical steps exactly correspond to the elements of the complement of the associated numerical set. This bijection is explained in detail in [21] and is related to the Dyck path construction in [10]. The number of parts in the partition is equal to the size of the complement of the numerical set, and the hook set can be easily calculated from the numerical set. Moreover, Keith and Nath use this bijection and basic facts about numerical sets to show the following.

Theorem 4 [21, Theorem 1] *Let $a_1, \dots, a_k$ be distinct positive integers. The number of simultaneous $(a_1, \dots, a_k)$-cores is finite if and only if* $\gcd\{a_1, \dots, a_k\} = 1$.

Another proof of this result is given in [29]. In Sect. 4, we show something stronger that when $\gcd\{a_1, \dots, a_k\} = 1$ the bijection φ takes the set of simultaneous $(a_1, \dots, a_k)$-cores to the lattice points of a rational polytope whose defining half-spaces we explicitly describe. In general, it is still an open problem to give formulas for the number of such points in terms of $a_1, \dots, a_k$. The particular cases of $(s, s+1, s+2)$-cores and $(s, s+1, \dots, s+k)$-cores have been addressed in [31] and [2], respectively.

We use results of Marzuola and Miller about the *atom monoid* associated to a numerical set [23] along with the Kunz coordinate vector of a numerical semigroup, described by Blanco and Puerto in [8], to give further bijections involving a-cores. The atom monoid of a numerical set T is defined by

$$A(T) = \{n \in \mathbb{N} : n + T \subseteq T\}.$$

Note that $A(T) \subseteq T$ since $0 \in T$, and $A(T) = T$ if and only if T is a numerical semigroup. The atom monoid is always closed under addition, so in some sense $A(T)$ is the underlying numerical semigroup of T. For a numerical semigroup S containing a, the associated *Apéry tuple* is $\mathrm{Ap}(S) = (x_1, \dots, x_{a-1}) \in \mathbb{N}^{a-1}$, where $ax_i + i$ is the smallest element of the numerical semigroup congruent to i mod a. The Apéry set of S is $\{0, ax_1 + 1, \dots, ax_{a-1} + a - 1\}$. The definition of $\mathrm{Ap}(S)$ depends on a, but the specific value of a we choose will always be clear from context. We can directly extend this definition to numerical sets T with $a \in A(T)$. We summarize our bijections as a proposition that we prove in Sect. 3.

Proposition 1 *The map φ described above gives a bijection between the set of a-core partitions and the set of numerical sets T with $a \in A(T)$. The map taking a numerical set T with $a \in A(T)$ to its Apéry tuple gives a bijection between these numerical sets and* $\mathbb{N}^{a-1}$.

We then use these correspondences to answer enumerative questions about hook sets of partitions. For example, we give another proof of the following result of Berg and Vazirani.

Proposition 2 *[7, Proposition 3.1.4] The number of a-cores with g parts is equal to the number of* $(a-1)$*-cores with less than or equal to g parts.*

Both Johnson and Chen, Huang, and Wang give proofs that the self-conjugate (a, b)-core partitions have the same average size as the set of all (a, b)-core partitions [11, 19]. It is natural to ask whether sizes of other subfamilies of (a, b)-cores have similar statistical properties. We focus on two particular cases; (a, b)-cores that correspond to numerical semigroups under the map φ, and the set of all partitions with a given hook set. Computational evidence suggests that the average size of an (a, b)-core corresponding to a numerical semigroup is not equal to the average size of all (a, b)-cores.

In this setting, we do not even have an analoge of Anderson's theorem on the number of these partitions. This is equivalent to asking for the number of semigroups containing a and b. For a set of nonnegative integers $n_1, \dots, n_t$, we define the numerical semigroup *generated by* them to be

$$\langle n_1, \dots, n_t \rangle = \left\{ \sum_{i=1}^{t} a_i n_i \mid a_i \in \mathbb{N} \right\}.$$

Note that any semigroup containing a and b also contains $\langle a, b \rangle$. A semigroup T containing a numerical semigroup S is called an *oversemigroup* of S. Let $O(S)$ denote the number of oversemigroups of S. Using a characterization due to Branco, García-García, García-Sanchez, and Rosales for when an Apéry tuple corresponds to a numerical semigroup we show that $O(\langle a, b \rangle)$ is equal to the number of lattice points in a certain rational polytope [9]. Hellus and Waldi have also studied this problem, giving formulas for small a and bounds for the general case [18]. We state their main result as Theorem 8.

We give our own calculations for $a \leq 4$ using using different methods.

Theorem 5 *If* $S = \langle 3,\ 6k + \ell \rangle$ *with* $\ell \in \{1, 2, 4, 5\}$*, then*

$$O(S) = (3k + \ell)(k + 1).$$

Theorem 6 *Suppose that* $S = \langle 4, 12k + \ell \rangle$ *with* $\ell \in \{1, 3, 5, 7, 9, 11\}$. *Then* $O(S)$ *is given by the following chart:*

ℓ	$O(S)$
1	$24k^3 + 30k^2 + 11k + 1$
3	$24k^3 + 42k^2 + 23k + 4$
5	$24k^3 + 54k^2 + 39k + 9$
7	$24k^3 + 66k^2 + 59k + 17$
9	$24k^3 + 78k^2 + 83k + 29$
11	$24k^3 + 90k^2 + 111k + 45$

It is not difficult using the bijection φ to show that the hook set of a partition is always the complement of a numerical semigroup. Let $\mathbb{N} \setminus S$ denote the complement of some numerical semigroup S. If a and b are not in $\mathbb{N} \setminus S$ then any partition with this hook set is a simultaneous (a, b)-core. In order to study statistical questions about sizes of partitions with a given hook set, we would first like to understand how many partitions have this hook set. We call this number $P(S)$. We investigate how the properties of S affect the behavior of this function, giving some results and suggesting questions for future work. The *Frobenius number* of a numerical set T is the largest element of its complement and is denoted $F(T)$. The size of the complement is called the *genus* of T, and the elements of the complement are called the *gaps* of T. These concepts play important roles in our analysis of this problem.

The study of the set of partitions with a given hook set fits in nicely with previous work of Chung and Herman [12], and of Craven [13], on partitions with equal hook multisets. In [12], the authors show that a partition is uniquely determined up to reflection by its *extended hook multiset*, in which hook lengths can take negative values. However, they also show that arbitrarily many distinct partitions can have the same hook multiset. This result has been vastly generalized by Craven.

Theorem 7 [13, Theorem 1.4] *Let k and ℓ be natural numbers. The for all sufficiently large n, there are k disjoint sets of ℓ partitions of n, such that all of the ℓ partitions in each set have the same multiset of hook numbers, and distinct sets contain partitions with different hook numbers, and moreover different products of hook numbers.*

Craven proves this result by defining certain classes of partitions called *enveloping partitions* that have the same hook multiset as many other partitions. It is natural to ask what are the properties of the numerical semigroups giving the underlying hook sets of these partitions that make them suitable for this construction. We focus on the opposite extreme. A numerical set T is called *symmetric* if for every $i \in [0, F(T)]$ exactly one of $i, F(T) - i$ is in T. We prove that there is a unique partition with a given hook set if and only if that hook set is the complement of a symmetric numerical semigroup. We also investigate the relationship between the function $P(S)$ and the number of *missing pairs* of S, that is, the number of pairs $i, F(S) - i$ in the complement of S with $i \in [0, F(S)/2]$.

We conclude the paper by discussing some asymptotic questions and conjectures based on computational evidence.

2 The Correspondence Between Numerical Sets and Partitions

We first explain the bijection φ introduced in the previous section connecting numerical sets to partitions and use it to find the relationship between atom monoids and hook sets. We begin with an example.

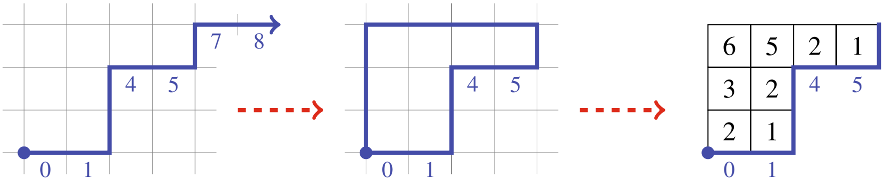
Example 1 Let $T = \{0, 1, 4, 5, 7, \rightarrow\}$, where "$\rightarrow$" means that T contains every integer greater than 7, as in the conventions of [16]. Clearly T is a numerical set with $F(T) = 6$, $g(T) = 3$, and $A(T) = \{0, 4, 5, 7, \rightarrow\}$.

Given a numerical set T, we construct a partition $\varphi(T)$ such that the map φ is a bijection from numerical sets to partitions. We construct $\varphi(T)$ by defining the profile of its Young diagram. We can think of this path as lying in $\mathbb{Z}^2$ with the bottom left corner of the Young diagram at the origin. Starting with $n = 0$:

- if $n \in T$ draw a line of unit length to the right,
- if $n \notin T$ draw a line of unit length up,
- repeat for $n + 1$.

For any n greater than the Frobenius number of T we draw a line to the right. As T is a numerical set this process ends with an infinite set of steps to the right. We disregard this section, forming the Young diagram with this profile walk, the line $x = 0$, and this horizontal line. The construction is understood most clearly with an example.

Example 2 If $T = \{0, 1, 4, 5, 7, \rightarrow\}$, then $\varphi(T) = (4, 2, 2)$:



Seeing that φ is a bijection is simple: to find the inverse image of a partition λ label the profile of the Young diagram as above, starting with 0. The complement of the numerical set $\varphi^{-1}(\lambda)$ consists of the positive integers labeling the vertical steps of the profile.

We give some basic properties of φ here, some of which might be evident from the example. These results are clear from [21], but we include them with proofs for completeness.

Proposition 3 *Given a numerical set T, the hook multiset of $\varphi(T)$ is*

$$\mathcal{H}(\varphi(T)) = \{n - t : n \notin T, t \in T, n > t\}.$$

Proof Consider a box B in the Young diagram of $\varphi(T)$ such that B is in the same column as the horizontal step on the profile associated to $t \in T$ in the construction of $\varphi(T)$, and the same row the vertical step associated to $n \notin T$.

Recall that the hook of B is the set of boxes to the right (the "arm"), the set of boxes below (the "leg"), and B itself. Counting steps along the profile shows that $n - t$ is the hook length of B. □

Proposition 4 *Given a numerical set T, the hook set of $\varphi(T)$ is the complement of its atom monoid: $H(\varphi(T)) = \mathbb{N} \setminus A(T)$.*

Proof By Proposition 3, this amounts to proving that $\mathbb{N} \setminus A(T) = D$, where $D = \{n - t : n \notin T, t \in T, n > t\}$.

Suppose $x \in \mathbb{N} \setminus A(T)$, so $x + t \notin T$ for some $t \in T$. This implies $(x + t) - t = x \in D$. Conversely, if $x \in D$ then $x = n - t$ for some $n \notin T$ and $t \in T$. This implies $x + t = n \notin T$, so $x \in \mathbb{N} \setminus A(T)$. □

Remark 1 In particular, since φ is bijective, Proposition 4 shows that the hook set of any partition is the complement of a numerical semigroup. This implies that a partition is an a-core if and only if a is not in its hook set, a simpler condition than having no hook lengths divisible by a.

3 The Correspondence of a-Cores and $\mathbb{N}^{a-1}$

In this section, we use the bijection between a-core partitions and $\mathbb{N}^{a-1}$ to prove several combinatorial results. This correspondence comes from taking the Apéry tuple of the numerical set associated to an a-core partition via the map φ. By Proposition 4, a partition λ is an a-core if and only if the atom monoid of $\varphi^{-1}(\lambda)$ contains a. For $a \in A(T)$, we have $n + a \in T$ for any $n \in T$, which means that if $x_i \in \mathrm{Ap}(T)$, $x_i + ka \in T$ for any $k \in \mathbb{N}$. This shows that the Apéry set and Apéry tuple uniquely determine a numerical set whose atom monoid contains a.

Consider $(x_1, \ldots, x_{a-1}) \in \mathbb{N}^{a-1}$ and the associated numerical set $T = \{ax_i + i + ma : m \in \mathbb{N}, 1 \leq i \leq a-1\}$. We see that $a \in A(T)$ and $\mathrm{Ap}(T) = (x_1, \ldots, x_{a-1})$. Hence $\mathbb{N}^{a-1}$ is in bijection with numerical sets whose atom monoid contains a.

In summary, we have the following one-to-one correspondences:

$$\left\{ \begin{array}{c} a\text{-core} \\ \text{partitions} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{numerical sets} \\ \text{whose atom monoid} \\ \text{contains } a \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{the tuples} \\ \text{of } \mathbb{N}^{a-1} \end{array} \right\},$$

completing the proof of Proposition 1. Note that the origin of $\mathbb{N}^{a-1}$ corresponds with the numerical set $T = \mathbb{N}$, which corresponds with the empty partition, an a-core for any a.

Recall that the Frobenius number $F(T)$ of the numerical set T is the maximum element of its complement. Note that $F(T) \notin A(T)$ but $n \in A(T)$ for any $n > F(T)$. By Proposition 4, the maximum hook length of $\varphi(T)$ is $F(T)$. Also, if $a \in A(T)$ and $\mathrm{Ap}(T) = (x_1, \ldots, x_{a-1})$, then $F(T) = \max\{ax_i + i - a\}$. The above bijection allows us to easily compute the number of a-core partitions by maximum hook length.

Proposition 5 *For* $1 \leq \ell \leq a - 1$*, the number of* a*-core partitions with maximum hook length* $ak + \ell$ *is* $(k+2)^{\ell-1}(k+1)^{a-\ell-1}$.

Proof The a-core partitions with maximum hook length $ak + \ell$ are those for which $\max\{ax_i + i - a\} = ax_\ell + \ell - a$ where $x_\ell = k + 1$. This implies $x_\ell > x_i$ for any $i > \ell$, and $x_\ell \geq x_i$ for any $i < \ell$. Therefore, such partitions are in bijection with choices for the x_i satisfying $x_i \in [0, k]$ for any $i \in [\ell + 1, a - 1]$ and $x_i \in [0, k + 1]$ for any $i \in [1, \ell - 1]$. □

Proposition 6 *For any* $k \in \mathbb{N}$*, the number of* a*-core partitions with maximum hook length less than* ak *is* $(k+1)^{a-1}$.

Proof An a-core partition λ has maximum hook length less than ak if and only if $\max\{ax_i + i - a\} < ak$, where $(x_1, \ldots, x_{a-1})$ is the Apéry tuple of the corresponding numerical set. This holds if and only if $x_i \leq k$ for each i. Therefore, the a-core partitions with maximum hook length less than ak are those which correspond with the lattice points of $[0, k]^{a-1} \subset \mathbb{N}^{a-1}$. □

We can similarly find the number of a-cores with a fixed number of parts. The construction of $\varphi(T)$ from T shows that the number of parts is equal to the size of $\mathbb{N} \setminus T$. So the number of parts of $\varphi(T)$ is equal to $g(T)$.

Proposition 7 *The number of a-core partitions with g parts is* $\binom{g+a-2}{a-2}$.

Proof Suppose $a \in A(T)$ and $\mathrm{Ap}(T) = (x_1, \dots, x_{a-1})$. If $n \equiv i \pmod{a}$ then $n \in T$ if and only if $n \geq ax_i + i$. Therefore, the genus of T is $x_1 + \cdots + x_{a-1}$, and the number of a-cores with g parts is equal to the number of points of the simplex $(x_1, \dots, x_{a-1}) \in \mathbb{N}^{a-1}$ such that $x_1 + \cdots + x_{a-1} = g$. It is well known that there are $\binom{g+a-2}{a-2}$ such points. □

Proposition 8 *The number of a-core partitions with less than or equal to g parts is* $\binom{g+a-1}{a-1}$.

Proof The number of numerical sets T with $a \in A(T)$ and genus less than or equal to g is the number of points $(x_1, \dots, x_{a-1}) \in \mathbb{N}^{a-1}$ such that $x_1 + \cdots + x_{a-1} \leq g$. Counting these points is the same as counting the number of points $(x_1, \dots, x_{a-1}, y) \in \mathbb{N}^a$ such that $x_1 + \ldots + x_{a-1} + y = g$. Therefore, there are $\binom{g+a-1}{a-1}$ such points. □

These two results together give another proof Berg and Vazirani's Proposition 2 stated in the introduction [7].

Since the conjugate of an a-core partition is also an a-core, the number of a-cores with g parts is equal to the number of a-cores with largest part g. Hence, Propositions 2, 7, and 8, may be restated with "largest part g" in place of "g parts."

We close this section by giving another interpretation of [19, Theorem 1.9] where Johnson relates the size of a partition corresponding to a quadratic function evaluated at the associated lattice point. Since our correspondence between core partitions and lattice points is different we get a different function, but the ideas are similar.

Proposition 9 *Let T be a numerical set with $a \in A(T)$ and* $\mathrm{Ap}(T) = (x_1, \dots, x_{a-1})$. *Then the size of the partition $\varphi(T)$ is*

$$\begin{aligned} F_a(x_1, \dots, x_{a-1}) &= \frac{a}{2}\sum_{i=1}^{a-1} x_i(x_i - 1) + \sum_{i=1}^{a-1} ix_i - \frac{1}{2}\left(\sum_{i=1}^{a-1} x_i\right)\left(-1 + \sum_{i=1}^{a-1} x_i\right) \\ &= \frac{a-1}{2}\sum_{i=1}^{a-1} x_i^2 + \sum_{i=1}^{a-1}\left(i - \frac{a-1}{2}\right)x_i - \sum_{1 \leq i < j \leq a-1} x_i x_j. \end{aligned}$$

Proof In the proof of Proposition 8, we noted that the genus of a numerical set T is the sum of the elements of the corresponding Apéry tuple. As noted above, the number of parts of $\varphi(T)$, which is equal to the number of rows of its Young diagram, is given by the genus of T. By Proposition 3, the hooks in the first column of the Young diagram are exactly the elements of $\mathbb{N} \setminus A(T)$. By the definition of a hook, the sum of these hook lengths is almost the size of $\varphi(T)$, except that we have overcounted the ith box from the top $i - 1$ times. This means we have overcounted $(g(T) - 1)g(T)/2$ boxes in the Young diagram and the size of $\varphi(T)$ is the sum of the gaps of T minus $(g(T) - 1)g(T)/2$.

If $ax_i + i$ is the smallest element of T congruent to $i \pmod a$, then the sum of gaps congruent to i is

$$\sum_{n=0}^{x_i-1} an + i = \frac{ax_i(x_i - 1)}{2} + ix_i.$$

Summing over all $i \in [1, a-1]$ and using $g(T) = \sum_{i=1}^{a-1} x_i$ completes the proof. □

4 The (a, b)-Core Polytope

In this section, we use the bijections of Proposition 1 to prove Theorem 4 and the stronger result that simultaneous $(a, b_1, \dots, b_m)$-core are in bijection with lattice points of a polytope that we define below. For now, we do not necessarily assume that $\gcd(a, b) = 1$ but we do assume that $a \nmid b$. Suppose that $b = ak + \ell$ where $\ell \in [1, a-1]$ and that T is a numerical set such that $\varphi(T)$ is an a-core partition with Apéry tuple $\mathrm{Ap}(T) = (x_1, \dots, x_{a-1})$. By the remark following Proposition 4, $\varphi(T)$ is a b-core partition if and only if $b \in A(T)$, which is true if and only if $ax_i + i + b \in T$ for all $i \in [1, a-1]$.

If $i + \ell < a$ then $ax_i + i + b \in T$ if and only if $ax_i + i + b \geq ax_{i+\ell} + (i + \ell)$. Similarly, if $i + \ell > a$ then $ax_i + i + b \in T$ if and only if $ax_i + i + b \geq ax_{i+\ell-a} + (i + \ell - a)$. Therefore, $\varphi(T)$ is a b-core if and only if $\mathrm{Ap}(T)$ satisfies the inequalities

$$\begin{aligned} x_\ell &\leq k, && \\ x_{i+\ell} &\leq k + x_i, && \text{if } i + \ell < a, \\ x_{i+\ell-a} &\leq k + x_i + 1, && \text{if } i + \ell > a, \\ x_i &\geq 0. && \end{aligned}$$

Let $\mathcal{P}_{a,b} \subseteq \mathbb{R}^{a-1}$ be the region defined by the intersection of these half-spaces. This is a rational polyhedral cone and is a rational polytope if and only if it is bounded, which is true if and only if $\gcd(a, b) = 1$. We now state and prove a more general result.

Proposition 10 *Suppose* $\gcd(a, b_1, \dots, b_m) = 1$ *where we write* $b_j = ak_j + \ell_j$ *for each* $j \in [1, m]$ *with* $\ell_j \in [1, a-1]$. *There is a bijection between* $(a, b_1, \dots, b_m)$-*core partitions and the integer points of the polytope defined by the following inequalities:*

$$\begin{aligned} x_{\ell_j} &\leq k_j, && && (1) \\ x_{i+\ell_j} &\leq k_j + x_i, && \textit{if } i + \ell_j < a, && (2) \\ x_{i+\ell_j-a} &\leq k_j + x_i + 1, && \textit{if } i + \ell_j > a, && (3) \\ x_i &\geq 0 && \textit{for } i \in [1, a-1], && (4) \end{aligned}$$

where we have one set of inequalities (1), (2), and (3) for each $j \in [1, m]$.

Proof Let $\mathcal{Q}$ be the intersection of the half-spaces defined by these inequalities and note that $\mathcal{Q} = \bigcap_{j=1}^{m} \mathcal{P}_{a,b_j}$. The lattice points of this region are in bijection with $(a, b_1, \ldots, b_m)$-cores, so we need only show that $\mathcal{Q}$ is bounded. Suppose $(x_1, \ldots, x_{a-1}) \in \mathcal{Q}$ with each x_i a nonnegative integer. We give an upper bound on each x_i that depends only on $a, b_1, \ldots, b_m$, which completes the proof.

Since $(x_1, \ldots, x_{a-1})$ satisfies (1) for each $j \in [1, m]$, we see $x_{\ell_j} \leq k_j$. After reindexing, (2) implies $x_i \leq k_j + x_{i-\ell_j}$ if $i > \ell_j$, and (3) implies $x_i \leq k_j + 1 + x_{i-\ell_j+a}$ if $i < \ell_j$. Hence

$$x_i \leq k_j + 1 + x_{i-\ell_j \pmod{a}}$$

, for each $j \in [1, m]$ where we write $x_{i \pmod{a}}$ as shorthand for $x_{i'}$ where $i' \in [1, a-1]$ and $i' \equiv i \pmod{a}$. Therefore

$$x_{\ell_{j_1}+\ell_{j_2} \pmod{a}} \leq k_{j_1} + 1 + x_{\ell_{j_2}} \leq k_{j_1} + k_{j_2} + 2.$$

Proceeding by induction, if $s = \sum_{j=1}^{m} y_j \ell_j$ for some $y_1, \ldots, y_m \in \mathbb{N}$ then

$$x_{s \pmod{a}} \leq \sum_{j=1}^{m} y_j (k_j + 1).$$

Since $\gcd(a, b_1, \ldots, b_m) = 1$, for each $i \in [1, a-1]$ there exist nonnegative integers $y_1, \ldots, y_m$ such that $\sum_{i=1}^{m} y_j \ell_j \equiv i \pmod{a}$. This gives an upper bound on each x_i depending only on $a, b_1, \ldots, b_m$. □

In particular, this proves Theorem 4. A formula for the number of integer points of the polytope $\mathcal{Q}$ is equivalent to a formula for the number of $(a, b_1, \ldots, b_m)$-cores. For example, giving such a formula in the $m = 1$ case is equivalent to Theorem 1 of Anderson. We note that several results in this area can be phrased in terms of counting integer points in special polytopes [2, 29, 31].

In general, it is difficult to give a formula for the number of integer points of a polytope in terms of the defining half-spaces but there are some circumstances in which the polytopes are particularly nice. For example, we can use this method to give another proof of Proposition 6.

Proof (Second proof of Proposition 6) The set of a-core partitions with maximum hook length less than ak is exactly the set of $(a,\ ak+1, \ldots, ak+(a-1))$-core partitions, since an (a, b)-core is also an $(a+b)$-core by Proposition 4. This set corresponds with the lattice points of the polytope $\mathcal{Q} = \bigcap_{i=1}^{a-1} \mathcal{P}_{a,\,ak+i}$. By (1) we have $\mathcal{Q} \subseteq [0, k]^{a-1}$, and by (2) and (3) we have $[0, k]^{a-1} \subseteq \mathcal{P}_{a,ak+i}$ for each i. Therefore $\mathcal{Q} = [0, k]^{a-1}$, which contains $(k+1)^{a-1}$ integer points, so there are $(k+1)^{a-1}$ a-core partitions with maximum hook length less than ak. □

We close this section with a suggestion for future research. Formulas for the number of integer points in families of rational polytopes can be be quite subtle, particularly when the polytope has vertices with large denominators. The volume of a polytope is often a good approximation for its number of integer points and is usually easier to find.

Problem 1 Give an approximation for the volume of the $(a, b_1, \dots, b_m)$-core polytope in terms of the integers $a, b_1, \dots, b_m$.

5 Counting (a, b)-Cores from Semigroups

In this section, we further investigate the correspondence between numerical sets with atom monoid containing a and a-core partitions. We focus on a natural subclass of these numerical sets, those that are actually numerical semigroups. Recall that a numerical set is a numerical semigroup if and only if it is closed under addition, or equivalently, it is equal to its atom monoid. We see that the bijection φ takes a numerical semigroup S to an a-core partition if and only if $a \in S$. Our main goal in this section is to describe the set of a-core partitions that come from numerical semigroups and to count the set of simultaneous (a, b)-cores that come from semigroups for certain pairs (a, b).

Recall from Theorem 1 that for positive integers $a, b \geq 2$ with $\gcd(a, b) = 1$ the total number of (a, b)-cores is

$$C(a, b) = \frac{1}{a+b}\binom{a+b}{a}.$$

We are interested in finding the proportion of these partitions that come from semigroups via the map φ. To do this, we first show that these partitions are in bijection with the lattice points of a polytope contained in the (a, b)-core polytope of the previous section.

A direct consequence of Proposition 4 is that for a numerical semigroup S the partition $\varphi(S)$ is an (a, b)-core if and only if $a, b \in S$. Since S is a semigroup it must also contain $\langle a, b\rangle$. Our goal is to give formulas for $O(\langle a, b\rangle)$ in terms of a and b and to investigate the ratio $O(\langle a, b\rangle)/C(a, b)$.

Hellus and Waldi have studied exactly this problem in [18]. They show that the set of oversemigroups of $\langle a, b\rangle$ are naturally in bijection with the set of integer points in a rational polytope. For a fixed and b increasing they show that computing $O(\langle a, b\rangle)$ is equivalent to counting lattice points in dilates of this polytope and that they can therefore use techniques from Ehrhart theory to study the behavior of $O(\langle a, b\rangle)$. This is notable because Ehrhart theory is also a major input of Johnson's proof of Armstrong's conjecture [19]. In particular, they prove the following result. A *quasipolynomial* of degree d is a function $f : \mathbb{N}^d \to \mathbb{C}$ of the form

$$f(n) = c_d(n)n^d + c_{d-1}(n)n^{d-1} + \cdots + c_0(n)$$

with periodic functions c_i having integer periods, $c_d \neq 0$.

Theorem 8 [18, Theorem 1.1] *Let $a \in \mathbb{N},\ a > 1$.*

1. *There is a quasipolynomial of degree $a - 1$ taking the value $O(\langle a, b\rangle)$ at each $b \in \mathbb{N}$ relatively prime to a.*
2. *The leading coefficient $c_{a-1}(n)$ of this quasipolynomial is constant and satisfies*

$$\frac{1}{(a-1)! \cdot a!} \leq c_{a-1}(n) \leq \frac{1}{(a-1) \cdot a!}.$$

3. *The function $O(\langle a, b\rangle)$ is increasing in both variables.*

Hellus and Waldi note that the upper and lower bounds of the second part of the statement coincide for $a = 2, 3$, that the upper bound is correct for $a = 4$, and that for $a = 5, 6, 7$ the correct value lies strictly between the upper and lower bound [22]. With the above theorem, finding the quasipolynomial $O(a, b)$ for fixed a can be done with a finite amount of computation. We also note that the idea of using Ehrhart theory to give quasipolynomial formulas for quantities associated to numerical semigroups also appears in [20].

We give our own calculations for $a \leq 4$, showing how to derive formulas of this type without prior knowledge that the answer is given by a quasipolynomial. We explicitly describe the $a - 1$ dimensional polytope whose integer points are in bijection with the oversemigroups of $\langle a, b\rangle$ and then divide it into $a - 2$ dimensional slices via parallel hyperplanes. We find exact formulas for the number of integer points in each slice.

Our first goal is to give defining inequalities for the polytope whose integer points are in bijection with oversemigroups of $\langle a, b\rangle$. This is equivalent to determining when an Apéry tuple $(x_1, \dots, x_{a-1})$ of a numerical set containing a actually corresponds to a numerical semigroup containing a. The following result of Branco, García-García, García-Sánchez, and Rosales, a slight variation of [9, Theorem 11], gives this characterization.

Theorem 9 [9, Theorem 11] *The map from a numerical semigroup to its Apéry tuple gives a one-to-one correspondence taking semigroups T containing a to solutions $(\ell_1, \dots, \ell_{a-1})$ of the system of inequalities*

$$x_i \in \mathbb{N} \ \textit{for all}\ i \in \{1, \dots, a-1\} \tag{5}$$

$$x_i + x_j \geq x_{i+j} \ \textit{for all}\ 1 \leq i \leq j \leq a-1,\ i + j \leq a - 1 \tag{6}$$

$$x_i + x_j + 1 \geq x_{i+j-a} \ \textit{for all}\ 1 \leq i \leq j \leq a-1,\ i + j > a. \tag{7}$$

Also, notice that $T \supseteq S$ if and only if

$$\ell_i \leq k_i \qquad \text{for all}\ i, \tag{8}$$

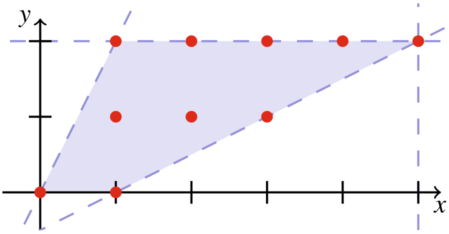
where $(k_1, \dots, k_{\{a-1\}})$ is the *Apéry tuple* of S. Therefore, the set of inequalities (6)–(8) gives necessary and sufficient conditions for T to be an oversemigroup of S.

These inequalities define an $a-1$ dimensional polytope in which the lattice points correspond exactly with the oversemigroups of S. In order to count the number of oversemigroups of S, we only need to count these lattice points. This polytope is of course contained in $\mathcal{P}_{a,b}$. Hellus and Waldi study a similar polytope, but phrase their results in terms of counting lattice paths and do not make a connection to general (a,b)-core partitions or numerical sets [18].

Example 3 Consider $S=\langle 3,8\rangle$. The inequalities (6)–(8) reduce to

$$\begin{aligned} 2x &\geq y \\ 2y+1 &\geq x \\ x &\leq 5 \\ y &\leq 2 \end{aligned}$$

which define the polytope:



Each lattice point (x,y) in this polytope uniquely corresponds to an oversemigroup of S, and thus with a $(3,8)$-core partition. There are 10 integer points in this polytope, so $O(\langle 3,8\rangle)=10$ and there are 10 simultaneous $(3,8)$-core partitions associated to numerical semigroups.

It seems difficult to give a general formula for $O(\langle a,b\rangle)$ so we begin by analyzing the cases $a=2,3,4$, finding explicit formulas for each. When $a=2$, it is clear that $O(\langle 2,2k+1\rangle)=k+1$ since any oversemigroup of $\langle 2,2k+1\rangle$ is determined uniquely by its smallest odd element. Our next goal is to prove Theorems 5 and 6 that were stated in the introduction. We note that both results express $O(\langle a,b\rangle)$ as a quasipolynomial in b of degree $a-1$, and agree with the calculations in [18].

In order to prove Theorem 5, we divide up the set of oversemigroups of $S=\langle 3,6k+\ell\rangle$ by genus. It is easy to show that the genus of $\langle a,b\rangle$ is $(a-1)(b-1)/2$ [16]. For each integer $n\in[0,6k+\ell-1]$, the genus of S we compute the number $O_n(S)$ of oversemigroups of S with genus n.

Lemma 1 *If* $S=\langle 3,\ 6k+\ell\rangle$ *with* $\ell\in\{1,2,4,5\}$ *then*

$$O_n(S)=\begin{cases} \lfloor \frac{n}{3}\rfloor+1 & 0\leq n\leq 3k+\frac{\ell}{2}-1 \\ \left\lfloor \frac{6k+\ell-1-n}{3}\right\rfloor+1 & 3k+\frac{\ell}{2}-1<n\leq 6k+\ell-1 \end{cases}.$$

Assuming Lemma 1 we prove Theorem 5.

Proof (of Theorem 5) For $3k \leq n < 3k + \ell$ Lemma 1 implies $O_n(S) = k + 1$, so we can rewrite the expression for $O_n(S)$ as

$$O_n(S) = \begin{cases} \lfloor \frac{n}{3} \rfloor + 1 & 0 \leq n < 3k \\ k + 1 & 3k \leq n < 3k + \ell \\ \left\lfloor \frac{6k+\ell-1-n}{3} \right\rfloor + 1 & 3k + \ell \leq n \leq 6k + \ell - 1 \end{cases} .$$

Now we find $O(S)$ by summing over n:

$$O(S) = \sum_{n=0}^{6k+\ell-1} O_n(S) = 2 \cdot 3 \cdot \frac{k(k+1)}{2} + \ell(k+1) = (3k+\ell)(k+1).$$

□

We now prove the lemma through a careful consideration of Apéry tuples.

Proof (of Lemma 1) Fix n, and suppose $T \supseteq S$ with $g(T) = n$. Write $\ell = 3i + j$ where $i \in \{0, 1\}$ and $j \in \{1, 2\}$. Let $m = 6k + \ell - 1 - n$. Since $m = g(T) - g(S)$ and $T \supseteq S$, T is the union of S together with m gaps of S. Let $6k + \ell - 3p$ be the smallest element of T that is congruent to ℓ modulo 3. We see that $p \geq 0$ because $6k + \ell \in S \subseteq T$. Since T is closed under addition it includes the elements $6k + \ell - 3p + 3t$ for all $t \geq 0$, so T includes at least p gaps of S.

Since we know the smallest element of T congruent to ℓ modulo 3, the remaining $m - p$ elements of $T \setminus S$ are all congruent to 2ℓ modulo 3. The smallest element of S congruent to 2ℓ modulo 3 is $12k + 2\ell$, and hence the smallest element of T congruent to 2ℓ must be $12k + 2\ell - 3(m - p)$ to account for the correct number of gaps. Therefore, the Apéry set of T is $\{0,\ 6k + \ell - 3p,\ 12k + 2\ell - 3(m - p)\}$.

Such a numerical set T is a numerical semigroup if and only if it satisfies the inequalities (6) - (7), which reduce to

$$\begin{aligned} 2(6k + \ell - 3p) &\geq 12k + 2\ell - 3(m - p) \\ 2(12k + 2\ell - 3(m - p)) &\geq 6k + \ell - 3p \end{aligned}$$

which in turn give

$$m \geq 3p \tag{9}$$

$$6k + \ell + 3p \geq 2m. \tag{10}$$

For fixed n each value of p gives a different numerical set T, and so $O_n(S)$ is equal to the number of values of p satisfying both (9) and (10).

For $0 \leq n \leq 3k + \frac{\ell}{2} - 1$ we have $3k + \frac{\ell}{2} \leq m \leq 6k + \ell - 1$. Since $m = 6k + \ell - 1 - n$, the above inequalities can be rewritten

$$\begin{aligned} 6k+\ell-1-n &\geq 3p \\ 6k+3p+\ell &\geq 12k+2\ell-2-2n \end{aligned}$$

which determine the interval

$$\frac{6k+\ell-2-2n}{3} \ \leq \ p \ \leq \ \frac{6k+\ell-1-n}{3}. \tag{11}$$

Since $n \leq 3k+\frac{\ell}{2}-1$, the lower bound for p is greater than or equal to 0. The distance between the bounds of p given in (11) is $\frac{n+1}{3}$. Considering each case of n modulo 3 shows that there are always $\lfloor \frac{n}{3} \rfloor + 1$ integers in this interval. Therefore, in this case $O_n(S) = \lfloor \frac{n}{3} \rfloor + 1$.

For $3k+\frac{\ell}{2}-1 < n \leq 6k+\ell-1$ we have $0 \leq m < 3k+\frac{\ell}{2}$. Because $2m < 6k+\ell \leq 6k+\ell+3p$ for any p, (10) holds for any $p \geq 0$. So we need only count integer solutions to (9). There are exactly $\lfloor m/3 \rfloor + 1$ integers p that satisfy (9), so in this case $O_n(S) = \lfloor \frac{m}{3} \rfloor + 1 = \left\lfloor \frac{6k+\ell-1-n}{3} \right\rfloor + 1$. □

Theorem 1 shows that for large k there are about $6k^2$ simultaneous $(3, 6k+\ell)$-cores. From Theorem 5, we know that about $3k^2$ of them are associated with semigroups. Therefore, as b approaches infinity, half of all $(3, b)$-cores correspond with numerical semigroups. We give another interpretation of this result in Theorem 10 in the next section.

The case of $a = 4$, stated as Theorem 6 in the introduction, is more complex but can be approached similarly. We give a proof of only the case $\ell = 1$ here since the other cases are very similar.

A first approach to prove this might be to count oversemigroup by genus as we did for $a = 3$. However, that approach does not work so nicely here; for example, the function that counts oversemigroups of $S = \langle 4, 12k+1 \rangle$ by genus is not unimodal. Instead, we count oversemigroups with Apéry tuple (x, n, y), where n is fixed. Let

$$O'_n(S) = \#\{T \supseteq S : T \text{ is a semigroup, } \operatorname{Ap}(T) = (x, n, y)\}.$$

Lemma 2 *If* $S = \langle 4,\ 12k+1 \rangle$ *then*

$$O'_n(S) = \begin{cases} (n+1)(6k - \frac{3n}{2} + 1) & 0 \leq n \leq 2k \\ (n+1)(3k - \lceil \frac{n}{2} \rceil + 1) + \frac{1}{2}(3k - \lfloor \frac{n}{2} \rfloor)(3k - \lfloor \frac{n}{2} \rfloor + 1) & 2k < n \leq 6k \end{cases}.$$

Using this lemma, we prove the $\ell = 1$ case of Theorem 6.

Proof (of Theorem 6 for $\ell = 1$) Suppose T is an oversemigroup of S with $\operatorname{Ap}(T) = (x, n, y)$. Since $6k \cdot 4 + 2 \in S$ we know $n \leq 6k$, which means $O(S) = \sum_{n=0}^{6k} O'_n(S)$. By Lemma 2, we have

$$\sum_{n=0}^{2k} O_n'(S) = \sum_{n=0}^{2k}(n+1)\left(6k - \frac{3}{2}n + 1\right) = 8k^3 + 14k^2 + 7k + 1,$$

by a standard induction argument.

We also have

$$\begin{aligned}\sum_{n=2k+1}^{6k} O_n'(S) &= \sum_{n=2k+1}^{6k} (n+1)(3k+1) - (n+1)\left\lceil \frac{n}{2} \right\rceil \\ &\quad + \frac{1}{2}\left(3k - \left\lfloor \frac{n}{2} \right\rfloor\right)\left(3k - \left\lfloor \frac{n}{2} \right\rfloor + 1\right) \\ &= (3k+1)(4k) + \left(\frac{6k(6k+1)}{2} - \frac{2k(2k+1)}{2}\right)(3k+1) \\ &\quad - (9k^2 + 3k - k^2 - k) \\ &\quad - \frac{1}{6}\,[3k(3k+1)(24k+1) - k(k+1)(8k+1)] \\ &\quad + \frac{1}{2}\Big[(9k^2+3k)(4k) - (9k^2-k^2) - 6k(9k^2-k^2) \\ &\qquad + \frac{1}{3}(3k(18k^2+1) - k(2k^2+1))\Big] \\ &= 16k^3 + 16k^2 + 4k, \end{aligned} \tag{12}$$

by a slightly more complicated induction argument.

Adding (5) and (12) gives $O(S) = 24k^3 + 30k^2 + 11k + 1$. □

We finish the argument by proving Lemma 2.

Proof (of Lemma 2) Suppose $T \supseteq S$ with $\mathrm{Ap}(T) = (x, n, y)$. This Apéry tuple must satisfy the inequalities (6)–(8), which means that the following inequalities must hold:

$$2x \geq n \tag{13}$$
$$2y + 1 \geq n \tag{14}$$
$$x + n \geq y \tag{15}$$
$$y + n + 1 \geq x \tag{16}$$
$$x \leq 3k \tag{17}$$
$$n \leq 6k \tag{18}$$
$$y \leq 9k. \tag{19}$$

First, consider the case where $0 \leq n \leq 2k$. If $x \leq y$ then $x = y - c$ for some $c \in \{0, 1, \ldots, n\}$. For any x that satisfies $\lceil \frac{n}{2} \rceil \leq x \leq 3k$, the inequalities (13)–(16) are satisfied, so for each value of c there are $3k - \lceil \frac{n}{2} \rceil + 1$ oversemigroups in this case.

If $x > y$ then $x = y + c$ for some $c \in \{1, \dots, n + 1\}$. The above inequalities are satisfied if and only if $\lfloor \frac{n}{2} \rfloor + c \leq x \leq 3k$. Therefore, for each c there are $3k - c - \lfloor \frac{n}{2} \rfloor + 1$ oversemigroups in this case.

Summing over all values of c for both $x \leq y$ and $x > y$, we see that

$$\begin{aligned} O'_n(S) &= (n+1)(3k - \lceil \tfrac{n}{2} \rceil + 1) + (n+1)(3k - \lfloor \tfrac{n}{2} \rfloor + 1) - \tfrac{(n+1)(n+2)}{2} \\ &= (n+1)(6k - \tfrac{3n}{2} + 1). \end{aligned}$$

Now consider the case where $2k < n \leq 6k$. If $x \leq y$ then $x = y - c$ for some $c \in \{0, 1, \dots, n\}$. As in the previous case, (x, n, y) is a valid Apéry tuple if and only if $\lceil \frac{n}{2} \rceil \leq x \leq 3k$, so for each c there are $3k - \lceil \frac{n}{2} \rceil + 1$ oversemigroups in this case.

If $x > y$ then $x = y + c$ for some $c \in \{1, 2, \dots, 3k - \lfloor \frac{n}{2} \rfloor\}$. Again, (13)–(16) are satisfied if and only if $\lfloor \frac{n}{2} \rfloor + c \leq x \leq 3k$. Therefore, for each value of c there are $3k - c - \lfloor \frac{n}{2} \rfloor + 1$ oversemigroups in this case.

Summing over all values of c for both $x \leq y$ and $x > y$, we obtain

$$\begin{aligned} O'_n(S) = {} & (n+1)(3k - \lceil \tfrac{n}{2} \rceil + 1) + (3k - \lfloor \tfrac{n}{2} \rfloor)(3k - \lfloor \tfrac{n}{2} \rfloor + 1) \\ & - (3k - \lfloor \tfrac{n}{2} \rfloor + 1)(3k - \lfloor \tfrac{n}{2} \rfloor + 1)/2 \\ = {} & (n+1)(3k - \lceil \tfrac{n}{2} \rceil + 1) + (3k - \lfloor \tfrac{n}{2} \rfloor)(3k - \lfloor \tfrac{n}{2} \rfloor + 1)/2. \end{aligned} \qquad \square$$

Comparing Theorem 6 with Theorem 1, we see that for large k there are approximately $72k^3$ simultaneous $(4,\ 12k + \ell)$-cores, of which about $24k^3$ are associated with semigroups. Thus, as b approaches infinity, one-third of all $(4, b)$-cores correspond with numerical semigroups.

We compare the behavior of $C(a, b)$—the total number of (a, b)-cores—with $O(\langle a, b \rangle)$ for large values of b:

a	$\lim_{b \to \infty} O(\langle a, b \rangle)/C(a, b)$
2	1
3	1/2
4	1/3

We can ask for the behavior of this ratio for larger values of a. As a degree $a - 1$ polynomial in b, the leading coefficient of $C(a, b)$ is $\frac{1}{a!}$. Theorem 8 of Hellus and Waldi shows that this ratio is between $\frac{1}{(a-1)!}$ and $\frac{1}{a-1}$. Therefore,

$$\lim_{a \to \infty} \lim_{b \to \infty} \frac{O(\langle a, b \rangle)}{C(a, b)} \leq \lim_{a \to \infty} \frac{1}{a - 1} = 0.$$

These results can be interpreted as special cases of Problem 1 since the leading coefficient of these quasipolynomials are closely related to the volumes of the (a, b)-core polytopes of the previous section.

6 Conjugate Partitions and Symmetric Numerical Sets

Recall that a numerical set T with Frobenius number F is symmetric if and only if for each $i \in [0, F]$ exactly one of i, $F - i$ is in T and that the conjugate of a partition λ is the partition $\widetilde{\lambda}$ that we get from interchanging the rows and columns of the Young diagram of λ. Our first goal is to relate these two concepts. We then focus on the particular case of 3-core partitions and their conjugates.

Proposition 11 *A numerical set T is symmetric if and only if $\varphi(T)$ is a self-conjugate partition.*

In order to prove this proposition, we give a characterization of the numerical set associated to $\widetilde{\lambda}$ under the bijection φ. The *dual* of a numerical set T with Frobenius number F is the numerical set $T^* = \{u \in \mathbb{Z}\ :\ F - u \notin T\}$. A numerical set and its dual have the same atom monoid, and it is clear that a numerical set is symmetric if and only if it is equal to its dual. For additional background on this concept, see [4, Section 1]. By considering pairs i, $F - i$ and whether or not they are elements of T we get the following characterization of T^*:

$$T^* = \{F - u\ : u \in \mathbb{Z} \setminus T\}.$$

Proposition 12 *Suppose T is a numerical set with Frobenius number F and $\varphi(T) = \lambda$. The numerical set associated with $\widetilde{\lambda}$ is T^*.*

Proof It is easy to see from the definition of hook length that $H(\lambda) = H(\widetilde{\lambda})$ and $F\left(\varphi^{-1}(\widetilde{\lambda})\right) = F\left(\varphi^{-1}(\lambda)\right)$. We now label the profile of λ in reverse order, starting with F and counting down. The up-steps of this labeling are of the form $F - u$ for $u \notin T$ and are exactly the right steps of $\widetilde{\lambda}$. □

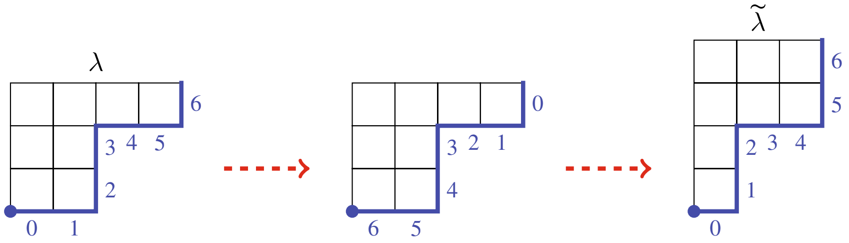
We use this characterization to prove Proposition 11. This is both a slight generalization of [16, Proposition 4.4] and a slight reframing of [23, Proposition 1], since it is now clear that a partition is self-conjugate if and only if the corresponding numerical set is equal to its dual.

Proof (of Proposition 11) Let F be the Frobenius number of T. We need only show that T is symmetric if and only if $T = \{F - u : u \in \mathbb{Z} \setminus T\}$.

First suppose T is symmetric and $u \in \mathbb{Z} \setminus T$. Then $F - u \in T$ by definition. If $F - u \in T$, then $F - (F - u) = u \notin T$.

Conversely, suppose $T = \{F - u : u \notin T\}$. If $x \in T$, then $x = F - u$ for some $u \notin T$. Now $u = F - x$ and we see that T is symmetric. □

We give an example to illustrate this process.



The conjugate of an a-core partition λ is also an a-core and we have seen how Apéry tuples map such partitions to $\mathbb{N}^{a-1}$ so it is natural to ask how conjugation acts on $\mathbb{N}^{a-1}$. In other words, we wish to find the Apéry tuple of $\varphi^{-1}(\widetilde{\lambda})$ given the Apéry tuple of $\varphi^{-1}(\lambda)$.

Proposition 13 *Suppose that λ is a partition with corresponding numerical set $T = \varphi^{-1}(\lambda)$ with Frobenius number F such that the Apéry tuple of T is* $\mathrm{Ap}(T) = (x_1, \dots, x_{a-1})$ *and $F \equiv \ell \pmod{a}$. Then the Apéry tuple of $T^* = \varphi^{-1}(\widetilde{\lambda})$ is* $\mathrm{Ap}(T^*) = (x'_1, \dots, x'_{a-1})$ *where*

$$x'_i = \begin{cases} x_\ell - x_{\ell-i} & i < \ell \\ x_\ell & i = \ell \\ x_\ell - x_{a+\ell-i} - 1 & i > \ell \end{cases}.$$

Proof Recall from Proposition 12 that $S = \{F - u : u \in \mathbb{Z} \setminus T\}$. Thus

$$\begin{aligned} ax'_i + i &= \min\{F - u : u \notin S, F - u \equiv i \pmod{a}\} \\ &= F - \max\{u \notin S : u \equiv \ell - i \pmod{a}\}. \end{aligned}$$

By the definition of the Apéry tuple,

$$\max\{u \notin S : u \equiv \ell - i \pmod{a}\} = \begin{cases} a(x_{\ell-i} - 1) + (\ell - i) & i < \ell \\ -a & i = \ell \\ ax_{a+\ell-i} + (\ell - i) & i > \ell \end{cases}.$$

Noting $F = a(x_\ell - 1) + \ell$ completes the proof. □

Proposition 13 allows us to prove a theorem unique to 3-core partitions that relates them to numerical semigroups.

Theorem 10 *Given a* 3*-core partition λ, either $\varphi^{-1}(\lambda)$ or $\varphi^{-1}(\widetilde{\lambda})$ is a numerical semigroup.*

Proof Let $T = \varphi^{-1}(\lambda)$ and $S = \varphi^{-1}(\widetilde{\lambda})$. Suppose that $\mathrm{Ap}(T) = (x_1, x_2)$. Recall that T is a numerical semigroup if and only if it satisfies the inequalities (6)–(7), which here reduce to

$$2x_1 \geq x_2 \tag{20}$$

$$2x_2 + 1 \geq x_1. \tag{21}$$

Notice that at least one of these must be true.

If (20) fails, then $x_1 < x_2$, and so by Proposition 13, $\mathrm{Ap}(S) = (x_2 - x_1, x_2)$. Using the fact that $2x_1 < x_2$ and $2x_2 + 1 \geq x_1$, we see that (6) and (7) are both satisfied for S, and hence S is a numerical semigroup.

If instead (21) fails, then $x_1 > x_2$, so by Proposition 13, $\mathrm{Ap}(S) = (x_1, x_1 - x_2 - 1)$. As before, we use the fact that $2x_1 \geq x_2$ and $2x_2 + 1 < x_1$ to show that (6) and (7) are satisfied for S. So S is again a numerical semigroup. □

From Theorem 5, we know the number of numerical semigroups containing $\langle 3, 6k + \ell\rangle$, so using Theorem 10 we can determine the number of these semigroups that are symmetric. A symmetric numerical semigroup is sent to a self-conjugate partition under φ, so this number is also equal to the number of self-conjugate $(3, 6k + \ell)$-core partitions associated to numerical semigroups.

Corollary 1 *The number of symmetric numerical semigroups containing* $\langle 3, 6k + \ell\rangle$ *is*

$$3k + \frac{3\ell}{2} - \frac{\ell^2}{6} - \frac{1}{3}.$$

Proof By Theorem 10, for any $(3, 6k + \ell)$-core partition λ, either $\varphi^{-1}(\lambda)$ or $\varphi^{-1}(\widetilde{\lambda})$ is a semigroup. Therefore, if we double count the number oversemigroups of S we will have counted every non-self-conjugate $(3, 6k + \ell)$-core exactly once, and we will have counted the number of self-conjugate $(3, 6k + \ell)$-cores twice. Therefore, the number of self-conjugate $(3, 6k + \ell)$-core partitions, which is the same as the number of symmetric oversemigroups of S by Theorem 10, is

$$2 \cdot O(\langle 3, 6k + \ell\rangle) - C(3, 6k + \ell) = 3k + \frac{3\ell}{2} - \frac{\ell^2}{6} - \frac{1}{3}.$$ □

7 Counting Partitions with a Given Hook Set

In much of this paper, we have studied statistical questions about distribution of sizes of the finite set of simultaneous $(a, b_1, \ldots, b_m)$-core partitions. In this section, we turn toward another finite collection of partitions, those which have the same hook set. By Proposition 4, the hook set of any partition is the complement of some numerical semigroup S. Our goal is to understand the set of partitions sharing a given hook set and what properties of the underlying semigroup influence the size of this set. Therefore, we rephrase this question as: Given a numerical semigroup S, for how many partitions λ is $H(\lambda) = \mathbb{N} \setminus S$? We call this number $P(S)$. By our discussion of the bijection φ in Sect. 2, this is equivalent to counting the number of numerical sets with atom monoid S.

This problem has been considered by Marzuola and Miller in [23] where they call it the *Anti-Atom Problem*. They give constraints on numerical sets sharing the same atom monoid S in terms of the dual numerical set S^*.

Proposition 14 *[23, Proposition 1] Suppose that S is a numerical semigroup and that T is a numerical set with $A(T) = S$. Then $S \subseteq T \subseteq S^*$.*

We note that the description of T^* given directly above Proposition 12 also gives a way to prove this fact in terms of partitions with a given hook set.

In cases where the gap between S and S^* is well-understood this result gives a strong characterization of the numerical sets with atom monoid S. A numerical semigroup S is *pseudosymmetric* if $F(S)$ is even and for every $i \in [0, F(S)/2)$ exactly one of i, $F(S) - i$ is in S.

Corollary 2 *[23, Corollary 2] A numerical monoid S with Frobenius number F is symmetric if and only if there is just one numerical set (which must be S itself) whose atom monoid is S. Equivalently, $P(S) = 1$ if and only if S is symmetric.*

If S is a pseudosymmetric numerical semigroup then there are precisely two numerical sets (which must be S and S^*) whose atom monoid is S. Equivalently, if S is pseudosymmetric then $P(S) = 2$.

The first part of the corollary is equivalent to Proposition 11. We will see below that the converse of the second statement does not hold. It seems difficult to give a complete classification of numerical semigroups S with $P(S) = 2$.

We give a bound for $P(S)$ in terms of how far away S is from being symmetric. A *missing pair* of S is a pair of elements i, $F(S) - i$ with $i \leq F(S) - i$ such that neither element is in S. Note that when $F(S)$ is even we have the degenerate missing pair consisting of the single element $F(S)/2$. Let $M(S)$ denote the union of the set of missing pairs of S.

Lemma 3 *For a numerical semigroup S we have $S^* = S \cup M(S)$.*

Proof Let F be the Frobenius number of S, which is also the Frobenius number of S^*. We need only consider elements less than F. We first recall that

$$S^* = \{F - u \ : u \in \mathbb{Z} \setminus S\}.$$

If $n, F - n$ is a missing pair of S then $F - n \in S^*$ since $n \notin S$, and $n = F - (F - n) \in S^*$ as $F - n \notin S$. Therefore $M(S) \subset S^*$.

For the reverse inclusion, suppose that $n = F - u \in S^*$, where $u \in \mathbb{N} \setminus S$. If $n \notin S$ then $u, n \in M(S)$. If $n \in S$ then $u = F - n \notin S^*$. We conclude that $S^* = S \cup M(S)$. □

We could replace every instance of $M(S)$ with $S^* \setminus S$, but we choose to keep the notation of missing pairs since it is more descriptive.

Corollary 3 *For a numerical semigroup S, $P(S) \leq 2^{|M(S)|}$.*

Proof A numerical semigroup T with hook set $\mathbb{N} \setminus S$ is the union of S with some subset of $M(S)$. □

Since $M(S)$ is empty for a symmetric semigroup and consists of a single element for a pseudosymmetric semigroup, this gives another proof of Corollary 2.

Now that we understand semigroups for which $|M(S)| \leq 1$, we consider those for which $|M(S)| = 2$.

Proposition 15 *For a numerical semigroup S with $|M(S)| = 2$, $P(S) \in \{2, 3\}$.*

Proof Let F be the Frobenius number of S and $a, F - a$ be the missing pair of S where $a < F/2$. Since S is not symmetric, $P(S) \geq 2$. By Corollary 3, we need only show that $P(S) \neq 4$. By the lemma above we need only show that $A\,(S \cup \{F - a\}) \neq S$.

We argue by contradiction. Suppose $A\,(S \cup \{F - a\}) = S$. Since $F - a \notin A(S \cup \{F - a\})$ there is some $n \in S \cup \{F - a\}$ such that $n + F - a \notin S$. Since $F - a > F/2$ we cannot have $n = F - a$. So $n \in S$ and $n \notin A\,(S \cup \{F - a\})$, which is a contradiction. □

We note that both cases $P(S) = 2$ and $P(S) = 3$ are possible. For example, $S = \{0, 4, \rightarrow\}$ has $A(S \cup \{1\}) = A(S \cup \{1, 2\}) = S$ and $P(S) = 3$, and $S = \{0, 3, 6, \rightarrow\}$ has $M(S) = \{1, 4\}$ and $P(S) = 2$.

Corollary 3 shows that if $|M(S)|$ is small then $P(S)$ is small. We give a family of semigroups showing that the converse does not necessarily hold.

Proposition 16 *For odd $N \in \mathbb{N}$ with $N \geq 11$, let R_N be the numerical semigroup*

$$R_N = \{0, \tfrac{N+1}{2}\} \cup E_N \cup \{N + 1,\ N + 2, \ldots\}.$$

where E_N is the set of even numbers in $\left(\frac{N+1}{2}, N - 1\right)$. We have that $P(R_N) = 2$ but $|M(R_N)| = 2\left\lceil \frac{N-1}{4} \right\rceil$.

Proof The statement about $M(R_N)$ follows easily from the fact that $F(R_N) = N$ and the observation that every missing pair except $\{1, N - 1\}$ is uniquely determined by an odd number in $\left(\frac{N+1}{2}, N - 1\right)$.

Since R_N is not symmetric we see that $R_N \subsetneq R_N^*$ and $A(R_N^*) = R_N$. Suppose $T \neq R_N$ is a numerical set with $A(T) = R_N$. By Proposition 14, we have $T \subseteq R_N^*$. We show that $P(R_N) = 2$ by showing that

$$R_N^* = \{N - u \ :\ u \in \mathbb{Z} \setminus R_N\} \subseteq T.$$

Notice $\mathbb{N} \setminus R_N = \{1, \ldots, \frac{N-1}{2}\} \cup O_N \cup \{N - 1\}$, where O_N is the set of odd numbers in $[\frac{N+3}{2}, N)$. Thus, $R_N^* = \{N - u : u \notin R_N\}$ is the set of even numbers in $[0, \frac{N-3}{2}]$ together with $\{1, \frac{N+1}{2}, \ldots, N - 1\}$ and $\{N + 1, N + 2, \ldots\}$.

Since $T \neq R_N$ there exists some $t \in \left(R_N^* \setminus R_N\right) \cap T$. Either $1 \in T$, T contains an even number in $(0, \frac{N-3}{2}]$, T contains $N-1$, or T contains an odd number in $(\frac{N+1}{2}, N-1)$. In each case we will show that $T = R_N^*$.

If $1 \in T$ then $A(T) = R_N$ implies that $\{\frac{N+1}{2}, \ldots, N-2\} \subset T$ since R_N contains the even numbers in this range. However, the odd numbers in this range are not in $A(T)$, meaning that for each $N-2k$ there is some $s_k \in T$ such that $N-2k+s_k \notin T$. The only possibility is that $N-2k+s_k = N$, so $s_k = 2k$, meaning T must contain the even numbers in $[0, \frac{N-3}{2}]$ and $T = R_N^*$.

Suppose T contains an even number $t \in (0, \frac{N-3}{2}]$. Since $A(T)$ contains all even numbers in $(\frac{N+1}{2}, N-1)$, we see that $N-1-t \in A(T)$. Since $t + (N-1-t) = N-1$, we have $N-1 \in T$. However, $N-1 \notin A(T)$, so there must exist $u \in T$ with $N-1+u \notin T$. The only possibility is $u = 1$, which by the argument of the previous paragraph shows $T = R_N^*$.

Now suppose that $N-1 \in T$. As in the previous paragraph, since $N-1 \notin A(T)$ we see that $1 \in T$ and $T = R_N^*$.

Finally, suppose T contains an odd number $t \in (\frac{N+1}{2}, N-1)$. Since $t \notin A(T)$ there exists $u \in T$ such that $t + u \notin T$. Since R_N contains all even numbers in $(\frac{N+1}{2}, N-1)$, we either have $t+u$ equal to an odd number in $(\frac{N+1}{2}, N-1)$ or equal to $N-1$. In the first case u is an even number in $(0, \frac{N-3}{2}]$, putting us in the situation described above, and we conclude $T = R_N^*$. If $t+u = N-1$ then u is an odd number in $(0, \frac{N-3}{2})$. Since $u \in R_N^*$ we have $u = 1$, putting us in the situation above. We conclude that $T = R_N^*$. □

We now use a main result of Marzuola and Miller [23] to study the opposite extreme, semigroups S for which $M(S)$ is as large as possible given the genus.

Proposition 17 *Let $S_N = \{0, N+1, N+2, \cdots\}$ be the numerical semigroup where $H(\varphi(S)) = \{1, 2, \cdots, N\}$. Then $P(S_N) \sim c \cdot 2^N$, where c is a constant approximately equal to* 0.2422.

Proof Let γ_N be the ratio of the number of numerical sets with atom monoid S_N to the number of numerical sets with Frobenius number N. One of the main results of [23] is that the sequence $\{\gamma_N\}$ is decreasing and converges to a number $\gamma \approx 0.4844$ with accuracy to within 0.0050. Numerical sets with Frobenius number N are in bijection with subsets of $\{1, \ldots, N-1\}$, so there are 2^{N-1} of them. Therefore, $P(S_N) = \gamma_N \cdot 2^{N-1}$, completing the proof. □

We end this section by giving a link between partitions with a given hook set and partitions that come from numerical semigroups under φ.

Proposition 18 *Let $S(N)$ be the number of partitions with maximum hook length N corresponding via φ to numerical semigroups and let $T(N)$ be the number of partitions with maximum hook length N. Then,*

$$\lim_{n \to \infty} \frac{S(N)}{T(N)} = 0.$$

We use a bound due to Backelin on the number of numerical semigroups with given Frobenius number.

Theorem 11 [6, Theorem 1.1] *The number of numerical semigroups S with Frobenius number N is at most* $4 \cdot 2^{\lfloor (N-1)/2 \rfloor}$.

Proof (of Proposition 18) Partitions with maximum hook length N are in bijection with numerical sets with Frobenius number N, so $T(N) = 2^{N-1}$. Similarly, $S(N)$ is the number of numerical semigroups with Frobenius number N. By Backelin's theorem,

$$\frac{S(N)}{T(N)} \leq 4 \cdot 2^{\lfloor (N-1)/2 \rfloor - (N-1)} \leq 4 \cdot 2^{-(N-1)/2},$$

and therefore

$$\lim_{n \to \infty} \frac{S(N)}{T(N)} = 0.$$

□

8 Further Questions

We begin by returning to Problem 1. The simultaneous $(a, b_1, \ldots, b_m)$-core partitions are in bijection with integer points in a certain polytope. We would like to be able to give formulas for the number of lattice points in this polytope and also for its volume. Understanding these questions gives one approach to determining the correct leading coefficient of the quasipolynomial given in the second part of Theorem 8 of Hellus and Waldi [18]. The size of a partition corresponding to a lattice point comes from evaluating the quadratic function $F_a(x_1, \ldots, x_{a-1})$ of Sect. 3. Under what circumstances can we give a nice description of the lattice point of this polytope on which this function takes its maximum value? When can we give a nice expression for the average value of this function taken over all of these lattice points or give even more detailed statistical information about this set of values? We would like to have a better understanding of how tools from Ehrhart theory can be used to study these problems.

It seems likely that most partitions are not associated to numerical semigroups by the bijection φ, as most numerical sets are not closed under addition. A subtle difficulty in addressing these types of questions comes from the fact that making statements about "most" partitions or "most" numerical sets requires an ordering. The most natural ordering on partitions, in our opinion, is by size. Proposition 18 shows that if we instead order partitions by the size of their maximum hook length our intuition is correct.

Conjecture 1 *Let $P(n)$ be the number of partitions of size at most n and let $S'(n)$ be the number of these that are associated to numerical semigroups under φ. Then*

$$\lim_{n\to\infty} \frac{S'(n)}{P(n)} = 0.$$

We ask a similar question for a-cores.

Problem 2 Let $a \geq 2$ be a positive integer, $P_a(n)$ be the number of a-core partitions of size at most n, and $S'_a(n)$ be the number of these partitions associated to numerical semigroups under φ. Determine

$$\lim_{n\to\infty} \frac{S'_a(n)}{P_a(n)}$$

as a function of a.

An easier subproblem would be to show that as a goes to infinity, this limit goes to zero. Consider the rational polyhedral cone giving the condition that an a-core comes from a semigroup and intersect it with the region where the quadratic function $F_a(x_1, \ldots, x_{a-1}) \leq 1$. It seems likely that techniques from Ehrhart theory combined with the volume of this set can be used to solve this problem.

We would also like to better understand how to use techniques from the first part of this paper to study $P(S)$. Suppose that S is a numerical semigroup containing a. Then every partition with hook set S corresponds to a point in $\mathbb{N}^{a-1}$ by taking the Apéry tuple of the corresponding numerical set. Can we say anything meaningful about the geometry of this finite set of points? We would also like to know the largest, smallest, and average size of a partition with hook set S.

We would also like to better understand the properties of S that control the size of $P(S)$. We have started to explore the link between the size of the set of missing pairs, $M(S)$, and the number of partitions with this hook set. We include some data related to this question. The *semigroup tree* allows us to visualize easily the relationship between numerical semigroups via their *effective generators*, the minimal generators greater than the Frobenius number. The tree is constructed as follows: the vertices of the tree are numerical semigroups, with the root as $\mathbb{N}$; for each vertex S in the tree, the children of this semigroup are the semigroups obtained from S by removing an effective generator. Each semigroup appears in the tree exactly once, and the distance between S and the root is exactly the genus of S. For more information about the semigroup tree, see [10].

Figure 2 shows the first 6 layers of the semigroup tree, in which each semigroup S is labeled with $|M(S)|$ and $P(S)$. Every semigroup generated by two elements is symmetric, so we see that these all satisfy $|M(S)| = 0$ and $P(S) = 1$. We also see that the semigroups $\langle g+1, g+2, \ldots, 2g+1\rangle$ are those which have the largest values of $P(S)$ at a given genus.

Lastly, throughout this paper we have explored the properties of hook sets of partitions, but have not really commented on hook multisets. We would like to better understand what properties of a multiset make it occur as the hook multiset of many different partitions. A good starting place might be a careful examination of the constructions given by Chung and Herman [12], and by Craven [13].

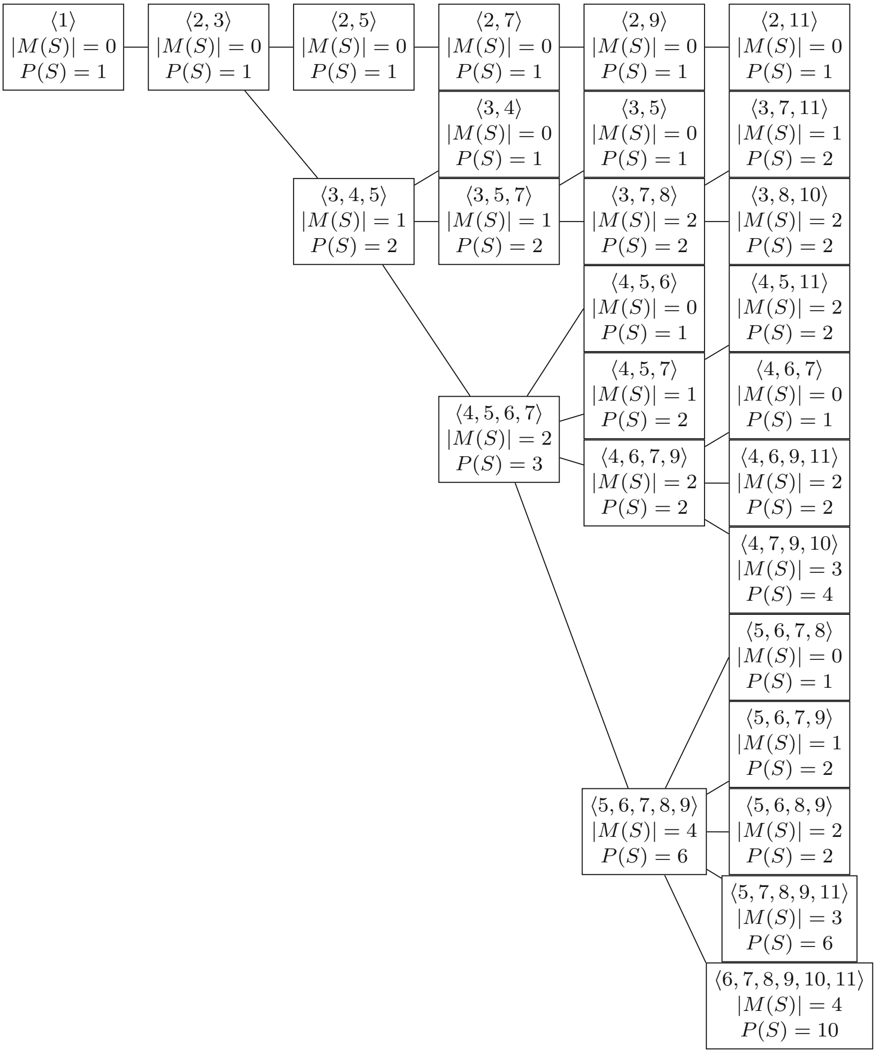

$\langle 1 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 2,3 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 2,5 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,4,5 \rangle$ $|M(S)| = 1$ $P(S) = 2$
$\langle 2,7 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,4 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,5,7 \rangle$ $|M(S)| = 1$ $P(S) = 2$
$\langle 4,5,6,7 \rangle$ $|M(S)| = 2$ $P(S) = 3$
$\langle 2,9 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,5 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,7,8 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 4,5,6 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 4,5,7 \rangle$ $|M(S)| = 1$ $P(S) = 2$
$\langle 4,6,7,9 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 5,6,7,8,9 \rangle$ $|M(S)| = 4$ $P(S) = 6$
$\langle 2,11 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 3,7,11 \rangle$ $|M(S)| = 1$ $P(S) = 2$
$\langle 3,8,10 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 4,5,11 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 4,6,7 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 4,6,9,11 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 4,7,9,10 \rangle$ $|M(S)| = 3$ $P(S) = 4$
$\langle 5,6,7,8 \rangle$ $|M(S)| = 0$ $P(S) = 1$
$\langle 5,6,7,9 \rangle$ $|M(S)| = 1$ $P(S) = 2$
$\langle 5,6,8,9 \rangle$ $|M(S)| = 2$ $P(S) = 2$
$\langle 5,7,8,9,11 \rangle$ $|M(S)| = 3$ $P(S) = 6$
$\langle 6,7,8,9,10,11 \rangle$ $|M(S)| = 4$ $P(S) = 10$

Fig. 2 The semigroup tree, with the root $\mathbb{N}$ on the left. Semigroups with a common genus are found in the same column, and each semigroup S is labeled with $|M(S)|$ and $P(S)$

Acknowledgements The third author thanks Mel Nathanson for organizing the 2010 CANT conference where he first learned about core partitions and their connection to numerical semigroups in a talk by William Keith. He also thanks Maria Monks Gillespie for helpful discussions on early parts of this project.

We would like to thank Florencia Orosz-Hunziker and Daniel Corey for their assistance throughout this project. Finally, we would like to thank the Summer Undergraduate Math Research at Yale program for organizing, funding, and supporting this project. SUMRY is supported in part by NSF grant CAREER DMS-1149054.

References

1. A. Aggarwal, Armstrong's conjecture for $(k, mk + 1)$-core partitions. Eur. J. Comb. **47**, 54–67 (2015)
2. T. Amdeberhan, E. Leven, Multi-cores, posets, and lattice paths. Adv. Appl. Math. **71**, 1–13 (2015)
3. J. Anderson, Partitions which are simultaneously t_1- and t_2-core. Discrete Math. **248**(1–3), 237–243 (2002)
4. E. Antokoletz, A. Miller, Symmetry and factorization of numerical sets and monoids. J. Algebra **247**, 636–671 (2002)
5. D. Armstrong, C.R.H. Hanusa, B.C. Jones, Results and conjectures on simultaneous core partitions. Eur. J. Comb. **41**, 205–220 (2014)
6. J. Backelin, On the number of semigroups of natural numbers. Math. Scand. **66**(2), 197–215 (1990)
7. C. Berg, M. Vazirani, $(\ell, 0)$-Carter partitions, a generating function, and their crystal theoretic interpretation. Electron. J. Comb. **15**(1), Research Paper 130, 23 pp. (2008)
8. V. Blanco, J. Puerto, An application of integer programming to the decomposition of numerical semigroups. SIAM J. Discrete Math. **26**(3), 1210–1237 (2012)
9. M. Branco, J. García-García, P.A. García-Sánchez, J.C. Rosales, Systems of inequalities and numerical semigroups. J. Lond. Math. Soc. **65**(2), 611–623 (2002)
10. M. Bras-Amorós, S. Bulygin, Towards a better understanding of the semigroup tree. Semigroup Forum **79**(3), 561–574 (2009)
11. W. Chen, H. Huang, L. Wang, Average size of a self-conjugate (s, t)-core partition. Proc. Am. Math. Soc. **144**(4), 1391–1399 (2016)
12. F. Chung, J. Herman, Some results on hook lengths. Discrete Math. **20**, 33–40 (1977)
13. D. Craven, Symmetric group character degrees and hook numbers. Proc. Lond. Math. Soc. (3) **96**(1), 26–50 (2008)
14. S. Ekhad, D. Zeilberger, Expressions for the variance and higher moments of the size of a simultaneous core partition and its limiting distribution. arXiv:1508.07637
15. J.S. Frame, G. de B. Robinson, R.M. Thrall, The hook graphs of the symmetric group. Canad. J. Math. **6**, 316–325 (1954)
16. P.A. García-Sánchez, J. Rosales, *Numerical Semigroups*, Developments in Mathematics, **20** (Springer, New York, 2009)
17. A. Granville, K. Ono, Defect zero p-blocks for finite simple groups. Trans. Am. Math. Soc. **348**(1), 331–347 (1996)
18. M. Hellus, R. Waldi, On the number of numerical semigroups containing two coprime integers p and q. Semigroup Forum **90**(3), 833–842 (2015)
19. P. Johnson, Lattice points and simultaneous core partitions. arXiv:1502.07934
20. N. Kaplan, Counting numerical semigroups by genus and some cases of a question of Wilf. J. Pure Appl. Algebra **216**(5), 1016–1032 (2012)
21. W. Keith, R. Nath, Partitions with prescribed hooksets. J. Comb. Number Theory **3**(1), 39–50 (2011)
22. E. Kunz, R. Waldi, Counting numerical semigroups. arXiv:1410.7150v1
23. J. Marzuola, A. Miller, Counting numerical sets with no small atoms. J. Comb. Theory Ser. A **117**(6), 650–667 (2010)
24. J.B. Olsson, D. Stanton, Block inclusions and cores of partitions. Aequationes Mathematicae **74**, 90–110 (2007)
25. R. Stanley, F. Zanello, The Catalan case of Armstrong's conjecture on simultaneous core partitions. SIAM J. Discrete Math. **29**(1), 658–666 (2015)
26. M. Thiel, N. Williams, Strange expectations and simultaneous cores. J. Algebraic Comb. **46**(1), 219–261 (2017)
27. A. Tripathi, On the largest size of a partition that is both s-core and t-core. J. Number Theory **129**, 1805–1811 (2009)

28. V. Wang, Simultaneous core partitions: parameterizations and sums. Electron. J. Comb. **23**(1), Paper 1.4, 34 pp. (2016)
29. H. Xiong, The number of simultaneous core partitions. arXiv:1409.7038
30. H. Xiong, Core partitions with distinct parts. arXiv:1508.07918
31. J. Yang, M. Zhong, R. Zhou, On the enumeration of $(s, s+1, s+2)$-core partitions. Eur. J. Combin. **49**, 203–217 (2015)

Pairs of Dot Products in Finite Fields and Rings

David Covert and Steven Senger

Abstract We obtain bounds on the number of triples that determine a given pair of dot products arising in a vector space over a finite field or a module over the set of integers modulo a power of a prime. More precisely, given $E \subset \mathbb{F}_q^d$ or $\mathbb{Z}_q^d$, we provide bounds on the size of the set

$$\{(u, v, w) \in E \times E \times E : u \cdot v = \alpha, u \cdot w = \beta\}$$

for units α and β.

Keywords Dot-product sets · Sum-product problem · Finite fields

1 Introduction

For a subset of a ring, $A \subset R$, the sumset and productset of A are defined as $A + A = \{a + a' : a, a' \in A\}$ and $A \cdot A = \{a \cdot a' : a, a' \in A\}$, respectively. The sum-product conjecture asserts that when $A \subset \mathbb{Z}$, then either $A + A$ or $A \cdot A$ is of large cardinality. For example, if we take $A \subset \mathbb{Z}$ to be a finite arithmetic progression of length n, you achieve $|A + A| = 2n - 1$, whereas $|A \cdot A| \geq cn^2/((\log n)^\delta \cdot (\log \log n)^{3/2})$ for some constant $c > 0$ and $\delta = 0.08607\ldots$ [7]. When $A \subset \mathbb{Z}$ is a geometric progression of length n, we have $|A \cdot A| = 2n - 1$, and yet it is easy to show that $|A + A| = \binom{n+1}{2}$. For subsets of integers, the following conjecture was made in [6].

Conjecture 1 *Let $A \subset \mathbb{Z}$ with $|A| = n$. For every $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ so that*

$$\max(|A + A|, |A \cdot A|) \geq C_\epsilon n^{2-\epsilon}.$$

D. Covert (✉)
University of Missouri, Saint Louis, USA
e-mail: covertdj@umsl.edu

S. Senger
Missouri State University, Springfield, USA

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_8

Much progress has been made on the sum-product problem. The best result to date belongs to Konyagin and Shkredov [11], wherein they demonstrated that for a sufficiently large constant C, we have the bound

$$\max(|A+A|, |A\cdot A|) \geq Cn^{4/3+c}$$

for any $c < \frac{5}{9813}$, whenever A is a set of real numbers with cardinality n. Work has also been done on analogues of the sum-product problem for general rings [12]. For example, the authors in [8] showed that if $E \subset \mathbb{F}_q^d$ is of sufficiently large cardinality, then we have

$$|\{(x,y)\in E\times E : x\cdot y=\alpha\}| = \frac{|E|^2}{q}(1+\underline{o}(1)),$$

for any $\alpha \in \mathbb{F}_q^*$. Here, $\mathbb{F}_q$ is the finite field with q elements, $\mathbb{F}_q^d$ is the d-dimensional vector space over $\mathbb{F}_q$, and $\mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$. As a corollary, they showed that $|dA^2| := |A\cdot A+\cdots+A\cdot A| \supset \mathbb{F}_q^*$, whenever $A \subset \mathbb{F}_q$ is such that $|A| \geq q^{\frac{1}{2}+\frac{1}{2d}}$. Much work has also been done to give such results when E has relatively small cardinality. See, for example, [10] and the references contained therein.

In [3], the second listed author and Daniel Barker studied pairs of dot products determined by sets $P \subset \mathbb{R}^2$. In addition to the applications toward the sum-product problem above, the problem of pairs of dot products has applications in coding theory, graph theory, and frame theory, among others [1, 2, 4]. The main results from [3] are as follows.

Theorem 1 *Suppose that $P \subset \mathbb{R}^2$ is a finite point set with cardinality $|P| = n$. Then, the set*

$$\Pi_{\alpha,\beta}(P) := \{(x,y,z)\in P\times P\times P : x\cdot y=\alpha, x\cdot z=\beta\}$$

satisfies the upper bound $|\Pi_{\alpha,\beta}(P)| \lesssim n^2$ whenever α and β are fixed, nonzero real numbers.

Note 1 Here and throughout, we use the notation $X \lesssim Y$ to mean that $X \leq cY$ for some constant $c > 0$. Similarly, we use $X \gtrsim Y$ for $Y \lesssim X$, and we use $X \approx Y$ if both $X \lesssim Y$ and $X \gtrsim Y$. Finally, we write $X \gtrapprox Y$ if for all $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ such that $X \gtrapprox C_\epsilon q^\epsilon Y$.

Theorem 1 is sharp, as shown in an explicit construction [3]. Additionally, they showed the following:

Theorem 2 *Suppose that $P \subset [0,1]^2$ is a set of n points that obey the separation condition*

$$\min(|p-q| : p,q\in P, p\neq q) \geq \epsilon.$$

Then, for $\epsilon > 0$ and fixed $\alpha, \beta \neq 0$, we have

$$|\Pi_{\alpha,\beta}(P)| \lesssim n^{4/3}\epsilon^{-1}\log\left(\epsilon^{-1}\right).$$

The purpose of this article is to study finite field and finite ring analogues of the results from [3]. Our main results are as follows.

Theorem 3 *Given a set, $E \subset \mathbb{F}_q^2$ or $\mathbb{Z}_q^d$, $|E| = n$, and fixed units α, β, we have the bound*

$$|\Pi_{\alpha,\beta}(E)| \lesssim n^2.$$

In general, for a set of n points, $E \subset \mathbb{F}_q^2$, one cannot expect to get an upper bound better than Theorem 3, as we will show via an explicit construction in Proposition 1. This proof and construction are similar to their analogues in [3]. However, if we view the separation condition from Theorem 2 as it relates to density (as is often the case for translating such results, such as in [9]), the previous proof techniques yield very little. It turns out that a discrepancy theoretic approach gives more information, as our second main result is for general subsets of $\mathbb{F}_q^d$, for $d \geq 2$, as opposed to just $d = 2$.

Theorem 4 *Let $d \geq 2$, $E \subset \mathbb{F}_q^d$, and suppose that $\alpha, \beta \in \mathbb{F}_q$. Then, we have the bound*

$$|\Pi_{\alpha,\beta}(E)| = \frac{|E|^3}{q^2}(1 + \underline{o}(1)),$$

for $|E| \gtrapprox q^{\frac{d+1}{2}}$ when $\alpha, \beta \in \mathbb{F}_q^*$, and for $|E| \gtrapprox q^{\frac{d+2}{2}}$ otherwise.

Note that Theorem 4 gives a quantitative version of Theorem 3 at least for sets $E \subset \mathbb{F}_q^2$ in the range $|E| \gtrapprox q^{3/2}$.

The proof of Theorem 4 relies on adapting the exponential sums found in the study of single dot products [8]. Since the results from [8] were extended to general rings $\mathbb{Z}_q^d$ in [5], Theorem 4 also easily extends to rings. Here and throughout, $\mathbb{Z}_q$ denotes the set of integers modulo q, $\mathbb{Z}_q^\times$ is the set of units in $\mathbb{Z}_q$, and $\mathbb{Z}_q^d = \mathbb{Z}_q \times \cdots \times \mathbb{Z}_q$ is the d-rank free module over $\mathbb{Z}_q$. For $E \subset \mathbb{Z}_q^d$, we define $\Pi_{\alpha,\beta}(E)$ exactly as before.

Theorem 5 *Suppose that $E \subset \mathbb{Z}_q^d$, where $q = p^\ell$ is the power of a prime $p \geq 3$. Then for units $\alpha, \beta \in \mathbb{Z}_q^\times$, we have*

$$|\Pi_{\alpha,\beta}(E)| = \frac{|E|^3}{q^2}(1 + \underline{o}(1))$$

whenever $|E| \gtrapprox q^{\frac{d(2\ell-1)}{2\ell} + \frac{1}{2\ell}}$. In particular,

$$|\Pi_{\alpha,\beta}(E)| \lesssim |E|^2$$

for sets $E \subset \mathbb{Z}_q^2$ of sufficiently large cardinality.

Remark 1 Notice that the proofs of Theorems 4 and 5 provide both a lower and upper bounds on the cardinality of $\Pi_{\alpha,\beta}(E)$, though we could achieve the upper bound $|\Pi_{\alpha,\beta}(E)| \lesssim q^{-2}|E|^3$ if we relaxed the condition $|E| \gtrapprox q^{\frac{d+1}{2}}$ to simply $|E| \gtrsim q^{\frac{d+1}{2}}$, for example.

2 Explicit Constructions

2.1 *Sharpness of Theorem 3*

We construct explicit sharpness examples for $\mathbb{F}_q^2$. The same constructions can be modified to yield sharpness in $\mathbb{Z}_q^2$ as well.

Proposition 1 *Given a natural number $n \lesssim q$ and elements $\alpha, \beta \in \mathbb{F}_q^*$, there is a set, $E \subset \mathbb{F}_q^2$ for which $|E| = n$ and*

$$|\Pi_{\alpha,\beta}(E)| \approx n^2.$$

Proof Let u be the point with coordinates $(1, 1)$. Now, distribute up to $\lceil \frac{n-1}{2} \rceil$ points along the line $y = \alpha - x$, and distribute the remaining up to $\lfloor \frac{n-1}{2} \rfloor$ points along the line $y = \beta - x$. If there are any points left over, put them anywhere not yet occupied.[1] Clearly, there are at least $|E|^2$ pairs of points (b, c), where q is chosen from the first line and r is chosen from the second. Notice that u contributes a triple to $\Pi_{\alpha,\beta}(E)$ for each such pair, giving us

$$|\Pi_{\alpha,\beta}(E)| \approx n^2.$$

2.2 *The Special Case $\alpha = \beta = 0$, $D = 2$*

Proposition 2 *There exists a set $E \subset \mathbb{F}_q^2$ of cardinality $|E| = n < 2q$ for which*

$$|\Pi_{0,0}(P)| \approx n^3.$$

Proof Select $\lceil \frac{n}{2} \rceil$ points with zero x-coordinate, and $\lfloor \frac{n}{2} \rfloor$ points with zero y-coordinate. Now, for each of the points with zero x-coordinate, there are about $\binom{\frac{n}{2}}{}\binom{\frac{n}{2}}{}$ pairs of points with zero y-coordinate. Notice that any point chosen with zero x-coordinate will have dot product zero with each point from the pair chosen with zero y-coordinate. Therefore, each of these $\frac{1}{8}n^3$ triples will contribute to $\Pi_{0,0}(E)$.

We can get just as many triples that contribute to $\Pi_{0,0}(E)$ by taking single points with zero y-coordinate and pairs of points with zero x-coordinate. In total, we get

$$|\Pi_{0,0}(P)| \approx \frac{1}{8}n^3 + \frac{1}{8}n^3 \approx n^3.$$

[1] This is just in the case that $(1, 1)$ is on one of the lines or $\alpha = \beta$.

3 Proofs of Main Results

3.1 Proof of Theorem 3

This proof is a modified version of the proof of Theorem 1 in [3], to which we refer to the reader for a more detailed exposition.

Proof We will simultaneously prove this for $E \subset \mathbb{F}_q^2$ and $E \subset \mathbb{Z}_q^2$. Here, we will use R_q to denote either $\mathbb{F}_q$ or $\mathbb{Z}_q$, and we will be more specific when necessary.

Our basic idea is to consider pairs of points $(v, w) \in E \times E$ and obtain a bound on the number of possible candidates for u to contribute a triple of the form (u, v, w) to $\Pi_{\alpha,\beta}(E)$. Consider $a = (a_x, a_y) \in R_q^2$, and notice that for a point $v \in E$, the set of points $L_\alpha(v)$ that determine the dot product α with v forms a line.

$$L_\alpha(v) = \left\{(x, y) \in R_q^2 : xv_x + yv_y = \alpha\right\}. \tag{1}$$

Also, v lies on a unique line containing the origin. We similarly define $L_\beta(v)$. Now, consider a second point $w \in E$. It is easy to see that if $|L_\alpha(v) \cap L_\beta(w)| > 1$, then v and w lie on the same line through the origin which implies that if v and w are on different lines through the origin, then $|L_\alpha(v) \cap L_\beta(w)| \leq 1$. We will use this dichotomy to decompose $E \times E$ into two sets:

$$\begin{aligned} A &= \{(v, w) \in E \times E : |L_\alpha(v) \cap L_\beta(w)| \leq 1, |L_\alpha(w) \cap L_\beta(v)| \leq 1\} \\ B &= (E \times E) \setminus A. \end{aligned}$$

Given $(v, w) \in A$, the pair can only be the last pair of at most one triple in $\Pi(E)$. This is of course only if $L_\alpha(v) \cap L_\beta(w)$ is a point in E. As there are no more than $|E|^2$ choices for pairs $(v, w) \in A$, the contribution to $\Pi(E)$ by point pairs in A is at most $|E|^2$

The analysis on the set of pairs in B is a bit more delicate. Consider an arbitrary pair, $(v, w) \in B$. Without loss of generality (possibly exchanging v with w or α with β) suppose $|L_\alpha(v) \cap L_\beta(w)| > 1$. Then, we get that

$$\begin{aligned} |L_\alpha(v) \cap L_\beta(w)| &> 1 \\ \left|\left\{(x, y) \in R_q^2 : xv_x + yv_y = \alpha\right\} \cap \left\{(x, y) \in R_q^2 : xw_x + yw_y = \beta\right\}\right| &> 1 \\ \left|\left\{(x, y) \in R_q^2 : xv_x + yv_y = \alpha \text{ and } xw_x + yw_y = \beta\right\}\right| &> 1. \end{aligned}$$

Namely, there will be more than one point with coordinates $(x, y) \in R_q^2$ satisfying

$$xv_x + yv_y = \alpha\left(\frac{xw_x + yw_y}{\beta}\right) = \frac{\alpha}{\beta}(xw_x + yw_y). \tag{2}$$

Note that β is a unit, and hence the quantity α/β is well defined. This restriction tells us that if $|L_\alpha(v) \cap L_\beta(w)| > 1$, then $|L_\alpha(v) \cap L_\beta(w')| = 0$, for any $w' \neq w$. This should not be surprising for if $\alpha = \beta$, then $L_\alpha(v) = L_\beta(w)$ forces $v = w$.

We pause for a moment to introduce an equivalence relation, say $\sim$, on the set of lines. Two lines $L_\alpha(v)$ and $L_\beta(w)$ are equivalent under $\sim$ if one can be translated to become a (possibly improper) subset of the other. It is clear that if $|L_\alpha(v) \cap L_\beta(w)| > 1$, then $L_\alpha(v) \sim L_\beta(w)$. The equivalence classes of $\sim$ keep track of the different "directions" that lines can have. So we can easily see that $L_\alpha(v) \sim L_\beta(v)$. Take note that if $R_q = \mathbb{Z}_q$, it is possible for two distinct lines to intersect in more than one point.

If $|L_\alpha(v) \cap L_\beta(w)| > 1$, then the pair (v, w) have no more than $\min\{|L_\alpha(v)|, |L_\beta(w)|\}$ possible choices for u to contribute a triple of the form (u, v, w) to $\Pi_{\alpha,\beta}(E)$. Now, we see that any other pair of points, say (v', w'), with $|L_\alpha(v') \cap L_\beta(w')| > 1$ and with $L_\alpha(v) \sim L_\alpha(v')$, will have $L_\alpha(v) \cap L_\alpha(v') = \emptyset$, and $L_\beta(w) \sim L_\beta(w')$, will have $L_\beta(w) \cap L_\beta(w') = \emptyset$. So any point u that contributes to a triple of the form $(u, v, w) \in \Pi_{\alpha,\beta}(E)$ can only contribute to a triple with a single pair (v, w) when $L_\alpha(v) \sim L_\beta(w)$.

Therefore, given any single equivalence class of $\sim$, there can be no more than $|E|$ choices for u to contribute a triple of the form (u, v, w) to $\Pi_{\alpha,\beta}(E)$ with $(v, w) \in B$. As there are no more than $|E|$ possible choices for equivalence classes of $L_\alpha(v)$ (as each point has only one associated equivalence class of $\sim$), there are no more than $|E|^2$ triples of the form $(u, v, w) \in \Pi_{\alpha,\beta}(E)$ with $(v, w) \in B$.

3.2 Proof of Theorem 4

Proof Let χ denote the canonical additive character of $\mathbb{F}_q$. By orthogonality, we have

$$\begin{aligned}|\Pi_{\alpha,\beta}(E)| &= |\{(x, y, z) \in E \times E \times E : x \cdot y = \alpha, x \cdot z = \beta\}\\ &= q^{-2} \sum_{s,t\in\mathbb{F}_q} \sum_{x,y,z\in E} \chi(s(x \cdot y - \alpha))\chi(t(\beta - x \cdot z))\\ &= q^{-2} \sum_{s,t\in\mathbb{F}_q} \sum_{x,y,z\in E} \chi(s\alpha)\chi(-t\beta)\chi(x \cdot (sy - tz))\\ &:= I + II + III,\end{aligned}$$

where I is the term with $s = t = 0$, II is the term with exactly one of s or t equal to zero, and III is the term with s and t both nonzero. Clearly

$$I = q^{-2} \sum_{s=t=0} \sum_{x,y,z\in E} \chi(s\alpha)\chi(-t\beta)\chi(x \cdot (sy - tz)) = |E|^3 q^{-2}.$$

For the second and third sums, we need the following known results.

Lemma 1 [8] *For any set $E \subset \mathbb{F}_q^d$, we have the bound*

$$\sum_{s \neq 0} \sum_{x,y \in E} \chi(s(x \cdot y - \gamma)) \leq |E| q^{\frac{d+1}{2}} \lambda(\gamma), \tag{3}$$

where $\lambda(\gamma) = 1$ for $\gamma \in \mathbb{F}_q^*$ and $\lambda(0) = \sqrt{q}$. Furthermore, we have

$$\sum_{s,s' \neq 0} \sum_{\substack{y,y' \in E \\ sy = s'y'}} \chi(\alpha(s' - s)) \leq |E| q \lambda(\gamma). \tag{4}$$

Note that the quantities in the above Lemma can be shown to be real numbers, so there is no need for absolute values. Now, separating the II term into two sums, each with exactly one of s or t zero,

$$II = q^{-2}|E| \left(\sum_{s \neq 0} \sum_{x,y \in E} \chi(s(x \cdot y - \alpha)) + \sum_{t \neq 0} \sum_{x,z \in E} \chi(t(x \cdot z - \beta)) \right)$$

From (3), it follows that $|II| \leq |E|^2 q^{\frac{d-3}{2}} (\lambda(\alpha) + \lambda(\beta))$. Finally, by the triangle-inequality, dominating a nonnegative sum over $x \in E$ by the same nonnegative sum over $x \in \mathbb{F}_q^d$, and applying Cauchy–Schwarz, we have

$$\begin{aligned}
|III| &\leq q^{-2} \sum_{x \in E} \left| \sum_{s \neq 0} \sum_{y \in E} \chi(s(x \cdot y - \alpha)) \right| \left| \sum_{t \neq 0} \sum_{z \in E} \chi(t(x \cdot z - \beta)) \right| \\
&\leq q^{-2} \sum_{x \in \mathbb{F}_q^d} \left| \sum_{s \neq 0} \sum_{y \in E} \chi(s(x \cdot y - \alpha)) \right| \left| \sum_{t \neq 0} \sum_{z \in E} \chi(t(x \cdot z - \beta)) \right| \\
&\leq q^{-2} \left(\sum_{x \in \mathbb{F}_q^d} \left| \sum_{s \neq 0} \sum_{y \in E} \chi(s(x \cdot y - \alpha)) \right|^2 \right)^{1/2} \\
&\quad \cdot \left(\sum_{x \in \mathbb{F}_q^d} \left| \sum_{t \neq 0} \sum_{z \in E} \chi(t(x \cdot z - \beta)) \right|^2 \right)^{1/2} \\
&=: q^{-2} III_\alpha \cdot III_\beta.
\end{aligned}$$

Now,

$$\begin{aligned}
III_\alpha^2 &= \sum_{x\in\mathbb{F}_q^d}\left|\sum_{s\neq 0}\sum_{y\in E}\chi(s(x\cdot y-\alpha))\right|^2\\
&=\sum_x\sum_{s,s'\neq 0}\sum_{y,y'\in E}\chi(s(x\cdot y-\alpha))\chi(-s'(x\cdot y'-\alpha))\\
&=\sum_x\sum_{s,s'\neq 0}\sum_{y,y'\in E}\chi(\alpha(s'-s))\chi(x\cdot(sy-s'y'))\\
&=q^d\sum_{s,s'\neq 0}\sum_{\substack{y,y'\in E\\ sy=s'y'}}\chi(\alpha(s'-s))\\
&\leq q^{d+1}|E|\lambda(\alpha)^2,
\end{aligned}$$

by (4). Similarly, we have $III_\beta \leq \sqrt{q^{d+1}|E|}\lambda(\beta)$. Combining these estimates yields

$$|III|\leq q^{d-1}|E|\lambda(\alpha)\lambda(\beta).$$

This completes the proof as we have

$$|\Pi_{\alpha,\beta}(E)|=\frac{|E|^3}{q^2}+R_{\alpha,\beta},$$

where

$$|R_{\alpha,\beta}|\leq |E|^2 q^{\frac{d-3}{2}}(\lambda(\alpha)+\lambda(\beta))+q^{d-1}|E|\lambda(\alpha)\lambda(\beta).$$

3.3 *Proof of Theorem 5*

The proof will imitate that of Theorem 4, so we omit some of the details. Let $\chi(\sigma)=\exp(2\pi i\sigma/q)$ be the canonical additive character of $\mathbb{Z}_q$, and identify E with its characteristic function. We use the following known bounds for dot-product sets in $\mathbb{Z}_q^d$.

Lemma 2 [5] *Suppose that $E\subset\mathbb{Z}_q^d$, where $q=p^\ell$ is the power of an odd prime. Suppose that $\gamma\in\mathbb{F}_q^\times$ is a unit. Then, we have the following upper bounds.*

$$\sum_{j\in\mathbb{Z}_q\setminus\{0\}}\sum_{x,y\in E}\chi(j(x\cdot y-\gamma))\leq 2|E|q^{\left(\frac{d-1}{2}\right)\left(2-\frac{1}{\ell}\right)+1} \tag{5}$$

and

$$\sum_{s,s'\in\mathbb{Z}_q\setminus\{0\}}\sum_{\substack{y,y'\in E\\ sy=s'y'}}\chi(\gamma(s'-s))\leq 2|E|q^{\frac{\ell d-d+1}{\ell}} \tag{6}$$

Note 2 The authors in [5] actually gave a slightly different bound than those in Lemma 2. For example in (5), they showed

$$\sum_{j\in\mathbb{Z}_q\setminus\{0\}}\sum_{x,y\in E}\chi(j(x\cdot y-\gamma))\leq\sum_{i=0}^{\ell-1}|E|q^{\left(\frac{d-1}{2}\right)\left(1+\frac{i}{\ell}\right)}\leq\ell|E|q^{\left(\frac{d-1}{2}\right)\left(2-\frac{1}{\ell}\right)+1}.$$

However, summing the geometric series removes the factor of ℓ in the estimate. Likewise, a factor of ℓ can be removed from the estimate in (6).

We proceed as before. Write

$$|\Pi_{\alpha,\beta}|=\frac{|E|^3}{q^2}+II+III,$$

where

$$II:=|E|q^{-2}\left(\sum_{s\neq 0}\sum_{x,y\in E}\chi(s\cdot(x\cdot y-\alpha))+\sum_{t\neq 0}\sum_{x,z\in E}\chi(s\cdot(x\cdot z-\beta))\right)$$

and

$$III:=q^{-2}\sum_{x\in E}\left(\sum_{s\neq 0}\sum_{y\in E}\chi(-s\alpha)\chi(s(x\cdot y))\right)\left(\sum_{t\neq 0}\sum_{z\in E}\chi(-t\beta)\chi(t(x\cdot z))\right).$$

Applying Lemma 2, we see that

$$|II|\leq 4|E|^2q^{-2}q^{\frac{d(2\ell-1)+1}{2\ell}},$$

while

$$|III|\leq q^{-2}\left(\sum_{x\in\mathbb{F}_q^d}\left|\sum_{s\neq 0}\sum_{y\in E}\chi(-s\alpha)\chi(s(x\cdot y))\right|^2\right)^{1/2}$$
$$\cdot\left(\sum_{x\in\mathbb{F}_q^d}\left|\sum_{t\neq 0}\sum_{z\in E}\chi(-t\beta)\chi(t(x\cdot z))\right|^2\right)^{1/2}$$
$$\leq 2|E|q^{-2}q^{\frac{\ell d-d+1}{\ell}}\leq 2|E|q^{-2}q^{\frac{d(2\ell-1)}{\ell}+\frac{1}{\ell}},$$

where in the last line, we reason as in the proof of Theorem 4, except with Lemma 2 in place of Lemma 1. This completes the proof.

References

1. J.A. Alvarez-Bermejo, J.A. Lopez-Ramos, J. Rosenthal, D. Schipani, Managing key multicasting through orthogonal systems. J. Discrete Math. Sci. Cryptogr
2. P. Bahls, Channel assignment on Cayley graphs. J. Graph Theory **67**, 169–177 (2011), https://doi.org/10.1002/jgt.20523
3. D. Barker, S. Senger, Upper bounds on pairs of dot products. J. Comb. Math. Comb. Comput
4. J.J. Benedetto, M. Fickus, Finite normalized tight frames. Adv. Comput. Math. **18**, 357–385 (2003)
5. D. Covert, A. Iosevich, J. Pakianathan, Geometric configurations in the ring of integers modulo p^{ℓ}. Indiana Univ. Math. J. **61**, 1949–1969 (2012)
6. P. Erdős, E. Szemerédi, in *On Sums and Products of Integers*. Studies in Pure Mathematics (Birkhäuser, Basel, 1983), pp. 213–218
7. K. Ford, Integers with a divisor in an interval. Ann. Math. **168**(2), 367–433 (2008)
8. D. Hart, A. Iosevich, D. Koh, M. Rudnev, Averages over hyperplanes, sum-product theory in vector spaces over finite fields and the Erdős-Falconer distance conjecture. Trans. Am. Math. Soc. **363**(6), 3255–3275 (2011)
9. A. Iosevich, S. Senger, Orthogonal systems in vector spaces over finite fields. Electron. J. Comb. **15** (2008)
10. N.H. Katz, C.Y. Shen, A slight improvement to Garaev's sum product estimate. Proc. Am. Math. Soc. **136**(137), 2499–2504 (2008)
11. S.V. Konyagin, I.D. Shkredov, New results on sums and products in $\mathbb{R}$. Proc. Steklov Inst. Math. **294**, 78 (2016), https://doi.org/10.1134/S0081543816060055
12. T. Tao, The sum-product phenomenon in arbitrary rings. Contrib. Discrete Math. **4**(2), 59–82 (2009)

Characteristic, Counting, and Representation Functions Characterized

Charles Helou

Abstract Given a set A of natural numbers, i.e., nonnegative integers, there are three distinctive functions attached to it, each of which completely determines A. These are the characteristic function $\chi_A(n)$ which is equal to 1 or 0 according as the natural number n lies or does not lie in A, the counting function $A(n)$ which gives the number of elements a of A satisfying $a \leq n$, and the representation function $r_A(n)$ which counts the ordered pairs (a, b) of elements $a, b \in A$ such that $a + b = n$. We establish direct relations between these three functions. In particular, we express each one of them in terms of each other one. We also characterize the representation functions by an intrinsic recursive relation which is a necessary and sufficient condition.

Keywords Representation function · Counting and characteristic functions

AMS Classification Numbers: 11B34 · 11B13 · 11B75

1 Introduction

Let A be a subset of $\mathbb{N} = \{0, 1, 2, \ldots\}$.

The characteristic function of A is defined by

$$\chi_A(n) = \begin{cases} 1 & \text{if } n \in A, \\ 0 & \text{if } n \notin A. \end{cases} \tag{1}$$

The counting function of A is defined by

$$A(n) = |A \cap [0, n]| = |\{a \in A : a \leq n\}| . \tag{2}$$

The representation function of A is defined by

C. Helou (✉)
Penn State Brandywine, 25 Yearsley Mill Rd, Media, PA 19063, USA
e-mail: cxh22@psu.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_9

$$r_A(n) = |\{(a,b) \in A \times A : a + b = n\}|. \tag{3}$$

Here $n \in \mathbb{N}$. But the three functions $chi_A(n)$, $A(n)$, $r_A(n)$, can be extended to all real numbers $x \in \mathbb{R}$, by simply replacing n by x in the above definitions.

Clearly, the functions $\chi_A(n)$ and $A(n)$ completely determine A, since the condition $n \in A$ is equivalent to either conditions: $\chi_A(n) = 1$ or $A(n) > A(n-1)$.

It is not as obvious that the function $r_A(n)$ completely determines A too, but it does, and several authors have written about this topic. In particular, the consequences of the equality, or of the partial equality from some point on, of the representation functions $r_A(n)$ and $r_B(n)$ of two sets of natural numbers A and B have been studied rather extensively [1, 2, 12–14, 17, 22, 23]. Other research has focused on studying the properties of representation functions, trying to characterize the class of representation functions and to determine which functions belong to this class. Also, many outstanding open problems and conjectures have been made in this respect [3–11, 15, 16, 18–21]. In particular, Melvyn B. Nathanson highlights in one of his papers [18] the following problem:

> What functions are representation functions?

The purpose of the present paper is twofold, first to establish relations between the three functions defined above, expressing each one of them in terms of each other one; and second, and more particularly, to attempt an answer to Nathanson's question. We thus give an intrinsic characterization of representation functions by proving that a function $f : \mathbb{N} \longrightarrow \mathbb{N}$ is the representation function of a subset A of $\mathbb{N}$ if and only if it satisfies the relation

$$f(n) = \frac{1}{2}\left(n + 1 - \sum_{k=0}^{n} (-1)^{f(2k)f(2(n-k))}\right),$$

for all $n \in \mathbb{N}$.

2 Preliminaries and Generating Series

We first note the following obvious relations between the characteristic and the counting functions of A:

$$A(n) = \sum_{k=0}^{n} \chi_A(k), \tag{4}$$

and

$$\chi_A(n) = A(n) - A(n-1), \tag{5}$$

for all $n \in \mathbb{N}$.

We then introduce the generating series of the three functions.

The generating series of $\chi_A(n)$ is

$$f_A(X) = \sum_{n=0}^{\infty} \chi_A(n) X^n = \sum_{a \in A} X^a, \tag{6}$$

also called the series associated with A.

The generating series of $r_A(n)$ is

$$g_A(X) = \sum_{n=0}^{\infty} r_A(n) X^n = \sum_{n=0}^{\infty} \left(\sum_{a,b \in A: a+b=n} 1 \right) X^n = \sum_{a,b \in A} X^{a+b} = \left(\sum_{a \in A} X^a \right)^2 = f_A(X)^2 \tag{7}$$

which is the square of the generating series $f_A(X)$ of $\chi_A(n)$.

The generating series of $A(n)$ is

$$h_A(X) = \sum_{n=0}^{\infty} A(n) X^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \chi_A(k) \right) X^n = \left(\sum_{n=0}^{\infty} X^n \right) \left(\sum_{n=0}^{\infty} \chi_A(n) X^n \right) = \frac{f_A(X)}{1-X}. \tag{8}$$

3 Relations Between the Counting and Representation Functions

Squaring the generating series of the counting function, we get

$$\frac{g_A(X)}{(1-X)^2} = \left(\frac{f_A(X)}{1-X} \right)^2 = \left(\sum_{n=0}^{\infty} A(n) X^n \right)^2 = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} A(k) A(n-k) \right) X^n. \tag{9}$$

On the other hand, as $g_A(X)$ is the generating series of $r_A(n)$, and as

$$\frac{1}{(1-X)^2} = \frac{d}{dX} \left(\frac{1}{1-X} \right) = \sum_{n=1}^{\infty} n X^{n-1} = \sum_{n=0}^{\infty} (n+1) X^n, \tag{10}$$

we also have

$$\begin{aligned} \frac{g_A(X)}{(1-X)^2} &= \left(\sum_{j=0}^{\infty} (j+1) X^j \right) \left(\sum_{k=0}^{\infty} r_A(k) X^k \right) = \sum_{n=0}^{\infty} \left(\sum_{j,k \in \mathbb{N}: j+k=n} (j+1) r_A(k) \right) X^n = \\ &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} (n-k+1) r_A(k) \right) X^n. \end{aligned} \tag{11}$$

Thus,

$$\frac{g_A(X)}{(1-X)^2} = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} A(k)A(n-k)\right)X^n = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}(n-k+1)r_A(k)\right)X^n. \tag{12}$$

This yields the following result, which gives the first relation between $A(n)$ and $r_A(n)$.

Proposition 1 *For every $n \in \mathbb{N}$, we have*

$$\sum_{k=0}^{n} A(k)A(n-k) = \sum_{k=0}^{n}(n-k+1)r_A(k) = \\ = (n+1)\sum_{k=0}^{n} r_A(k) - \sum_{k=0}^{n} kr_A(k), \tag{13}$$

i.e.,

$$\sum_{k=0}^{n}(n-k+1)r_A(k) = 2\sum_{0\le k<\frac{n}{2}} A(k)A(n-k) + \chi_{\mathbb{N}}\left(\frac{n}{2}\right)A\left(\frac{n}{2}\right)^2. \tag{14}$$

Corollary 1 *For $n \in \mathbb{N}$, we have*

$$r_A(n) = 2\sum_{0\le k<\frac{n}{2}} A(k)A(n-k) + \chi_{\mathbb{N}}\left(\frac{n}{2}\right)A\left(\frac{n}{2}\right)^2 - \sum_{k=0}^{n-1}(n-k+1)r_A(k). \tag{15}$$

Example 1 Applying the relation in the Corollary to $n = 0, 1, 2, \ldots$ in increasing order and back-substituting in terms of the $A(n)$'s alone, we get

$r_A(0) = A(0)^2,$

$r_A(1) = 2A(0)A(1) - 2A(0)^2,$

$r_A(2) = A(0)^2 - 4A(0)A(1) + 2A(0)A(2) + A(1)^2,$

$r_A(3) = 2A(0)A(1) - 4A(0)A(2) + 2A(0)A(3) - 2A(1)^2 + 2A(1)A(2)$

$$r_A(4) = 2A(0)A(2) - 4A(0)A(3) + 2A(0)A(4) + A(1)^2 - 4A(1)A(2) + \\ + 2A(1)A(3) + A(2)^2.$$

Proposition 2 *For any $n \in \mathbb{N}$, we have*

$$r_A(n) = \sum_{k=0}^{n} A(k)\left(A(n-k) - 2A(n-k-1) + A(n-k-2)\right) = \\ = \sum_{j=0}^{n}\left(A(j) - 2A(j-1) + A(j-2)\right)A(n-j). \tag{16}$$

Proof Using the generating series for $r_A(n)$ and for $A(n)$, we get

$$\begin{aligned}\sum_{n=0}^{\infty} r_A(n)X^n &= g_A(X) = \left(\frac{f_A(X)}{1-X}\right)^2 \left(1-2X+X^2\right) = \left(\sum_{n=0}^{\infty} A(n)X^n\right)^2 \left(1-2X+X^2\right)\\ &= \left(\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} A(k)A(n-k)\right)X^n\right)\left(1-2X+X^2\right)\\ &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} A(k)A(n-k)\right)X^n - 2\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} A(k)A(n-k)\right)X^{n+1}\\ &+ \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} A(k)A(n-k)\right)X^{n+2}\\ &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} (A(k)A(n-k) - 2A(k)A(n-1-k) + A(k)A(n-2-k))\right)X^n.\end{aligned}$$

Hence, for all $n \in \mathbb{N}$,

$$\begin{aligned} r_A(n) &= \sum_{k=0}^{n} A(k)\cdot(A(n-k) - 2A(n-k-1) + A(n-k-2))\\ &= \sum_{j=0}^{n} (A(j) - 2A(j-1) + A(j-2))\cdot A(n-j).\end{aligned}$$

Corollary 2 *For any $n \in \mathbb{N}$, we have*

$$r_A(n) = \sum_{\substack{j,k\in\mathbb{N}:\\ j+k\le n}} c_n(j,k)A(j)A(k), \tag{17}$$

where

$$c_n(j,k) = \begin{cases} 1, & \textit{if } j+k = n \textit{ or } n-2 \\ -2, & \textit{if } j+k = n-1 \\ 0, & \textit{otherwise}\end{cases}. \tag{18}$$

Proof By Proposition 2,

$$\begin{aligned} r_A(n) &= \sum_{k=0}^{n} A(k)A(n-k) - 2\sum_{k=0}^{n} A(k)A(n-k-1) + \sum_{k=0}^{n} A(k)A(n-k-2) =\\ &= \sum_{j+k=n} A(j)A(k) - 2\sum_{j+k=n-1} A(j)A(k) + \sum_{j+k=n-2} A(j)A(k) =\\ &= \sum_{j+k\le n} c_n(j,k)A(j)A(k).\end{aligned}$$

Corollary 3 *For any $n \in \mathbb{N}$, we have*

$$r_A(n) = \sum_{j=0}^{n} (\chi_A(j) - \chi_A(j-1))\, A(n-j) =$$

$$= \sum_{a \in A \smallsetminus (1+A)} A(n-a) - \sum_{b \in (1+A) \smallsetminus A} A(n-b), \qquad (19)$$

the last two sums being obviously restricted to $a \le n$ and $b \le n$.

Proof For any $k \in \mathbb{N}$, we have

$$A(k) = A(k-1) + \chi_A(k). \qquad (20)$$

Hence, for any $j \in \mathbb{N}$,

$$A(j) - 2A(j-1) + A(j-2) = A(j) - A(j-1) - (A(j-1) - A(j-2)) =$$

$$= \chi_A(j) - \chi_A(j-1) = \begin{cases} 0, \text{ if } j-1, j \in A \text{ or } j-1, j \notin A \\ 1, \text{ if } j \in A, j-1 \notin A \\ -1, \text{ if } j \notin A, j-1 \in A \end{cases}. \qquad (21)$$

This, in conjunction with Proposition 2, implies

$$r_A(n) = \sum_{j=0}^{n} (A(j) - 2A(j-1) + A(j-2)) \cdot A(n-j) =$$

$$= \sum_{j=0}^{n} (\chi_A(j) - \chi_A(j-1))\, A(n-j) =$$

$$= \sum_{\substack{0 \le j \le n: \\ j \in A,\ j-1 \notin A}} A(n-j) - \sum_{\substack{0 \le j \le n: \\ j-1 \in A,\ j \notin A}} A(n-j) =$$

$$= \sum_{a \in A \smallsetminus (1+A)} A(n-a) - \sum_{b \in (1+A) \smallsetminus A} A(n-b).$$

Example 2 By Corollary 2, $r_A(5) = \sum_{\substack{j,k \in \mathbb{N}: \\ j+k \le 5}} c_5(j,k) A(j) A(k)$, where

$$c_5(j,k) = \begin{cases} 1, & \text{if } j+k = 5 \text{ or } 3 \\ -2, & \text{if } j+k = 4 \\ 0, & \text{otherwise} \end{cases}.$$

Hence,

$$\begin{aligned} r_A(5) = &2A(0)A(3) - 4A(0)A(4) + 2A(0)A(5) + 2A(1)A(2) - 4A(1)A(3) \\ &+ 2A(1)A(4) - 2A(2)^2 + 2A(2)A(3). \end{aligned}$$

Similarly,

$$\begin{aligned} r_A(6) = &2A(0)A(4) - 4A(0)A(5) + 2A(0)A(6) + 2A(1)A(3) - 4A(1)A(4) \\ &+ 2A(1)A(5) + A(2)^2 - 4A(2)A(3) + 2A(2)A(4) + A(3)^2. \end{aligned}$$

$$\begin{aligned} r_A(7) = &2A(0)A(5) - 4A(0)A(6) + 2A(0)A(7) + 2A(1)A(4) - 4A(1)A(5) \\ &+ 2A(1)A(6) + 2A(2)A(3) - 4A(2)A(4) + 2A(2)A(5) - 2A(3)^2 + 2A(3)A(4). \end{aligned}$$

Remark 1 It follows from Corollary 2 that, for $n \in \mathbb{N}$,

$$\sum_{\substack{j,k\in\mathbb{N}: \\ j+k\leq n}} c_n(j,k) = \begin{cases} 0, \text{if } \ n \geq 1 \\ 1, \text{if } \ n = 0. \end{cases} \tag{22}$$

Indeed, for $n \geq 1$, we have

$$\begin{aligned} \sum_{\substack{j,k\in\mathbb{N}: \\ j+k\leq n}} c_n(j,k) &= \sum_{h=0}^{n} \sum_{j+k=h} c_n(j,k) = \sum_{j+k=n} c_n(j,k) + \sum_{j+k=n-1} c_n(j,k) + \sum_{j+k=n-2} c_n(j,k) = \\ &= \sum_{j+k=n} 1 + \sum_{j+k=n-1} (-2) + \sum_{j+k=n-2} 1 = \sum_{j=0}^{n} 1 - \sum_{j=0}^{n-1} 2 + \sum_{j=0}^{n-2} 1 = \\ &= (n+1) - 2n + (n-1) = 0. \end{aligned}$$

For $n = 0$, the sum reduces to $c_0(0,0) = 1$. □

Remark 2 For $n \geq 1$, we have

$$(1+A)(n) = A(n-1) = A(n) - \chi_A(n) = \begin{cases} A(n), & \text{if } n \notin A \\ A(n) - 1, & \text{if } n \in A. \end{cases} \tag{23}$$

So the last two sums in Corollary 3 have the same number of terms each if $n \notin A$, while the first of the two sums has one more term than the second one if $n \in A$.

Let $I = A \cap (1+A)$, $B = A \setminus (1+A) = A \setminus I$, and $C = (1+A) \setminus A = (1+A) \setminus I$. Then,

$$C(n) = (1+A)(n) - I(n) = A(n) - I(n) - \chi_A(n) = B(n) - \chi_A(n). \tag{24}$$

Let $B[n] = B \cap [0,n] = \{b_1 < b_2 < \cdots < b_{h-1} < b_h\}$ and $C[n] = C \cap [0,n] = \{c_1 < c_2 < \cdots < c_{h-1} \leq c_h\}$, where $c_h = c_{h-1}$ if $n \in A$, and $c_{h-1} < c_h$ if $n \notin A$. In view of Corollary 3,

$$r_A(n) = \sum_{b\in B[n]} A(n-b) - \sum_{c\in C[n]} A(n-c), \tag{25}$$

i.e.,

$$\begin{aligned} r_A(n) &= \sum_{k=1}^{h} A(n-b_k) - \sum_{k=1}^{h-1} A(n-c_k) - (1-\chi_A(n))\, A(n-c_h) = \\ &= \sum_{k=1}^{h} (A(n-b_k) - A(n-c_k)) + \chi_A(n)A(n-c_h). \end{aligned} \tag{26}$$

Note also that $b_k < c_k$ for $1 \le k < h$ (and for $k = h$ if $n \notin A$), since if $A = \{a_1 < a_2 < \cdots < a_n < \cdots\}$, then $1 + A = \{a_1 + 1 < a_2 + 1 < \cdots < a_n + 1 < \cdots\}$, and B (resp. C) is obtained from A (resp. $1 + A$) by removing the same set I. It follows that $n - b_k > n - c_k$ and therefore $A(n - b_k) \ge A(n - c_k)$ for $1 \le k < h$ (and for $k = h$ if $n \notin A$).

Remark 3 We have

$$f_A(X) = \frac{1}{1-X}\left(\sum_{a\in A\setminus(1+A)} X^a - \sum_{b\in(1+A)\setminus A} X^b\right). \tag{27}$$

Indeed, letting $I = A \cap (1 + A)$, so that $A \setminus (1 + A) = A \setminus I$ and $(1 + A) \setminus A = (1 + A) \setminus I$, we have

$$\begin{aligned} (1-X)f_A(X) = f_A(X) - Xf_A(X) &= \sum_{a\in A} X^a - \sum_{a\in A} X^{a+1} = \sum_{a\in A} X^a - \sum_{b\in 1+A} X^b = \\ &= \sum_{a\in I} X^a + \sum_{a\in A\setminus I} X^a - \left(\sum_{b\in I} X^b + \sum_{b\in(1+A)\setminus I} X^b\right) = \\ &= \sum_{a\in A\setminus I} X^a - \sum_{b\in(1+A)\setminus I} X^b. \end{aligned}$$

4 Relations Between the Characteristic and the Representation Functions

Just as the counting function $A(n)$ determines A, the number of representations function $r_A(n)$ also determines A.

Indeed, as

$$\sum_{n=0}^{\infty} r_A(n)X^n = g_A(X) = f_A(X)^2,$$

we have

$$f_A(X) = \sum_{n=0}^{\infty} \chi_A(n) X^n = g_A(X)^{1/2},$$

where if

$$A = \{a_1 < a_2 < \cdots < a_n < \cdots\},$$

then

$$f_A(X) = X^{a_1} \sum_{n=1}^{\infty} X^{a_n - a_1} = X^{a_1}(1 + Xv(X)),$$

with the power series $v(X) = \sum_{n=2}^{\infty} X^{a_n - a_1 - 1}$, so that

$$g_A(X) = X^{2a_1}(1 + Xu(X)),$$

with a power series $u(X)$, and therefore

$$f_A(X) = X^{a_1}(1 + Xu(X))^{1/2} = X^{a_1} \sum_{k=0}^{\infty} \binom{1/2}{k} X^k u(X)^k$$

is well defined. Moreover, replacing A by $-a_1 + A = \{a_n - a_1 : n \geq 1\}$, we may assume that $0 \in A$, so that $r_A(0) = 1$, and

$$g_A(X) = \sum_{n=0}^{\infty} r_A(n) X^n = 1 + Xu(X),$$

with

$$u(X) = \sum_{n=1}^{\infty} r_A(n) X^{n-1} = \sum_{n=0}^{\infty} r_A(n+1) X^n.$$

Then,

$$f_A(X) = \sum_{n=0}^{\infty} \chi_A(n) X^n = g_A(X)^{1/2} = (1 + Xu(X))^{1/2} = \sum_{k=0}^{\infty} \binom{1/2}{k} X^k u(X)^k, \tag{28}$$

where $\binom{x}{k}$ denotes the binomial coefficient, defined by

$$\binom{x}{k} = \frac{x(x-1)(x-2)\cdots(x-k+1)}{k!}, \tag{29}$$

for integers $k \geq 1$, while $\binom{x}{0} = 1$.

Proposition 3 *Assuming that* $0 \in A$, *we have, for* $n \geq 1$,

$$\chi_A(n) = \sum_{k=1}^{n} \binom{1/2}{k} \sum_{\substack{(n_1,\ldots,n_k)\in(\mathbb{N}^*)^k: \\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k) =$$

$$= \sum_{k=1}^{n} (-1)^{k-1} \frac{1\cdot 3\cdot 5\cdot \cdots \cdot (2k-3)}{k!2^k} \sum_{\substack{(n_1,\ldots,n_k)\in(\mathbb{N}^*)^k: \\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k). \quad (30)$$

Proof We have $f_A(X) = 1 + \sum\limits_{n=1}^{\infty} \chi_A(n) X^n$, and

$$g_A(X) = f_A(X)^2 = 1 + \sum_{n=1}^{\infty} r_A(n) X^n = 1 + Xu(X),$$

where $u(X)$ is a power series with nonnegative integer coefficients. So

$$f_A(X) = (1 + Xu(X))^{1/2} = 1 + \sum_{k=1}^{\infty} \binom{1/2}{k} X^k u(X)^k$$

$$= 1 + \sum_{k=1}^{\infty} \binom{1/2}{k} \left(\sum_{n=1}^{\infty} r_A(n) X^n \right)^k. \quad (31)$$

Moreover, for a positive integer $k \in \mathbb{N}^*$, we have

$$\left(\sum_{n=1}^{\infty} r_A(n) X^n \right)^k = \sum_{(n_1,\ldots,n_k)\in\mathbb{N}^{*k}} r_A(n_1)\cdots r_A(n_k) X^{n_1+\cdots+n_k} =$$

$$= \sum_{n=k}^{\infty} (\sum_{k=1}^{n} \sum_{\substack{(n_1,\ldots,n_k)\in\mathbb{N}^{*k}: \\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k)) X^n. \quad (32)$$

Hence,

$$f_A(X) = 1 + \sum_{n=1}^{\infty} \chi_A(n) X^n = 1 + \sum_{k=1}^{\infty} \binom{1/2}{k} \left(\sum_{n=1}^{\infty} r_A(n) X^n \right)^k =$$

$$= 1 + \sum_{k=1}^{\infty} \binom{1/2}{k} \sum_{n=k}^{\infty} (\sum_{k=1}^{n} \sum_{\substack{(n_1,\ldots,n_k)\in\mathbb{N}^{*k}: \\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k)) X^n =$$

$$= 1 + \sum_{n=1}^{\infty} (\sum_{k=1}^{n} \binom{1/2}{k} \sum_{\substack{(n_1,\ldots,n_k)\in\mathbb{N}^{*k}: \\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k)) X^n. \quad (33)$$

Thus, for $n \geq 1$,

$$\chi_A(n) = \sum_{k=1}^{n} \binom{1/2}{k} \sum_{\substack{(n_1,\ldots,n_k)\in\mathbb{N}^{*k}:\\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k).$$

Furthermore, for $k \geq 1$,

$$\begin{aligned}\binom{1/2}{k} &= \frac{(1/2)\,(-1/2)\,(-3/2)\cdots(1/2-k+1)}{k!} = \\ &= (-1)^{k-1}\frac{1\cdot 3\cdot 5\cdot\cdots\cdot(2k-3)}{k!2^k}.\end{aligned} \tag{34}$$

Therefore, for $n \geq 1$,

$$\chi_A(n) = \sum_{k=1}^{n} (-1)^{k-1}\frac{1\cdot 3\cdot 5\cdot\cdots\cdot(2k-3)}{k!2^k} \sum_{\substack{(n_1,\ldots,n_k)\in(\mathbb{N}^*)^k:\\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k).$$

Example 3 $\chi_A(1) = \frac{1}{2}r_A(1)$,

$$\begin{aligned}\chi_A(2) &= \frac{1}{2}r_A(2) - \frac{1}{8}r_A(1)^2,\\ \chi_A(3) &= \frac{1}{2}r_A(3) - \frac{1}{4}r_A(1)r_A(2) + \frac{1}{16}r_A(1)^3,\\ \chi_A(4) &= \frac{1}{2}r_A(4) - \frac{1}{4}r_A(1)r_A(3) - \frac{1}{8}r_A(2)^2 + \frac{3}{16}r_A(1)^2r_A(2) - \frac{5}{128}r_A(1)^4,\end{aligned}$$

$$\begin{aligned}\chi_A(5) =& \frac{1}{2}r_A(5) - \frac{1}{4}r_A(1)r_A(4) - \frac{1}{4}r_A(2)r_A(3) + \frac{3}{16}r_A(1)^2r_A(3) + \frac{3}{16}r_A(1)r_A(2)^2\\ &- \frac{5}{32}r_A(1)^3r_A(2) + \frac{7}{256}r_A(1)^5.\end{aligned}$$

Corollary 4 *For a subset A of $\mathbb{N}$ containing 0, the counting function of A is given by*

$$A(x) = \sum_{\substack{n\in\mathbb{N}\\ n\leq x}} \chi_A(n) = 1 + \sum_{\substack{n\in\mathbb{N}^*\\ n\leq x}} \sum_{k=1}^{n} \binom{1/2}{k} \sum_{\substack{(n_1,\ldots,n_k)\in(\mathbb{N}^*)^k:\\ n_1+\cdots+n_k=n}} r_A(n_1)\cdots r_A(n_k), \tag{35}$$

for $x \geq 0$.

Example 4 For $n \geq 6$,

$$
\begin{aligned}
\chi_A(n) = &\frac{1}{2}r_A(n) - \frac{1}{8}\sum_{n_1=1}^{n-1} r_A(n_1)r_A(n-n_1)+\\
&+\frac{1}{16}\sum_{n_1=1}^{n-2} r_A(n_1)\sum_{n_2=1}^{n-n_1-1} r_A(n_2)r_A(n-n_1-n_2)\\
&-\frac{5}{2^7}\sum_{n_1=1}^{n-3} r_A(n_1)\sum_{n_2=1}^{n-n_1-2} r_A(n_2)\sum_{n_3=1}^{n-n_1-n_2-1} r_A(n_3)r_A(n-n_1-n_2-n_3)\\
&+\cdots\\
&+(-1)^{n-1}\frac{1\cdot 3\cdot 5\cdot\cdots\cdot(2n-3)}{n!2^n}r_A(1)^n. \qquad (36)
\end{aligned}
$$

Remark 4 Conversely, $r_A(n)$ can be written in terms of $\chi_A(n)$. Indeed, using Proposition 2 and the relations between the characteristic and the counting functions, we get, for $n \in \mathbb{N}$,

$$
\begin{aligned}
r_A(n) = &\sum_{k=0}^{n} A(k)\left(A(n-k) - 2A(n-k-1) + A(n-k-2)\right) =\\
&= \sum_{k=0}^{n}\left(\sum_{j=0}^{k}\chi_A(j)\right)\left(\sum_{j=0}^{n-k}\chi_A(j) - 2\sum_{j=0}^{n-k-1}\chi_A(j) + \sum_{j=0}^{n-k-2}\chi_A(j)\right) =\\
&= \sum_{k=0}^{n}\left(\sum_{j=0}^{k}\chi_A(j)\right)\left(\chi_A(n-k) - \chi_A(n-k-1)\right),
\end{aligned}
$$

i.e.,

$$
r_A(n) = \sum_{k=0}^{n}\left(\chi_A(n-k) - \chi_A(n-k-1)\right)\sum_{j=0}^{k}\chi_A(j), \qquad (37)
$$

for all $n \in \mathbb{N}$.

Alternatively,

$$
\begin{aligned}
r_A(n) &= |\{0 \le j \le n : j, n-j \in A\}| = |\{0 \le j \le n : \chi_A(j) = \chi_A(n-j) = 1\}|\\
&= \sum_{j=0}^{n}\chi_A(j)\chi_A(n-j). \qquad (38)
\end{aligned}
$$

So

$$
r_A(n) = \sum_{j=0}^{n}\chi_A(j)\chi_A(n-j) = 2\sum_{\substack{0\le j<k\le n:\\ j+k=n}}\chi_A(j)\chi_A(k) + \chi_A\left(\frac{n}{2}\right). \qquad (39)
$$

5 Characterization of the Representation Function

Lemma 1 *For any $n \in \mathbb{N}$, we have*

$$r_A(2n) \equiv \chi_A(n) \pmod 2. \tag{40}$$

Proof In view of (39),

$$r_A(2n) = 2 \sum_{\substack{0 \le j < k \le n: \\ j+k=2n}} \chi_A(j)\chi_A(k) + \chi_A(n) \equiv \chi_A(n) \pmod 2.$$

Corollary 5 *For any $n \in \mathbb{N}$, we have*

$$n \in A \iff r_A(2n) \equiv 1 \pmod 2. \tag{41}$$

Definition 1 For an integer $a \in \mathbb{Z}$, let $res_2(a)$ denote the least nonnegative residue of a modulo 2, i.e.,

$$res_2(a) = \begin{cases} 0, & \text{if } a \equiv 0 \pmod 2 \\ 1, & \text{if } a \equiv 1 \pmod 2. \end{cases} \tag{42}$$

Remark 5 It is easy to verify that, for $a, b \in \mathbb{Z}$, we have

$$res_2(a) = \frac{1-(-1)^a}{2}, \tag{43}$$

$$res_2(ab) = res_2(a)res_2(b), \tag{44}$$

$$res_2(a^n) = res_2(a)^n = res_2(a), \quad \text{for } n \ge 1, \tag{45}$$

$$res_2(a+b) = res_2(a) + (-1)^a res_2(b) = res_2(b) + (-1)^b res_2(a). \tag{46}$$

$$res_2(-a) = res_2(a), \quad res_2(a-b) = res_2(a+b), \tag{47}$$

Remark 6 It follows from Lemma 1 and from Remark 5 that, for any $n \in \mathbb{N}$, we have

$$\chi_A(n) = res_2(r_A(2n)) = \frac{1-(-1)^{r_A(2n)}}{2}. \tag{48}$$

Hence,

$$f_A(X) = \sum_{n=0}^{\infty} \chi_A(n)X^n = \sum_{n=0}^{\infty} \frac{1-(-1)^{r_A(2n)}}{2} X^n = \frac{1}{2}\left(\frac{1}{1-X} - \sum_{n=0}^{\infty}(-1)^{r_A(2n)}X^n\right). \tag{49}$$

Moreover, for any $n \in \mathbb{N}$,

$$A(n) = \sum_{k=0}^{n} \chi_A(k) = \sum_{k=0}^{n} res_2(r_A(2k)) = \sum_{k=0}^{n} \frac{1-(-1)^{r_A(2k)}}{2} = \frac{1}{2}\left(n+1-\sum_{k=0}^{n}(-1)^{r_A(2k)}\right), \tag{50}$$

and

$$\begin{aligned} r_A(n) &= \sum_{k=0}^{n} \chi_A(k)\chi_A(n-k) = \sum_{k=0}^{n} res_2(r_A(2k))res_2(r_A(2(n-k))) \\ &= \sum_{k=0}^{n} res_2(r_A(2k)r_A(2(n-k))) = \sum_{k=0}^{n} \frac{1-(-1)^{r_A(2k)r_A(2(n-k))}}{2} \\ &= \frac{1}{2}\left(n+1-\sum_{k=0}^{n}(-1)^{r_A(2k)r_A(2n-2k)}\right). \end{aligned} \tag{51}$$

Thus, the values of $r_A(2n) \pmod 2$ completely determine A and therefore completely determine all values of $r_A(n)$. In other words, the representation function r_A of A is completely determined by the parity of its values at the even natural numbers.

Moreover, the relation (51) characterizes the representation function, as seen from the following Theorem.

Theorem 1 *Let $f : \mathbb{N} \longrightarrow \mathbb{N}$ be a function from the set of nonnegative integers $\mathbb{N}$ into itself, satisfying the relation*

$$f(n) = \frac{1}{2}\left(n+1-\sum_{k=0}^{n}(-1)^{f(2k)f(2(n-k))}\right), \quad \textit{for all } n \in \mathbb{N}. \tag{52}$$

Then $f = r_A$ is the representation function of the subset A of $\mathbb{N}$ defined by

$$A = \{n \in \mathbb{N} : f(2n) \equiv 1 \pmod 2\}. \tag{53}$$

Proof For any $n \in \mathbb{N}$, we have

$$\sum_{k=0}^{n}(-1)^{f(2k)f(2(n-k))} = \sum_{k\in I} 1 - \sum_{k\in J} 1 = |I| - |J|,$$

where

$$I = \{k \in \mathbb{N},\ 0 \le k \le n : f(2k) \equiv 0 \pmod 2 \ \text{ or } \ f(2(n-k)) \equiv 0 \pmod 2\}$$

and

$$J = \{k \in \mathbb{N},\ 0 \le k \le n : f(2k) \equiv f(2(n-k)) \equiv 1 \pmod 2\}.$$

Now, by definition of A, we have

$$I = \{k \in \mathbb{N},\ 0 \le k \le n : k \notin A \ \text{ or } \ n-k \notin A\}$$

and

$$J = \{k \in \mathbb{N},\ 0 \le k \le n : k \in A \ \text{ and } \ n-k \in A\}.$$

Clearly,

$$I \cup J = \{k \in \mathbb{N} :\ 0 \le k \le n\}, \quad \text{and} \quad I \cap J = \emptyset,$$

so that

$$|I| + |J| = |I \cup J| = |\{k \in \mathbb{N} :\ 0 \le k \le n\}| = n+1$$

and

$$\sum_{k=0}^{n} (-1)^{f(2k)f(2(n-k))} = |I| - |J| = n+1-2|J|.$$

It follows from this, and from the defining relation (52) of f, that

$$f(n) = \frac{1}{2}\left(n+1-\sum_{k=0}^{n}(-1)^{f(2k)f(2(n-k))}\right) = |J|.$$

Moreover,

$$|J| = |\{k \in \mathbb{N},\ 0 \le k \le n : k \in A \ \text{ and } \ n-k \in A\}| =$$
$$= \sum_{k \in A \text{ and } (n-k) \in A} 1 = \sum_{k=0}^{n} \chi_A(k)\chi_A(n-k) = r_A(n),$$

in view of (38).

Thus,

$$f(n) = r_A(n),$$

for all $n \in \mathbb{N}$.

Corollary 6 *A function $f : \mathbb{N} \longrightarrow \mathbb{N}$ is the representation function of a subset A of $\mathbb{N}$ if and only if it satisfies the relation*

$$f(n) = \frac{1}{2}\left(n+1-\sum_{k=0}^{n}(-1)^{f(2k)f(2(n-k))}\right), \quad \textit{for all } n \in \mathbb{N}.$$

Proof This follows from (51) in Remark 6 and from Theorem 1.

Remark 7 Corollary 6 provides a characterization of representation functions. It is easier to characterize the characteristic and the counting functions. Indeed, any function $f : \mathbb{N} \longrightarrow \{0, 1\}$ is the characteristic function of a unique subset A of $\mathbb{N}$, namely of $A = f^{-1}(1)$. Also, any increasing (not necessarily strictly increasing) function $f : \mathbb{N} \longrightarrow \mathbb{N}$ is the counting function of a unique subset A of $\mathbb{N}$, namely of $A = \{n \in \mathbb{N} : f(n) > f(n-1)\}$, where we set, by definition, $f(-1) = 0$.

References

1. Y.-G. Chen, M. Tang, Partitions of natural numbers with the same representation functions. J. Number Theory **129**(11), 2689–2695 (2009)
2. J. Cilleruelo, M.B. Nathanson, Dense sets of integers with prescribed representation functions. European J. Comb. **34**(8), 1297–1306 (2013)
3. A. Dubickas, A basis of finite and infinite sets with small representation function. Electron. J. Comb. **19**(1), Paper 6, 16 pp (2012)
4. A. Dubickas, On the supremum of the representation function of a sumset. Quaest. Math. **37**(1), 1–8 (2014)
5. P. Erdős, W.H.J. Fuchs, On a problem of additive number theory. J. Lond. Math. Soc. **31**, 67–73 (1956)
6. P. Erdős, P. Turán, On a problem of Sidon in additive number theory. J. Lond. Math. Soc. **16**, 212–215 (1941)
7. G. Grekos, L. Haddad, C. Helou, J. Pihko, On the Erdős-Turán conjecture. J. Number Theory **102**, 339–352 (2003)
8. G. Grekos, L. Haddad, C. Helou, J. Pihko, The class of Erdős-Turán sets. Acta Arith. **117**, 81–105 (2005)
9. G. Grekos, L. Haddad, C. Helou, J. Pihko, Representation functions, Sidon sets and bases. Acta Arith. **130**(2), 149–156 (2007)
10. G. Grekos, L. Haddad, C. Helou, J. Pihko, Supremum of representation functions. Integers **11**, A30, 14 pp (2011)
11. H. Halberstam, K.F. Roth, *Sequences* (Clarendon Press, Oxford, 1966)
12. J. Lee, Infinitely often dense bases for the integers with a prescribed representation function. Integers **10**(A24), 299–307 (2010)
13. V.F. Lev, Reconstructing integer sets from their representation functions. Electron. J. Comb. **11**(1), Research Paper 78, 6 pp (2004)
14. M.B. Nathanson, Representation functions of sequences in additive number theory. Proc. Am. Math. Soc. **72**, 16–20 (1978)
15. M.B. Nathanson, in *The inverse problem for representation functions of additive bases*. Number Theory. (Springer, New York, 2004), pp. 253–262
16. M.B. Nathanson, Representation functions of additive bases for abelian semigroups. Int. J. Math. Math. Sci. 29–32, 1589–1597 (2004)
17. M.B. Nathanson, Every function is the representation function of an additive basis for the integers. Port. Math. (N.S.) **62**(1), 55–72 (2005)
18. M.B. Nathanson, Inverse problems for representation functions in additive number theory. Surveys in number theory, 89–117, Dev. Math. **17** (Springer, New York, 2008)
19. C. Sándor, Range of bounded additive representation functions. Period. Math. Hungar. **42**(1–2), 169–177 (2001)
20. C. Sándor, Partitions of natural numbers and their representation functions. Integers **4**, A18, 5 pp (2004)
21. A. Sárközy, V.T. Sós, On additive representation functions. The mathematics of Paul Erdős, I, 129–150, Algorithms Comb. **13** (Springer, Berlin, 1997)

22. M. Tang, Partitions of the set of natural numbers and their representation functions. Discrete Math. **308**(12), 2614–2616 (2008)
23. M. Tang, Dense sets of integers with a prescribed representation function. Bull. Aust. Math. Soc. **84**(1), 40–43 (2011)

Partitions into Parts Simultaneously Regular, Distinct, And/or Flat

William J. Keith

Abstract We explore partitions that lie in the intersection of several sets of classical interest: partitions with parts indivisible by m, appearing fewer than m times, or differing by less than m. We find results on their behavior and generating functions: more results for those simultaneously regular and distinct, fewest for those distinct and flat. We offer some conjectures in the area.

Keywords Partitions

1 Introduction

A partition of n is a nonincreasing sequence of positive integers, which sums to n, i.e., $\lambda \vdash n$ if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$ and $\lambda_1 + \cdots + \lambda_k = n$. Their study was initiated by Euler, who proved the usual first result seen by a student of the area, namely

Theorem 1 *The number of partitions of n in which all parts are odd equals the number of partitions of n in which parts are distinct.*

The theorem was proved by a hands-on combinatorial mapping found by J. J. Sylvester and then generalized to all moduli by a more general mapping given by his student Glaisher:

Theorem 2 *The number of partitions of n in which no part is divisible by m equals the number of partitions of n in which parts appear fewer than m times.*

The map of Glaisher's proof can be extended to a general mapping on all partitions: if, given j not divisible by m, the part jm^k appears $\sum_{\ell=0}^{\infty} a_{k,\ell} m^\ell$ times in λ, written in the base m expansion, then in $\phi(\lambda)$, write the part jm^ℓ appearing $a_{k,\ell} m^k$ times for each nonzero $a_{k,\ell}$. If no part in λ is divisible by m ($a_{k,\ell} = 0$ for $k > 0$), then in $\phi(\lambda)$, no part will appear m or more times, and vice versa.

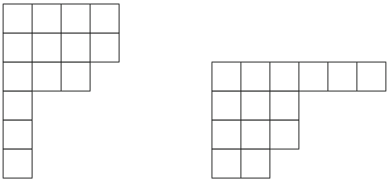
W. J. Keith (✉)
Michigan Technological University, Houghton, MI, USA
e-mail: wjkeith@mtu.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_10

The fixed points of the map are precisely those partitions in which parts are not divisible by m (called *m-regular* partitions) and in which no part appears m or more times (a partition with the latter property we will call *m-distinct*).

The fixed points of an interesting map ought to be of interest, but a search of the literature suggests that little work has been done with these partitions, with the strong exception of the $m = 2$ case, partitions into distinct odd parts. Denoting the number of such partitions of n by $p_{2,2}(n)$, we have that $p_{2,2}(n) \equiv p(n) \pmod 2$, and since the parity of $p(n)$ is a long-standing question of great interest, $p_{2,2}(n)$ has been much studied, often for its parity properties.

Equal in number with these subsets of partitions of n is the set of those in which the differences between consecutive parts are less than m, and the smallest part is less than m. The proof is by *conjugation*, which is defined in terms of the *Ferrers diagram* of a partition; a set of unit squares justified to the origin in the fourth quadrant, in which the ith row below the x-axis has λ_i squares. The conjugate of λ, λ' is the partition with Ferrers diagram given by the reflection of the diagram of λ across the diagonal. An example is:



$$\lambda = (4, 4, 3, 1, 1, 1) \vdash 14, \ \ \lambda' = (6, 3, 3, 2)$$

Now it is easy to see that partitions with parts appearing fewer then m times conjugate to partitions with differences less than m and smallest part less than m. For the remainder of this paper, we will call the latter *m-flat* partitions, after [1].

Remark 1 : A direct map between m-flat and m-regular partitions was developed in [1], translated from the German in an appendix to [2]. (Rather, several involutions on all partitions were constructed, some of which restrict to a map between these sets.) The fixed points are, again, those that simultaneously satisfy both conditions.

Conjugation does not fix those partitions simultaneously m-flat and m-distinct, but it does fix the class. In fact, the fixed points of conjugation are in bijection with partitions into distinct odd parts (read vertical-to-horizontal hook lengths). It might be of interest to develop an involution on partitions which does fix this class; given the utility of conjugation as a theorem-proving tool, its other properties might be of great use. (If the involution fixes all m-flat, m-distinct partitions, it will necessarily have at least some other fixed points, as the parity of this subset does not necessarily match that of the number of partitions of n.)

In the remainder of the paper, we explore those partitions that simultaneously satisfy two of these three conditions, generalizing the question to moduli not necessarily equal for the two conditions. In Sect. 2, we discuss partitions simultaneously s-regular and t-distinct; we can say the most about these. In Sect. 3, we discuss

s-regular, t-flat partitions; we can say a few things about these, mostly when $s|t$. In Sect. 4, we discuss s-distinct, t-flat partitions; about these we can say little, despite the fact that they have the simplest diagrammatic interpretation. In the last section, we close with some comments and possible lines of future investigation.

2 Regular and Distinct

The generating function for partitions which are s-regular and t-distinct is easy to write down: it is

Theorem 3

$$P_{R,D}^{(s,t)}(q) = \sum_{n=0}^{\infty} p_{R,D}^{(s,t)}(n)q^n = \prod_{k=1}^{\infty} \frac{(1-q^{sk})(1-q^{tk})}{(1-q^k)(1-q^{stk})}.$$

$P_{R,D}^{(s,t)}(q)$ is an η-quotient, i.e., (up to a factor of a power of q) a quotient of functions of the form $\eta(z) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n)$, $q = e^{2\pi i z}$. By work of Stephanie Treneer [3], it is known that all such functions are weakly holomorphic modular forms, and so it is likely that they will exhibit many congruences. Numerical experimentation quickly finds many. For instance,

Theorem 4 *For $n \geq 0$,*

$$p_{R,D}^{(2,2)}(125n+99) \equiv 0 \pmod 5 \quad \textit{(Rødseth)} \tag{1}$$

$$p_{R,D}^{(3,3)}(4n+2) \equiv 0 \pmod 2 \tag{2}$$

$$p_{R,D}^{(2,5)}(4n+3) \equiv 0 \pmod 2 \quad \textit{and} \tag{3}$$

$$\sum_{n=0}^{\infty} p_{R,D}^{(2,5)}(4n+1)q^n \equiv f_5 \pmod 2. \tag{4}$$

Here and in the remainder of the paper, we employ the shorthand notation f_k for

$$f_k = \prod_{n=1}^{\infty}(1-q^{nk}) = q^{-k/24}\eta(kz) = (q^k;q^k)_{\infty}.$$

Furthermore, when we state for two power series $f(q) = \sum_{n=n_0}^{\infty} a(n)q^n$ and $g(q) = \sum_{n=n_1}^{\infty} b(n)q^n$ that $f(q) \equiv_p g(q)$, we mean that $a(n) \equiv b(n) \pmod p$ for all n.

Proof As noted, the first clause of Theorem 4 was proved by Øystein Rødseth [4], who was studying the properties of $p_{2,2}(n)$.

To prove the other clauses, we use several identities that dissect various η-products. All of the ones we use here can be found in [5]. In addition, it is useful to note that $f_k^p \equiv_p f_{kp}$ for p any prime.

For clause (2), we will need:

$$\frac{f_3{}^3}{f_1} = \frac{f_4^3 f_6^2}{f_2^2 f_{12}} + q\frac{f_{12}^3}{f_4} \tag{5}$$

$$\frac{1}{f_1 f_3} = \frac{f_8^2 f_{12}^5}{f_2^2 f_4 f_6^4 f_{24}^2} + q\frac{f_4^5 f_{24}^2}{f_2^4 f_6^2 f_8^2 f_{12}}. \tag{6}$$

Now observe that

$$P_{R,D}^{(3,3)}(q) = \frac{f_3^2}{f_1 f_9} = \frac{f_3^3}{f_1} \cdot \frac{1}{f_3 f_9} = \left(\frac{f_4^3 f_6^2}{f_2^2 f_{12}} + q\frac{f_{12}^3}{f_4}\right)\left(\frac{f_{24}^2 f_{36}^5}{f_6^2 f_{12} f_{18}^4 f_{72}^2} + q^3\frac{f_{12}^5 f_{72}^2}{f_6^4 f_{18}^2 f_{24}^2 f_{36}}\right).$$

Expanding out the multiplication and reducing modulo 2 where possible, we find

$$P_{R,D}^{(3,3)}(q) \equiv_2 \frac{f_4^3 f_{24}^2 f_{36}^5}{f_2^2 f_{12}^2 f_{18}^4 f_{72}^2} + q\,(\ldots) + q^3\,(\ldots) + q^4\frac{f_{12}^8 f_{72}^2}{f_4 f_6^4 f_{18}^2 f_{24}^2 f_{36}}.$$

The elided terms are all of the form q^{2n+1} and so are irrelevant to the theorem. Furthermore, neither of the other summands contains powers of the form q^{4n+2} with odd coefficients, since all factors of f_2, f_4, and f_{18} are raised to even powers, and we may invoke $f_2^2 \equiv_2 f_4$. Hence only powers q^{4n} in these summands may have noneven coefficients, and hence any coefficient $p_{R,D}^{(3,3)}(4n+2) \equiv 0 \pmod 2$.

For clauses (3) and (4), we additionally require the identity

$$\frac{f_5}{f_1} = \frac{f_8 f_{20}^2}{f_2^2 f_{40}} + q\frac{f_4^3 f_{10} f_{40}}{f_2^3 f_8 f_{20}}.$$

Thus,

$$P_{R,D}^{(2,5)}(q) = \frac{f_2 f_5}{f_1 f_{10}} = \frac{f_8 f_{20}^2}{f_2 f_{10} f_{40}} + q\frac{f_4^3 f_{40}}{f_2^2 f_8 f_{20}}.$$

Again, since $f_2^2 \equiv_2 f_4$, no term in the latter summand has a noneven coefficient on a power q^{4n+3}, and so claim (3) holds. Further using this identity to reduce the summand, we find that

$$P_{R,D}^{(2,5)}(q) \equiv_2 \cdots + q f_{20},$$

where the elided terms are even powers. Extracting terms of the form q^{4n+1} and making the substitution $q^4 \to q$, we obtain clause (4), and the theorem holds.

Many other such congruences can easily be found and proved through similar methods.

2.1 Symmetry

Another observation of interest is the symmetry of the generating function, which yields the immediate result

Theorem 5 *The number of partitions of n which are s-regular and t-distinct equals the number of partitions of n which are t-regular and s-distinct.*

It is then reasonable to ask for a map that realizes this equality; as it turns out, if s and t are coprime, a double use of Glaisher's bijection does the job. Denote by ϕ_m Glaisher's involution with modulus m. Then we have the following.

Theorem 6 *If s and t are coprime, then $\phi_s\phi_t$ maps s-regular, t-distinct partitions to t-regular, s-distinct partitions.*

Although this could have been observed earlier, we will see in the midst of this proof that

Corollary 1 *If s and t are coprime, the number of s-regular, t-distinct partitions are equal to the number of partitions simultaneously s-regular and t-regular.*

However, the s-distinct, t-distinct partitions are merely the s-distinct partitions assuming s is the smaller of the two values.

Proof If s and t are coprime, then let λ be an s-regular, t-distinct partition. The first step ϕ_t replaces parts of sizes jt^k with appearances of the part j; since jt^k was not divisible by s, neither is j, and so the result is also an s-regular, t-regular partition; all such partitions can arise this way (ϕ_s or ϕ_t reverses the map in the desired direction) and so the corollary follows. At this point, applying ϕ_s produces an s-distinct partition which is still t-regular, since j is not divisible by t and js^k is also not divisible by t, as these are coprime.

Example Consider $\lambda = (4, 2, 1)$ as a 7-regular, 2-distinct partition. Then $\phi_2(4, 2, 1) = (1, 1, 1, 1, 1, 1, 1)$, which is both 7-regular and 2-regular. Then $\phi_7(1, 1, 1, 1, 1, 1, 1) = (7)$, which is 2-regular and 7-distinct.

If s and t are not coprime, then, during a visit to Michigan Tech, it was conjectured by Bridget Tenner of DePaul University that

Conjecture 1 *Iteration of the previous map suffices to produce a bijection. That is, there exists ℓ, varying with λ, such that $(\phi_s\phi_t)^\ell$ maps an s-regular, t-distinct partition λ to a unique s-distinct, t-regular partition, with no intervening $(\phi_s\phi_t)^k$ being s-regular and t-distinct.*

Since ϕ_s and ϕ_t are involutions and the set of partitions of n is finite, the sequence of images $(\phi_s\phi_t)^\ell(\lambda)$ eventually cycles for any λ; the claim then becomes that such a sequence starting at an s-regular, t-distinct partition will encounter a t-regular, s-distinct partition before encountering another s-regular, t-distinct partition.

(The author must retract a claim made during the presentation at CANT 2016 that the proof of this conjecture is nontrivial but straightforward. For an indication of the curious behavior that such a sequence can display, the reader might examine the behavior of (50, 50, 50, 50, 50, 50) as a 6-regular, 10-distinct partition; the map works, but requires 65 iterations, and actually passes through (50, 50, 50, 50, 50, 50) again halfway through the 63rd step.)

2.2 McKay-Thompson Series

For a final comment on the $P_{R,D}^{(s,t)}$ partitions, there is a connection which may be spurious but could be very interesting if it is true in any depth.

To first give some background, recall the j-invariant

$$j(\tau) = \frac{1}{q} + 196884q + 21493760q^2 + \dots.$$

Monstrous Moonshine [6] is the conjecture, now theorem [7], that the coefficients of this function are sums of the dimensions of irreducible representations of the Monster group M: $1 = 1, 196884 = 196883 + 1, 21296876 + 196883 + 1$, etc. That is, there is an ∞-dimensional graded representation of M whose graded dimension is given by these coefficients, and whose lower-weight pieces decompose into irreps of dimension 1, 196883, 21296876, etc., which sum in fairly simple ways to the coefficients of j. The graded dimension is the graded trace of the identity element $e \in M$; the McKay-Thompson series T_g is the generating function for the graded traces of nontrivial elements $g \in M$.

If we search the invaluable Online Encyclopedia of Integer Sequences [8] for the coefficients of the generating function $P_{(R,D)}^{(3,3)}$, we find that they match OEIS sequence A112194 [9]: "McKay-Thompson series of class 54c for the Monster group." McKay-Thompson series are often of the form $\frac{f_s f_t}{f_1 f_{st}}$, usually shifted by a power of q and with a substitution $q \to q^\ell$; for instance, the generating function for this McKay-Thompson series is actually $\frac{1}{q} P_{(R,D)}^{(3,3)}(q^6)$. With a little more searching, we find many of these in the OEIS: $(s, t) = (2, 5)$ gives class 60F; $(s, t) = (3, 4)$ gives 48h; $(s, t) = (5, 7)$ gives class 35B, but $(s, t) = (3, 7)$ is not there.

So one wonders: is there a simple, partition-theoretic interpretation of these generating functions in terms of the dimensions being counted? That is:

Question 1: Are there structures in M or its representations which are in bijection with partitions into, say, partitions into parts not divisible by 2 and appearing less than 5 times, which yield the graded traces of elements in the apparently associated conjugacy classes?

Since any (s, t) is a permissible pair for $P_{R,D}^{(s,t)}$, but McKay-Thompson series are restricted by the Monster group itself, such combinatorial descriptions might be "coincidental"; but, given the great interest in the structure of the Monster group and its subgroups, even descriptions in a few cases might be valuable and interesting in their own right.

3 Regular and Flat

In this section, we discuss partitions simultaneously s-regular and t-flat. For these, we can write down the generating function in some restricted cases: namely, when $s|t$, much more easily if $s = t$.

We defined $(q;q)_\infty$ earlier; it now becomes useful for us to generalize to the notation $(a;q)_n = \prod_{i=0}^{n-1}(1-aq^i)$, in which case $(q;q)_\infty = \lim_{n\to\infty}(q;q)_n$. The empty product is 1, so $(a;q)_0 = 1$.

3.1 *t-regular, t-flat Partitions*

When $s = t$, our task is easiest.

Theorem 7 *The generating function for partitions both t-regular and t-flat is*

$$P_{R,F}^{(t,t)} = \sum_{j=0}^{\infty}\sum_{i=0}^{j} \frac{(-1)^i q^{\binom{i+1}{2}t+j-i}(q^{(i+1)t};q^t)_{j-i}}{(q;q)_{j-i}}.$$

Proof The proof strategy is to note that a t-regular partition can be broken into its flat part, plus differences of multiples of t:

$$\begin{array}{l} a_1\ t\ t\ t \\ a_2\ t\ t \\ a_3\ t\ t \\ a_4\ t\ t \\ a_5 \end{array}$$

where the a_i are nonzero residues modulo t, and each t represents t added to the part. If $a_{i+1} \le a_i$, then the number of t units in the flat part of λ_i equals the number of such units in λ_{i+1}, whereas if $a_{i+1} > a_i$, the number of t units in λ_i is 1 greater than the number in λ_{i+1}. For example, if the above diagram represents the 5-regular partition $(17, 13, 11, 11, 4)$, then the flat part of the partition is $(12, 8, 6, 6, 4)$. An amount 5 was added to parts 1 through 4. Notice that the t-flat part of a t-regular partition is still t-regular; more generally, the s-flat part of a t-regular partition is still t-regular if t divides s.

The amounts added will be multiples of t of sizes up to t times the number of parts of the partition; thus, the generating function for t-regular partitions with exactly j

parts equals the generating function for t-flat, t-regular partitions with exactly j parts, times the generating function for partitions into multiples of t no larger than jt.

Thus, suppressing the t for now and referring only to the generating functions for partitions of the desired type into exactly j parts, we have

$$P_{R,F}^{(\text{j parts})}(q) \times \frac{1}{(q^t;q^t)_j} = P_R^{(\text{j parts})}.$$

Next, we must determine the generating function for t-regular partitions into exactly j parts. We do so by considering all partitions of inclusion–exclusion on the number of sizes of parts of λ divisible by t, obtaining the following:

Lemma 1

$$P_R^{(j\ parts)} = \sum_{i=0}^{j} \frac{q^{j-i}}{(q;q)_{j-i}} (-1)^i q^{\binom{i+1}{2}t} \frac{1}{(q^t;q^t)_i}.$$

The argument is as follows: begin with $j-i$ guaranteed parts of size 1 and add any desired amount; add exactly i sizes of part divisible by t, from t to it; finally, add additional multiples of t to these parts alone. Count those in which we guaranteed at least i different sizes of part divisible by t with $(-1)^i$; by inclusion–exclusion, the resulting sum counts exactly those partitions with no part divisible by t.

So, combining identities,

$$P_{R,F}^{(j\text{ parts})}(q) \times \frac{1}{(q^t;q^t)_j} = P_R^{(j\text{ parts})} = \sum_{i=0}^{j} \frac{q^{j-i}}{(q;q)_{j-i}} (-1)^i q^{\binom{i+1}{2}t} \frac{1}{(q^t;q^t)_i}.$$

Multiplying through, we obtain

$$P_{R,F}^{(j\text{ parts})}(q) = \sum_{i=0}^{j} \frac{q^{j-i}}{(q;q)_{j-i}} (-1)^i q^{\binom{i+1}{2}t} (q^{(i+1)t};q^t)_{j-i}.$$

Summing over numbers of parts j, we complete the proof.

An alternative version of this generating function has more terms but is combinatorially interesting. Observe that, given a vector ρ of nonzero residues modulo t, the t-flat partition with residues equal to ρ when read in order is uniquely given. The number of units of size t below residue ρ_i is precisely the number of pairs (ρ_k, ρ_{k+1}) with $k \geq i$ for which $\rho_k < \rho_{k+1}$, i.e., the number of ascents in the multiset permutation, identified by ρ, of the multiset of residues listed.

Example Suppose that $t = 3$ and that ρ consists of two 1 s and 2 s each. The possible partitions are:

$$\begin{array}{l}2\\2\\1\\1\end{array},\ \begin{array}{l}2\ 3\\1\ 3\\2\\1\end{array},\ \begin{array}{l}2\ 3\\1\ 3\\1\ 3\\2\end{array},\ \begin{array}{l}1\ 3\\2\\2\\1\end{array},\ \begin{array}{l}1\ 3\ 3\\2\ 3\\1\ 3\\2\end{array},\ \begin{array}{l}1\ 3\\1\ 3\\2\\2\end{array}$$

The t-*complement* ρ^c of ρ is the vector $(t+1-\rho_1, \ldots, t+1-\rho_k)$; since ascents in ρ map to descents in ρ^c, the number of t units depending from the residue vector is easily seen to be the major index of ρ^c. It is well-known (see for instance [10]) that $maj(\rho^c)$ is equidistributed with $maj(\rho)$ over all permutations ρ of the same multiset, and that if ρ contains i_1 ones, i_2 twos, …, and i_{t-1} residues $t-1$, then the q-multinomial coefficient

$$\begin{bmatrix} i_1 + \cdots + i_{t-1} \\ i_1, \ldots, i_{t-1} \end{bmatrix}_q := \frac{(q;q)_{i_1+\cdots+i_{t-1}}}{(q;q)_{i_1} \ldots (q;q)_{i_{t-1}}}$$

is the generating function for the major index over all multiset permutations of ρ, i.e.,

$$\begin{bmatrix} i_1 + \cdots + i_{t-1} \\ i_1, \ldots, i_{t-1} \end{bmatrix}_q = \sum_{\sigma(\rho)} q^{maj(\sigma(\rho))}$$

where summation is over all multiset permutations of ρ.

Since the units are of size t, we find that the generating function for the t-regular, t-flat partitions with residue vector some permutation of ρ, which we may denote by $P_{R,F}^{(t,t;\rho)}(q)$, is given by

Theorem 8

$$P_{R,F}^{(t,t;\rho)} = q^{i_1+\cdots+(t-1)i_{t-1}} \begin{bmatrix} i_1 + \cdots + i_{t-1} \\ i_1, \ldots, i_{t-1} \end{bmatrix}_{q^t}.$$

Finally, we note that if our partitions are s-regular and t-flat with s dividing t, a small variation of the previous argument suffices; we are restricted to a subset of the possible residues modulo t. In the first form of the generating function, when constructing $P_R^{(j\text{ parts})}$, we additionally include–exclude parts with residues divisible by s, producing additional summations. For instance, if $2s = t$, we have

$$P_R^{(j\text{ parts})} = \sum_{i,k} \frac{q^{j-i-k}}{(q;q)_{j-i-k}} (-1)^i q^{\binom{i+1}{2}t} (-1)^i q^{\binom{i}{2}t+ks} \frac{1}{(q^t;q^t)_i (q^t;q^t)_k}$$

and hence,

$$P_{R,F}^{(j\text{ parts})} = \sum_{i,k} \frac{q^{j-i-k}}{(q;q)_{j-i-k}} (-1)^i q^{\binom{i+1}{2}t} (-1)^i q^{\binom{i}{2}t+ks} \frac{(q^t;q^t)_j}{(q^t;q^t)_i (q^t;q^t)_k}.$$

Other than restricting the permissible residue vectors ρ, the second form of the generating function is unchanged.

3.2 Other Observations

Unlike the other two classes discussed in this paper, simple calculation shows us that s-regular, t-flat partitions are not symmetric in s and t. For instance, (1, 1, 1) is 3-regular and 2-flat, and also 2-regular and 3-flat; (2, 1) is 3-regular and 2-flat, but not 2-regular; and (3) is in neither class. Comparatively, it appears to be the case that the number of s-regular, t-flat partitions grows faster when $s < t$ than when $s > t$. An extreme example is the 2-regular, t-flat partitions, which are partitions into odd parts not differing by too much, whereas the s-regular, 2-flat partitions can only be partitions into consecutive parts up to size $s - 1$. The asymptotics of these partitions is unexplored, however.

Letting $P_{R,F}^{(s,t;k)}(q)$ be the generating function for s-regular, t-flat partitions with largest part at most k, we have that

$$P_{R,F}^{(s,t;k)}(q) = P_{R,F}^{(s,t;k-1)}(q) + \frac{\chi(s \nmid k)q^k}{1-q^k}\left(P_{R,F}^{(s,t;k-1)} - P_{R,F}^{(s,t;k-t)}\right)$$

where $\chi(T)$ is the indicator function of the truth of statement T.

Not many of these generating functions are in the OEIS. The 2-regular (i.e., partitions into odd parts), 3-flat partitions are partitions into odd parts with consecutive (among odds) sizes, starting with a minimum size of 1; these constitute the mock theta function $\psi(q)$, OEIS sequence A053251. The 2-regular, 4-flat partitions are the same, except that a 1 need not appear (a 3 always will), and hence, $p_{R,F}^{(2,3)}(n) = p_{R,F}^{(2,4)}(n-1)$ for $n > 0$. As mentioned earlier, the s-regular, 2-flat partitions are just the partitions into consecutive parts from 1 to $s - 1$, such as OEIS sequence A014591.

4 Distinct and Flat

For some reason, we can say very little about Partitions simultaneously distinct and flat; except in the most restricted cases, we do not even have a generating function written down for these partitions. Such observations as can be made are collected below.

Recalling the definition of the Ferrers diagram of a partition, we see that partitions into parts s-distinct and t-flat can be described geometrically; they are the partitions in which the vertical segments of the outer boundary of the Ferrers diagram—the *profile* of the partition—are of length less than s, and horizontal segments are of length less than t.

It is easy to see from this form that the generating function of the s-distinct, t-flat partitions is symmetric in s and t: the s-distinct, t-flat partitions of n are in bijection with the t-distinct, s-flat partitions of n by conjugation. One notes that the class of t-distinct, t-flat partitions is preserved, but not the partitions themselves; since the number of t-distinct, t-flat partitions of n is not necessarily of the same parity as the number of partitions of n, it is too much to hope for an involution that has only the t-distinct, t-flat partitions as its fixed points, but one wonders if there is an involution which at least fixes all of these.

Despite the existence of this simple geometric description, it has been difficult to assert any general form of the generating function. The s-distinct, 2-flat partitions are simply those in which all parts from 1 to some k appear, but at most $s-1$ times. Their generating function is

$$P_{D,F}^{(s,2)} = \sum_{k=0}^{\infty} q^{\binom{k+1}{2}} \frac{(q^{s-1};q^{s-1})_k}{(q;q)_k}.$$

In particular, the 3-distinct, 2-flat partitions are counted by OEIS sequence A053261, the mock theta function $\psi_1(q)$.

More generally, one can write down various recurrences. For instance, if $P_{D,F}^{(s,t;k)}(q)$ is the generating function for s-distinct, t-flat partitions in which the largest part is at most k, then

$$P_{D,F}^{(s,t;k)}(q) = P_{D,F}^{(s,t;k-1)}(q) + \left(q^k \frac{1-q^{(s-1)k}}{1-q^k}\right)\left(P_{D,F}^{(s,t;k-1)}(q) - P_{D,F}^{(s,t;k-t)}(q)\right)$$

with appropriate initial conditions. The standard techniques for solving generating functions, however, do not seem to solve this recurrence very well.

By taking $q \to 1$ in the previous recurrences, we obtain a solvable difference equation, which can tell us something about the number of such partitions with largest part at most k. For instance, if $s=t=3$, the simplest case not covered by the generating function above, we are considering partitions in which parts differ by no more than 2 and repeat no more than twice. Letting $f(k) = P_{D,F}^{(3,3;k)}(1)$, we find that we have the difference equation

$$f(k) = 3f(k-1) - 2f(k-3),$$

with initial conditions $f(0)=1$, $f(1)=3$, $f(2)=9$, which yields OEIS sequence A077846, $(1,3,9,25,69,189,517,\dots)$. At the OEIS entry, we find the expression $f(n) = \sum_{i,j=0}^{n} 2^j \binom{j}{i-j}$; this is sometimes suggestive of a form for the generating function for a combinatorial expression when one replaces $\binom{N}{M}$ by $\begin{bmatrix} N \\ M \end{bmatrix}_{q^k}$ for some useful k, but nothing obvious seems to work along these lines for this problem.

The *hooklength* of a square in the Ferrers diagram, identified as position (i,j) when the lower right-hand corner of the square is at (x,y) coordinates $(-i,-j)$

where the upper-left corner is the origin, is the sum of the number of squares directly right of and below the square at (i, j), plus 1. The hooklengths in the partition $(4, 4, 3, 1, 1, 1)$ are illustrated below.

9	5	4	2
8	4	3	1
6	2	1	
3			
2			
1			

A partition is *t-core* if t is not among its hooklengths. The partition above is 7-core or t-core for $t > 9$. Since a partition in which parts differ by t or appear t or more times would automatically have t among the hooklengths in its outermost squares, the t-core partitions perforce form a subset of the t-distinct, t-flat partitions of n. In the case of $t = 2$, the sets are equal, as the partitions involved are just the triangular partitions $(n, n-1, \ldots, 2, 1)$. It might have been hoped that this observation would be useful in producing generating functions, but investigation along this line did not pan out.

5 Further Observations and Questions

Clearly since little can be said about s-distinct, t-flat partitions, less can possibly be said about partitions simultaneously r-regular, s-distinct, and t-flat. Those that are 2-regular, 2-distinct, and 3-flat are partitions consisting of consecutive odd numbers starting from 1, so their generating function is the Jacobi theta function $\sum_{n=0}^{\infty} q^{n^2}$. Those that are 2-regular, 3-distinct, and 3-flat permit an additional appearance of each odd part, and these are the 5th order mock theta function $\phi_0(q)$, OEIS entry A053258.

Several interesting open questions can be posed:

1. The fact that mock theta functions arise in numerous contexts related to these partitions might be spurious, but after all, a mock theta function has coefficients that do not grow "too fast," and the combination of flatness and another condition restricts partitions rather heavily; while it is perhaps a bit much to hope that the s-regular or s-distinct and t-flat partitions all qualify as mock theta functions, perhaps there is a closer connection here.
2. A full and careful proof of Tenner's conjecture on $\phi_s \phi_t$ for s and t not coprime should be interesting to produce.
3. What is the generating function for partitions with profile segments of length less than 2, that is, into parts appearing not more than twice, with parts differing by at most 2, including starting with 1 or 2?

4. It is easy to show based on Ramanujan's congruences that the number of 5-regular, 5-distinct partitions of $5n + 4$ is divisible by 5. Dyson's rank and the crank do not realize this congruence; is there another natural statistic on this subset which does so?

For item 3, the set of partitions involved is of natural interest, the property is invariant under the most natural involution on partitions, and it has at least a potential relation to the much-studied 3-core partitions, and yet the simple question of writing down the generating function for the set seems to elude any of the basic techniques for doing so. It would certainly be interesting to see this function written down, and more generally that for the s-distinct, t-flat partitions.

Item 4 is of interest regarding congruences for the partition function such as $p(5n + 4) \equiv 0 \pmod 5$. One observes that if $p(An + B) \equiv 0 \pmod C$ for all n, it must also hold that the $p_{A,A}(An + B)$, the number of A-regular, A-distinct partitions of $An + B$, possesses this congruence, i.e., $p_{A,A}(An + B) \equiv 0 \pmod C$. This follows since one may write a recurrence, perhaps a complicated one but still having integer coefficients, for $p_{A,A}(n)$ in terms of $p(n)$, $p(n - A)$, $p(n - 2A)$, etc., and if the latter are all divisible by C, then $p_{A,A}(n)$ will be as well. Since $p_{5,5}(5n + 4)$ shares the congruence but the currently constructed statistics fail to realize the congruence, perhaps another statistic exists that does so—and perhaps, due to the set being considered—is somewhat more natural and susceptible to simpler proof of its properties than the rank and crank. A really elementary combinatorial proof of Ramanujan's congruences does not yet exist in the literature.

There are certainly many other questions to be explored with these partitions; it is somewhat surprising that they have escaped serious notice for so long, and it is hoped that this paper will spur some interest in this area.

Acknowledgements The author warmly thanks Melvyn B. Nathanson and all other organizers and staffers of CANT 2016 for the opportunity to speak and the production of this proceedings volume, and for the lively and interesting discussions and problem sessions which surround the presentations at the conference.

References

1. D. Stockhofe, Bijektive Abbildungen auf der Menge der Partitionen einer Naturlichen Zahl. Bayreuth. Math. Schr. **10**, 1–59 (1982)
2. W. Keith, Ranks of partitions and Durfee symbols. Ph.D. Thesis, Pennsylvania State University, (June 2007). http://etda.libraries.psu.edu/theses/approved/WorldWideIndex/ETD-2026/index.html
3. S. Treneer, Congruences for the coefficients of weakly holomorphic modular forms. Proc. Lond. Math. Soc. **93**, 304–324 (2006)
4. Ø. Rødseth, Dissections of the generating functions of $q(n)$ and $q_0(n)$. Arbok University Bergen Mat. Nat. 1969, 12 (1970), 3–12. MR0434959 (55:7922)
5. E.X.W. Xia, O.X.M. Yao, Analogues of Ramanujan's partition identities. Ramanujan J. **31**, 373–396 (2013). https://doi.org/10.1007/s11139-012-9439-x
6. http://mathworld.wolfram.com/MonstrousMoonshine.html

7. R.E. Borcherds, Monstrous moonshine and monstrous Lie superalgebras. Invent. Math. **109**, 405–444 (1992)
8. The On-Line Encyclopedia of Integer Sequences, published electronically at https://oeis.org Dec 2016
9. McKay-Thompson series of class 54c for the Monster group. https://oeis.org/A112194
10. Stanley, R. Enumerative Combinatorics, vol. 1, Cambridge Studies in advanced Mathematics vol. 49, eds. by Fulton, Garling, Ribet, and Walters (Cambridge University Press, NY, 1997)

White's Theorem

An Exposition of White's Characterization of Empty Lattice Tetrahedra

Mizan R. Khan and Karen M. Rogers

Abstract We give an exposition of White's characterization of empty lattice tetrahedra. In particular, we describe the second author's proof of White's theorem that appeared in her doctoral thesis (Rogers in Doctoral dissertation 1993) [7].

Keywords Lattice tetrahedron · Empty lattice polyhedron

1 Introduction

The motivating example is the *lattice* tetrahedron with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and $(1, 1, c)$ with c being an arbitrary positive integer. We denote this tetrahedron as $T_{1,1,c}$. Regardless of the size of c (and consequently the volume of $T_{1,1,c}$), $T_{1,1,c}$ does not contain any lattice points other than its vertices. This is in surprising contrast to the situation in $\mathbb{R}^2$ where a *lattice* triangle does not contain any lattice points, other than its vertices, if and only if it has area 1/2. (To see this we invoke Pick's theorem.)

Reeve [4] posed the problem of characterizing such tetrahedra. Some years later, White [10] solved this problem. Over the years, different authors have given proofs of White's theorem (see [1, 3, 5, 6, 8]). The second author gave a proof of White's theorem in her doctoral dissertation [7]. In this article, we give a detailed exposition of this proof.

Before stating the relevant theorems, we establish some notation and definitions. Let $a, b, c \in \mathbb{Z}$ with $0 \le a, b < c$. We will use d to denote the integer

$$d = (1 - a - b) \mod c, 0 \le d < c.$$

M. R. Khan (✉)
Department of Mathematical Sciences, Eastern Connecticut State University,
Willimantic, CT 06226, USA
e-mail: khanm@easternct.edu

K. M. Rogers
Department of Mathematics and Computer Science, Oxford College at Emory University,
Oxford, GA 30054, USA
e-mail: karen.m.rogers@emory.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_11

Furthermore, $T_{a,b,c}$ will denote the lattice tetrahedron with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and (a, b, c).

Definition 1 Following Reznick [6], we call a lattice polyhedron that does not contain any lattice points other than its vertices an *empty lattice polyhedron*. Such a polyhedron belongs to a larger set of lattice polyhedra that do not contain any lattice points on their boundary other than the vertices. We call such polyhedra *clean lattice polyhedra*.

We insert a warning about the the terminology, particularly in the case of tetrahedra. Other names in the literature for empty tetrahedra are *fundamental*, *primitive*, *Reeve*.

Definition 2 An *affine unimodular map* is an affine map

$$L : \mathbb{R}^3 \to \mathbb{R}^3 \text{of the form} L(\mathbf{x}) = M\mathbf{x} + \mathbf{u},$$

where $M \in GL_3(\mathbb{Z})$, $\det(M) = \pm 1$ and $\mathbf{u} \in \mathbb{Z}^3$.

We now state the two theorems that we will prove.

Theorem 1 *Let T be an empty lattice tetrahedron. Then there is an affine unimodular map L such that $L(T) = T_{a,b,c}$, with* $0 \leq a, b < c$ *and* $\gcd(a, c) = \gcd(b, c) = \gcd(d, c) = 1$.

Theorem 2 (**White**) *The lattice tetrahedron $T_{a,b,c}$ is empty if and only if* $\gcd(a, c) = \gcd(b, c) = \gcd(d, c) = 1$ *and at least one of the following hold:*

$$a = 1, b = 1, c = 1, d = 1.$$

We now state definitions and background results that will be used to prove the two theorems.

Definition 3 A set of lattice points $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ in $\mathbb{Z}^n$ is said to be *primitive* if it is a basis for the sublattice $\mathbb{Z}^n \cap (\mathbb{R}\mathbf{v}_1 \oplus \cdots \oplus \mathbb{R}\mathbf{v}_k)$. Geometrically, this means that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is primitive if and only if the parallelepiped spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is empty.

The following is a list of standard results we will use. The proofs can be found in [9, Lectures V, VIII]. However, we have rephrased some of the statements. Consequently, the reader who consults [9] may need to read the relevant material carefully.

Theorem 3 *Every lattice has an integral basis.*

Theorem 4 *The property of being a lattice basis is preserved under the action of any unimodular transformation, that is, if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis for $\mathbb{Z}^n$ and $T : \mathbb{R}^n \to \mathbb{R}^n$ is an unimodular transformation, then $T(\mathbf{v}_1), T(\mathbf{v}_2), \ldots, T(\mathbf{v}_n)$ is also a basis of $\mathbb{Z}^n$. Furthermore, given two lattice bases, there is an unimodular transformation that maps one basis into the other.*

Theorem 5 *Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a linearly independent set of elements of $\mathbb{Z}^n$, and let $H = \mathbb{Z}\mathbf{v}_1 \oplus \ldots \oplus \mathbb{Z}\mathbf{v}_n$. Then the order of the quotient group $\mathbb{Z}^n/H$ equals*

$$\#(\mathbb{Z}^n/H) = |\det(\mathbf{v}_1, \ldots, \mathbf{v}_n)|.$$

Theorem 6 *Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ be a primitive set of $\mathbb{Z}^n$. Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ can be extended to a basis of $\mathbb{Z}^n$.*

We mention an interesting fact that emerges in the course of proving White's theorem. From Theorem 5, it follows that if $T_{a,b,c}$ is empty, then the parallelepiped spanned by $(1, 0, 0)$, $(0, 1, 0)$, (a, b, c) contains $(c - 1)$ lattice points in its interior. In the course of proving Theorem 2, we will find that *all of these points are coplanar*! More precisely, we have the following.

Corollary 1 *Let $P_{a,b,c}$ denote the parallelepiped spanned by* $(1, 0, 0)$, $(0, 1, 0)$*, and* (a, b, c)*. If $T_{a,b,c}$ is empty, then $P_{a,b,c}$ contains* $(c - 1)$ *lattice points in its interior. If $a = 1$, then all of these lattice points lie on the plane $x = 1$; if $b = 1$, then all of these lattice points lie on the plane $y = 1$; if $d = 1$, then all of these lattice points lie on the plane $x + y - z = 1$.*

Warning: The co-planarity of these lattice points was mentioned in an article of the first author [2, Theorem 3.2]. Unfortunately, the description of the planes in [2] is completely incorrect! The author should have done his homework and not just relied on his faulty visualization skills!!

2 Proofs

We begin with some notation. Let $\mathbf{u} = (u_1, u_2, u_3) \in \mathbb{Z}^3$. We will denote the integer $\gcd(u_1, u_2, u_3)$ by $\gcd(\mathbf{u})$. Occasionally, we will use e_1, e_2, and e_3 to denote the vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

Proposition 1 *Let $\mathbf{u}, \mathbf{v}$ be two linearly independent elements in $\mathbb{Z}^3$. The following statements are equivalent.*

1. *P, the parallelogram spanned by* $\mathbf{u}$ *and* $\mathbf{v}$ *is an empty parallelogram.*
2. *T, the triangle spanned by* $\mathbf{u}$ *and* $\mathbf{v}$ *is an empty triangle.*
3. $\gcd(\mathbf{u} \times \mathbf{v}) = 1$.

Proof Clearly (1) ⇒ (2). We prove the contrapositive to demonstrate that (2) ⇒ (1). We assume that P contains a lattice point $\mathbf{x}$ that is not a vertex of P. Then either $\mathbf{x}$ or $(\mathbf{u} + \mathbf{v} - \mathbf{x})$ lies in T. Since neither lattice point can be a vertex of T, we conclude that T is not an empty triangle.

We now turn to proving that (1) and (3) are equivalent.

(3) ⇒ (1): Since $\gcd(\mathbf{u} \times \mathbf{v}) = 1$, there exists, by the Extended Euclidean algorithm, $\mathbf{w} \in \mathbb{Z}^3$ such that $(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w} = 1 = \det(\mathbf{u}, \mathbf{v}, \mathbf{w})$. By Theorem 5, $\mathbf{u}, \mathbf{v}, \mathbf{w}$ is a

basis of $\mathbb{Z}^3$, and consequently they span an empty parallelepiped. We conclude that P is an empty parallelogram.

(1) $\Rightarrow$ (3): Since P is an empty parallelogram, $\{\mathbf{u}, \mathbf{v}\}$ is a primitive set of $\mathbb{Z}^3$, and consequently by Theorem 6 there is a lattice point $\mathbf{w}$ such that $\mathbf{u}, \mathbf{v}, \mathbf{w}$ is a basis of $\mathbb{Z}^3$. Consequently, $|\det(\mathbf{u}, \mathbf{v}, \mathbf{w})| = 1$. Since $\det(\mathbf{u}, \mathbf{v}, \mathbf{w}) = (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}$, we conclude that $\gcd(\mathbf{u} \times \mathbf{v}) = 1$.

Corollary 2 *The tetrahedron $T_{a,b,c}$ is clean if and only if* $\gcd(a, c) = \gcd(b, c) = \gcd(d, c) = 1$.

Proof Let $\triangle_1, \triangle_2, \triangle_3, \triangle_4$ denote the faces of $T_{a,b,c}$ where $\triangle_1$ is the triangle spanned by e_1 and e_2; $\triangle_2$ is the triangle spanned by e_1 and (a, b, c); $\triangle_3$ is the triangle spanned by e_2 and (a, b, c); and $\triangle_4$ is the triangle spanned by $(e_2 - e_1)$ and $((a, b, c) - e_1)$. $T_{a,b,c}$ is a clean tetrahedron if and only if $\triangle_1, \triangle_2, \triangle_3$, and $\triangle_4$ are all empty lattice triangles. Clearly $\triangle_1$ is an empty triangle. By Proposition 1, the triangles $\triangle_2, \triangle_3, \triangle_4$ are empty if and only if

$$\gcd(e_1 \times (a, b, c)) = \gcd(e_2 \times (a, b, c)) = \gcd((e_2 - e_1) \times ((a, b, c) - e_1)) = 1,$$

that is, $\gcd(b, c) = \gcd(a, c) = \gcd(d, c) = 1$.

Proof (Proof of Theorem 1) Let T be an empty lattice tetrahedron in $\mathbb{R}^3$. Without loss of generality we may assume that the origin is one of the vertices and the other 3 vertices are $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$. Since the triangle spanned by $\mathbf{u}$ and $\mathbf{v}$ is empty, by Proposition 1, the same holds for the parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$. Therefore, $\{\mathbf{u}, \mathbf{v}\}$ is a primitive set of $\mathbb{Z}^3$, and by Theorem 6 can be extended to a basis of $\mathbb{Z}^3$, $\mathbf{u}, \mathbf{v}, \mathbf{x}$. Now by Theorem 4, we have a unimodular transformation L_1 such that $L_1(\mathbf{u}) = e_1$, $L_2(\mathbf{v}) = e_2$, and $L_3(\mathbf{x}) = e_3$. Under this transformation, we see that the tetrahedron T is equivalent to the tetrahedron T_1 with vertices $0, e_1, e_2$ and (A, B, c) where $A, B, c \in \mathbb{Z}$ and $\mathrm{vol}(T) = |c/6|$. If $c < 0$, we can compose L_1 with the unimodular transformation

$$L_2((x, y, z)) = (x, y, -z).$$

Consequently, we can assume that $c > 0$. We now use the division algorithm to express

$$A = q_1 c + a \text{ and } B = q_2 c + b, 0 \leq a, b < c.$$

By acting on T_1 by the unimodular transformation

$$L_3((x, y, z)) = (x - q_1 z, y - q_2 z, z)$$

we get that T is equivalent to the tetrahedron T_2 with vertices $0, e_1, e_2$ and (a, b, c). Since T_2 is a clean tetrahedron, we invoke Corollary 2 to conclude that $\gcd(a, c) = \gcd(b, c) = \gcd(d, c) = 1$.

We now turn to the proof of White's theorem. Our proof is arranged in four parts. These are as follows:

Part 1: We prove that the tetrahedron $T_{a,b,c}$ is empty if and only if a system of equations involving a, b, d hold.

Part 2: This system of equations give an immediate proof of the ($\Leftarrow$) direction of White's theorem.

Part 3: The proof of the ($\Rightarrow$) direction of White's theorem is considerably more involved. We first develop a slight modification of the system of equations. This then leads us to define a finite set of arithmetic functions f_n. We then state and prove certain properties of these functions.

Part 4: We use the properties of f_n to complete the proof.

We will invoke the following identity in several places

Lemma 1 *Let $x \in \mathbb{R}$. If $x \notin \mathbb{Z}$, then*

$$\langle -x \rangle = 1 - \langle x \rangle. \tag{1}$$

We will typically invoke this identity in the following form:

$$\left\langle \frac{kl}{c} \right\rangle + \left\langle \frac{k(c-l)}{c} \right\rangle = 1 \tag{2}$$

for $0 < l < c, \gcd(l, c) = 1,$ *and* $k = 1, \ldots, c-1.$

Proposition 2 *Let $c \in \mathbb{Z}$ with $c > 1$ and let $T_{a,b,c}$ be a clean lattice tetrahedron. Then, $T_{a,b,c}$ is empty if and only if*

$$\left\langle \frac{ka}{c} \right\rangle + \left\langle \frac{kb}{c} \right\rangle + \left\langle \frac{kd}{c} \right\rangle - \frac{k}{c} = 1 \tag{3}$$

holds for $k = 1, \ldots, c-1.$

Proof (Proof of Part 1) Let P denote the parallelepiped spanned by e_1, e_2 and (a, b, c). Since volume$(P) = c$ and the faces of P are empty lattice parallelograms, we infer that P contains $(c-1)$ lattice points in its interior. These lattice points are

$$\left\langle \frac{k(c-a)}{c} \right\rangle (1, 0, 0) + \left\langle \frac{k(c-b)}{c} \right\rangle (0, 1, 0) + \frac{k}{c}(a, b, c) \tag{4}$$

with $k = 1, \ldots, c-1$.

$T_{a,b,c}$ is empty if and only if

$$1 < \left\langle \frac{k(c-a)}{c} \right\rangle + \left\langle \frac{k(c-b)}{c} \right\rangle + \frac{k}{c} < 2,$$

for $k = 1, \ldots, c-1$. Some algebraic manipulation in conjunction with identity (1) gives the system of inequalities

$$0 < \left\langle \frac{ka}{c} \right\rangle + \left\langle \frac{kb}{c} \right\rangle - \frac{k}{c} < 1,$$

for $k = 1, \ldots, c-1$. We now observe that

$$\left\langle \frac{ka}{c} \right\rangle + \left\langle \frac{kb}{c} \right\rangle - \frac{k}{c} \equiv \left\langle \frac{k(a+b-1)}{c} \right\rangle \pmod{\mathbb{Z}}, \tag{5}$$

for $k = 1, \ldots, c-1$. Since both sides of the congruence are between 0 and 1, we conclude that we have a system of *equalities*

$$\left\langle \frac{ka}{c} \right\rangle + \left\langle \frac{kb}{c} \right\rangle - \frac{k}{c} = \left\langle \frac{k(a+b-1)}{c} \right\rangle,$$

for $k = 1, \ldots, c-1$. After a little more algebraic manipulation, we conclude that $T_{a,b,c}$ is empty if and only if

$$\left\langle \frac{ka}{c} \right\rangle + \left\langle \frac{kb}{c} \right\rangle + \left\langle \frac{kd}{c} \right\rangle - \frac{k}{c} = 1$$

for $k = 1, \ldots, c-1$.

We can now easily prove (⇐) direction of White's theorem. The system of equations (3) in conjunction with the system of identities (2) allow us to conclude that the following tetrahedra are empty.

Corollary 3 *Let* $\gcd(a, c) = 1$. *Then the tetrahedra* $T_{1,a,c}$ *and* $T_{a,c-a,c}$ *are empty.*

To prove the (⇒) direction of White's theorem, we will work with a modification of (3). Define a set of arithmetic functions f_n for $n \in \mathbb{Z}^+$, $n < c$ and $\gcd(n, c) = 1$,

$$f_n : \{1, \ldots, c-2\} \to \{0, 1\}$$

via

$$f_n(k) = \left\langle \frac{kn}{c} \right\rangle - \left\langle \frac{(k+1)n}{c} \right\rangle + \frac{n}{c} = \left[\frac{(k+1)n}{c} \right] - \left[\frac{kn}{c} \right]. \tag{6}$$

From (3), we obtain the system of equations

$$f_a(k) + f_b(k) + f_d(k) + \frac{1}{c} = \frac{a+b+d}{c}, \tag{7}$$

for $k = 1, \ldots, c-2$. We now look at the case of $k = 1$ in (3) which shows that

$$\frac{a+b+d}{c} = 1 + \frac{1}{c}.$$

Thus, we can rewrite (7) as the system of equations

$$f_a(k) + f_b(k) + f_d(k) = 1, \tag{8}$$

for $k = 1, \ldots, c-2$. We will work with this system (8) in conjunction with the properties of f_n to arrive at a proof of White's theorem.

Proposition 3 *The function f_n has the following properties.*

(i) $f_1^{-1}(\{1\}) = \emptyset$.

(ii) *For $n > 1$,*

$$f_n^{-1}(\{1\}) = \{\, [kc/n] : k = 1, \ldots, n-1 \}\,.$$

(iii) $f_{c-n} = 1 - f_n$.

Proof For $k = 1, \ldots, c-2$,

$$f_1(k) = \left[\frac{k+1}{c}\right] - \left[\frac{k}{c}\right] = 0 - 0 = 0,$$

which proves (i).

We now prove statement (ii). If $l \in f_n^{-1}(\{1\})$ then there exists $k \in \mathbb{Z}^+$ such that

$$\frac{ln}{c} < k < \frac{(l+1)n}{c}.$$

It follows that $l = [kc/n]$. Conversely, if $l = [kc/n]$ for some integer k, with $1 \leq k \leq n-1$, then we have that

$$l < \frac{kc}{n} < l+1.$$

We now obtain that

$$\frac{ln}{c} < k < \frac{(l+1)n}{c}$$

and consequently $l \in f_n^{-1}(\{1\})$.

Statement (iii) is a consequence of identity (2).

$$\begin{aligned} f_{c-n}(k) &= \left\langle \frac{k(c-n)}{c} \right\rangle - \left\langle \frac{(k+1)(c-n)}{c} \right\rangle + \frac{c-n}{c} \\ &= 1 - \left\langle \frac{kn}{c} \right\rangle - 1 + \left\langle \frac{(k+1)n}{c} \right\rangle + 1 - \frac{n}{c} \\ &= 1 - f_n(k). \end{aligned}$$

We now complete the proof of White's theorem.

Proof (*Proof of Part 3*) Let $T_{a,b,c}$ be an empty tetrahedron with $c \geq 2$. We want to prove that either $a = 1$ or $b = 1$ or $d = 1$. We will argue by contradiction. So we assume that $a, b, d \geq 2$. Consequently none of the sets $f_a^{-1}(\{1\}), f_b^{-1}(\{1\}), f_d^{-1}(\{1\})$ are empty. Since

$$f_a + f_b + f_d = 1,$$

can infer that a, b and d are distinct integers, and the sets

$$f_a^{-1}(\{1\}), f_b^{-1}(\{1\}), f_d^{-1}(\{1\})$$

are pairwise disjoint. (**Spoiler alert**: Our argument hinges crucially on the fact that $f_b^{-1}(\{1\}) \cap f_d^{-1}(\{1\}) = \emptyset$.) Without loss of generality, we can assume that $a > b > d$. It follows that $1 \in f_a^{-1}(\{1\})$, and consequently $1 \notin \left(f_b^{-1}(\{1\}) \cup f_d^{-1}(\{1\})\right)$. We now have that

$$f_b + f_d = f_{c-a}$$

and consequently

$$\left(f_b^{-1}(\{1\}) \cup f_d^{-1}(\{1\})\right) = f_{c-a}^{-1}(\{1\}),$$

that is,

$$\{\,[kc/b] : k = 1, \ldots, b-1\} \cup \{\,[kc/d] : k = 1, \ldots, d-1\}$$

$$= \{\,[kc/(c-a)] : k = 1, \ldots, (c-a-1)\}\,.$$

We now compare the smallest and largest elements in each of the 3 sets. Since $b > d \geq 2$ and $1 \notin f_{c-a}^{-1}(\{1\})$, we have that

$$2 \leq \left[\frac{c}{c-a}\right] = \left[\frac{c}{b}\right] < \left[\frac{c}{d}\right] \leq \left[\frac{(d-1)c}{d}\right] < \left[\frac{(b-1)c}{b}\right] = \left[\frac{(c-a-1)c}{c-a}\right].$$

We remark that the strict inequalities occur since

$$f_b^{-1}(\{1\}) \cap f_d^{-1}(\{1\}) = \emptyset.$$

Let s be the positive integer such that

$$\left[\frac{c}{d}\right] = \left[\frac{sc}{c-a}\right].$$

We now obtain that

$$\left[\frac{(s-1)c}{c-a}\right] = \left[\frac{(s-1)c}{b}\right] \text{ and } \left[\frac{(s+1)c}{c-a}\right] \leq \left[\frac{sc}{b}\right].$$

Combining these two observations, we get

$$\left[\frac{(s+1)c}{c-a}\right]-\left[\frac{(s-1)c}{c-a}\right]\leq\left[\frac{sc}{b}\right]-\left[\frac{(s-1)c}{b}\right],$$

which implies the inequality

$$2\left[\frac{c}{c-a}\right]\leq\left[\frac{c}{b}\right]+1.$$

This leads to the contradiction that

$$\left[\frac{c}{c-a}\right]\leq 1,$$

and consequently our assumption that $a, b, d \geq 2$ is false.

Proof (*Proof of Corollary* 1) Let $T_{a,b,c}$ be empty, with $c > 1$. By Theorem 2, we have that either $a = 1$ or $b = 1$ or $b = c - a$. If $a = 1$, then by replacing a by 1 in (4), we see that the x co-ordinate of each lattice point inside $P_{1,b,c}$ equals 1. The same argument works if $b = 1$. The only case that needs a little more work is, if $b = c - a$. In this case, (4) becomes

$$\left\langle\frac{k(c-a)}{c}\right\rangle(1,0,0)+\left\langle\frac{ka}{c}\right\rangle(0,1,0)+\frac{k}{c}(a,c-a,c). \tag{9}$$

If we now add the x and y co-ordinates and subtract the z co-ordinate, we get

$$\left\langle\frac{k(c-a)}{c}\right\rangle+\left\langle\frac{ka}{c}\right\rangle+\frac{ka}{c}+\frac{k(c-a)}{c}-k=\left\langle\frac{k(c-a)}{c}\right\rangle+\left\langle\frac{ka}{c}\right\rangle.$$

We now invoke the identities (2) to conclude that the RHS equals 1.

Acknowledgements This article is a variant of an earlier manuscript drafted and submitted in the early 1990s. It was written in collaboration with Therese Hart who did the initial calculations that led us to discover White's result. We withdrew the article after discovering that White had anticipated our main discovery three decades previously. We would like to express our gratitude to Mel Nathanson who encouraged us to write this expository piece.

References

1. F. Breuer, F. von Heymann, Staircases in $\mathbb{Z}^2$. Integers **10**, 807–847 (2010)
2. M.R. Khan, A counting formula for primitive tetrahedra in $\mathbb{Z}^3$. Am. Math. Mon. **106**(6), 525–533 (1999)
3. D.R. Morrison, G. Stevens, Terminal Quotient Singularities in dimensions three and four. Proc. Amer. Math. Soc. **90**(1), 15–20 (1984)
4. J.E. Reeve, On the volume of lattice tetrahedra. Proc. London, Math. Soc. **7**(3) 378–395 (1957)
5. B. Reznick, Lattice point simplices. Discrete Math. **60**, 219–242 (1986)

6. B. Reznick, Clean lattice tetrahedra (preprint)
7. K.M. Rogers, Primitive simplices in $\mathbb{Z}^3$ and $\mathbb{Z}^4$, Doctoral dissertation, Columbia University, 1993
8. H.E. Scarf, Integral polyhedra in three space. Math. Oper. Res. **10**, 403–438 (1985)
9. C.L. Siegel, *Lectures on the Geometry of Numbers*, (Springer-Verlag, 1989)
10. G.K. White, Lattice Tetrahedra. Canad. J. Math **16**, 389–396 (1964)

A Misère-Play $\star$-Operator

Matthieu Dufour, Silvia Heubach and Urban Larsson

Abstract We study the $\star$-operator (Larsson et al. in Theoret. Comp. Sci. 412:8–10, 729–735, 2011) of impartial vector subtraction games (Golomb in J. Combin. Theory 1:443–458, 1965). Here, we extend the operator to the misère-play convention and prove convergence and other properties; notably, more structure is obtained under misère-play as compared with the normal-play convention (Larsson in Theoret. Comput. Sci. 422:52–58, 2012).

Keywords Combinatorial game · Game convergence · Game creation operator Impartial game · Misére play · Star operator · Sum-free set

1 Introduction

The notion of *vector subtraction games* was introduced by Golomb [4], motivated by methods in computer science. Then, much later the game family reappeared [3] under a different name (invariant subtraction games) and now the motivation was a conjecture in number theory.

The proposed problem was solved [5] by introducing the normal-play $\star$*-operator* on the class of games, and subsequently, some very general properties of this $\star$-operator were discovered [6]. All this work was done using the so-called normal-play convention for impartial combinatorial games [1]. Here, we introduce the $\star$-operator under the *misère-play* convention and prove some general properties. Let

M. Dufour
Department of Mathematics, Université du Québec à Montréal, Montréal,
Québec H3C 3P8, Canada
e-mail: dufour.matthieu@uqam.ca

S. Heubach
Department of Mathematics, California State University Los Angeles,
Los Angeles, CA 90032, USA
e-mail: sheubac@calstatela.edu

U. Larsson (✉)
The Faculty of Industrial Engineering and Management, Technion-Israel
Institute of Technology, 3200003 Haifa, Israel
e-mail: urban031@gmail.com

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_12

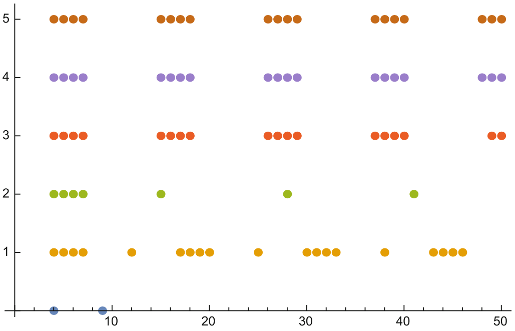


Fig. 1 The move sets of a sequence of games arising from the initial game with move set $\mathscr{G}^0 = \{4, 9\}$. The values at level i represent the P-positions of the game whose moves are listed on level $i - 1$

us begin by using an example of a game in one dimension (those are usually just called subtraction games).

Imagine two players who alternate in removing tokens from a single heap, subject to the rules that either 4 or 9 tokens be removed, and that if you cannot move, you win (*misère-play*). In this particular game, the first player to move wins if there are less than 4 tokens in the pile, because these positions are terminal, and if there are between 4 and 7 tokens, then the other player wins. By a recursive procedure, one computes the pattern of P-positions (these are the positions from which the current player cannot win given optimal play). The initial pattern of P-positions is shown in the first line of Fig. 1, and the sequence is periodic as illustrated (on the 0th line, we show the allowed moves of this game).

Since the underlying structure of the moves and the P-positions is the same (the nonnegative integers), one can play a new game where the P-positions of the first game are used as moves in the new game. The new set of moves is then

$$\mathscr{G}^1 = \{4, 5, 6, 7, 12, 17, 18, 19, 20, 25, \ldots\}.$$

The P-positions of this new game are shown on the second row of Fig. 1. By iterating this process, we get a sequence of games where the moves in the next game consist of the P-positions of the previous game (this is the $\star$-operator to be defined formally below).

In Fig. 1, the games shown on rows 4 and 5 have the same moves, and one of the results in this paper is that the sequence of games converges to a limit game, for any choice of the initial set of moves, and in any dimension. Moreover, the limit game is *reflexive* (its set of P-positions is identical with the move set), and we show that it can be defined (non-recursively) by a simple ‘sumset’ rule. We also show that the limit game is the same for any two games (in the same dimension) if the set of smallest (in the natural partial order) moves is the same. This is the third main result of this paper, which concludes Sect. 2.

In one dimension, we obtain precise structure results for the class of reflexive games (Sect. 3), and we use them to discuss some applications, for example, on so-called *maximal sum-free sets* in relation with reflexive games (Sect. 4). Moreover, we find the set of all games which have a reflexive set S as its set of P-positions (the notion of an *S-complete set* of games is introduced in Definition 6).

We have obtained some preliminary structure results in two dimensions, which are illustrated in Sect. 5, where we also discuss directions for future work.

The remaining part of the introduction concerns some basic concepts and definitions, and then in Sect. 1.2, we use an example of a game on one heap to illustrate our method of proof for convergence.

1.1 Basic Concepts

Let $\mathbb{N}$ denote the positive integers, and $\mathbb{N}_0$ the nonnegative integers. Unless otherwise stated, $\mathscr{M}$ will be a misére-play game on $d \in \mathbb{N}$ heaps (dimension d), and we use calligraphy notation for sets when we want to indicate that we think of a subset of vectors as a game. All games we consider are impartial and of the following form, e.g., [3, 4].

Definition 1 Let $d \in \mathbb{N}$, and let $\mathscr{M} \subseteq \mathbb{N}_0^d$ be the set of moves. In the *d-dimensional vector subtraction game* $\mathscr{M}$, a player can move from position $x \in \mathbb{N}_0^d$ to position $y \in \mathbb{N}_0^d$ if $x - y \in \mathscr{M}$. A position y for which $x - y \in \mathscr{M}$ is called an *option of* x. We consider the misère-play version of the game, that is, a player who cannot move wins.

Note that when we talk about a game, we refer to its rule set or *subtraction set*, and we are interested to determine for each position whether it is a P- or N-positon. We are not concerned about finding efficient strategies from starting positions, but rather want to investigate the patterns of the set of P-positions.

Since our games are multidimensional, we use the natural partial order on $\mathbb{N}_0^d$, namely $x \preceq y$ if and only if $x_i \leqslant y_i$ for $i = 1, \ldots, d$, and $x \prec y$ if and only if $x \preceq y$ with strict inequality holding for at least one component.

Definition 2 A nonempty subset I of a partially ordered set $(X, \preceq)$ is a *lower ideal* if for every $x \in I$, $y \preceq x$ implies that $y \in I$.

We denote the set of terminal positions of the game $\mathscr{M}$ by $T_{\mathscr{M}}$. By definition of a vector subtraction game, $T_{\mathscr{M}}$ is the set of all x smaller than or unrelated to every $m \in \mathscr{M}$, that is, $T_{\mathscr{M}} = \{x \not\succeq m \mid m \in \mathscr{M}\}$. Of course, if $0 \in \mathscr{M}$ then $T_{\mathscr{M}} = \varnothing$. Moreover, since we play the misère version, we have the following observation.

Note 1 For any game $\mathscr{M}$, in any dimension, the set of terminal positions is a lower ideal, and, $T_{\mathscr{M}} \subseteq N(\mathscr{M})$.

It is well known that for impartial games without cycles (that is, no repeated game positions), there are exactly two *outcome classes*, called N and P [1]. In misère-play, they are characterized as follows: a position is an N-position if it has no option, or if there is at least one P-position in its set of options. Otherwise, a position is a P-position. In other words, a position is a P-position if and only if its set of options is a nonempty set of N-positions. We denote the set of N-positions of a misère-play game $\mathscr{M}$ by $N(\mathscr{M})$ and the set of P-positions by $P(\mathscr{M})$.

Note that in Definition 1, we allow $\mathscr{M} = \varnothing$ and also the case $0 \in \mathscr{M}$ (that is, a pass move is allowed). If $0 \in \mathscr{M}$, then each position can be repeated so the outcome is a draw, and hence $P(\mathscr{M}) = \varnothing$. This trivial draw game was originally included in the definition of normal-play vector subtraction games by Golomb [4].[1] It is not very interesting from a game player's perspective, but from a theoretical point of view, as we will see, there is no reason to exclude it. Similarly, if $\mathscr{M} = \varnothing$, then $P(\mathscr{M}) = \varnothing$, because all positions are N-positions due to the misère convention.

On the other hand, if $0 \notin \mathscr{M}$, then we get a recursive definition of the outcomes of all positions from the characterization of N- and P-positions above, and by Note 1, recurrence starts with N-positions. Moreover, observe that any smallest move $0 \neq m \in \mathscr{M}$ is a P-position, so in this case $P(\mathscr{M}) \neq \varnothing$. In fact, each game $\mathscr{M}$ has a unique set of minimal elements which we denote by $\min(\mathscr{M})$[2], and we have the following fundamental observation.

Note 2 For any game $\mathscr{M}$, in any dimension, if $0 \neq \min(\mathscr{M})$, then $\min(\mathscr{M}) \subseteq P(\mathscr{M})$.

Since the underlying structure of moves and P-positions is the same (sets of integer vectors), we can iteratively create new games [5, 6].

Definition 3 Let $\mathscr{M}$ be a game in any dimension. Then, $\mathscr{M}^\star$ is the game with subtraction set $\mathscr{M}^\star = P(\mathscr{M})$.

This defines the *misère-play $\star$-operator*[3] which acts on impartial subtraction games.[4] A P-position in game $\mathscr{M}$ becomes a move in game $\mathscr{M}^\star$ (and an N-position in $\mathscr{M}$ becomes a non-move in game $\mathscr{M}^\star$). We can now study properties of sequence of games created by repeated applications of the $\star$-operator. First, we define special sequences of games, obtained by the fixed points of the operator.

Definition 4 The game $\mathscr{M} \subseteq \mathbb{N}_0^d$ is *reflexive* if $\mathscr{M} = \mathscr{M}^\star$.

Definition 5 Let $\mathscr{M}^0 = \mathscr{M}$ be a game in any dimension, and let $\mathscr{M}^i = (\mathscr{M}^{i-1})^\star$ for $i > 0$. The sequence of games $\mathscr{M}^i$ *converges* (with respect to $\star$) if $\mathscr{M}^\infty = \lim_{i\to\infty} \mathscr{M}^i$ exists.

Note that due to the recursive definition of the outcomes of an impartial combinatorial game, the notion of convergence is point-wise. The following lemma is immediate from the definition of reflexivity.

[1] He also restricted the set of terminal positions to contain only 0, a definition not used in connection with the $\star$-operator.

[2] In one dimension, $\min(\mathscr{M})$ consists of a single value and we sometimes abuse notation and write the minimal number instead of the set. If $\mathscr{M} = \varnothing$ then we define $\min \mathscr{M} = \varnothing$.

[3] Note that the $\star$-operator under misère rules is the same as the $\star$-operator in normal-play [5, 6]. However, since in misère-play 0 is never a P-position, the definition simplifies in this case.

[4] The $\star$-operator is in fact an infinite class of operators, one operator for each dimension. However, we will refer to 'the' $\star$-operator because the operator acts in the same way in each dimension.

Lemma 1 *The game $\mathscr{M} \subseteq \mathbb{N}_0^d$ is reflexive if and only if there is a game $\mathscr{X}$ such that $\mathscr{M} = \mathscr{X}^\infty$.*

Proof If $\mathscr{M}$ is reflexive, then we may take $\mathscr{X} = \mathscr{M}$, because $\mathscr{M} = \mathscr{M}^\star = \cdots = \mathscr{M}^\infty$. If $\mathscr{M} = \mathscr{X}^\infty$, for some game $\mathscr{X}$, then by definition of a limit game, $\mathscr{M}$ is reflexive. □

Note 3 We have that $P(\mathscr{M}) = \varnothing$ if and only if $0 \in \mathscr{M}$ or $\mathscr{M} = \varnothing$. Consequently, if $0 \in \mathscr{M}$ or $\mathscr{M} = \varnothing$, then $\mathscr{M}^\infty = \varnothing$.

Vector subtraction games that have the same sets of P-positions have been studied before (see e.g., [7]). We will be particularly interested in games for which the set of P-positions is a reflexive game, which motivates the following definition.

Definition 6 Given misère or normal-play convention, we call a set of games $G = \{\mathscr{G}_i\}$ *S-solvable* if, for all i, $P(\mathscr{G}_i) = S$. If G contains all such games (that is, G is S-solvable and $\mathscr{H} \notin G$ implies $P(\mathscr{H}) \neq S$), then we say that the set of games G is *S-complete*.

1.2 One Heap Examples

We begin by illustrating our results on reflexive games and their limit behavior via the following examples of play on one heap.

Figure 2 shows the result of applying the ⋆-operator five times to two different games. On the left, the move set is $\mathscr{H}^0 = \{4, 7, 11\}$, while on the right, it is $\mathscr{G}^0 = \{4, 9\}$ (same as in Fig. 1). Note that both sets have the same minimal move, $k = 4$. Figure 2 suggests that both games converge to the same limit game, which exhibits a periodic structure: it consists of groups of k consecutive integers, and the smallest values in consecutive groups differ by $13 = 3 \cdot 4 - 1 = 3 \cdot k - 1$. We will show that all games, under the misère-play ⋆-operator, have a limit game, and that the limit game is uniquely determined by the set of smallest elements.

In proving the convergence result, the approach is to show that the outcome class (move or non-move) of the smallest position with differing outcome class in consecutive games will become 'fixed' in subsequent iterations. Therefore, the set of positions whose outcome class remains unchanged from iteration to iteration increases in each step, and any values already in the set of 'fixed' positions cannot become 'unfixed.' Figure 3 shows the first five iterations of the game $\mathscr{G}^0 = \{4, 9\}$. The rectangles identify the smallest elements that differ when comparing $\mathscr{G}^i$ and $\mathscr{G}^{i+1}$. For example, for games $\mathscr{G}^0$ and $\mathscr{G}^1$, the smallest differing element is $x = 5$. For $\mathscr{G}^0$, $x = 5$ is not a move, but for $\mathscr{G}^1$ (and all subsequent games) it is. Similarly, the smallest differing element when comparing $\mathscr{G}^1$ and $\mathscr{G}^2$ is $x = 12$, which is a move in $\mathscr{G}^1$, but then becomes fixed as a non-move in $\mathscr{G}^2$ and subsequent games. For the game $\mathscr{G}^0 = \{4, 9\}$, the initial set of *outcome-fixed* positions is $\{1, 2, 3, 4\}$ (the terminal positions and the smallest move), $\{1, 2, \ldots, 11\}$ after the first iteration, then $\{1, 2, \ldots, 15\}$, and finally $\{1, 2, \ldots, 47\}$.

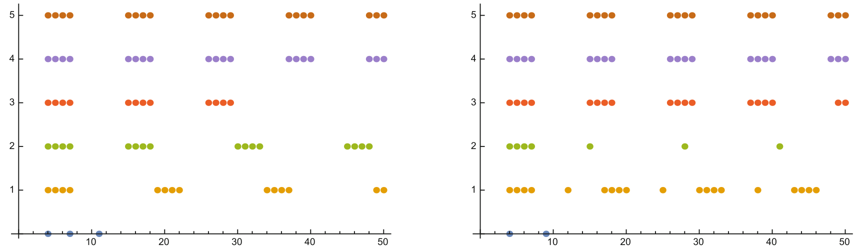
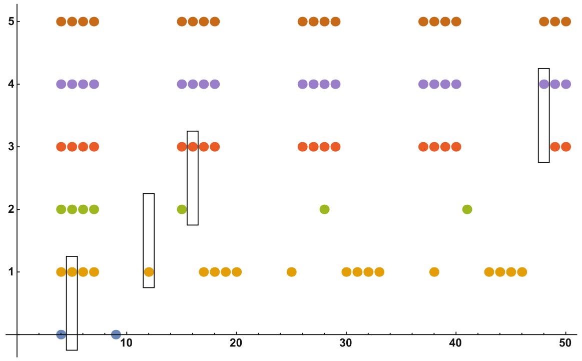


Fig. 2 The behavior of the $\star$-operator for two different games, $\mathscr{H}^0 = \{4, 7, 11\}$ and $\mathscr{G}^0 = \{4, 9\}$, that have the same minimal move. The values at level i represent the move sets $\mathscr{H}^i$ and $\mathscr{G}^i$, respectively

Fig. 3 Rectangles identifying the smallest elements with differing outcome class in the ith and the $(i + 1)$st iteration of the game $G^0 = \{4, 9\}$



2 Convergence and Reflexivity

As we have seen, the definition of the $\star$-operator does not depend on the given dimension, and as we will see, neither does its most notable property, convergence to a fixed point, the class of reflexive games being the fixed points of the operator. The following lemma makes this property conceivable.

Lemma 2 *If* $0 \notin \mathscr{M}$*, then* $\min(\mathscr{M}) = \min(P(\mathscr{M}))$*, and consequently,* $\min(\mathscr{M}) = \min(\mathscr{M}^i)$ *for all* $i \geq 0$*, and* $\min(\mathscr{M}) = \min(\mathscr{M}^\infty)$ *if the limit exists.*

Proof If $\mathscr{M} = \varnothing$, then $P(\mathscr{M}) = \varnothing = \mathscr{M}$, so the conclusion holds. If $\mathscr{M} \neq \varnothing$, let $m \in \min(\mathscr{M})$. Then by Note 2, $m \in \mathscr{P}(\mathscr{M})$. Also, for $x \prec m$, $x \in T_{\mathscr{M}} \subseteq N(\mathscr{M})$, and therefore, $m \in \min(P(\mathscr{M}))$. Thus, $\min(\mathscr{M}) \subseteq \min(P(\mathscr{M}))$. On the other hand, for $m' \in \min(P(\mathscr{M}))$, assume $m' \notin \min(\mathscr{M})$.

There are two possibilities. First, if for $m \in \min(\mathscr{M})$, $m' \succ m$, then Note 2 contradicts that $m' \in \min(P(\mathscr{M}))$ (because $m \in \mathscr{P}(\mathscr{M})$). Second, if $m' \not\succ m \in \min(\mathscr{M})$, then $m' \in T_{\mathscr{M}}$, which contradicts $m' \in \mathscr{P}(\mathscr{M})$. Therefore, $m' \in \min(\mathscr{M})$, which implies that $\min(P(\mathscr{M})) \subseteq \min(\mathscr{M})$, so $\min(P(\mathscr{M})) = \min(\mathscr{M})$. By definition of the $\star$-operator, we have that $\min(\mathscr{M}) = \min(\mathscr{M}^i)$ for all $i \geq 0$, and the last statement follows from the definition of the limit game. □

For the ⋆-operator, its definition as well as most of its important properties are independent of the dimension, and it is the main purpose of this section to study these general properties. To emphasize the type of behavior, we introduce the class of *accumulation-point operators*.

Definition 7 Let Ω be a (totally ordered) set, and let $f : \Omega^d \to \Omega^d$ be an operator defined in any dimension $d \in \mathbb{N}$. Then f is an *accumulation-point operator* (associated with Ω) if, for any dimension $d \in \mathbb{N}$ and any $X \subseteq \Omega^d$, $\lim_{n\to\infty} f^n X$ exists, where $f^n X = f f^{n-1} X$, for $n > 0$ and $f^0 = f$.

In the context of our vector subtraction games, recall that $\Omega = \mathbb{N}_0$ includes the case of pass moves.

Theorem 1 *The misère-play ⋆-operator is an accumulation-point operator associated with* $\mathbb{N}_0$. *That is, for any* $d \in \mathbb{N}$, *each game* $\mathscr{M} \subseteq \mathbb{N}_0^d$ *converges to a (reflexive) limit game* $\mathscr{M}^\infty$.

Proof If either $\mathscr{M} = \varnothing$ or $0 \in \mathscr{M}$, then by Note 3, $\mathscr{M}^\infty$ exists.

Now let $\varnothing \neq \mathscr{M} \subseteq \mathbb{N}_0^d \setminus \{0\}$. Assume that for some $i \geqslant 0$,

$$(\mathscr{M}^i \setminus P(\mathscr{M}^i)) \cup (P(\mathscr{M}^i) \setminus \mathscr{M}^i) \neq \varnothing,$$

since otherwise $\mathscr{M}^0 = \mathscr{M}^1 = \mathscr{M}^\infty$ (by the definition of the star-operator). Let

$$X(i) = \min((\mathscr{M}^i \setminus P(\mathscr{M}^i)) \cup (P(\mathscr{M}^i) \setminus \mathscr{M}^i))$$

be the set of minimal differing elements among moves and P-positions at the i^{th} iteration. Note that by definition of $X(i)$, if $z \not\succeq x$ for all $x \in X(i)$, then $z \in (\mathscr{M}^i \cap P(\mathscr{M}^i)) \cup ((\mathscr{M}^i)^c \cap (P(\mathscr{M}^i))^c)$, so either $z \in \mathscr{M}^j$ for all $j > i$ or $z \notin \mathscr{M}^j$ for all $j > i$. Now for each $x \in X(i)$, we consider the following two cases:

Case 1: Suppose $x \in \mathscr{M}^i \setminus P(\mathscr{M}^i)$. It suffices to show that $x \notin P(\mathscr{M}^{i+1})$, since then $x \notin \mathscr{M}^j$ for all $j > i$. Note that x is not a terminal position because x is a move. Also, because x is not a P-position, there is a move $m \in \mathscr{M}^i$ such that

$$x - m = z \in P(\mathscr{M}^i). \tag{1}$$

However, since $0 \prec m, z \prec x$, then, by definition of $X(i)$, $m \in \mathscr{M}^{i+1}$ and $z \in P(\mathscr{M}^{i+1})$, which, by equation (1), implies that $x \notin P(\mathscr{M}^{i+1})$, as desired.[5]

Case 2: Suppose that $x \in P(\mathscr{M}^i) \setminus \mathscr{M}^i$. It suffices to show that $x \in P(\mathscr{M}^{i+1})$, since then $x \in \mathscr{M}^j$ for all $j > i$. Let's assume to the contrary that $x \notin P(\mathscr{M}^{i+1})$. Then there exists a move $m \in \mathscr{M}^{i+1}$ such that $x - m = z \in P(\mathscr{M}^{i+1})$. But then $m \in P(\mathscr{M}^i) = \mathscr{M}^{i+1}$, by definition of the ⋆-operator, and $z \in \mathscr{M}^i \cap P(\mathscr{M}^i)$, by definition of $X(i)$. Therefore, in the game $\mathscr{M}^i$ we have a move z from a P-position x to the P-position m, a contradiction, so $x \in P(\mathscr{M}^{i+1})$.[6] □

[5] An example of this case is $x = 12 \in X(1)$ in Fig. 3.

[6] Examples of this case are $x = 5 \in X(0)$, $x = 16 \in X(2)$, and $x = 48 \in X(3)$ in Fig. 3.

We now characterize reflexive games via a ‘sumset’ property.

Definition 8 Suppose that $A, B \subseteq \mathbb{N}_0^d$. Then $A + B = \{a + b \mid a \in A, b \in B\}$.

The following result on so-called sum-free sets is also discussed in Sect. 4.

Theorem 2 *Let $A \subseteq \mathbb{N}_0^d$. Then the game $\mathscr{A}$ with move set A is reflexive if and only if*

$$A + A = A^c \setminus T_{\mathscr{A}}, \tag{2}$$

where A^c denotes the complement of A with respect to $\mathbb{N}_0^d$.

Proof If $A = \varnothing$, then all positions are terminal N-positions, so $T_{\mathscr{A}} = \mathbb{N}_0^d$ and $P(\mathscr{A}) = \varnothing = A$. Thus $\mathscr{A}$ is reflexive and (2) holds.

Next we assume that A is nonempty. If $0 \in A$, then because $0 \notin P(\mathscr{A})$, $\mathscr{A}$ is not reflexive. On the other hand, $0 \in A + A$, but $0 \notin \mathscr{A}^c$, so (2) does not hold and the claim is true in this case also.

Now we assume that A is nonempty and $0 \notin A$. Note that for any such game $\mathscr{A}$, we have that for any nonterminal position $x \in N(\mathscr{A}) \setminus T_{\mathscr{A}}$ there is a move $m = x - z \in A$ that leads to a P-position $z \in P(\mathscr{A})$. Therefore, $N(\mathscr{A}) \setminus T_{\mathscr{A}} \subseteq A + P(\mathscr{A})$.

On the other hand, since a move from a P-position cannot result in a P-position, for any $z \in P(\mathscr{A})$ and any move $m \in A$, $m + z = x \in N(\mathscr{A}) \setminus T_{\mathscr{A}}$. Thus, $A + P(\mathscr{A}) \subseteq N(\mathscr{A}) \setminus T_{\mathscr{A}}$, and we have

$$A + P(\mathscr{A}) = N(\mathscr{A}) \setminus T_{\mathscr{A}}. \tag{3}$$

We now prove that $\mathscr{A}$ is reflexive if and only if (2) holds.
‘$\Rightarrow$’ If A is a reflexive, then $P(\mathscr{A}) = A$ and $N(\mathscr{A}) = A^c$, so (3) reduces to (2).
‘$\Leftarrow$’ Let $B = P(\mathscr{A})$, so we need to prove that $B = A$. Assume to the contrary that there is $x \in A^c \cap B$. Then $B \subseteq A^c$, because otherwise, there would exist $z \in A \cap B$ and, by (2), a move $m = x - z \in A$ from P-position $x \in A^c \cap B \subset A^c \setminus T_{\mathscr{A}}$ to P-position $z \in A \cap B$. So, $B \subseteq A^c$, or equivalently, $A \subseteq B^c = N(\mathscr{A})$. However, in misère-play, a smallest move (which exists by assumption) is always a P-position, which contradicts that $A \subseteq N(\mathscr{A})$, and so $A^c \cap B = \varnothing$. Therefore, $B \subseteq A$.

It remains to prove that $A \subseteq B$, or equivalently, $B^c \subseteq A^c$. Let $x \in B^c$. Note that $T_{\mathscr{A}} \subseteq A^c \cap B^c$ because terminal positions are neither moves nor P-positions. Thus, we assume without loss of generality that $x \in B^c \setminus T_{\mathscr{A}}$, that is, x is a nonterminal N-position. By (3), there is a move $m = x - z \in A$ from x to $z \in B \subseteq A$. Since both z and m are in A, then by assumption (2) we have that $m + z = x \in A^c \setminus T_{\mathscr{A}}$. Since $x \notin T_{\mathscr{A}}$, we must have that $x \in A^c$, which completes the proof. □

Using the sumset property of Theorem 2, we completely characterize the limit games. There is exactly one reflexive limit game for each set of minimal moves, that is, the set of minimal elements uniquely determines the limit game.

Theorem 3 *Let $\mathscr{M}$ and $\mathscr{G}$ be nonempty games. Then $\mathscr{M}^\infty = \mathscr{G}^\infty \iff \min(\mathscr{M}) = \min(\mathscr{G})$.*

Proof '⇒' If $\mathscr{M}^\infty = \mathscr{G}^\infty = \varnothing$, then by Note 3, $\{0\} = \min(\mathscr{M}) = \min(\mathscr{G})$ since both $\mathscr{M}$ and $\mathscr{G}$ are nonempty. If $\mathscr{M}^\infty$ and $\mathscr{G}^\infty$ are nonempty games, then by Note 2, $0 \notin \mathscr{M} \cap \mathscr{G}$, and by Lemma 2, we have $\min(\mathscr{M}) = \min(\mathscr{M}^\infty) = \min(\mathscr{G}^\infty) = \min(\mathscr{G})$ as claimed.

'⇐' If $\{0\} = \min(\mathscr{M}) = \min(\mathscr{G})$, then $\mathscr{M}^\infty = \mathscr{G}^\infty = \varnothing$ by Note 3. If $\{0\} \neq \min(\mathscr{M}) = \min(\mathscr{G})$, then by Lemma 2, $\min(\mathscr{M}^\infty) = \min(\mathscr{G}^\infty)$ and $T_{\mathscr{M}^\infty} = T_{\mathscr{G}^\infty}$. We need to show that $\mathscr{M}^\infty = \mathscr{G}^\infty$. Assume to the contrary that there is a smallest differing element

$$x = \min(\mathscr{G}^\infty \setminus \mathscr{M}^\infty \cup \mathscr{M}^\infty \setminus \mathscr{G}^\infty).$$

Without loss of generality we may assume that $x \in \mathscr{G}^\infty \setminus \mathscr{M}^\infty$. Be definition of x, $x \notin \mathscr{M}^\infty$. Also, $x \succeq m \in \min(\mathscr{G}^\infty) = \min(\mathscr{M}^\infty)$, so $x \notin T_{\mathscr{M}^\infty}$, that is, $x \in (\mathscr{M}^\infty)^c \setminus T_{\mathscr{M}^\infty}$. Since $\mathscr{M}^\infty$ is reflexive, by Theorem 2, there must be $0 \neq y, z \in \mathscr{M}^\infty$ such that $y + z = x$. However, since $y, z \prec x$, by minimality of x, we have $y, z \in \mathscr{G}^\infty$. Applying Theorem 2 to $\mathscr{G}^\infty$ now implies that $x \in (\mathscr{G}^\infty)^c \setminus T_{\mathscr{G}^\infty}$, a contradiction. Thus $\mathscr{M}^\infty = \mathscr{G}^\infty$. □

Theorems 1 and 3 confirm what was suggested in Fig. 2; the games converge to the same limit game. Now the question becomes: what do limit games 'look like'? We will completely answer this question in the next section for games on one heap, and then in the final section, we sketch some of the observed behavior for two heaps (see also [2]).

Both the misère-play ⋆-operator and the normal-play ⋆⋆-operator converge in any dimension, but the properties of the fixed points are not the same. Our results imply that the misère-play convergence is *stable* in the following sense.

Corollary 1 *Let $\mathscr{M}$ be a reflexive game in any dimension, and let Y be a finite set of vectors in the same dimension. For almost all* perturbations *of the form $\mathscr{M}_Y = (\mathscr{M} \setminus Y) \cup (Y \setminus \mathscr{M})$, $\mathscr{M}^\infty = \mathscr{M}_Y{}^\infty$.*

Proof This is a consequence of Theorems 1 and 3. □

3 A Characterization of Limit Games in One Dimension

We first consider $d = 1$, that is, play on a single heap. Motivated by the structure of the limiting game in Fig. 1, for any $k \in \mathbb{N}$, we define the period $p_k = 3k - 1$ and let $\mathscr{M}_k$ denote the set

$$\mathscr{M}_k = \{ip_k + k, \dots, ip_k + 2k - 1 \mid i \in \mathbb{N}_0\},$$

with $\mathscr{M}_0 = \varnothing$. Note that $k = \min(\mathscr{M}_k)$ for $k \geqslant 1$. By Theorem 3, the games in Example 1.2 have the same limit game, and we will see in Theorem 4 and Corollary 2, that $\mathscr{H}^\infty = \mathscr{G}^\infty = \mathscr{M}_4$.

Since the set $\mathscr{M}_k$ is periodic with period p_k, we find it convenient to make our using arithmetic modulo p_k. We denote the set of residuals modulo p of elements of a set A by $[A]_p$. With this notation, it follows from the definition of $\mathscr{M}_k$ that for $k \geqslant 1$,

$$\begin{aligned} &[\mathscr{M}_k]_{p_k} = \{k, k+1, \ldots, 2k-1\} \text{ and} \\ &[\mathscr{M}_k^c]_{p_k} = \{0, 1, \ldots, k-1, 2k, \ldots, 3k-2\} \equiv_{p_k} \{-(k-1), \ldots, k-1\}. \end{aligned} \tag{4}$$

Theorem 4 *The game $\mathscr{M} \subseteq \mathbb{N}_0$ is reflexive if and only if $\mathscr{M} = \mathscr{M}_k$, for some $k \in \mathbb{N}_0$.*

Proof '$\Leftarrow$' If $k = 0$, then $\mathscr{M} = \mathscr{M}_0 = \varnothing$, which is reflexive by Note 3. Suppose next that $\mathscr{M} = \mathscr{M}_k$ is nonempty and let $k = \min(\mathscr{M}_k) \geq 1$. We show that the game $\mathscr{M}_k$ is reflexive using Theorem 2. Note that by (4),

$$\begin{aligned} [\mathscr{M}_k + \mathscr{M}_k]_{p_k} &= [\{2k, \ldots, 4k-2\}]_{p_k} \\ &= \{2k, \ldots, 3k-2, 0, \ldots, k-1\} = [\mathscr{M}_k^c]_{p_k}. \end{aligned}$$

Since for any element $m \in \mathscr{M}_k$, $m + m \geq 2k$ and the terminal positions are given by $T_{\mathscr{M}_k} = \{0, \ldots, k-1\}$, we have that $\mathscr{M}_k + \mathscr{M}_k \subseteq \mathscr{M}_k^c \setminus T_{\mathscr{M}_k}$. On the other hand, let $z \in \mathscr{M}_k^c \setminus T_{\mathscr{M}_k}$, so $z = i \cdot p_k + r$ with $r \in [\mathscr{M}_k^c]_{p_k}$. If $0 \leq r \leq k-1$, then $i \geqslant 1$ (because z is not a terminal position), and we can write $z = x + y$ with $x = (i-1)p_k + k + r \in \mathscr{M}_k$ and $y = 2k - 1 \in \mathscr{M}_k$. If $2k \leq r \leq 3k-2$, then $z = x + y$ with $x = i \cdot p_k + k \in \mathscr{M}_k$ and $y = r - k \in \mathscr{M}_k$. Thus $\mathscr{M}_k^c \setminus T_{\mathscr{M}_k} \subseteq \mathscr{M}_k + \mathscr{M}_k$, so $\mathscr{M}_k$ is reflexive by Theorem 2.

'$\Rightarrow$' We show that if $\mathscr{M} \neq \mathscr{M}_k$, then $\mathscr{M}$ is not reflexive. Let $k = \min(\mathscr{M})$. If $k = 0$, then $\mathscr{M} \neq \mathscr{M}_\ell$ for any ℓ, and furthermore, by Note 3, $\mathscr{M}$ is not reflexive. Now assume that $k > 0$, so $k \in P(\mathscr{M})$ by Note 2. Assume that there is a positive integer $x = \min(\mathscr{M}_k \setminus \mathscr{M} \cup \mathscr{M} \setminus \mathscr{M}_k)$, that is, x is the smallest value that differs between $\mathscr{M}$ and $\mathscr{M}_k$. Necessarily, $x > k$.

Suppose first that $x \in \mathscr{M}_k \setminus \mathscr{M}$. Because $x \notin \mathscr{M}$, it suffices to prove that $x \in P(\mathscr{M}) = \mathscr{M}^\star$ to show that $\mathscr{M}$ is not reflexive. Since $x > k$, there exists $y \in \mathscr{M}_k \cap \mathscr{M} \supseteq \{k\}$ such that $y < x$. For any such y, $y \in P(\mathscr{M}_k)$ by reflexivity of $\mathscr{M}_k$. By minimality of x, $y \in P(\mathscr{M})$ because the same moves are available from y in both $\mathscr{M}$ and $\mathscr{M}_k$. Since $x, y \in \mathscr{M}_k$, we have $x = i \cdot p_k + r$ and $y = j \cdot p_k + s$ for some $0 \leqslant j \leqslant i$ and $k \leqslant r, s \leqslant 2k-1$. Thus $z = x - y = (i-j) \cdot p_k + (r-s)$ with $-k + 1 \leqslant r - s \leqslant k-1$, so $z \notin \mathscr{M}_k$, and by minimality of x, $z \notin \mathscr{M}$. This implies that there is no move in $\mathscr{M}$ from x to a P-position y, so $x \in P(\mathscr{M})$, which completes this case.

Suppose next that $x \in \mathscr{M} \setminus \mathscr{M}_k$. It suffices to prove that $x \notin P(\mathscr{M})$ to show that $\mathscr{M}$ is not reflexive. By the minimality of x, it suffices to find an option z of x with $z \in P(\mathscr{M})$, that is $z = x - y$ for some $y \in \mathscr{M}$. Because $y, z < x$, we have $y, z \in \mathscr{M}_k \cap \mathscr{M}$ due to the minimality of x. Since $x \notin \mathscr{M}_k$, $x = i \cdot p_k + r$ for some $i \geq 0$ and $r \in [\mathscr{M}_k^c]_{p_k}$. If $r \in \{0, \ldots, k-1\}$, let $y = (i-1)p_k + (2k-1) \in \mathscr{M}_k$, otherwise choose $y = i \cdot p_k + k \in \mathscr{M}_k$. In each case, $[x - y]_{p_k} \in [\mathscr{M}_k]_{p_k}$. This shows that there is a move from x to a P-position $z \in P(\mathscr{M})$, so $x \notin P(\mathscr{M})$, which implies that $\mathscr{M}$ is not reflexive either in this case. Overall, the game $\mathscr{M}$ is reflexive if and only if $\mathscr{M}$ is of the form $\mathscr{M}_k$. □

Now that we have identified a family of games that are reflexive, we will show that these games are the only ones that can occur as limit games.

Corollary 2 *Let $\mathscr{M} \subseteq \mathbb{N}_0$ and let $k = \min(\mathscr{M})$ if $\mathscr{M} \neq \varnothing$, and $k = 0$ otherwise. Then $\lim_{i\to\infty} \mathscr{M}^i = \mathscr{M}_k$.*

Proof Since the limit game is reflexive, Theorem 4 applies, and $\mathscr{M}^\infty = \mathscr{M}_j$ for some $j \in \mathbb{N}$. If $\mathscr{M} = \varnothing$ or $0 \in \mathscr{M}$, then $\mathscr{M}^\infty = \varnothing = \mathscr{M}_0$, so the claim is true. If $\mathscr{M}$ is nonempty and $0 \notin \mathscr{M}$, then by Lemma 2, $k = \min(\mathscr{M}) = \min(\mathscr{M}^\infty)$. Since $\min(\mathscr{M}_j) = j$ for $j > 0$, the minimum uniquely determines $\mathscr{M}_j$, so we have that $\mathscr{M}^\infty = \mathscr{M}_k$. □

In conclusion, in one dimension we understand the structure of any limit game—it is periodic and is completely determined by the minimal move. This result is quite surprising in its simplicity, especially since in the case of normal-play, general formulas for limit games are rare in any dimension, the exceptions consisting of a few ‘immediately’ reflexive game families [5, 6].

Now that we have identified the sets $\mathscr{M}_k$ as the only possible limit games, we answer which games have $\mathscr{M}_k$ as their set of P-positions.

Theorem 5 *Let $k \in \mathbb{N}$ and $A_k = \{k, 2k-1\}$. Then $P(\mathscr{X}) = M_k$ if and only if $A_k \subseteq X \subseteq M_k$. That is, the set of games $\{\mathscr{X} \mid A_k \subseteq X \subseteq M_k\}$ is M_k-solvable and also M_k-complete.*

Proof We begin by proving that $P(\mathscr{A}_k) = M_k$. Clearly, $T_{\mathscr{A}_k} = \{0, \ldots, k-1\} \subset N(\mathscr{A}_k)$. We compute modulo $p_k = 3k-1$ and use (4) to justify that for each $x \in M_k^c \setminus \{0, \ldots, k-1\}$, $x - k \in M_k$, or $x - (2k-1) \in M_k$. Indeed, if $x \in \{0, \ldots, k-1\} \pmod{p_k}$, then $x - (2k-1) \in M_k$, and otherwise $x - k \in M_k$. For the other direction we must show that for all $x \in M_k$, both $x - k \in M_k^c$ and $x - (2k-1) \in M_k^c$, and this follows directly by (4). Thus $P(\mathscr{A}_k) = M_k$.

To prove the statement for a general set X with $A_k \subseteq X \subseteq M_k$, we use that $P(\mathscr{M}_k) = M_k$ (by Theorem 4). Hence, no move in $\mathscr{X}$ connects any two candidate P-positions in M_k. Moreover, since $A_k \subseteq X$, for each candidate N-position we find a move to a candidate P-position using the moves k or $2k-1$.

It remains to prove that no other sets X have the property $P(\mathscr{X}) = M_k$, that is, we need to show that if there is $x \in A_k \setminus X$ or $x \in X \setminus M_k$, then $P(\mathscr{X}) \neq M_k$. Suppose that there is a smallest $x \in A_k \setminus X$, with $M_k = P(\mathscr{X})$. Then $x = k$ or $x = 2k-1$; in the first case, if k is not a move, then $P(\mathscr{X}) = M_k$ implies that $x < k$ is a terminal N-position, so k as a non-move is also terminal and hence an N-position, a contradiction. Hence assume k is a move, but $2k-1$ is not. Then, there is no move from $4k-2 \in M_k^c$ to a P-position in $M_k = P(\mathscr{X})$, contradicting that M_k is the set of P-positions.

Suppose next that there is a smallest move $x \in X \setminus M_k$ with $P(\mathscr{X}) = M_k$. If $x \in T_{\mathscr{M}_k}$, then $\mathscr{X}$ and $\mathscr{M}_k$ do not have the same P-positions (since x is a P-position in $\mathscr{X}$, but an N-position in $\mathscr{M}_k$). Hence, we must have $x \notin T_{\mathscr{M}_k}$ and $x \in \{-(k-1), \dots, k-1\} \pmod{p_k}$. But, for each such x, we find two P-positions $y, z \in \{k, \dots, 2k-1\} \pmod{p_k}$ such that $y - z = x$, which contradicts x being a move. □

Given a game $\mathscr{M}$ (in any dimension), we denote the number of iterations of the misère-play $\star$-operator until the limit game appears for the first time by $\varphi(\mathscr{M}) = \min\{i \mid \mathscr{M}^i = \mathscr{M}^\infty\} \in \mathbb{N}_0 \cup \{\infty\}$. For the game $\mathscr{M} = \{k\}$, we derive $\varphi(\mathscr{M})$.

Lemma 3 *Let $\mathscr{M} = \{k\}$ with $k \geqslant 2$. Then*

1. $\mathscr{M}^1 = \{x \mid x \equiv k, \dots, 2k-1 \pmod{2k}\} = [\{k, \dots, 2k-1\}]_{2k}$.
2. $\mathscr{M}^2 = \{k, \dots, 2k-1\} \cup \{4k-1, 6k-1, \dots\}$.
3. $\mathscr{M}^3 = \{k, \dots, 2k-1\} \cup \{4k-1, \dots, 5k-2\} \cup \{7k-2, 9k-2, \dots\}$.
4. $\mathscr{M}^4 = \mathscr{M}_k \cap \{0, \dots, 10k-3\}$.
5. $\mathscr{M}^5 = \mathscr{M}_k$ *for any* k.

Figure 4 illustrates Lemma 3 for $\mathscr{M} = \{4\}$.

Proof 1. Let $S = [\{k, \dots, 2k-1\}]_{2k}$. The terminal positions of $\mathscr{M}$ are given by $T_{\mathscr{M}} = \{0, 1, \dots, k-1\} \subset S^c$. For any position $x \in S$, the position $x - k \notin S$. Also, for $x \notin S$, the position $x - k \in S$, so $S = P(\mathscr{M}) = \mathscr{M}^1$.

2. Let $S = \{k, \dots, 2k-1\} \cup \{4k-1, 6k-1, \dots\}$. The allowed moves are of the form $m = i \cdot 2k + r$ with $k \leqslant r \leqslant 2k-1$ and $i \geq 0$. Since $\mathscr{M}^1 \cap \mathscr{M}_k = \{0, \dots, 3k-1\}$, these moves are already fixed as P-positions. If $3k \leqslant x \leqslant 4k-2$, then $x - (2k-1) \in S$, so $x \in N(\mathscr{M}^1)$. If $x = j \cdot 2k - 1$ with $j \geq 2$, then $x - m \in \{0, \dots, k-1\} \subset S^c$. Also, for any $x > 4k-1$ with $x \notin S$, $x = j \cdot 2k + r$ with $0 \leq r \leq 2k-2$ and $j \geq 2$. Then for $0 \leq r < k-1$, $x - m \in S$ for $m = (j-1) \cdot$

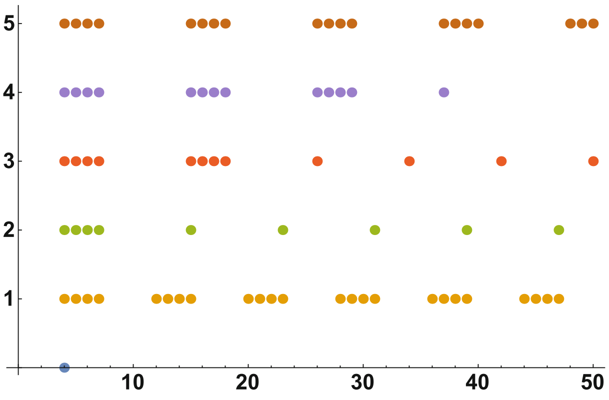


Fig. 4 The iterations of the misère-play $\star$-operator for $\mathscr{M} = \{4\}$

$2k + k + r$, and for $k \le r \le 2k - 2$, $x - (r + 1) \in S$.

3. Let $S = \{k, \dots, 2k - 1\} \cup \{4k - 1, \dots, 5k - 2\} \cup \{7k - 2, 9k - 2, \dots\}$. Note that $\mathscr{M}^2 \cap \mathscr{M}_k = \{0, \dots, 4k - 1\}$. If $x \in \{4k - 2, \dots, 5k - 2\}$, then the possible moves from x are of the form $m \in \{k, \dots, 2k - 1\} \cup \{4k - 1\}$, which gives $m - x \in \{1, \dots, k - 1\} \cup \{2k + 1, \dots, 4k - 2\} \subset N(\mathscr{M}^2)$. Suppose next that $x \in \{5k - 1, \dots, 7k - 3\}$. Then, there is a move $m \in \{k, \dots, 2k - 1\}$ to a position in the set $\{4k - 1, \dots, 5k - 2\} \subset P(\mathscr{M}^2)$, so $x \in N(\mathscr{M}^2)$. If $x \in \{7k - 2, 9k - 2, \dots\}$, then one can easily check that there is no move to any $y \in S$. If $7k - 2 \le x \notin \{7k - 2, 9k - 2, \dots\}$, then for $(2i - 1)k - 1 \le x \le (2i)k - 2$ and $i \ge 4$, the move $m = 2(i - 1)k - 1$ will lead to a P-position in $\{k, \dots, 2k - 1\}$, while for $(2i)k - 1 \le x \le (2i + 1)k - 1$, the move leading to a P-position is $m = 2(i - 2)k - 1$.

4. Note that $\mathscr{M}^3$ is identical with $\mathscr{M}_k$ for positions $x \le 7k - 2$, so it remains to investigate the case $x > 7k - 2$. Here, the argument is similar to 3.

5. This follows from Theorem 5. □

Corollary 3 *For $\mathscr{M} = \{k\}$, convergence to the limit set $\mathscr{M}^\infty = \mathscr{M}_k$ occurs in a finite number of steps. In particular, $\varphi(\{0\}) = \varphi(\{1\}) = 1$, and for $k \geqslant 2$, $\varphi(\{k\}) = 5$.*

Proof For $k = 0$, it follows from Note 3 that $\mathscr{M}^1 = \varnothing = \mathscr{M}_0$. For $k = 1$, let $S = \{1, 3, 5, \dots\} = \mathscr{M}_1$. Then for $x \in S$, $y = x - 1 \in S^c$, and likewise, for $x \in S^c$, $y = x - 1 \in S$, so $P(\{1\}) = \mathscr{M}_1$. In both cases, $\varphi(\{k\}) = 1$. For $k \ge 2$, $\varphi(\{k\}) = 5$ follows by Lemma 3.

We do not yet understand $\varphi(\mathscr{M})$ for any other case than the one described in Corollary 3. We have some experimental suggestions in the two-dimensional case, presented in Sect. 5.

4 Sum-Free Sets and Reflexivity

A set $A \subset \mathbb{N}$ is *sum-free* if the equation $a + b = c$ has no solution with $a, b, c \in A$. A sum-free set $A \subset \mathbb{N}$ is *maximal* if $A \cup \{x\}$ sum-free, with $x \in \mathbb{N}$, implies that $x \in A$. A sum-free set $A \subset \mathbb{N}$ is *perfect* if $\{a + b \mid a, b \in A\} = \mathbb{N} \setminus A$. For example, the set of odd positive numbers is a perfect sum-free set. Each perfect sum-free set is also maximal, but a maximal set need not be perfect.

For example, for $k > 1$, the set M_k (from Sect. 3) is maximal but not perfect. We can remedy the situation by studying instead the sets $\mathbb{N}_k = \{k, k + 1, \dots\}$, for $k \in \mathbb{N}$. We say that a sum-free set $A \subset \mathbb{N}_k$ is *k-min perfect* if $\{a + b \mid a, b \in A\} = \mathbb{N}_k \setminus A$ and $\min(A) = k$. We get the following result.

Theorem 6 *Let $k \in \mathbb{N}$. A sum-free set $A \subset \mathbb{N}_k$ is k-min perfect if and only if the misère-play subtraction game $\mathscr{A}$ is reflexive. Hence, the only k-min perfect sets are the sets M_k.*

Proof These are direct consequences of Theorems 2 and 4. □

If the density of a set $X \subset \mathbb{N}$ exists, then it is

$$\delta(X) = \lim_{n\to\infty} \frac{|X \cap \{1, \ldots, n\}|}{n}.$$

Let $A \subset \mathbb{N}_k$. Notice that the set of odd numbers greater than $k > 1$ is no longer maximal. Since M_k is maximal, we have a lower bound for how 'dense' a maximal set with smallest number k can be, namely

$$\delta(M_k) = \frac{k}{3k-1}.$$

Can we do better? (Of course, if we relax 'maximal' to be only 'sum-free', then the odd numbers $\geq k$ suffice for any k.)

What is the maximum of

$$\limsup_{n\to\infty} \frac{|A \cap \{1, \ldots, n\}|}{n}$$

for maximal sets A with $\min(A) = k$?

5 Structures in Two Dimensions

This section is intended as an overview of the behavior in two dimensions and should be regarded as an informal exposition. We indicate experimental similarities and differences with the known structures in one dimension.

In one dimension, all reflexive games have the same geometrical structure up to rescaling (as demonstrated in Sect. 3). In two dimensions, the geometrical structures of the reflexive games vary much more, even though for certain classes of games we still obtain similar rescaled structures. At the very least, our experiments show that we must distinguish classes of games according to where the minimal moves occur, as they must have different behavior due to Theorem 3. That the conjectured behavior is the same within each class is harder to prove in general but possible to be shown in certain cases. The following classification scheme is the least required:

1. The game has only one minimal move
   a. on one of the axes
   b. not on an axis
2. The game has exactly two minimal moves
   a. none of the minimal moves is on an axis
   b. exactly one of the minimal moves is on an axis

  c. both minimal moves are on the axes

3. The game has at least three minimal moves

   a. none of the minimal moves is on an axis
   b. exactly one of the minimal moves is on an axis
   c. there is a minimal move on each axis

The class 2(c) most closely resembles the one-dimensional case, as the two-dimensional limit game inherits some of its structure from the respective one-dimensional limit games. Figure 5 shows the iterations for a game of the form $\mathscr{M} = \min(\mathscr{M}) = \{(k, 0), (0, \ell)\}$, the simplest form of case 2(c), for $k = 4$ and $\ell = 3$. It appears that this game converges to a limit game after seven steps. In addition, after five steps, the behavior along the axes is as described in Theorem 4.

Informally, we define $\mathscr{M}_{k,\ell}$ as the type of limit game shown in Fig. 5 (for $k = 4$ and $\ell = 3$). It can be defined in a periodic manner based on k, ℓ, and the one-dimensional associated periods. We are in the process of proving this game to be reflexive [2]. Due to the periodic structure of $\mathscr{M}_{k,\ell}$, we know the limit game to be periodic along half lines of rational slopes. The structure of the limit game is generic, but the number of iterations until convergence can vary for this class.

Computer explorations for games in the other classes (see, for example, Figs. 7 and 8) suggest that all limit games have some type of periodic structure, which leads to the following conjecture.

Conjecture 1 Limit games for all two-dimensional vector subtraction games under the misère-play ⋆-operator are ultimately periodic along any line of rational slope.

Returning to class 2(c), one can ask which games $\mathscr{A}_{k,\ell}$ have the property that $P(\mathscr{A}_{k,\ell}) = \mathscr{M}_{k,\ell}$ (see Theorem 5 for the one-dimensional equivalent), and more

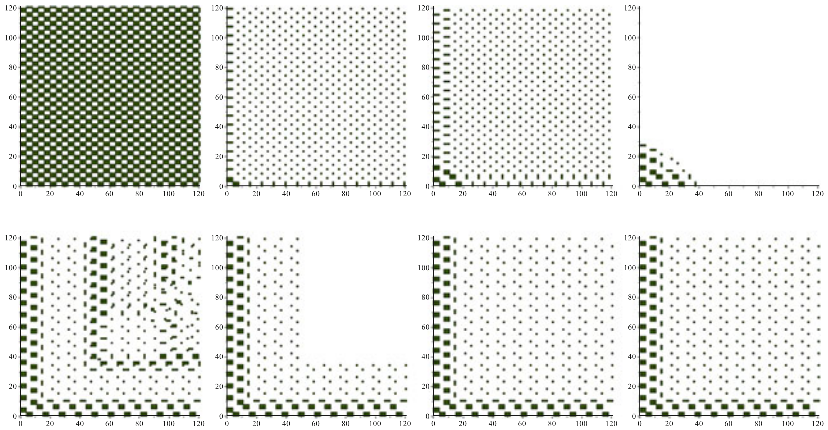


Fig. 5 Iterations of the misère-play ⋆-operator for the game $\mathscr{M} = \{(4, 0), (0, 3)\}$ where the game shown in the upper left is $\mathscr{M}^{\star}$. The limit game is reached after seven steps in this case

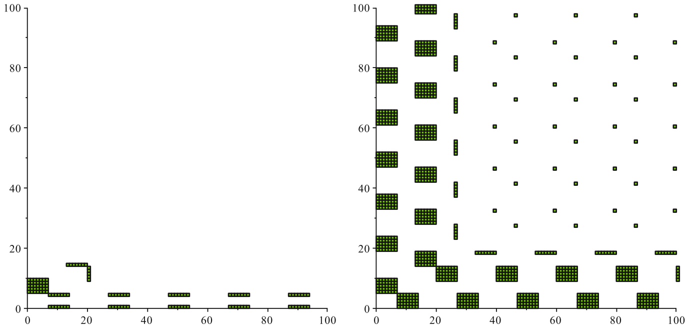


Fig. 6 The graph on the right represents the P-positions of the game $\mathscr{A}_{k,l}$ shown on the left, with $(k, \ell) = (7, 5)$

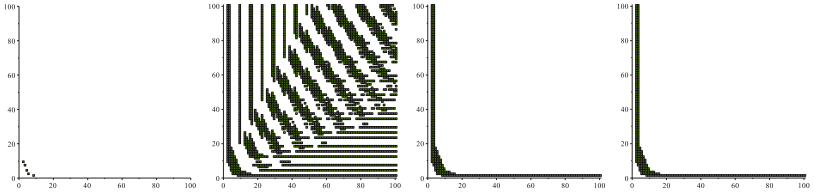


Fig. 7 The graphs show convergence after two iterations for the game $\mathscr{M} = \{(2, 9), (3, 7), (4, 4), (5, 2), (8, 1)\}$; case 3(a)

specifically, whether there is a smallest such game. In Fig. 6, we display the 'smallest' game discovered so far that has the reflexive game $\mathscr{M}_{k,\ell}$ as its set of P-positions.

Question 1 Is the game depicted on the left in Fig. 6 a generic description of a smallest game with a reflexive game of type 2(c) as its set of P-positions?

We conclude this section with some cases when there are at least three minimal moves. Suppose that $\min(\mathscr{M}) \cap \{(0, x), (x, 0) \mid x \in \mathbb{N}\} = \varnothing$, so we are in class 3(a). Then $\varphi(\mathscr{M}) = 2$, that is, $\mathscr{M}^{\star\star}$ is reflexive. It is not hard (but somewhat technical) to prove this statement by an explicit description of the generic description of the right-most graph in Fig. 6. Note also that this 'penultimate lower ideal' is already a subset of the second graph.

By comparison, the case 3(c) has most variation, and we do not yet know if each games in this class converges in a finite number of steps. We conclude by showing behavior of four games of the form

$$\mathscr{M}_x = \{(0, 5), (x, x), (5, 0)\},$$

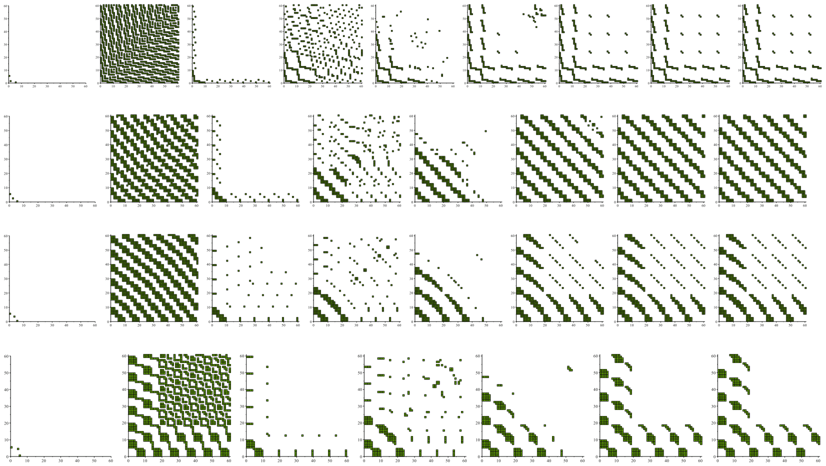
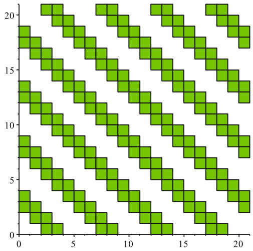


Fig. 8 Iterations of the ⋆-operator for four type 3c games $\mathscr{M} = \{(0, 5), (x, x), (5, 0)\}$, for $x = 1, 2, 3, 4$

Fig. 9 A reflexive game with diagonal shaped moves



for $x = 1, 2, 3, 4$. Based on Fig. 8, we hypothesize that $\varphi(\mathscr{M}_1) = 7, \varphi(\mathscr{M}_2) = 6, \varphi(\mathscr{M}_3) = 6, \varphi(\mathscr{M}_4) = 5$. Note that some limit games have generalized 'L-shapes', while others have 'negative-slope-diagonal-stripes,' and yet others appear to be a blend of the two.

The simplest nontrivial game whose limit game has 'diagonal stripes of negative slopes' is $\mathscr{M} = \{(0, 2), (1, 1), (2, 0)\}$. It converges in five steps to the game in Fig. 9. It generalizes the game $\{(0, 1), (1, 0)\}$, which trivially converges in one step to a checkerboard pattern.

We have performed many computer experiments in two dimensions but have not (yet) found any limit game with 'random' or 'chaotic' behavior. This is quite different from reflexivity in normal-play, where the crystal-like patterns so common in misère-play are rare.

Acknowledgements Thanks to Lydia Ievins and Michale Bergman for making the trip to CANT 2016 possible.

References

1. E.R. Berlekamp, J.H. Conway, R.K. Guy, Winning Ways, 2nd edn. (1–2 Academic Press, London 1982). 1–4. A. K. Peters, Wellesley/MA (2001/03/03/04)
2. C. Bloomfield, M. Dufour, S. Heubach, U. Larsson, Properties for the $\star$-operator of vector subtraction games (in preparation)
3. E. Duchêne, M. Rigo, Invariant games. Theoret. Comput. Sci. **411**(34–36), 3169–3180 (2010)
4. S.W. Golomb, A mathematical investigation of games of "take-away". J. Combin. Theory **1**, 443–458 (1966)
5. U. Larsson, P. Hegarty, A.S. Fraenkel, Invariant and dual subtraction games resolving the Duchêne-Rigo Conjecture. Theoret. Comp. Sci. **412**(8–10), 729–735 (2011)
6. U. Larsson, The $\star$-operator and invariant subtraction games. Theoret. Comput. Sci. **422**, 52–58 (2012)
7. U. Larsson, J. Wästlund, From heaps of matches to the limits of computability. Electron. J. Combin. **20**, 41 (2013)

A New Proof of Khovanskiĭ's Theorem on the Geometry of Sumsets

Jaewoo Lee

Abstract Khovanskiĭ studied how iterated sumsets grow geometrically, and provided the growth polynomial for sumsets as well as an approximation to lattice points inside polytopes. In this paper, we present a new proof of the theorem about geometric growth of sumsets.

Keywords Sumsets · Geometry of numbers · Geometry of sumsets
Growth of sumsets · Lattice points

1 Introduction and Notation

We denote by $\mathbf{Z}^n$ the group of lattice points in $\mathbf{R}^n$. Let A be a set of n- dimensional lattice points. For any positive integer h, we define the *h-fold sumset* $hA = \{a_1 + a_2 + \cdots + a_h : a_1, a_2, \ldots, a_h \in A\}$, and the *dilation* $h * A = \{ha : a \in A\}$.

A *hyperplane* H is the set $\{x \in \mathbb{R}^n : (x, u) = \alpha\}$ for a nonzero $u \in \mathbb{R}^n$ and a number α, where $(.,.)$ indicates an inner product in $\mathbb{R}^n$. The vector u is called a *normal vector* to H. A hyperplane divides $\mathbb{R}^n$ into two closed half-spaces H^+ and H^- where

$$H^+ = \{x \in \mathbb{R}^n : (x, u) \geq \alpha\},$$

$$H^- = \{x \in \mathbb{R}^n : (x, u) \leq \alpha\}.$$

We write $d(x, y)$ to denote the distance between two points $x, y \in \mathbb{R}^n$. If $S, T \subseteq \mathbb{R}^n$, then

$$d(x, S) = \inf_{s \in S} d(x, s)\,,$$

and

$$d(S, T) = \inf_{s \in S, t \in T} d(s, t).$$

J. Lee (✉)
Borough of Manhattan Community College (CUNY), New York, NY, USA
e-mail: jalee@bmcc.cuny.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_13

In particular, the distance from a point $x \in \mathbb{R}^n$ to a hyperplane H where $x \notin H$ is given by the length of the perpendicular line segment from x to H.

If two different hyperplanes H_1 and H_2 are parallel, we may write $H_1 = \{x : (x, u) = \alpha_1\}$ and $H_2 = \{x : (x, u) = \alpha_2\}$. Take any $x \in H_1$. Then, $d(x, H_2)$ is given by the perpendicular line segment. To calculate the distance between H_1 and H_2, note that $x + tu$ where $t \in \mathbb{R}$ gives the perpendicular ray from x to H_2. If the ray meets H_2 at $t = t_2$, then $t_2 = (\alpha_2 - \alpha_1)/|u|^2$. Thus, $d(x, H_2) = |t_2 u| = (\alpha_2 - \alpha_1)/|u|$, which is independent of the choice of x. Therefore, when H_1 and H_2 are parallel, $d(H_1, H_2)$ is given by the length of any perpendicular line segment joining them.

A *polytope* is the convex hull of a finite set of points in $\mathbb{R}^n$ (the algebraic definition), or equivalently, the bounded set which is an intersection of finitely many closed half-spaces (the geometric definition). Let $\Delta = \mathrm{conv}(A)$, the convex hull of A, where $A = \{a_1, a_2, \ldots, a_m\}$ is a finite set of lattice points in $\mathbb{R}^n$. Then, define the *dilation of* Δ, $h * \Delta$, as

$$\begin{aligned} h * \Delta &= \{hx : x \in \Delta\} \\ &= \left\{ \sum_{i=1}^{m} \lambda_i a_i : \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = h \right\} \\ &= \mathrm{conv}(ha_1, \ldots, ha_m). \end{aligned}$$

In [3], Khovanskiĭ showed that $|hA|$ eventually becomes a polynomial. Furthermore, he studied the geometric growth of hA in [3] as well.

Theorem 1 (Khovanskiĭ's Theorem) *Suppose $\mathbb{Z}^n(A) = \mathbb{Z}^n$ where $\mathbb{Z}^n(A)$ is the group generated by the difference set $A - A = \{a - a' : a, a' \in A\}$. There exists a constant ρ with the following property: For any positive integer h, every lattice point of $h * \Delta$, whose distance to the boundary $\partial(h * \Delta)$ is more than ρ, belongs to the sumset hA.*

Theorem 1 implies that the sumset hA in $\mathbb{R}^n$ takes over the central region of dilated polytopes, with fringes relatively getting smaller as h grows, giving a way to approximate lattice points inside polytopes. Using ideas in [3] and the algebraic definition of polytopes, the author [4] proved Theorem 1 conditionally. In this paper, we prove Theorem 1 unconditionally using ideas in [3] and the geometric definition of polytopes.

2 Lemmas and Proof of Theorem

We start with some lemmas that Khovanskiĭ proved in [3]. Let A be a finite subset of $\mathbb{Z}^n$, $A = \{a_1, \ldots, a_m\}$, with $|A| = m$ and $\Delta = \mathrm{conv}(A)$. Also, assume that A generates $\mathbb{Z}^n$ as a group.

Lemma 1 *There exists a constant C with the following property: For every linear combination $\sum \lambda_i a_i$ of $a_i \in A$ with real coefficients λ_i such that $\sum \lambda_i a_i$ is a lattice point, there exists a linear combination $\sum n_i a_i$ of a_i with integer coefficients n_i such that $\sum n_i a_i = \sum \lambda_i a_i$ and $\sum |n_i - \lambda_i| < C$.*

Proof Let $X = \{x : x \in \mathbb{Z}^n, x = \sum \lambda_i a_i, \text{ with } 0 \leq \lambda_i \leq 1\}$, which is a finite set. Since A generates $\mathbb{Z}^n$, each $x \in X$ can be written as $x = \sum_{i=1}^m n_i(x) a_i$, where $n_i(x) \in \mathbb{Z}$. So for each $x \in X$, we fix one representation $\sum_{i=1}^m n_i(x) a_i$ with $n_i(x) \in \mathbb{Z}$. Let $q = \max_{x \in X} \sum_{i=1}^m |n_i(x)|$ and let $C = m + q$, a positive integer. Then, for any $z = \sum \lambda_i a_i \in \mathbb{Z}^n$, $x = z - \sum [\lambda_i] a_i \in X$. So $x = \sum_{i=1}^m n_i(x) a_i$ with $n_i(x) \in \mathbb{Z}$ and $z = \sum_{i=1}^m \big(n_i(x) + [\lambda_i]\big) a_i = \sum_{i=1}^m \lambda_i a_i$ with $\sum |n_i(x) + [\lambda_i] - \lambda_i| < \sum_{i=1}^m \big(|n_i(x)| + 1\big) \leq q + m = C$. □

Let h be a positive integer and assume $0 \in A$. Then,

$$\Delta = \left\{ \sum \lambda_i a_i : \lambda_i \geq 0, \sum \lambda_i \leq 1 \right\}$$

and

$$h * \Delta = \left\{ \sum \lambda_i a_i : \lambda_i \geq 0, \sum \lambda_i \leq h \right\} .$$

Define

$$\Delta(h, C) = \left\{ \sum \lambda_i a_i : \lambda_i \geq C, \sum \lambda_i \leq h - C \right\}$$

with C as in Lemma 1.

Then, if $x = \sum \lambda_i a_i \in \Delta(h, C)$, let $\lambda_i = \alpha_i + C,\ \alpha_i \geq 0$. So

$$\begin{aligned} \Delta(h, C) &= \left\{ \sum (\alpha_i + C) a_i : \alpha_i \geq 0, \ \sum \alpha_i \leq h - C - mC \right\} \\ &= C \sum a_i + \left\{ \sum \alpha_i a_i : \alpha_i \geq 0, \ \sum \alpha_i \leq h - C - mC \right\} \\ &= C \sum a_i + (h - C - mC) * \Delta. \end{aligned}$$

Note $\Delta(h, C)$ is an empty set when $h < C + mC$, a single point $C \sum a_i$ when $h = C + mC$, and a dilation of Δ translated by a lattice point contained in $h * \Delta$ when $h \geq C + mC + 1$.

Lemma 2 *If $\mathbb{Z}^n(A) = \mathbb{Z}^n$ and $0 \in A$, then every lattice point in $\Delta(h, C)$ belongs to the sumset hA.*

Proof Let z be a lattice point in $\Delta(h, C)$. Then,

$$z = \sum \lambda_i a_i, \quad \lambda_i \geq C, \sum \lambda_i \leq h - C.$$

By Lemma 1, $z = \sum n_i a_i,\ n_i \in \mathbb{Z},\ \sum |n_i - \lambda_i| < C$. If $n_i < 0$ for some i, then $|n_i - \lambda_i| > C$. Therefore, all n_i must be nonnegative. And $\sum n_i = \sum |n_i| = \sum |n_i - \lambda_i + \lambda_i| \leq \sum |n_i - \lambda_i| + \sum |\lambda_i| < C + h - C = h$.

Thus, $z = \sum n_i a_i,\ n_i \geq 0,\ \sum n_i < h$. Since $0 \in A$,

$$hA = \left\{ \sum n_i a_i : n_i \geq 0, \sum n_i \leq h \right\} .$$

Therefore, $z \in hA$. □

Now, we prove Theorem 1.

Proof Take any hyperplane

$$H = \{x : (x, u) = \alpha\}.$$

Then, for a positive integer h,

$$h * H = \{x : (x, u) = h\alpha\},$$

so the dilation of a hyperplane results in another hyperplane which is parallel to the original one. And

$$H - b = \{x : (x, u) = \alpha - (b, u)\}$$

where $b \in \mathbb{R}^n$, so the translation of a hyperplane is a hyperplane that is parallel to the original one as well.

Now, let us calculate the distance between

$$H_1 = h * H$$

and

$$H_2 = g * H - b ,$$

where $h > g$ and h, g are positive integers with $b \in \mathbb{R}^n$. Then, $H_1 = \{x : (x, u) = h\alpha\}$, $H_2 = \{x : (x, u) = g\alpha - (b, u)\}$, so H_1 is parallel to H_2. Take any point $x_1 \in H_1$. Then, $x_1 + tu,\ t \in \mathbb{R}$ is a ray perpendicular to both H_1 and H_2. Let us say $x_1 + tu \in H_2$ when $t = t_2$. Then,

$$t_2 = \frac{(g - h)\alpha - (b, u)}{|u|^2},$$

$$d(H_1, H_2) = |t_2 u| = \frac{|(g - h)\alpha - (b, u)|}{|u|} .$$

Without loss of generality, we may assume $0 \in A$ because, if not, take any $a \in A$ which is also a vertex of Δ. Then, take $\bar{\Delta} = \Delta - a$ so that $0 \in \bar{\Delta}$. Then, $\bar{\Delta} = \text{conv}(A - a)$ and $h * \bar{\Delta} = h * \text{conv}(A - a) = h * \Delta - ha$. And, for any positive integer h, if $x \in (h * \Delta) \cap \mathbb{Z}^n$ with $d(x, \partial(h * \Delta)) > \rho$, then $x - ha \in h * \bar{\Delta}$ and $d(x - ha, \partial(h * \bar{\Delta})) > \rho$ since a translation does not change the distance. Thus, $x - ha \in h(A - a) = hA - ha$. So $x \in hA$, proving our claim.

Let $h \geq C + mC + 1$. Recall

$$\Delta(h, C) = C \sum a_i + (h - C - mC) * \Delta \, .$$

Let $\Delta = G_1^+ \cap \cdots \cap G_l^+$ where G_is are hyperplanes $\{x : (x, u_i) = \alpha_i\}$ with $G_i \cap \Delta \neq \emptyset$. Then, $h * \Delta = H_1^+ \cap \cdots \cap H_l^+$ and $\Delta(h, C) = H_1^{'+} \cap \cdots \cap H_l^{'+}$ where $H_i = h * G_i$, $H_i^{'} = (h - C - mC) * G_i + C \sum a_i$. And for all $h \geq C + mC + 1$,

$$d(H_i, H_i^{'}) = \frac{|(-C - mC)\alpha_i + (C \sum a_i, u_i)|}{|u_i|}$$

for $i = 1, \ldots, l$, using the result above on the distance between hyperplanes. Thus, for all $i = 1, \ldots, l$, the distance $d(H_i, H_i^{'})$ remains the same for all $h \geq C + mC + 1$.

Therefore, fix any $h \geq C + mC + 1$. Define

$$\rho = \max \{ \delta\big((C + mC) * \Delta\big), \; d(H_i, H_i^{'}), i = 1, \ldots, l \}$$

where $\delta(S)$ represents the diameter of the set S. Then, ρ is independent of h.

Let $z \in h * \Delta$ be a lattice point with $d(z, \partial(h * \Delta)) > \rho$. Note that if $h \leq C + mC$, then by the definition of ρ, such z does not exist.

Let $F_i = H_i \cap (h * \Delta) \neq \emptyset$ be a face of $h * \Delta$ and $F_i^{'} = H_i^{'} \cap \Delta(h, C) \neq \emptyset$ be a face of $\Delta(h, C)$. Assume $z \in H_1^{'-} \setminus H_1^{'}$. Then, $d(z, H_1) < d(H_1^{'}, H_1) \leq \rho$, but $d(z, F_1) > \rho$. Thus, the perpendicular ray from z to H_1 does not intersect F_1. Every compact convex body in $\mathbb{R}^n$ with nonempty interior is homeomorphic to the closed n-ball, and its boundary is homeomorphic to the $(n - 1)$-sphere (see, e.g., [1, p. 56]). So, $\partial(h * \Delta)$ is homeomorphic to the $(n - 1)$-sphere. Thus, the perpendicular ray from z to H_1 intersects $\partial(h * \Delta)$ somewhere, say, at z_2 which is a point of a face F_2, $F_2 \neq F_1$. Then, $z_2 \in F_2 \subseteq h * \Delta$, so $z_2 \in H_1^+$. Then, $d(z, F_2) \leq d(z, z_2) \leq d(z, H_1) < \rho$, a contradiction. Therefore, $z \in H_1^{'+}$. Similarly, z belongs to other $H_i^{'+}$ for $i = 2, \ldots, l$ as well. Thus, $z \in H_1^{'+} \cap \cdots \cap H_l^{'+} = \Delta(h, C)$. Then, by Lemma 2, $z \in hA$. □

Han [2] showed that for $A \subseteq \mathbb{R}^2$ satisfying some conditions, the cardinality of hA in the fringe region of dilated polytopes is a linear function of h when h is sufficiently large. It will be interesting if we can tell something more about the density or the distribution of sumsets in the fringe region.

References

1. G.E. Bredon, *Topology and Geomtry* (Springer, New York, 1993)
2. S.S. Han, The boundary structure of the sumset in $\mathbb{Z}^2$, in *Number Theory, New York, 2003* (Springer, New York, 2004), pp. 201–218
3. A.G. Khovanskiĭ, The Newton polytope, the Hilbert polynomial and sums of finite sets. (Russian) Funktsional. Anal. i Prilozhen 26, 57–63, 96 (1992); translation. Funct. Anal. Appl. **26**, 276–281 (1992)
4. J. Lee, Algebraic proof for the geometric structure of sumsets. Integers **11**, 477–486 (2011)

Initial Sums of the Legendre Symbol: Is min + max ≥ 0 ?

Kieren MacMillan and Jonathan Sondow

Abstract Dirichlet famously proved that for primes p of the form $4n + 3$, the half-interval $(0, \frac{1}{2}p)$ contains more quadratic residues modulo p than nonresidues. An elementary argument then uses this to prove an inequality for an initial sum of the Legendre symbol $\left(\frac{a}{p}\right)$ for any odd prime p, namely $\sum_{0<a<\frac{1}{2}p}\left(\frac{a}{p}\right) \geq 0$, with strict inequality if and only if $p \equiv 3 \pmod 4$. From computations with the first 25000 primes, Sondow conjectured that

$$\min_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) + \max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) \geq 0,$$

also with strict inequality if and only if $p \equiv 3 \pmod 4$. In this note, we prove that equality holds when $p \equiv 1 \pmod 4$, and that if $3 \neq p \equiv 3 \pmod 4$ then $\max\limits_{0<k<p} \sum_{a=1}^{k}\left(\frac{a}{p}\right)$ exceeds the class number $h(-p)$. We also give extensions to the Jacobi and Kronecker symbols $\left(\frac{a}{n}\right)$.

Keywords Quadratic residue · Legendre symbol · Class number
Jacobi symbol · Kronecker symbol

1 Introduction: The Legendre Symbol

Let p be an odd prime number. For any integer a, the *Legendre symbol* $\left(\frac{a}{p}\right)$ is defined as

K. MacMillan (✉)
55 Lessard Avenue, Ontario, Toronto M6S 1X6, Canada
e-mail: kieren@alumni.rice.edu

J. Sondow
209 West 97th Street, New York, NY 10025, USA
e-mail: jsondow@alumni.princeton.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_14

Fig. 1 Johann Peter Gustav Lejeune Dirichlet (1805–1859)

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{if } p \mid a, \\ +1 & \text{if } p \nmid a \text{ and the congruence } x^2 \equiv a \pmod{p} \text{ has a solution,} \\ -1 & \text{if } p \nmid a \text{ and the congruence } x^2 \equiv a \pmod{p} \text{ has no solution.} \end{cases}$$

In the second case, we say that a is a *quadratic residue* modulo p, and in the third case that a is a *quadratic nonresidue* modulo p (Fig. 1).

In 1839, as a by-product of his investigations on binary quadratic forms [4], Dirichlet established the remarkable theorem that *for primes p of the form* $4n+3$, *the half-interval* $(0, \frac{1}{2}p)$ *contains more quadratic residues modulo p than nonresidues*. (No elementary proof is known, but the proofs in [1, 5, 9] avoid quadratic forms by using complex variables or Fourier series.) On the other hand, an elementary argument shows that *if p is a prime of the form* $4n+1$, *then* $(0, \frac{1}{2}p)$ *contains exactly as many quadratic residues modulo p as nonresidues*. The two cases may be stated together using the Legendre symbol.

Theorem 1 (Dirichlet) *For any odd prime p, the Legendre symbol* $\left(\frac{a}{p}\right)$ *satisfies*

$$\sum_{0<a<\frac{1}{2}p} \left(\frac{a}{p}\right) \begin{cases} = 0 & \text{if } p \equiv 1 \pmod{4}, \\ > 0 & \text{if } p \equiv 3 \pmod{4}. \end{cases} \tag{1}$$

Sondow made a related prediction after calculations with the first 25000 primes.

Conjecture 1 (Sondow) *For any odd prime p, the minimum and maximum initial sums of the Legendre symbol* $\left(\frac{a}{p}\right)$ *satisfy*

$$\min_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) + \max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) \begin{cases} = 0 & \text{if } p \equiv 1 \pmod{4}, \\ > 0 & \text{if } p \equiv 3 \pmod{4}. \end{cases} \tag{2}$$

For brevity, when p is an odd prime, we denote

$$m = m(p) := \min_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right), \qquad M = M(p) := \max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right). \tag{3}$$

Table 1 Initial sums of the Legendre symbol for $p = 5$ and $p = 7$

p	$p \pmod 4$	a	$\left(\frac{a}{p}\right)$	$\sum_{a=1}^{k}\left(\frac{a}{p}\right)$	$m + M$
5	1	**1**, 2, 3, **4**	$1, -1, -1, 1$	$0, 1, 0, -1, 0$	$-1 + 1 = 0$
7	3	**1**, **2**, 3, **4**, 5, 6	$1, 1, -1, 1, -1, -1$	$0, 1, 2, 1, 2, 1, 0$	$0 + 2 > 0$

As $m \le 0 \le M$, the inequality $m + M \ge 0$ holds if and only if it is true that $\max(|m|, M) = M$. Thus, if Conjecture 1 is true, then for $p = 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, \ldots$, we have

$$\max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) = \max_{0<k<p} \left| \sum_{a=1}^{k} \left(\frac{a}{p}\right) \right| = 1, 1, 2, 3, 2, 2, 3, 5, 3, 6, 4, 4, 5, \ldots$$

(see [7, Sequence A177865]). Indeed, computing both maxima led to Conjecture 1.

Table 1 shows the examples $p = 5$ and 7 with quadratic residues in **bold** and the values of $\sum_{a=1}^{k}\left(\frac{a}{p}\right)$ listed for $k = 0, 1, \ldots, p-2, p-1$ to exhibit their (anti)symmetry.

In Sect. 2, we prove the first case of Conjecture 1. In Sect. 3, we conjecture that in the second case the class number $h(-p)$ is a lower bound and prove a special case. In Sects. 4 and 5, we discuss extensions to the Jacobi and Kronecker symbols $\left(\frac{a}{n}\right)$.

2 The Case $p \equiv 1 \pmod 4$

We prove the first case of Conjecture 1.

Theorem 2 *For primes* $p \equiv 1 \pmod 4$*, we have the equality*

$$\min_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) + \max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) = 0.$$

Proof At first, we let p be any odd prime. Denote by $\mathcal{S}_k = \mathcal{S}_k(p) := \sum_{a=1}^{k}\left(\frac{a}{p}\right)$ the kth initial sum of $\left(\frac{a}{p}\right)$, so that the empty sum is $\mathcal{S}_0 = 0$.

We first show that $\mathcal{S}_{p-1} = 0$. A reduced residue system modulo p consists of $(p-1)/2$ quadratic residues congruent to the numbers $1^2, 2^2, \ldots, ((p-1)/2)^2$, and $(p-1)/2$ quadratic nonresidues (see, e.g., [8, p. 80]). Thus half of the Legendre symbols $\left(\frac{a}{p}\right)$ are $+1$ and half are -1, so $\mathcal{S}_{p-1} = 0$.

We next prove that, for $k = 0, 1, \ldots, p-1$, the (anti)symmetry

$$\mathcal{S}_{p-1-k} = (-1)^{\frac{p+1}{2}} \mathcal{S}_k \tag{4}$$

holds. *Euler's criterion* (see, e.g., [8, p. 81]) asserts that

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}.$$

From this and the definition of the Legendre symbol, we infer that

$$\begin{gathered}\left(\frac{aa'}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{a'}{p}\right), \qquad \left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}, \\ a \equiv a' \pmod{p} \implies \left(\frac{a}{p}\right) = \left(\frac{a'}{p}\right).\end{gathered} \tag{5}$$

Hence $\left(\frac{p-a}{p}\right) = \left(\frac{-a}{p}\right) = (-1)^{\frac{p-1}{2}}\left(\frac{a}{p}\right)$. Now, by re-indexing, we can write

$$\mathcal{S}_{p-k-1} = \sum_{k<a<p} \left(\frac{p-a}{p}\right) = (-1)^{\frac{p-1}{2}} \sum_{k<a<p} \left(\frac{a}{p}\right) = (-1)^{\frac{p-1}{2}} (\mathcal{S}_{p-1} - \mathcal{S}_k)$$

and as $\mathcal{S}_{p-1} = 0$ we get (4).

Finally, assume that $p \equiv 1 \pmod 4$ and let the maximum initial sum be $M = \mathcal{S}_{k_0}$. Then by (4), for $k = 0, 1, \ldots, p-1$ we have

$$\mathcal{S}_k \le \mathcal{S}_{k_0} \quad \implies \quad -\mathcal{S}_k \ge -\mathcal{S}_{k_0} \quad \implies \quad \mathcal{S}_{p-k-1} \ge \mathcal{S}_{p-k_0-1}.$$

Thus the minimum initial sum is $m = \mathcal{S}_{p-k_0-1} = -\mathcal{S}_{k_0} = -M$. This proves the theorem. □

Notice that for a prime $p \equiv 1 \pmod 4$, setting $k = \frac{1}{2}(p-1)$ in (4) yields $\mathcal{S}_{\frac{1}{2}(p-1)} = 0$. This proves the first case of (1) and strengthens its connection with the first case of (2).

3 The Case $p \equiv 3 \pmod 4$

For the second case of Theorem 1, Dirichlet actually proved an exact formula. Its statement involves a quantity denoted $h(-p)$ called either "the ideal class number of the imaginary quadratic field $\mathbb{Q}(\sqrt{-p})$" or, equivalently, "the class number of binary quadratic forms of discriminant $-p$" (see [1, 5, 9]). In the present note, the only property of the class number $h(-p)$ we use is that it is a positive integer.

Theorem 3 (Dirichlet) *For any prime $p \ge 7$ with $p \equiv 3 \pmod 4$, let $h = h(-p)$. Then*

$$\sum_{0<a<\frac{1}{2}p} \left(\frac{a}{p}\right) = \left(2 - \left(\frac{2}{p}\right)\right) h = \begin{cases} 3h & \text{if } p \equiv 3 \pmod 8, \\ h & \text{if } p \equiv 7 \pmod 8. \end{cases} \tag{6}$$

Table 2 Theorem 3 and Conjecture 2 for $p = 7$ and $p = 11$

p	$p \pmod 8$	$\left(\frac{a}{p}\right)$	$\sum_{a=1}^{k}\left(\frac{a}{p}\right)$	$m + M$	h
7	7	$1, 1, -1, 1, -1, -1$	$0, 1, 2, 1, 2, 1, 0$	2	1
11	3	$1, -1, 1, 1, 1, -1, -1, -1, 1, -1$	$0, 1, 0, 1, 2, 3, 2, 1, 0, 1, 0$	3	1

Table 3 Theorem 3 and Conjecture 2 up to $p = 157$

$p =$	7	11	19	23	31	43	47	59	67	71	79	83	103	107	127	131	139	157
$p \pmod 8 =$	7	3	3	7	7	3	7	3	3	7	7	3	7	3	7	3	3	7
$m =$	0	0	-1	0	0	-2	0	0	-3	0	0	0	0	-2	-1	0	-3	0
$M =$	2	3	3	5	6	5	8	9	6	10	10	9	10	9	10	15	9	14
$m + M =$	2	3	2	5	6	3	8	9	3	10	10	9	10	7	9	15	6	14
$h =$	1	1	1	3	3	1	5	3	1	7	5	3	5	3	5	5	3	7

The second equality follows from the quadratic reciprocity law supplement

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}. \tag{7}$$

The computations which led to Conjecture 1 suggest more precisely that, in the second case, the class number is a lower bound. This leads to our second conjecture.

Conjecture 2 *For any prime $p \geq 7$ with $p \equiv 3 \pmod 4$, we have the strict inequality*

$$\min_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) + \max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) > h(-p).$$

For example, when $p = 19$ it becomes $-1 + 3 > 1$. Thus Conjecture 2 is sharp.

Using the notation (3), we illustrate Theorem 3 and Conjecture 2 with a table for $p = 7$ and $p = 11$, a table for all $p \equiv 3 \pmod 4$ with $7 \leq p \leq 157$, and a plot for $p = 163$ (Tables 2, 3 and Fig. 2).

For primes $p \neq 3$ with $p \equiv 3 \pmod 4$, Dirichlet's class number formula (6) implies the lower bound $M \geq h$. We sharpen it slightly, confirming a special case of Conjecture 2.

Theorem 4 *For any prime $p \geq 7$ with $p \equiv 3 \pmod 4$, we have the strict inequality*

$$\max_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) > h(-p).$$

In particular, Conjecture 2 holds true when $\min\limits_{0<k<p} \sum_{a=1}^{k}\left(\frac{a}{p}\right)$ vanishes (Fig. 2).

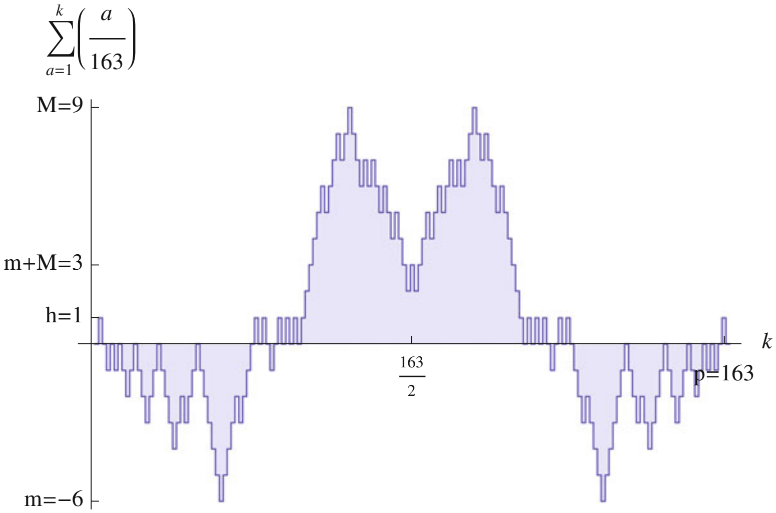


Fig. 2 Theorem 3 and Conjecture 2 for $p = 163$

Proof First let p be any odd prime. Using the notation in Sect. 2, we see by changing the order of summation and applying the formula $\mathcal{S}_{p-1} = 0$, that

$$\sum_{0<k<p} \mathcal{S}_k = \sum_{0<k<p} \sum_{a=1}^{k} \left(\frac{a}{p}\right) = \sum_{0<a<p} \sum_{k=a}^{p-1} \left(\frac{a}{p}\right) = \sum_{0<a<p} \left(\frac{a}{p}\right)(p-a) = -\sum_{0<a<p} \left(\frac{a}{p}\right) a.$$

Since for primes $p \neq 3$ with $p \equiv 3 \pmod 4$, Dirichlet proved the class number formula

$$h(-p) = -\sum_{0<a<p} \left(\frac{a}{p}\right) \frac{a}{p}$$

(see [5, Eq. (3)], [6, p. 219, Eq. (25)]), it follows that the mean of the numbers $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{p-1}$ is $\frac{1}{p}\sum_{k=0}^{p-1} \mathcal{S}_k = h(-p)$. As $\mathcal{S}_0 = 0$ is less than $h(-p)$, some $\mathcal{S}_k$ must be greater than $h(-p)$. The theorem follows. □

Here is an easy application.

Corollary 1 *Let $p \neq 3$ be a prime of the form $4n + 3$. Then on some interval $[1, k]$ with $2 \leq k \leq p - 3$, the excess in the number of quadratic residues modulo p over the number of nonresidues exceeds the class number $h(-p)$. Moreover, if $p \equiv 7 \pmod 8$, then $a \neq \frac{1}{2}(p-1)$.*

Proof Since $\mathcal{S}_k \leq 1 \leq h(-p)$ for $k = 1, p-2, p-1$, the first assertion follows from Theorem 4. The second then follows from Theorem 3. □

Thus Conjecture 2, and hence also Conjecture 1, holds at least for those primes p for which all initial sums $\sum_{a=1}^{k}\left(\frac{a}{p}\right)$ are nonnegative, i.e., for $p =$7, 11, 23, 31, 47, 59, 71, 79, 83, 103, 131, 151, 167, 191, 199, 239, 251, 263, 271, 311, 359, 383, ... (see [7, Sequence A095102]). Of course, the case $p \equiv 3 \pmod 8$ follows immediately from (6), since $M \geq 3h > h$. Similarly, it might be easier to prove Conjecture 2 for $p \equiv 3 \pmod 8$ than for $p \equiv 7 \pmod 8$.

4 An Extension: The Jacobi Symbol

To extend our results from odd prime modulus p to odd composite modulus n, we replace the Legendre symbol $\left(\frac{a}{p}\right)$ with the *Jacobi symbol* $\left(\frac{a}{n}\right)$, where n is any odd positive integer. If it has prime factorization $n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$, with each exponent $e_i \geq 1$, then $\left(\frac{a}{n}\right)$ is defined as the product of the Legendre symbols

$$\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right)^{e_1} \left(\frac{a}{p_2}\right)^{e_2} \cdots \left(\frac{a}{p_s}\right)^{e_s}. \tag{8}$$

Thus $\left(\frac{a}{1}\right) = 1$ is the empty product, $\left(\frac{a}{n}\right) = 0$ if $\gcd(a, n) > 1$, and $\left(\frac{a}{n}\right) = \pm 1$ if $\gcd(a, n) = 1$. Also, $a \equiv a' \pmod n$ implies $\left(\frac{a}{n}\right) = \left(\frac{a'}{n}\right)$.

If $\left(\frac{a}{n}\right) = -1$, then $\left(\frac{a}{p_i}\right) = -1$ for some $p_i \mid n$, and so the congruence $x^2 \equiv a \pmod n$ has no solution. But the converse is false: there may or may not be a solution if $\left(\frac{a}{n}\right) = +1$. For example, 2 is not a square modulo 15, even though $\left(\frac{2}{15}\right) = \left(\frac{2}{3}\right)\left(\frac{2}{5}\right) = (-1)(-1) = +1$.

A generalization of Conjecture 1 to the Jacobi symbol $\left(\frac{a}{n}\right)$ would also be false; for instance, if $n = 423 \equiv 3 \pmod 4$, then—using the notation (3) with n in place of p—the sum $m + M = -5 + 4 = -1$ is negative. However, Theorem 2 extends in the following way.

Theorem 5 *For odd n, the initial sums of the Jacobi symbol* $\left(\frac{a}{n}\right)$ *satisfy*

$$\min_{0<k<n} \sum_{a=1}^{k} \left(\frac{a}{n}\right) + \max_{0<k<n} \sum_{a=1}^{k} \left(\frac{a}{n}\right) = \begin{cases} 0 & \textit{if } n \equiv 1 \pmod 4 \textit{ is not a,} \quad \square \\ 1 + \phi(n) & \textit{if } n > 1 \;\; \textit{is} \;\; a, \quad \square \end{cases}$$

where $\phi(n) := \#\{a : 1 \leq a \leq n,\ \gcd(a, n) = 1\}$ *is Euler's totient function.*

Proof If $n = r^2$ is a square, then $\left(\frac{a}{n}\right) = \left(\frac{a}{r^2}\right) = \left(\frac{a}{r}\right)^2 = 1$ when $\gcd(a, n) = 1$, and $\left(\frac{a}{n}\right) = 0$ when $\gcd(a, n) > 1$. Now $n > 1$ implies $m = 1$ and $M = \phi(n)$, proving the second case.

To prove the first case, it suffices (as in the proof of Theorem 2) to show (A) that $\left(\frac{n-a}{n}\right) = \left(\frac{a}{n}\right)$ if $n \equiv 1 \pmod 4$, and (B) that $\mathcal{S}_{n-1} := \sum_{a=1}^{n-1} \left(\frac{a}{n}\right) = 0$ if n is not a square.

For (A), if $n \equiv 1 \pmod 4$, then n has an even number $\nu \geq 0$ of prime factors $\equiv 3 \pmod 4$, counted with multiplicity. Hence $\left(\frac{-1}{n}\right) = (-1)^\nu = 1$ and $\left(\frac{n-a}{n}\right) = \left(\frac{-a}{n}\right) = \left(\frac{a}{n}\right)$.

For (B), we first find an integer b with $\left(\frac{b}{n}\right) = -1$. If odd $n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$ is not a square, where $p_1 < p_2 < \cdots < p_s$ are primes, then some exponent e_i is odd. From $\mathcal{S}_{p-1} = 0$ with $p = p_i$, there exists c with $\left(\frac{c}{p_i}\right) = -1$. By the Chinese Remainder Theorem, the system of congruences

$$x \equiv c \pmod{p_i}, \qquad x \equiv 1 \pmod{p_j} \quad (1 \le j \le s,\ j \ne i)$$

has a solution $x = b$. As e_i is odd, we infer from (5) and (8) that $\left(\frac{b}{n}\right) = -1$.

Hence $\gcd(b, n) = 1$, and so the numbers $b, 2b, 3b, \dots, (n-1)b$ represent the $n - 1$ nonzero congruence classes modulo n. Therefore

$$\mathcal{S}_{n-1} = \sum_{a=1}^{n-1} \left(\frac{ab}{n}\right) = \left(\frac{b}{n}\right) \sum_{a=1}^{n-1} \left(\frac{a}{n}\right) = -\mathcal{S}_{n-1}$$

so that $\mathcal{S}_{n-1} = 0$. The first case of the theorem follows (as in Sect. 2) and we are done. □

To shorten the proof, one could use the theory of *Dirichlet characters* (see [2, pp. 40–41], [3, pp. 27–30]). Namely, the Jacobi symbol $\left(\frac{a}{n}\right)$ is a character modulo n, and any character χ modulo n has period sum $\sum_{a=0}^{n-1} \chi(a) = 0$ if χ is nonprincipal, i.e., if $\chi(a) \ne 0, 1$ for some a.

5 A Further Extension: The Kronecker Symbol

In attempting to extend our results to even numbers n, we may try using the *Kronecker symbol* $\left(\frac{a}{n}\right)$, where n is any positive integer. If $n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$ with $p_1 < p_2 < \cdots < p_s$, then $\left(\frac{a}{n}\right)$ is defined by the product (8) together with, in the case $p_1 = 2$, the values (compare (7))

$$\left(\frac{a}{2}\right) := \begin{cases} 0 & \text{if } a \text{ is even,} \\ \left(\dfrac{2}{|a|}\right) & \text{if } a \text{ is odd.} \end{cases}$$

However, for even n it is not true that $a \equiv a' \pmod{n}$ implies $\left(\frac{a}{n}\right) = \left(\frac{a'}{n}\right)$; for example, $21 \equiv 1 \pmod{10}$ but $\left(\frac{21}{10}\right) = \left(\frac{21}{2}\right)\left(\frac{21}{5}\right) = (-1)(+1) = -1 \ne +1 = \left(\frac{1}{10}\right)$. Thus the last paragraph in the proof of Theorem 5 would not be valid if n were even. For the same reason, the Kronecker symbol $\left(\frac{a}{n}\right)$ is not always a Dirichlet character modulo n.

For $n = 60, 142, 240, 423, 963 \equiv 4, 6, 0, 7, 3 \pmod 8$, the sum $m + M = -1$, $-4, -1, -1, -1$ is negative. Thus, while the inequalities of Conjecture 1 extend from odd primes p to composite numbers $n \equiv 1$ or $5 \pmod 8$ as in Theorem 5, the inequalities do not extend to composite numbers $n \equiv 0, 3, 4, 6,$ or $7 \pmod 8$.

For the remaining case $n \equiv 2 \pmod 8$, we offer the following conjecture, which is supported by numerical experiments.

Conjecture 3 *For any positive integer* $n \equiv 2 \pmod 8$*, the initial sums of the Kronecker symbol* $\left(\frac{a}{n}\right)$ *satisfy the inequality*

$$\min_{0<k<n} \sum_{a=1}^{k} \left(\frac{a}{n}\right) + \max_{0<k<n} \sum_{a=1}^{k} \left(\frac{a}{n}\right) \geq 0.$$

Acknowledgements We are grateful to Patrick Gallagher for stimulating discussions on quadratic residues and to Tauno Metsänkylä for information on class numbers.

References

1. B.C. Berndt, Classical theorems on quadratic residues. Enseign. Math. **2**(22), 261–304 (1976)
2. R. Crandall, C. Pomerance, *Prime Numbers: A Computational Perspective*, 2nd edn. (Springer, New York, 2005)
3. Davenport, H.: *Multiplicative Number Theory*, vol. 74, 2nd edn., ed. by H.L. Montgomery. Graduate Texts in Mathematics (Springer, New York, 1980)
4. P.G.L. Dirichlet, Recherches sur diverses applications de l'analyse infinitésimale à la théorie des nombres. J. Reine Angew. Math. **19**, 324–369 (1839), https://doi.org/10.1515/crll.1840.21.1
5. K. Girstmair, A "popular" class number formula. Amer. Math. Monthly **101**, 997–1001 (1994)
6. P. Ribenboim, *My Numbers, My Friends: Popular Lectures on Number Theory* (Springer, New York, 2000)
7. N.J.A. Sloane, The on-line encyclopedia of integer sequences (2015), http://oeis.org/
8. I.M. Vinogradov, *Elements of Number Theory* (trans. S. Kravetz), 5th revised ed. (Dover, New York, 1954)
9. A.L. Whiteman, Theorems on quadratic residues. Math. Mag. **23**, 71–74 (1949)

A Second Wave of Expanders in Finite Fields

Brendan Murphy and Giorgis Petridis

Abstract This is an expository survey on recent sum-product results in finite fields. We present a number of sum-product or "expander" results that say that if $|A| > p^{2/3}$, then some set determined by sums and product of elements of A is nearly as large as possible, and if $|A| < p^{2/3}$, then the set in question is significantly larger than A. These results are based on a point-plane incidence bound of Rudnev and are quantitatively stronger than a wave of earlier results following Bourgain, Katz, and Tao's breakthrough sum-product result. In addition, we present two geometric results: an incidence bound due to Stevens and de Zeeuw, and bound on collinear triples, and an example of an expander that breaks the threshold of $p^{2/3}$ required by the other results. We have simplified proofs wherever possible and hope that this survey may serve as a compact guide to recent advances in arithmetic combinatorics over finite fields. We do not claim originality for any of the results.

Keywords Sum product problem · Incidence bounds · Collinear triples
Arithmetic combinatorics

1 Introduction

This is an expository survey of recent results related to the *sum-product problem over finite fields*. Roughly speaking, the sum-product problem is to show that a finite subset of a field cannot have both additive and multiplicative structure (unless it is essentially a subfield). For instance, if p is prime and A is a subset of the field $\mathbb{F}_p$ with p elements, then we would expect the set

B. Murphy (✉)
University of Bristol, School of Mathematics and Heilbronn Institute
for Mathematical Research, Bristol, England
e-mail: brendan.murphy@bristol.ac.uk

G. Petridis
Department of Mathematics, University of Georgia, Athens, USA
e-mail: giorgis@cantab.net

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_15

$$A + AA := \{a_1 + a_2a_3 \colon a_1, a_2, a_3 \in A\}$$

to be much larger than $|A|$, since $\mathbb{F}_p$ has no non-trivial subfields.

In general, we will consider polynomials $f \in \mathbb{Z}[x_1, \ldots, x_n]$ and ask if there is a $\delta > 0$ such that

$$|f(A, \ldots, A)| \geq |A|^{1+\delta}$$

for all "small" subsets A of $\mathbb{F}_p$. We will call such polynomials *expanding polynomials* or *expanders*.

Explicit examples of expanding polynomials were first given in characteristic zero [7, 8]. The arguments employed here typically use topological properties of the underlying field—for instance, the order of the integers or reals. Over finite fields, such as $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, such properties are unavailable, and expansion results are more difficult to prove. Using Fourier analysis in $\mathbb{F}_p$, Garaev [10] showed that for $A \subseteq \mathbb{F}_p$

$$\max(|A + A|, |AA|) \gg \min\left(\sqrt{p|A|}, \frac{|A|^2}{p^{1/2}}\right), \tag{1}$$

which is optimal for $|A| > p^{2/3}$ and trivial for $|A| < p^{1/2}$.

Bourgain et al. [3] proved the first non-trivial sum-product estimate for "small" subsets of finite fields. They showed that if A is a subset of the prime field $\mathbb{F}_p$ such that $p^\alpha < |A| < p^{1-\alpha}$ for some $\alpha > 0$, then there is some $\epsilon > 0$ depending on α such that

$$\max(|A + A|, |AA|) \gg |A|^{1+\epsilon}. \tag{2}$$

The bounds on $|A|$ rule out the possibility that $|A \cap \mathbb{F}| \gg p^{-\alpha} \max(|A|, |F|)$ for any subfield $\mathbb{F}$ of $\mathbb{F}_p$ (i.e., for $\mathbb{F} = \{0\}, \mathbb{F} = \mathbb{F}_p$); in general, it is true that there is a non-trivial sum-product estimate for $A \subseteq \mathbb{F}_q$ as long as A is not "roughly equivalent" to a subfield. The estimate (30) still holds when the lower bound on $|A|$ is dropped—this is due to Glibichuk and Konyagin [11].

Garaev [9] found the first explicit value of ϵ, which was then improved by several authors [2, 14, 26], finally resulting in the lower bound

$$\max(|A + A|, |AA|) \gg |A|^{1+1/11}(\log |A|)^{-4/11}.$$

The method behind these early sum-product results for finite fields is called the *pivot method*. The pivot method is essentially algebraic; it is a flexible method, but it is quantitatively inefficient.

Recently, a new *geometric* method for proving sum-product results in finite fields was discovered. This geometric method is based on a point-plane incidence bound of Rudnev [27]. Rudnev's bound has ushered in a new wave of expander results.

For instance, Roche-Newton et al. [25] applied Rudnev's bound to show that if A is a subset of $\mathbb{F}_p$ with $|A| < p^{2/3}$, then

$$\max(|A + A|, |AA|) \gg |A|^{1+1/5}.$$

Even more impressive is their lower bound for the mixed sum-product set $A + AA$: For $A \subseteq \mathbb{F}_q$

$$|A + AA| \gg \min(|A|^{3/2}, p) \tag{3}$$

where again p is the characteristic of the field $\mathbb{F}_q$. For $|A| < p^{2/3}$, this bound matches what can be proved directly by the Szemerédi-Trotter incidence bound over $\mathbb{R}$, namely

$$|A + AA| \gg |A|^{3/2} \tag{4}$$

for all finite subsets $A \subseteq \mathbb{R}$. The bound (4) has only been slightly improved over $\mathbb{R}$ [29]; thus, Rudnev's point-plane incidence bound allows us to prove expander results that nearly match those known over the real numbers.

A number of similar results have followed from Rudnev's point-plane bound. These results are often of the form $|f(A^k)| \gg \min(|A|^{3/2}, p)$ for some polynomial $f \in \mathbb{Z}[x_1, \dots, x_k]$; thus if $|A| > p^{2/3}$, then $|f(A^k)| \gg p$. We say that these results are at the "$p^{2/3}$ threshold":

1. $|AA + AA| \gg \min(p, |A|^{3/2})$ (Rudnev [27])
2. $|(A - A)(A - A)| \gg p$ if $|A| > p^{2/3}$ (Bennett et al. [1], see also [13])
3. $|(A - A)^2 + (A - A)^2| \gg \min(p, |A|^{3/2})$ (Petridis [20], see also [5])
4. $|A + AA| \gg \min(p, |A|^{3/2})$ (Roche-Newton et al. [25])
5. $|A(A + A)| \gg \min(p, |A|^{3/2})$ (Aksoy-Yazici et al. [32])

In the last section of the paper, we present an expander result below the $p^{2/3}$ threshold. Namely, that if $|A| > p^{5/8}$, then

$$|(A - A)(A - A)| \gg p. \tag{5}$$

This result is due to the second author [21]. As an expander result, this says that the polynomial $f(x, y, z, w) = (x - y)(z - w)$ satisfies $|f(A^4)| \gg p$ whenever $|A| > p^{5/8}$.

In this survey, we take Rudnev's point-plane incidence bound as a black box and use it to prove a variety of sum-product estimates. We have tried to present the cleanest possible proofs and have chosen results that illustrate how to apply the point-plane incidence bound in a variety of situations. We do not claim originality for any of the results.

In Sect. 2, we introduce Rudnev's point-plane incidence bound and use it prove that $|A + AA| \gg \min(p, |A|^{3/2})$. This method of proof will be a model for many later arguments. The section ends with a generalization of the method, due to [32], phrased in terms of certain "energies" $E(Q; A)$ or $E(L, A)$, where $A \subseteq \mathbb{F}_p$, $Q \subseteq \mathbb{F}_p^2$, and L is a collection of lines in $\mathbb{F}_p^2$.

This generalized argument will be applied in Sect. 3 to prove two further expander results and in Sect. 4 to prove two geometric results: an incidence bound due to Stevens and de Zeeuw, and a bound on "collinear triples" due to Aksoy-Yazici et al. [32].

The final section of paper contains a proof of the expansion result (5), which seems to be the first such result below the $p^{2/3}$ threshold.

2 A Geometric Approach to Sum-Product Problems in Finite Fields

In this section, we present a proof of (3) based on Rudnev's point-plane incidence bound, which will serve as a prototype for further applications. We then generalize the method of proof; this generalized formulation will be applied to a variety of applications in the remaining sections.

2.1 Rudnev's Point-Plane Incidence Bound

Rudnev's incidence bound is the following.

Theorem 1 (Rudnev [27]) *Let $\mathbb{F}$ denote a field, and let p denote the characteristic of $\mathbb{F}$. Let P be a set of points in $\mathbb{F}^3$, and let Π be a set of planes in $\mathbb{F}^3$ with $|P| \leq |\Pi|$. If $p > 0$, assume that $|P| \ll p^2$. Let k denote the maximum number of points of P contained in a line. Then,*

$$I(P, \Pi) \ll |P|^{1/2}|\Pi| + k|P|.$$

Theorem 1 is strongest when $|P| = |\Pi|$. See [6] for a short proof of Theorem 1, due to de Zeeuw.

For convenience, we combine Theorem 1 with an incidence bound for large subsets of $\mathbb{F}_p^2$.

Corollary 1 *Let p be an odd prime, let P be a collection of points in $\mathbb{F}_p^3$, and let Π be a collection of planes in $\mathbb{F}_p^3$.*

Suppose that $|P| = |\Pi| = N$ and that at most k points of P are collinear. Then, the number of point-plane incidences satisfies

$$I(P, \Pi) \ll \frac{N^2}{p} + N^{3/2} + kN.$$

The advantage of Corollary 1 over Theorem 1 is that we do not need to bound the size of the point set and the collection of planes before applying the bound.

Proof By [16] (see also [12, 15, 31]), we have

$$I(P, \Pi) \leq \frac{|P||\Pi|}{p} + p\sqrt{|P||\Pi|} = \frac{N^2}{p} + pN.$$

Thus if $N > p^2$, then

$$I(P, \Pi) \ll \frac{N^2}{p}.$$

On the other hand, if $N < p^2$, then by Theorem 1 we have

$$I(P, \Pi) \ll |P|^{1/2}|\Pi| + k|P| = N^{3/2} + kN.$$

2.2 A Lower Bound for $|A + AA|$

In this section, we prove the following theorem, due to Roche-Newton et al. [25].

Theorem 2 *For all subsets A of $\mathbb{F}_p$, we have*

$$|A + AA| \gg \min(p, |A|^{3/2}).$$

The proof of Theorem 2 will serve as a model for the rest of the results in this section.

Proof First, we apply Cauchy-Schwarz. Let

$$r_{A+AA}(x) = |\{(a, b, c) \in A^3 \colon a + bc = x\}|.$$

The support of r_{A+AA} is $|A + AA|$ and

$$\sum_x r_{A+AA}(x) = |A|^3,$$

and thus by Cauchy-Schwarz

$$|A|^6 = \left(\sum_x r_{A+AA}(x)\right)^2 \leq |A + AA| \sum_x r^2_{A+AA}(x).$$

To show that

$$|A + AA| \gg \min(p, |A|^{3/2})$$

it suffices to show that

$$\sum_x r^2_{A+AA}(x) \ll \max\left(\frac{|A|^6}{p}, |A|^{9/2}\right).$$

Next, we reduce the problem to a point-plane incidence problem. The second moment of $r_{A+AA}(x)$ counts the number of solutions to the equation

$$a + bc = a' + b'c' \tag{6}$$

with a, b, c, a', b', c' in A.

To bound the number of solutions to this equation, we will realize each solution as an incidence between a certain point and a certain plane. Let $\pi_{a,b,c'}$ denote the set of points (x, y, z) such that

$$a = x - by + c'z.$$

The point $(x, y, z) = (a', c, b')$ is incident to the plane $\pi_{a,b,c'}$ precisely when (6) is satisfied: If

$$a = a' - bc + c'b',$$

then

$$a + bc = a' + b'c'.$$

Finally, we apply Rudnev's point-plane incidence bound, in the form of Corollary 1. Let $P = \{(a', c, b') \in A^3\}$ and let $\Pi = \{\pi_{a,b,c'} : (a, b, c') \in A^3\}$. Then, $|P| = |\Pi| = |A|^3$. Thus by Corollary 1, we have

$$I(P, \Pi) \ll \frac{|A|^6}{p} + |A|^{9/2} + k|A|^3.$$

This yields the desired upper bound on the second moment of $r_{A+AA}(x)$, provided that the number k of collinear points of $P = A \times A \times A$ is not too large.

It is not hard to show that $k \leq |A|$: If ℓ is parallel to the x-axis, then $|P \cap \ell| \leq |A|$, while if ℓ is not parallel to the x-axis, then ℓ may be parameterized in terms of y or z, which again implies that $|P \cap \ell| \leq |A|$.

Since $k|A|^3 \leq |A|^4 \leq |A|^{9/2}$, we have

$$\begin{aligned}\sum_x r^2_{A+AA}(x) = I(P, \Pi) &\ll \frac{|A|^6}{p} + |A|^{9/2} + k|A|^3 \\ &\ll \max\left(\frac{|A|^6}{p}, |A|^{9/2}\right),\end{aligned}$$

as desired.

2.3 Generalizing the Method

In this section, we will generalize the method used to count solutions to (6). This generalization first appeared in [32]; below we present simplification of the original argument.

In order to form the set of points and planes associated with Eq. (6)

$$a + bc = a' + b'c'$$

it was essential that (a, c) was independent from b and (a', c') was independent from b'. While we also knew that a and c were independent, we do not make use of this in forming the points and planes.

Given a set of pairs $Q \subseteq \mathbb{F}_p^2$ and a set $A \subseteq \mathbb{F}_p$, let $E(Q; A)$ denote the number of solutions to

$$ma + b = m'a' + b' \tag{7}$$

with (m, b), (m', b') in Q and a, a' in A.

Theorem 3

$$E(Q; A) \ll \frac{|Q|^2|A|^2}{p} + (|Q||A|)^{3/2} + k|Q||A|,$$

where

$$k \leq \max\left(|A|, \max_{\ell \text{ line in } \mathbb{F}^2} |Q \cap \ell|\right).$$

Proof For each (m, b) in Q and a in A, form a plane

$$\pi_{(m,b),a'} = \{(x, y, z) \in \mathbb{F}_q^3 : mx + b = ya' + z\}.$$

Eq. (7) holds if and only if $(a, m', b') \in \pi_{(m,b),a'}$.

If we let $P = A \times Q$ and let Π denote the set of all planes $\pi_{(m,b),a'}$ with (m, b) in Q and a' in A. Then, $|P| = |\Pi|$, so we have

$$I(P, \Pi) \ll \frac{|P|^2}{p} + |P|^{3/2} + k|P|.$$

To bound k, we argue as before: If the x-coordinate of ℓ is not constant, then $|P \cap \ell| \leq |A|$, since we may parameterize ℓ in terms of x, and $P = A \times Q$. If the x-coordinate of ℓ is constant (say equal to a_0), then

$$|P \cap \ell| \leq |\{a_0\} \times Q \cap \ell| \leq \max_{\ell \text{ line in } \mathbb{F}^2} |Q \cap \ell|.$$

2.4 A Bound for the Energy of Affine Transformations Acting on the Line

In [32], the points in Q were associated with lines by duality. There is a natural interpretation of this dual quantity; however, the proof is more convoluted. Now that we have the bound for (7) in hand, we can give the dual version quite easily.

To each point (m, b) in Q, we associate an affine transformation $\ell_{m,b}$ defined by $\ell_{m,b}(x) = mx + b$. We let L_Q denote the set of all $\ell_{m,b}$ with (m, b) in Q. With this notation, equation (7) counts the number of solutions to

$$\ell(a) = \ell'(a') \tag{8}$$

with ℓ, ℓ' in L_Q and a, a' in A. We use $E(L, A)$ to denote the number of solutions to (8).

Corollary 2 *Let L be a set of lines in* $\mathbb{F}_p^2$, *and let A be a subset of* $\mathbb{F}_p$. *Let* κ *denote the size of the largest pencil of lines in L; that is,* κ *is maximum size of a subset* $L' \subseteq L$ *such that all of the lines of* L' *are parallel or pass through a common point.*

Then,

$$E(L, A) \ll \frac{|L|^2|A|^2}{p} + (|L||A|)^{3/2} + k|L||A|,$$

where $k \leq \max(|A|, \kappa)$.

Proof Let Q be such that $L = L_Q$. Then,

$$E(L, A) = E(Q; A)$$

and k is the maximum of $|A|$ and the maximum number of points of Q lying on a line, which is precisely maximum number of lines in a pencil.

The quantity $E(L, A)$, which is the number of solutions to

$$\ell(a) = \ell'(a') \qquad \ell, \ell' \in L, a, a' \in A,$$

is analogous to the *multiplicative energy* $E^{\times}(B, A)$ of a set B and a set A, which is the number of solutions to

$$ba = b'a' \qquad b, b' \in B, a, a' \in A.$$

3 Expansion Results at the $p^{2/3}$ Threshold

3.1 A Lower Bound for $|A(A + A)|$

Theorem 4 *For any subset A of* $\mathbb{F}_p$, *we have*

$$|A(A + A)| \gg \min(p, |A|^{3/2}).$$

Proof Without loss of generality, suppose that A does not contain 0.

By Cauchy-Schwarz, we have

$$|A|^6 \le |A(A+A)|\,|\{(a,b,c,a',b',c') \in A^6 \colon a(b+c) = a'(b'+c')\}|. \tag{9}$$

We wish to bound the number of solutions to

$$a(b+c) = a'(b'+c') \tag{10}$$

with $a, \ldots, c'$ in A.

Since we can write $a(b+c) = ab + ac$, if we let $Q = \{(a, ac) \colon a, c \in A\}$, then the number of solutions to (10) is $E(Q; A)$. The map $(a, c) \mapsto (a, ac)$ is injective, as long as $a \neq 0$, so $|Q| = |A|^2$. At most $|A|$ elements of Q lie on a single line, so by Theorem 3, the number of solutions to (10) is

$$|\{(a,b,c,a',b',c') \in A^6 \colon a(b+c) = a'(b'+c')\}| \ll \frac{|A|^6}{p} + |A|^{9/2}.$$

Combining this bound with (9) yields the desired lower bound on $|A(A+A)|$.

Note 1 The set of points $Q = \{(a, ac) \colon a, c \in A\}$ is projectively equivalent to $A \times A$, which immediately implies that $|Q \cap \ell| \le |A|$ for any line ℓ. In general, if Q is projectively equivalent to $B \times C$, then we have $k \le \max(|A|, |B|, |C|)$.

The following example, suggested by Roche-Newton, can be proved by a similar argument.

Exercise 1 Let

$$A(AA+1) = \{a(bc+1) \colon a, b, c \in A\}.$$

Show that

$$|A(AA+1)| \gg \min(p, |A|^{3/2}).$$

3.2 A Lower Bound for $|(A-A)^2 + (A-A)^2|$

In this section, we show that there is a point (u, v) in $A \times A$ such that

$$|(A-u)^2 + (A-v)^2| \gg \min(p, |A|^{3/2}). \tag{11}$$

This result is due to the second author [20].

Geometrically, Eq. (11) says that the product set $P = A \times A$ determines $\gg \min(p, |P|^{3/4})$ distances to the point $(u, v) \in P$.

Proof To prove a lower bound for $|(A-u)^2 + (A-v)^2|$, we will bound the number of solutions to

$$(a-u)^2+(b-v)^2=(c-u)^2+(d-v)^2 \qquad a,b,c,d,u,v \in A. \tag{12}$$

Then, we will pigeonhole over u and v and apply a Cauchy-Schwarz energy-type argument.

To bound the number of solutions to (12), we rearrange the equation

$$(a-u)^2-(c-u)^2=(d-v)^2-(b-v)^2$$

and simplify

$$a^2-c^2-2(a-c)u=d^2-b^2-2(d-b)v. \tag{13}$$

Equation (13) is linear in u, and u is independent from a, c, similarly for v, b, d, so we might hope to apply Theorem 3.

Let

$$Q=\{(-2(a-c),a^2-c^2)\colon a,c\in A\}.$$

Then, the number of solutions to (13) is $E(Q;A)$.

Note that $|Q|=|A|^2$, since the map

$$(a,c)\mapsto(-2(a-c),a^2-c^2)$$

is invertible.

Further, at most $2|A|$ points of Q are contained in a single line, since for fixed α,β,γ, the number of solutions to

$$\alpha[-2(a-c)]+\beta(a^2-c^2)=\gamma$$

is bounded by the maximum number of pairs (a,c) of $A\times A$ that are contained in the quadratic curve

$$-\alpha(x-y)+\beta(x^2-y^2)=\gamma.$$

Given any x, there are at most two solutions for y.

Thus by Theorem 3, the number of solutions to (12) is at most

$$E(Q;A)\ll\frac{|A|^6}{p}+(|A|^3)^{3/2}+2|A|^3\ll\frac{|A|^6}{p}+|A|^{9/2}.$$

By the pigeonhole principle, it follows that there is a pair (u,v) in $A\times A$ such that the number of solutions to

$$(a-u)^2+(b-v)^2=(c-u)^2+(d-v)^2 \qquad a,b,c,d\in A$$

is at most $O(|A|^4/p+|A|^{7/2})$.

By Cauchy-Schwarz, we have

$$|A|^4 \ll |(A-u)^2 + (A-v)^2| \cdot \max(|A|^4/p, |A|^{7/2}),$$

which implies the desired lower bound.

See [23] for a generalization of this result to higher dimensions, as well as a general result on expanding quadratic polynomials.

4 Incidence Results for Points and Lines in $\mathbb{F}_p^2$

4.1 An Incidence Bound for Cartesian Product Point Sets $P = A \times B$

The following incidence bound is due to Stevens and de Zeeuw [30].

Theorem 5 *Let A and B be subsets of $\mathbb{F}_p$ with $|A| \leq |B|$. If $P = A \times B$ and L is a set of lines in $\mathbb{F}_p^2$, then*

$$I(P, L) \ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + |P|^{2/3}|L|^{2/3} + |L|.$$

In particular, if $|A||L| \leq p^2$, then

$$I(P, L) \ll |A|^{3/2}|B|^{1/2}|L|^{3/4} + |P| + |L|. \tag{14}$$

Before we prove Theorem 5, we prove a lemma that gives the correct leading terms.

Lemma 1 *For $P = A \times B$, as above, and any set of lines L, we have*

$$I(P, L) \leq |B|^{1/2} E(L, A)^{1/2}.$$

Thus,

$$I(P, L) \ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + k(|A||B||L|)^{1/2}.$$

A priori, we have no control over k, so Theorem 5 does not follow immediately from Lemma 1.

Proof (*Proof of Lemma* 1) We have

$$I(P,L) = |\{(a,b,\ell) \in A \times B \times L \colon b = \ell(a)\}| = \sum_{b \in B} |\{(a,\ell) \in A \times L \colon b = \ell(a)\}|.$$

Thus by Cauchy-Schwarz,

$$I(P,L) \leq |B|^{1/2} \left(\sum_b |\{(a,\ell) \in A \times L \colon b = \ell(a)\}|^2 \right)^{1/2}.$$

The sum over all b in $\mathbb{F}_p$ is equal to $E(L,A)$; that is, it is equal to the number of solutions to

$$\ell(a) = \ell'(a')$$

with ℓ, ℓ' in L and a, a' in A. Thus,

$$I(P,L) \leq |B|^{1/2} E(L,A)^{1/2}.$$

To apply Lemma 1, we need to make sure that not too many lines of L lie in a pencil.

Proof (*Proof of Theorem* 5) Let $k > 0$ be a parameter that we will choose later.

We begin by pruning large pencils of lines from L. Suppose that L contains a pencil P_1 with more than k lines. This pencil contributes at most $|A||B| + |P_1|$ incidences. Let $L_1 = L \setminus P_1$. We continue pruning pencils until we reach a set of lines L' that contains no pencils of size greater than k. This process takes at most $|L|/k$ steps; hence, the lines removed contribute at most

$$\sum_{i=1}^{|L|/k} (|A||B| + |P_i|) = \frac{|A||B||L|}{k} + |L|$$

incidences.

By Lemma 1 and Corollary 2, we have

$$\begin{aligned} I(P,L') \leq |B|^{1/2} E(L,A)^{1/2} &\ll |B|^{1/2} \left(\frac{|L|^2|A|^2}{p} + (|L||A|)^{3/2} + k|L||A| \right)^{1/2} \\ &\ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + \sqrt{k|A||B||L|}. \end{aligned}$$

Since $I(P,L) = I(P,L') + I(P, L \setminus L')$, we have

$$I(P,L) \ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + \sqrt{k|A||B||L|} + \frac{|A||B||L|}{k} + |L|.$$

Setting $k = (|A||B||L|)^{1/3}$ yields

$$I(P, L) \ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + |P|^{2/3}|L|^{2/3} + |L|.$$

To prove (14) note that we have

$$I(A \times B, L) \leq |A||L| + |A||B|,$$

since vertical lines contribute at most $|A||B|$ incidences and non-vertical lines can be incident to at most $|A|$ points of $A \times B$. Thus if $|A||L| \leq |B|^2$, we have

$$I(A \times B, L) \leq |A||L| + |P| \leq |A|^{3/4}|B|^{1/2}|L|^{3/4} + |P|.$$

On the other hand, if $|B|^2 \leq |A||L| \leq p^2$, then

$$\begin{aligned} I(A \times B, L) &\ll \frac{|A||B|^{1/2}|L|}{p^{1/2}} + |A|^{3/4}|B|^{1/2}|L|^{3/4} + |P|^{2/3}|L|^{2/3} + |L| \\ &\ll |A|^{3/4}|B|^{1/2}|L|^{3/4} + |L|. \end{aligned}$$

Exercise 2 Theorem 5 can be used to prove a number of sum-product results using Elekes' method [7].

1. Use the lines $\ell_{a,b}(t) = a(t + b)$ with $a, b \in A$ and the point set $P = A \times A(A + A)$ to show that
$$|A(A + A)| \gg \min(p, |A|^{3/2}).$$
2. Use the lines $\ell_{a,b}(t) = at + b$ with $a, b \in A$ and the point set $P = A \times (A + AA)$ to show that
$$|A + AA| \gg \min(p, |A|^{3/2}).$$
3. Use the lines $\ell_{a,b}(t) = t/a + b$ or $\ell_{a,b}(t) = a(t - b)$ and a point set of the form $P = AA \times (A + A)$ or $P = (A + A) \times AA$ to show that
$$\max(|A + A|, |AA|) \gg \min(p^{1/3}|A|^{2/3}, |A|^{6/5}).$$

The last part of the exercise implies that if $|A| \leq p^{5/8}$, then

$$\max(|A + A|, |AA|) \gg |A|^{6/5}.$$

Since $|A|^2/p^{1/2} > p^{1/3}|A|^{2/3}$ when $|A| > p^{5/8}$, the best known sum-product results in $\mathbb{F}_p$ can be summarized as

$$\max(|A + A|, |AA|) \gg \min(\sqrt{p|A|}, |A|^2/p^{1/2}, |A|^{6/5}). \tag{15}$$

In [30], Stevens and de Zeeuw use Theorem 5 in conjunction with a clever induction argument to prove a point-line incidence bound for general point sets $P \subseteq \mathbb{F}_p^2$. Namely, that for any set of lines L in $\mathbb{F}_p^2$ such that $|P|^{7/8} < |L| < |P|^{8/7}$ and $|L|^{13} \ll p^{15}|P|^2$,

$$I(P, L) \ll |P|^{11/15}|L|^{11/15}.$$

Further applications of this bound and Theorem 5 may be found in [30].

4.2 A Bound for the Number of Collinear Triples in $P = A \times A$

Given a subset A of $\mathbb{F}_p$, let $T(A)$ denote the number of *collinear triples* of points in $P = A \times A$.

For any set A, we have $T(A) \ll |A|^5$, which we may see as follows. Three points (a, a'), (b, b'), (c, c') in $P = A \times A$ are collinear if

$$\det \begin{pmatrix} 1 & 1 & 1 \\ a & b & c \\ a' & b' & c' \end{pmatrix} = 0. \tag{16}$$

Evaluating the determinant yields the equation

$$(b - a)(c' - a') = (b' - a')(c - a). \tag{17}$$

Since we have six variables in $|A|^6$ and one equation, we have $\ll |A|^5$ solutions.

Recall that to find lower bounds for $|A + AA|$ and $|A(A + A)|$, we found upper bounds for six-variable energy-type equations. It turns out that (17) can be bounded in a similar way, leading to the following bound, due to [32], see also [19, 22].

Theorem 6 *Let A be a subset of $\mathbb{F}_p$. If $|A| \ll p^{2/3}$, then*

$$T(A) \ll \frac{|A|^6}{p} + |A|^{9/2}.$$

Proof If $a, b \neq c$ and $a', b' \neq c$, then Eq. (17) reduces to

$$\frac{b - a}{c - a} = \frac{b' - a'}{c' - a'}. \tag{18}$$

Since the number of collinear triples where $a = c$, $b = c$, $a' = c'$, or $b' = c'$ is $O(|A|^4)$, we have

$$T(A) = \left|\left\{(a, \ldots, c') \in A^6 \colon \frac{b-a}{c-a} = \frac{b'-a'}{c'-a'} \neq 0, \infty\right\}\right| + O(|A|^4). \tag{19}$$

Thus to bound $T(A)$, it suffices to count the number of solutions to (18) with a, b, c, a', b', c' in A. We apply Theorem 3 to (18).

Let

$$Q = \{(1/(c-a), -a/(c-a)) \colon a, c \in A\}.$$

By (19) and our definition of Q, it follows that $T(A) = E(Q; A) + O(|A|^4)$. The proposition will follow from Theorem 3 if we can show that $|Q| = |A|^2$ and $k \leq |A|$, since then

$$E(Q; A) \ll \frac{|A|^6}{p} + (|A|^3)^{3/2} + |A|^4 \ll \frac{|A|^6}{p} + |A|^{9/2}.$$

First $|Q| = |A|^2$, since every $(x, y) \in Q$ corresponds to a unique pair (c, a) in $A \times A$, where

$$a = -\frac{y}{x} \quad \text{and} \quad c = \frac{1}{x} - \frac{y}{x}.$$

Second, to show that $k \leq |A|$ we must show that at most k points of Q are collinear. Consider the linear equation $\alpha x + \beta y = \gamma$ with α, β, and γ fixed; suppose one of α, β equals 1. Plugging in $x = 1/(c-a)$ and $y = -a/(c-a)$ yields the equation

$$\alpha - \beta a = \gamma(c-a),$$

which has at most $|A|$ solutions (a, c), as required.

The number of collinear triples $T(A)$ can be expressed in terms of the multiplicative energy of shifts of A:

$$T(A) = \sum_{a, a' \in A} E^\times(A - a, A - a'). \tag{20}$$

This is easy to see from (17). We first learned of Eq. (20) in [24], and the proof there inspired the proof of Theorem 6.

The following easy corollary was used in [32] to prove an incidence bound for points and lines (which has since been subsumed by Theorem 5).

Corollary 3 *Let A be a subset of $\mathbb{F}_p$ with $|A| < p^{2/3}$, and let L_k denote the set of lines containing at least k points of $P = A \times A$. If $k > 3$, then*

$$|L_k| \ll \frac{|A|^{9/2}}{k^3}.$$

Proof We have

$$\binom{k}{3}|L_k| \leq \sum_{\ell \in L_k} \binom{|P \cap \ell|}{3} \ll T(A) \ll |A|^{9/2}.$$

Since $k > 3$, we have $\binom{k}{3} \gg k^3$, so the bound follows.

Theorem 5 implies that

$$|L_k| \ll \frac{|A|^5}{k^4} \tag{21}$$

for $k > |A|^{3/2}/p^{1/2}$. In Lemma 4, we show that the same bound actually holds whenever $k > 2|A|^2/p$. The bound (21) is essentially equivalent to the statement that for $|A| < p^{2/3}$, the point set $A \times A$ determines $\ll |A|^5 \log(|A|)$ collinear *quadruples*. Given such a bound for collinear quadruples, we may recover (21) by the same method used to prove Corollary 3. See [19] for further discussion.

5 An Expander Below the $p^{2/3}$ Threshold

In this section, we prove the following theorem due to the second listed author [21]:

Theorem 7 *Let p be a prime, and let A be a subset of $\mathbb{F}_p$. Then, the number of solutions to*

$$(a-b)(c-d) = (a'-b')(c'-d') \quad \textit{with } a, b, c, d, a', b', c', d' \textit{ in } A \tag{22}$$

is $|A|^8/p + O(p^{2/3}|A|^{16/3})$.

Hence if $|A| \gg p^{5/8}$, then the number of solution is $O(|A|^8/p)$, and hence,

$$|(A-A)(A-A)| \gg p.$$

This result is that it says that $|(A-A)(A-A)|$ is nearly as large as possible when $|A|$ is at least $p^{5/8}$, which is lower than the $p^{2/3}$ threshold. Subsequently, Rudnev et al. [28] proved that

$$\left|\left\{\frac{ab-c}{a-d} : a, b, c, d \in A\right\}\right| \gg p$$

whenever $|A| \gg p^{25/42-o(1)}$, which also breaks the $p^{2/3}$ threshold. Recently, the authors, together with Roche-Newton et al. [17], have proved several results that pass the $p^{2/3}$ threshold. For instance,

$$|R[A]| = \left|\left\{\frac{b-a}{c-a} : a, b, c \in A\right\}\right| \gg p$$

whenever $|A| \geq p^{3/5}$, and

$$|R[A]| \gg \frac{|A|^{8/5}}{\log^2(|A|)}$$

whenever $|A| \leq p^{5/12}$.

Proof (Proof of Theorem 7) As before, we use an energy-type argument: Let $r(x) = r_{(A-A)(A-A)}(x)$. Then, $r(x)$ is supported on $(A - A)(A - A)$ and $\sum_x r(x) = |A|^4$, and thus

$$|A|^8 \leq |(A - A)(A - A)| \sum_x r^2(x).$$

The second moment of $r(x)$ counts solutions to Eq. (22).

There are $O(|A|^6)$ solutions where either side of (22) is zero; thus, we have

$$\sum_x r^2(x) = \left|\left\{ \frac{a-b}{a'-b'} = \frac{c-d}{c'-d'} \neq 0, \infty \right\}\right| + O(|A|^6). \tag{23}$$

We can write this quantity as a second moment of a different function, which we will call Q_ξ:

$$Q_\xi := \left|\left\{ (a, b, c, d) \in A^4 : \frac{a-b}{c-d} = \xi \right\}\right|. \tag{24}$$

Then by (23) and (24), we have

$$\sum_x r^2(x) = \sum_{\xi \neq 0} Q_\xi^2 + O(|A|^6). \tag{25}$$

The following lemma provides the necessary bound for the second moment of Q_ξ:

Lemma 2

$$\sum_{\xi \neq 0} Q_\xi^2 \leq \frac{|A|^8}{p} + O(p^{2/3}|A|^{16/3}).$$

We defer the proof of Lemma 2 and finish the proof of Theorem 7.

Combining (25) with Lemma 2 yields

$$\sum_x r^2(x) \leq \frac{|A|^8}{p} + O(|A|^6 + p^{2/3}|A|^{16/3}).$$

Since $|A|^6 \ll p^{2/3}|A|^{16/3}$ for all A, we have

$$\sum_x r^2(x) \leq \frac{|A|^8}{p} + O(p^{2/3}|A|^{16/3}), \tag{26}$$

as claimed.

If $|A| \geq p^{5/8}$, then $\sum_x r^2(x) \ll |A|^8/p$, so $|(A-A)(A-A)| \gg p$.

Now we prove Lemma 2.

Proof (ProofofLemma 2) To begin, we record some basic facts about Q_ξ and introduce a related quantity, E_ξ. For $\xi \neq 0$, we have

$$Q_\xi = |\{(a,b,c,d) \in A^4 \colon a - \xi c = b - \xi d,\ a \neq b, c \neq d\}| = E^+(A, \xi A) - |A|^2. \tag{27}$$

Since

$$\sum_{\xi \neq 0} Q_\xi = |A|^2(|A|-1)^2,$$

we have

$$\sum_{\xi \in X} E^+(A, \xi A) = \sum_{\xi \in X} \left(Q_\xi + |A|^2\right) \leq |A|^4 + |X||A|^2. \tag{28}$$

It follows from (28) that if we set

$$E_\xi = E^+(A, \xi A) - \frac{|A|^4}{p},$$

then

$$\sum_{\xi \neq 0} E_\xi \leq p|A|^2. \tag{29}$$

The quantity E_ξ is useful because it is *nonnegative*: By Cauchy-Schwarz,

$$E^+(A, \xi A) \geq \frac{|A|^4}{|A \pm \xi A|} \geq \frac{|A|^4}{p}.$$

Now we will estimate the second moment of Q_ξ. To begin, we replace one power of Q_ξ by E_ξ and estimate the error:

$$\begin{aligned}
\sum_{\xi \neq 0} Q_\xi^2 &= \sum_{\xi \neq 0} Q_\xi \left(E^+(A, \xi A) - |A|^2\right) \\
&= \sum_{\xi \neq 0} Q_\xi \left(E_\xi + \frac{|A|^4}{p} - |A|^2\right) \\
&\leq \sum_{\xi \neq 0} Q_\xi E_\xi + \frac{|A|^4}{p} \sum_{\xi \neq 0} Q_\xi \\
&\leq \frac{|A|^8}{p} + \sum_{\xi \neq 0} Q_\xi E_\xi.
\end{aligned}$$

Thus by (25),

$$\sum_x r^2(x) = \sum_{\xi \neq 0} Q_\xi^2 + O(|A|^6) \leq \frac{|A|^8}{p} + \sum_{\xi \neq 0} Q_\xi E_\xi + O(|A|^6). \tag{30}$$

Now, to estimate the sum over ξ, we divide into two cases. Let $B_K = \{\xi \neq 0 \colon Q_\xi > |A|^3/K\}$. Then,

$$\sum_{\xi \neq 0} Q_\xi E_\xi \leq \sum_{\xi \in B_K} Q_\xi E_\xi + \frac{|A|^3}{K} \sum_{\xi \neq 0} E_\xi = I + II. \tag{31}$$

We bound second term by (29):

$$II = \frac{|A|^3}{K} \sum_{\xi \neq 0} E_\xi \leq \frac{p|A|^5}{K}. \tag{32}$$

To bound the first term, we use the trivial bound $|Q_\xi| \leq |A|^3$ to find

$$I = \sum_{\xi \in B_K} Q_\xi E_\xi \leq |A|^3 \sum_{\xi \in B_K} E_\xi \leq |A|^3 \sum_{\xi \in B_K} E^+(A, \xi A). \tag{33}$$

To bound this last sum, we use the following lemma, which we will prove in the next section.

Lemma 3 *If $|A| \ll p^{2/3}$, then for any $X \subseteq \mathbb{F}_p$ such that $|X| \leq |A|^3$,*

$$\sum_{\xi \in X} E^+(A, \xi A) \ll |A|^3 |X|^{2/3}.$$

Since

$$\frac{|A|^3}{K} |B_K| < \sum_{\xi \in B_K} Q_\xi \leq |A|^4$$

and $K \leq |A|$, we have

$$|B_K| \leq |A|^2.$$

Thus, we may apply Lemma 3 with $X = B_K$.

By Lemma 3 and (33),

$$I \ll |A|^6 |B_K|^{2/3}. \tag{34}$$

Now we use Lemma 3 again to bound $|B_K|$:

$$\frac{|A|^3}{K}|B_K| \leq \sum_{\xi \in B_K} E^+(A, \xi A) \ll |A|^3|B_K|^{2/3},$$

and hence, $|B_K| \ll K^3$.

Combining the bounds for I and II with the bound $|B_K| \ll K^3$, we have

$$\sum_{\xi \neq 0} Q_\xi E_\xi \ll K^2|A|^6 + \frac{p|A|^5}{K}.$$

To balance the terms on the right-hand side of the previous equation, we set $K = (p/|A|)^{1/3}$:

$$\sum_{\xi \neq 0} Q_\xi E_\xi \ll p^{2/3}|A|^{16/3}. \tag{35}$$

This completes the proof of Lemma 2, pending the proof of Lemma 3.

***Proof of Lemma* 3**

Recall that Lemma 3 states that if $|A| \ll p^{2/3}$, then for any set $X \subseteq \mathbb{F}_p$ such that $|X| \leq |A|^3$, we have

$$\sum_{\xi \in X} E^+(A, \xi A) \ll |A|^3|X|^{2/3}.$$

This is an explicit version of Bourgain's Theorem C from [4]. Similar results were proved over $\mathbb{R}$ in [18] by the Szemerédi-Trotter incidence bound. We use the same approach as [18], but we use the following lemma in place of the Szemerédi-Trotter theorem.

Lemma 4 *Let A be a subset of $\mathbb{F}_p$, and let L_t denote the set of lines in $\mathbb{F}_P^2$ that contain at least t points of $P = A \times A$. If $t > \min(2|A|^2/p, 1)$, then*

$$|L_t| \ll \frac{|A|^5}{t^4}.$$

The proof of Lemma 4 requires the following bound, which is implicit in the work of Bourgain et al. [3] and appears explicitly in [16]:

$$\sum_{\text{all lines } \ell} \left(i(\ell) - \frac{|A|^2}{p} \right)^2 \leq p|A|^2, \tag{36}$$

where $i(\ell) = |(A \times A) \cap \ell|$.

Proof (Proof of Lemma 4) For a line ℓ in $\mathbb{F}_p^2$, let $i(\ell) = |P \cap \ell|$, where $P = A \times A$. Thus if $\ell \in L_t$, then $i(\ell) \geq t$.

Since $t > 2|A|^2/p$, we have

$$i(\ell) - \frac{|A|^2}{p} \geq \frac{t}{2}$$

for all ℓ in L_t. Thus,

$$\frac{|L_t|t^2}{4} \leq \sum_{\ell \in L_t} \left(i(\ell) - \frac{|A|^2}{p} \right)^2.$$

On the other hand, by Eq. (36) the right-hand side of the previous equation is at most $p|A|^2$, so

$$|L_t| \ll \frac{p|A|^2}{t^2}.$$

Now we consider two cases. If $t \leq c|A|^{3/2}/p^{1/2}$, we have

$$|L_t| \leq \frac{c^2|A|^3}{pt^2}|L_t| \ll \frac{|A|^5}{t^4}.$$

If $t \geq c|A|^{3/2}/p^{1/2}$, then we will apply Theorem 5. Since

$$t|L_t| \leq I(P, L_t),$$

by Theorem 5, we have

$$t|L_t| \ll \frac{|A|^{3/2}|L_t|}{p^{1/2}} + |A|^{5/4}|L_t|^{3/4} + |A|^2.$$

Since $t \geq c|A|^{3/2}/p^{1/2}$, if c is sufficiently large (depending on the implicit constants in Theorem 5), we have

$$t|L_t| \ll |A|^{5/4}|L_t|^{3/4} + |A|^2,$$

and hence,

$$|L_t| \ll \frac{|A|^5}{t^4} + \frac{|A|^2}{t} \ll \frac{|A|^5}{t^4}.$$

(The last inequality follows because $t \leq |A|$.)

Finally, note that if $1 < t \ll 1$, then $|L_t| \ll |A|^5/t^4$ is trivial, since $|L_t| \leq |A|^4$.

Now we proceed to the proof of the main result of this section.

Proof (Proof of Lemma 3) To show that

$$S := \sum_{\xi \in X} E^+(A, \xi A) \ll |A|^3|X|^{2/3},$$

we first write

$$S = \sum_{\xi \in X} \sum_{y} r^2_{A+\xi A}(y).$$

Let Z_j denote the set of pairs $\{(\xi, y) : r_{A+\xi A}(y) > \Delta 2^j\}$. Then,

$$S \ll \Delta |X||A|^2 + \sum_{j \geq 0} |Z_j|(\Delta 2^j)^2. \tag{37}$$

On the other hand, for each pair (ξ, y) in Z_j, we may associate the line $\ell_{\xi,y} = \{(a, b) : a + \xi b = y\}$. Since the line $\ell_{\xi,y}$ contains at least $\Delta 2^j$ points of $A \times A$, by Lemma 3 we have

$$|Z_j| \leq |L_j| \ll \frac{|A|^5}{(\Delta 2^j)^4}, \tag{38}$$

whenever $\Delta 2^j \geq \min(2|A|^2/p, 1)$. (We do not need strict inequality because it is included in the definition of Z_j.)

Assume for now that $\Delta \geq \min(2|A|^2/p, 1)$; at the end of the argument, we will prove that our choice of Δ satisfies this condition whenever $|A| \ll p^{2/3}$. By (37) and (38), we have

$$S \ll \Delta |X||A|^2 + \sum_{j \geq 0} (\Delta 2^j)^2 \frac{|A|^5}{(\Delta 2^j)^4},$$

Thus,

$$S \ll \Delta |X||A|^2 + \frac{|A|^5}{\Delta^2}.$$

Choosing $\Delta = |A|/|X|^{1/3}$ yields

$$S \ll |A|^3 |X|^{2/3},$$

as desired.

Now we will check that $\Delta = |A|/|X|^{1/3}$ is at least $2|A|^2/p$ whenever $|A| \ll p^{2/3}$:

$$\Delta = \frac{|A|}{|X|^{1/3}} \geq \frac{2|A|^2}{p} \iff |X| \ll \frac{p^3}{|A|^3}.$$

On the other hand, if $|A| \ll p^{2/3}$, then $p^3/|A|^3 \gg p \geq |X|$. Finally, $|X| \leq |A|^3$ implies $\Delta \geq 1$.

Acknowledgements We thank Olly Roche-Newton, Misha Rudnev, and Sophie Stevens for several helpful suggestions. We would also like to thank Mel Nathanson for inviting us to write this survey for the proceedings of CANT 2015/2016.

References

1. M. Bennett, D. Hart, A. Iosevich, J. Pakianathan, M. Rudnev, Group actions and geometric combinatorics in $\mathbb{F}_q^d$. 11 2013
2. J. Bourgain, M.Z. Garaev, On a variant of sum-product estimates and explicit exponential sum bounds in prime fields. Math. Proc. Camb. Philos. Soc. **146**(1), 1–21 (2009)
3. J. Bourgain, N. Katz, T. Tao, A sum-product estimate in finite fields, and applications. Geom. Funct. Anal. **14**(1), 27–57 (2004)
4. J. Bourgain, Multilinear exponential sums in prime fields under optimal entropy condition on the sources. Geom. Funct. Anal. **18**(5), 1477–1502 (2009)
5. J. Chapman, M. Burak Erdogan, D. Hart, A. Iosevich, D. Koh, *Pinned Distance Sets, k-simplices, Wolff's Exponent in Finite Fields and Sum-Product Estimates* (2009)
6. F. de Zeeuw, A short proof of Rudnev's point-plane incidence bound (2016). arXiv:1612.02719
7. G. Elekes, On the number of sums and products. Acta Arith. **81**(4), 365–367 (1997)
8. P. Erdos, E. Szemerédi, On sums and products of integers. Stud. Pure Math. 213–218 (1983)
9. M.Z. Garaev, An explicit sum-product estimate in $\mathbb{F}_p$. Int. Math. Res. Not. IMRN, (11):Art. ID rnm035, 11 (2007)
10. M.Z. Garaev, The sum-product estimate for large subsets of prime fields. Proc. Am. Math. Soc. **136**(8), 2735–2739 (2008)
11. A.A. Glibichuk, S.V. Konyagin, Additive properties of product sets in fields of prime order, in *Additive Combinatorics*. CRM Proceedings and Lecture Notes, vol. 43 (2007), pp. 279–286
12. D. Hart, A. Iosevich, D. Koh, M. Rudnev, Averages over hyperplanes, sum-product theory in vector spaces over finite fields and the Erdős-Falconer distance conjecture. Trans. Am. Math. Soc. **363**(6), 3255–3275 (2011)
13. D. Hart, A. Iosevich, J. Solymosi, Sum-product estimates in finite fields via Kloosterman sums. Int. Math. Res. Not. IMRN, (5):Art. ID rnm007, 14 (2007)
14. N.H. Katz, C.-Y. Shen, A slight improvement to Garaev's sum product estimate. Proc. Am. Math. Soc. **136**(7), 2499–2504 (2008)
15. B. Lund, S. Saraf, *Incidence Bounds for Block Designs* (2014)
16. B. Murphy, G. Petridis, A point-line incidence identity in finite fields, and applications. Mosc. J. Comb. Number Theory **6**(1), 64–95 (2016)
17. B. Murphy, G. Petridis, O. Roche-Newton, M. Rudnev, I.D. Shkredov, New results on sum-product type growth in positive characteristic (2017)
18. B. Murphy, O. Roche-Newton, I. Shkredov, Variations on the sum-product problem. SIAM J. Discrete Math. **29**(1), 514–540 (2015)
19. G. Petridis, Collinear triples and quadruples for Cartesian products in $\mathbb{F}_p^2$ (2016)
20. G. Petridis, Pinned algebraic distances determined by Cartesian products in $\mathbb{F}_p^2$ (2016)
21. G. Petridis, Products of differences in prime order finite fields (2016)
22. G. Petridis, I.E. Shparlinski, Bounds on trilinear and quadrilinear exponential sums (2016)
23. T. Pham, L.A. Vinh, F. de Zeeuw, Three-variable expanding polynomials and higher-dimensional distinct distances (2016). arXiv:1612.09032
24. O. Roche-Newton, A short proof of a near-optimal cardinality estimate for the product of a sum set (2015)
25. O. Roche-Newton, M. Rudnev, I.D. Shkredov, New sum-product type estimates over finite fields. Adv. Math. **293**, 589–605 (2016)
26. M. Rudnev, An improved sum-product inequality in fields of prime order. Int. Math. Res. Not. IMRN **16**, 3693–3705 (2012)
27. M. Rudnev, On the number of incidences between planes and points in three dimensions. Combinatorica (2014) (To appear)
28. M. Rudnev, I.D. Shkredov, S. Stevens, On the energy variant of the sum-product conjecture (2016)
29. I.D. Shkredov, On a question of A. Balog. Pacific J. Math. **280**(1), 227–240 (2016)
30. S. Stevens, F. de Zeeuw, An improved point-line incidence bound over arbitrary fields (2016)

31. L.A. Vinh, The Szemerédi-trotter type theorem and the sum-product estimate in finite fields. Eur. J. Combin. **32**(8), 1177–1181 (2011)
32. E.A. Yazici, B. Murphy, M. Rudnev, I. Shkredov, Growth estimates in positive characteristic via collisions. Int. Math. Res. Not, IMRN (2016)

Sumsets Contained in Sets of Upper Banach Density 1

Melvyn B. Nathanson

Abstract Every set A of positive integers with upper Banach density 1 contains an infinite sequence of pairwise disjoint subsets $(B_i)_{i=1}^{\infty}$ such that B_i has upper Banach density 1 for all $i \in \mathbf{N}$ and $\sum_{i \in I} B_i \subseteq A$ for every nonempty finite set I of positive integers.

Keywords Sumsets · Banach density · Additive number theory · Ramsay theory

2010 Mathematics Subject Classification: 11A05 · 11B05 · 11B13 · 11B75

1 Upper Banach Density

Let $\mathbf{N}$, $\mathbf{N}_0$, and $\mathbf{Z}$ denote, respectively, the sets of positive integers, nonnegative integers, and integers. Let $|S|$ denote the cardinality of the set S. We define the *interval of integers*

$$[x, y] = \{n \in \mathbf{N} : x \leq n \leq y\}.$$

Let A be a set of positive integers. Let $n \in \mathbf{N}$. For all $u \in \mathbf{N}_0$, we have

$$|A \cap [u, u + n - 1]| \in [0, n]$$

and so

$$f_A(n) = \max_{u \in \mathbf{N}_0} |A \cap [u, u + n - 1]|$$

exists. The *upper Banach density* of A is

Supported in part by a grant from the PSC-CUNY Research Award Program.

M. B. Nathanson (✉)
Department of Mathematics, Lehman College (CUNY),
Bronx, NY 10468, USA
e-mail: melvyn.nathanson@lehman.cuny.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_16

$$\delta(A) = \limsup_{n\to\infty} \frac{f_A(n)}{n}.$$

Let $n_1, n_2 \in \mathbf{N}$. There exists $u_1^* \in \mathbf{N}_0$ such that, with $u_2^* = u_1^* + n_1$,

$$\begin{aligned} f_A(n_1+n_2) &= \left|A \cap [u_1^*, u_1^* + n_1 + n_2 - 1]\right| \\ &= \left|A \cap [u_1^*, u_1^* + n_1 - 1]\right| + \left|A \cap [u_1^* + n_1, u_1^* + n_1 + n_2 - 1]\right| \\ &= \left|A \cap [u_1^*, u_1^* + n_1 - 1]\right| + \left|A \cap [u_2^*, u_2^* + n_2 - 1]\right| \\ &\le f_A(n_1) + f_A(n_2). \end{aligned}$$

It is well known, and proved in the Appendix, that this inequality implies that

$$\delta(A) = \lim_{n\to\infty} \frac{f_A(n)}{n} = \inf_{n\in\mathbf{N}} \frac{f_A(n)}{n}.$$

2 An Erdős Sumset Conjecture

About 40 years ago, Erdős conjectured that if A is a set of positive integers of positive upper Banach density, then there exist infinite sets B and C of positive integers such that $B + C \subseteq A$. This conjecture has not yet been verified or disproved.

The *translation* of the set X by t is the set

$$X + t = \{x + t : x \in X\}.$$

Let B and C be sets of integers. For every integer t, if $B' = B + t$ and $C' = C - t$, then

$$B' + C' = (B + t) + (C - t) = B + C.$$

In particular, if C is bounded below and $t = \min(C)$, then $0 = \min(C')$ and $B' \subseteq B' + C'$. It follows that if B and C are infinite sets such that $B + C \subseteq A$, then, by translation, there exist infinite sets B' and C' such that $B' \subseteq A$ and $B' + C' \subseteq A$.

However, a set A with positive upper Banach density does not necessarily contain infinite subsets B and C with $B + C \subseteq A$. For example, let A be any set of odd numbers. For all sets B and C of odd numbers, the sumset $B + C$ is a set of even numbers, and so $A \cap (B + C) = \emptyset$. Of course, in this example, we have $B + C \subseteq A + 1$.

In this note, we prove that if A is a set of positive integers with upper Banach density $\delta(A) = 1$, then for every $h \ge 2$ there exist pairwise disjoint subsets $B_1, \ldots, B_h$ of A such that $\delta(B_i) = 1$ for all $i = 1, \ldots, h$ and

$$B_1 + \cdots + B_h \subseteq A.$$

Indeed, Theorem 2 states an even stronger result.

There are sets A of upper Banach density 1 for which no infinite subset B of A satisfies $2B \subseteq A + t$ for any integer t. A simple example is

$$A = \bigcup_{i=1}^{\infty} \left[4^i, 4^i + i - 1 \right].$$

The set A is the union of the infinite sequence of pairwise disjoint intervals

$$A_i = \left[4^i, 4^i + i - 1 \right].$$

Let $t \in \mathbf{N}_0$. There exists $i_0(t)$ such that $4^i - i > t$ for all $i \geq i_0(t)$. If $b_i \in A_i$ for some $i \geq i_0(t)$, then

$$4^i + i + t < 2 \cdot 4^i \leq 2b_i < 2 \cdot 4^i + 2i < 4^{i+1} - 2t \leq 4^{i+1} - t$$

and so $2b_i \notin 2A \pm t$. If B is an infinite subset of A, then for infinitely many i, there exist integers $b_i \in B \cap A_i$, and so $2B \not\subseteq A + t$ for all $t \in \mathbf{Z}$.

There are very few results about the Erdős conjecture. In 1980, Nathanson [9] proved that if $\delta(A) > 0$, then for every n there is a finite set C with $|C| = n$ and a subset B of A with $\delta(B) > 0$ such that $B + C \subseteq A$. In 2015, Di Nasso et al. [3] used nonstandard analysis to prove that the Erdős conjecture is true for sets A with upper Banach density $\delta(A) > 1/2$. They also proved that if $\delta(A) > 0$, then there exist infinite sets B and C and an integer t such that

$$B + C \subseteq A \cup (A + t).$$

It would be of interest to have purely combinatorial proofs of the results of Di Nasso et al.

For related work, see Di Nasso [1, 2], Gromov [4], Hegyvári [5, 6], Hindman [7], and Jin [8].

3 Results

The following result is well known.

Lemma 1 *A set of positive integers has upper Banach density 1 if and only if, for every d, it contains infinitely many pairwise disjoint intervals of d consecutive integers.*

Proof Let A be a set of positive integers. If, for every positive integer d, the set A contains an interval of d consecutive integers, then

$$\max_{u \in \mathbf{N}_0} \left(\frac{|A \cap [u, u+d-1]|}{d} \right) = 1$$

and so

$$\delta(A) = \lim_{d \to \infty} \max_{u \in \mathbf{N}_0} \left(\frac{|A \cap [u, u+d-1]|}{d} \right) = 1.$$

Suppose that, for some integer $d \geq 2$, the set A contains no interval of d consecutive integers. For every $u \in \mathbf{N}_0$, we consider the interval $I_{u,n} = [u, u+n-1]$. By the division algorithm, there are integers q and r with $0 \leq r < d$ such that

$$|I_{u,n}| = n = qd + r$$

and

$$q = \frac{n-r}{d} > \frac{n}{d} - 1.$$

For $j = 1, \ldots, q$, the intervals of integers

$$I_{u,n}^{(j)} = [u + (j-1)d, u + jd - 1]$$

and

$$I_{u,n}^{(q+1)} = [u + qd, u + n - 1]$$

are pairwise disjoint subsets of $I_{u,n}$ such that

$$I_{u,n} = \bigcup_{j=1}^{q+1} I_{u,n}^{(j)}.$$

We have

$$A \cap I_{u,n} = \bigcup_{j=1}^{q+1} (A \cap I_{u,n}^{(j)})$$

If A contains no interval of d consecutive integers, then, for all $j \in [1, q]$, at least one element of the interval $I_{u,n}^{(j)}$ is not an element of A, and so

$$|A \cap I_{u,n}^{(j)}| \leq |I_{u,n}^{(j)}| - 1.$$

It follows that

$$\begin{aligned}
|A \cap I_{u,n}| &= \sum_{j=1}^{q+1} \left|A \cap I_{u,n}^{(j)}\right| \leq \sum_{j=1}^{q} \left(\left|I_{u,n}^{(j)}\right| - 1\right) + \left|I_{u,n}^{(q+1)}\right| \\
&= \sum_{j=1}^{q+1} \left|I_{u,n}^{(j)}\right| - q = |I_{u,n}| - q = n - q \\
&< n - \frac{n}{d} + 1 = \left(1 - \frac{1}{d}\right) n + 1.
\end{aligned}$$

Dividing by $n = |I_{u,n}|$, we obtain

$$\max_{u \in \mathbf{N}_0} \frac{|A \cap I_{u,n}|}{n} \leq 1 - \frac{1}{d} + \frac{1}{n}.$$

and so

$$\delta(A) = \lim_{n \to \infty} \max_{u \in \mathbf{N}_0} \frac{|A \cap I_{u,n}|}{n} \leq 1 - \frac{1}{d} < 1$$

which is absurd. Therefore, A contains an interval of d consecutive integers for every $d \in \mathbf{N}$.

To prove that A contains infinitely many intervals of size d, it suffices to prove that if $[u, u + d - 1] \subseteq A$, then $[v, v + d - 1] \subseteq A$ for some $v \geq u + d$. Let $d' = u + 2d$. There exists $u' \in \mathbf{N}$ such that

$$[u', u' + d' - 1] = [u', u' + u + 2d - 1] \subseteq A.$$

Choosing $v = u' + u + d$, we have $v \geq u + d$ and

$$[v, v + d - 1] \subseteq [u', u' + u + 2d - 1] \subseteq A.$$

This completes the proof.

Let $\mathcal{F}(S)$ denote the set of all finite subsets of the set S, and let $\mathcal{F}^*(S)$ denote the set of all nonempty finite subsets of S. We have the fundamental binomial identity

$$\mathcal{F}^*([1, n + 1]) = \mathcal{F}^*([1, n]) \cup \{\{n + 1\} \cup J : J \in \mathcal{F}([1, n])\}. \tag{1}$$

Theorem 1 *Let A be a set of positive integers that has upper Banach density 1. For every sequence $(\ell_j)_{j=1}^{\infty}$ of positive integers, there exists a sequence $(b_j)_{j=1}^{\infty}$ of positive integers such that*

$$b_{j+1} \geq b_j + \ell_j$$

for all $j \in \mathbf{N}$*, and*

$$\sum_{j \in J} [b_j, b_j + \ell_j - 1] \subseteq A$$

for all $J \in \mathcal{F}^*(\mathbf{N})$.

Proof We shall construct the sequence $(b_j)_{j=1}^{\infty}$ by induction. For $n = 1$, choose $b_1 \in A$ such that $[b_1, b_1 + \ell_1 - 1] \subseteq A$.

Suppose that $(b_j)_{j=1}^{n}$ is a finite sequence of positive integers such that $b_{j+1} \geq b_j + \ell_j$ for $j \in [1, n-1]$ and

$$\sum_{j \in J} [b_j, b_j + \ell_j - 1] \subseteq A \tag{2}$$

for all $J \in \mathcal{F}^*([1, n])$. By Lemma 1, there exists $b_{n+1} \in A$ such that

$$b_{n+1} \geq b_n + \ell_n$$

and

$$\left[b_{n+1}, \sum_{j=1}^{n+1} (b_j + \ell_j) - 1 \right] \subseteq A.$$

It follows that

$$\left[b_{n+1}, b_{n+1} + \ell_{n+1} - 1 \right] \subseteq A.$$

Let $J \in \mathcal{F}([1, n])$. If

$$\begin{aligned} a &\in \sum_{j \in \{n+1\} \cup J} [b_j, b_j + \ell_j - 1] \\ &= \left[b_{n+1}, b_{n+1} + \ell_{n+1} - 1 \right] + \sum_{j \in J} [b_j, b_j + \ell_j - 1] \end{aligned}$$

then

$$\begin{aligned} b_{n+1} \leq a &\leq (b_{n+1} + \ell_{n+1} - 1) + \sum_{j \in J} \left(b_j + \ell_j - 1 \right) \\ &\leq \sum_{j=1}^{n+1} (b_j + \ell_j) - 1 \end{aligned}$$

and so $a \in A$ and

$$\sum_{j\in\{n+1\}\cup J}[b_j, b_j+\ell_j-1] \subseteq \left[b_{n+1}, \sum_{j=1}^{n+1}(b_j+\ell_j)-1\right] \subseteq A. \tag{3}$$

Relations (1), (2), and (3) imply that

$$\sum_{j\in J}[b_j, b_j+\ell_j-1] \subseteq A$$

for all $J \in \mathcal{F}^*([1, n+1])$. This completes the induction.

Theorem 2 *Every set A of positive integers that has upper Banach density 1 contains an infinite sequence of pairwise disjoint subsets $(B_i)_{i=1}^{\infty}$ such that B_i has upper Banach density 1 for all $i \in \mathbf{N}$ and*

$$\sum_{i\in I} B_i \subseteq A$$

for all $I \in \mathcal{F}^*(\mathbf{N})$.

Proof Let $(\ell_j)_{j=1}^{\infty}$ be a sequence of positive integers such that $\lim_{j\to\infty}\ell_j = \infty$, and let $(b_j)_{j=1}^{\infty}$ be a sequence of positive integers that satisfies Theorem 1. (For simplicity, we can let $\ell_j = j$ for all j.) Let $(X_i)_{i=1}^{\infty}$ be a sequence of infinite sets of positive integers that are pairwise disjoint. For $i \in \mathbf{N}$, let

$$B_i = \bigcup_{j\in X_i}[b_j, b_j+\ell_j-1].$$

The set B_i contains intervals of ℓ_j consecutive integers for infinitely many ℓ_j, and so B_i has upper Banach density 1.

Let $I \in \mathcal{F}^*(\mathbf{N})$. If

$$a \in \sum_{i\in I} B_i \subseteq A$$

then for each $i \in I$ there exists $a_i \in B_i$ such that $a = \sum_{i\in I} a_i$. If $a_i \in B_i$, then there exists $j_i \in X_i$ such that

$$x_i \in \left[b_{j_i}, b_{j_i}+\ell_{j_i}-1\right].$$

We have $J = \{j_i : i \in I\} \in \mathcal{F}^*(\mathbf{N})$ and

$$a \in \sum_{j_i\in J}\left[b_{j_i}, b_{j_i}+\ell_{j_i}-1\right] \subseteq A.$$

This completes the proof.

Theorem 3 *Let A be a set of integers that contains arbitrarily long finite arithmetic progressions with bounded differences. There exist positive integers m and r, and an infinite sequence of pairwise disjoint sets $(B_i)_{i=1}^{\infty}$ such that B_i has upper Banach density 1 for all $i \in \mathbf{N}$ and*

$$m * \sum_{i \in I} B_i + r \subseteq A$$

for all $I \in \mathcal{F}^*(\mathbf{N})$.

Proof If the differences in the infinite set of finite arithmetic progressions contained in A are bounded by m_0, then there exists a difference $m \leq m_0$ that occurs infinitely often. It follows that there are arbitrarily long finite arithmetic progressions with difference m. Because there are only finitely many congruence classes modulo m, there exists a congruence class $r \pmod{m}$ such that A contains arbitrarily long sequences of consecutive integers in the congruence class $r \pmod{m}$. Thus, there exists an infinite set A' such that

$$m * A' + r \subseteq A$$

and A' contains arbitrarily long sequences of consecutive integers. Equivalently, A' has Banach density 1. By Theorem 2, the sequence A' contains an infinite sequence of pairwise disjoint subsets $(B_i)_{i=1}^{\infty}$ such that B_i has upper Banach density 1 for all $i \in \mathbf{N}$ and

$$\sum_{i \in I} B_i \subseteq A'$$

for all $I \in \mathcal{F}^*(\mathbf{N})$. It follows that

$$m * \sum_{i \in I} B_i + r \subseteq m * A' + r \subseteq A$$

for all $I \in \mathcal{F}^*(\mathbf{N})$. This completes the proof.

Appendix: Subadditivity and Limits

A real-valued arithmetic function f is *subadditive* if

$$f(n_1 + n_2) \leq f(n_1) + f(n_2) \tag{4}$$

for all $n_1, n_2 \in \mathbf{N}$.

The following result is sometimes called *Fekete's lemma*.

Lemma 2 *If f is a subadditive arithmetic function, then* $\lim_{n\to\infty} f(n)/n$ *exists, and*

$$\lim_{n\to\infty} \frac{f(n)}{n} = \inf_{n\in\mathbf{N}} \frac{f(n)}{n}.$$

Proof It follows by induction from inequality (4) that

$$f(n_1 + \cdots + n_q) \le f(n_1) + \cdots + f(n_q)$$

for all $n_1, \ldots, n_q \in \mathbf{N}$. Let $f(0) = 0$. Fix a positive integer d. For all $q, r \in \mathbf{N}_0$, we have

$$f(qd + r) \le qf(d) + f(r).$$

By the division algorithm, every nonnegative integer n can be represented uniquely in the form $n = qd + r$, where $q \in \mathbf{N}_0$ and $r \in [0, d-1]$. Therefore,

$$\frac{f(n)}{n} = \frac{f(qd + r)}{n} \le \frac{qf(d)}{qd} + \frac{f(r)}{n} = \frac{f(d)}{d} + \frac{f(r)}{n}.$$

Because the set $\{f(r) : r \in [0, d-1]\}$ is bounded, it follows that

$$\limsup_{n\to\infty} \frac{f(n)}{n} \le \limsup_{n\to\infty} \left(\frac{f(d)}{d} + \frac{f(r)}{n} \right) = \frac{f(d)}{d}$$

for all $d \in \mathbf{N}$, and so

$$\limsup_{n\to\infty} \frac{f(n)}{n} \le \inf_{d\in\mathbf{N}} \frac{f(d)}{d} \le \liminf_{d\to\infty} \frac{f(d)}{d} = \liminf_{n\to\infty} \frac{f(n)}{n}.$$

This completes the proof.

References

1. M. Di Nasso, An elementary proof of Jin's theorem with a bound. Electron. J. Combin. **21**(2), Paper 2.37, 7 (2014)
2. M. Di Nasso, Embeddability properties of difference sets, Integers **14**, Paper No. A **27**, 24 (2014)
3. M. Di Nasso, I. Goldbring, R. Jin, S. Leth, M. Lupini, K. Mahlburg, On a sumset conjecture of Erdős. Canad. J. Math. **67**(4), 795–809 (2015)
4. M.L. Gromov, Colorful categories. Uspekhi Mat. Nauk **70**(4), 424, 3–76 (2015)
5. N. Hegyvári, On the dimension of the Hilbert cubes. J. Num. Theor. **77**(2), 326–330 (1999)
6. N. Hegyvári, On additive and multiplicative Hilbert cubes. J. Combin. Theor. Ser. A **115**(2), 354–360 (2008)

7. N. Hindman, Ultrafilters and combinatorial number theory, Number theory, Carbondale, Proceedings of Southern Illinois Conference, Southern Illinois University, Carbondale, Ill., Lecture Notes in Mathematics, vol. 751. Springer, Berlin **1979**, 119–184 (1979)
8. R. Jin, Standardizing nonstandard methods for upper Banach density problems, Unusual Applications of Number Theory, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 4, Amer. Math. Soc., Providence, RI, 109–124 (2004)
9. M.B. Nathanson, Sumsets contained in infinite sets of integers. J. Combin. Theor. Ser. A **28**(2), 150–155 (1980)

The Erdős Paradox

Melvyn B. Nathanson

Prologue

The great Hungarian mathematician Paul Erdős was born in Budapest on March 26, 1913. He died alone in a hospital room in Warsaw, Poland, on Friday afternoon, September 20, 1996. It was sad and ironic that he was alone, because he probably had more friends in more places than any mathematician in the world. He was in Warsaw for a conference. Vera Sós had also been there, but had gone to Budapest on Thursday and intended to return on Saturday with András Sárközy to travel with Paul to a number theory meeting in Vilnius. On Thursday night, Erdős felt ill and called the desk in his hotel. He was having a heart attack and was taken to a hospital, where he died about 12 hours later. No one knew he was in the hospital. When Paul did not appear at the meeting on Friday morning, one of the Polish mathematicians called the hotel. He did not get through, and no one tried to telephone the hotel again for several hours. By the time it was learned that Paul was in the hospital, he was dead.

Vera was informed by telephone on Friday afternoon that Paul had died. She returned to Warsaw on Saturday. It was decided that Paul should be cremated. This was contrary to Jewish law, but Paul was not an observant Jew and it is not known what he would have wanted. Nor was he buried promptly in accordance with Jewish tradition. Instead, four weeks later, on October 18, there was a secular funeral service in Budapest, and his ashes were buried in the Jewish cemetery in Budapest.

Erdős strongly identified with Hungary and with Judaism. He was not religious, but he visited Israel often, and established a mathematics prize and a postdoctoral fellowship there. He also established a prize and a lectureship in Hungary. He told me that he was happy whenever someone proved a beautiful theorem, but that he was especially happy if the person who proved the theorem was Hungarian or Jewish.

M. B. Nathanson (✉)
Department of Mathematics, Lehman College (CUNY),
Bronx, NY 10468, USA
e-mail: melvyn.nathanson@lehman.cuny.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_17

Mathematicians from the USA, Israel, and many European countries traveled to Hungary to attend Erdos's funeral. The following day a conference, entitled "Paul Erdős and his Mathematics," took place at the Hungarian Academy of Sciences in Budapest, and mathematicians who were present for the funeral were asked to lecture on different parts of Erdős's work. I was asked to chair one of the sessions and to begin with some personal remarks about my relationship with Erdős and his life and style.

This paper is in two parts. The first is the verbatim text of my remarks at the Erdős memorial conference in Budapest on October 19, 1996. A few months after the funeral and conference, I returned to Europe to lecture in Germany. At Bielefeld, someone told me that my eulogy had generated controversy, and indeed, I heard the same report a few weeks later when I was back in the USA. Eighteen years later, on the 100th anniversary of his birth, it is fitting to reconsider Erdős's life and work.

1 Eulogy, Delivered in Budapest on October 19, 1996

I knew Erdős for 25 years, half my life, but still not very long compared to many people in this room. His memory was much better than mine; he often reminded me that we proved the theorems in our first paper in 1972 in a car as we drove back to Southern Illinois University in Carbondale after a meeting of the Illinois Number Theory Conference in Normal, Illinois. He visited me often in Carbondale and even more often after I moved to New Jersey. He would frequently leave his winter coat in my house when he left for Europe in the spring, and retrieve it when he returned in the fall. I still have a carton of his belongings in my attic. My children Becky and Alex, who are five and seven years old, would ask, "When is Paul coming to visit again?" They liked his silly tricks for kids, like dropping a coin and catching it before it hit the floor. He was tolerant of the dietary rules in my house, which meant, for example, no milk in his espresso if we had just eaten meat.

He was tough. "No illegal thinking," he would say when we were working together. This meant no thinking about mathematical problems other than the ones we were working on at that time. In other words, he knew how to enforce party discipline.

Erdős loved to discuss politics, especially Sam and Joe, which, in his idiosyncratic language, meant the USA (Uncle Sam) and the Soviet Union (Joseph Stalin). His politics seemed to me to be the politics of the 30s, much to the left of my own. He embraced a kind of naive and altruistic socialism that I associate with idealistic intellectuals of his generation. He never wanted to believe what I told him about the Soviet Union as an "evil empire." I think he was genuinely saddened by the fact that the demise of communism in the Soviet Union meant the failure of certain dreams and principles that were important to him.

Erdős's cultural interests were narrowly focused. When he was in my house, he always wanted to hear "noise" (that is, music), especially Bach. He loved to quote Hungarian poetry (in translation). I assume that when he was young he read literature (he was amazed that Anatole France is a forgotten literary figure today), but I don't think he read much anymore.

I subscribe to many political journals. When he came to my house, he would look for the latest issue of *Foreign Affairs*, but usually disagreed with the contents. Not long ago, an American historian at Pacific Lutheran University published a book entitled *Ordinary Men*,[1] a study of how large numbers of "ordinary Germans," not just a few SS, actively and willingly participated in the murder of Jews. He found the book on my desk and read it, but did not believe or did not want to believe it could be true, because it conflicted with his belief in the natural goodness of ordinary men.

He had absolutely no interest in the visual arts. My wife was a curator at the Museum of Modern Art in New York, and we went with her one day to the museum. It has the finest collection of modern art in the world, but Paul was bored. After a few minutes, he went out to the sculpture garden and started, as usual, to prove and conjecture.

Paul's mathematics was like his politics. He learned mathematics in the 1930s in Hungary and England, and England at that time was a kind of mathematical backwater. For the rest of his life, he concentrated on the fields that he had learned as a boy. Elementary and analytic number theory, at the level of Landau, graph theory, set theory, probability theory, and classical analysis. In these fields, he was an absolute master, a virtuoso.

At the same time, it is extraordinary to think of the parts of mathematics he never learned. Much of contemporary number theory, for example. In retrospect, probably the greatest number theorist of the 1930s was Hecke, but Erdős knew nothing about his work and cared less. Hardy and Littlewood dominated British number theory when Erdős lived in England, but I doubt they understood Hecke.

There is an essay by Irving Segal[2] in the current issue of the *Bulletin of the American Mathematical Society*. He tells the story of the visit of another great Hungarian mathematician, John von Neumann, to Cambridge in the 1930s. After his lecture, Hardy remarked, "Obviously a very intelligent young man. But was that *mathematics*?"

A few months ago, on his last visit to New Jersey, I was telling Erdős something about p-adic analysis. Erdős was not interested. "You know," he said about the p-adic numbers, "they don't really exist."

Paul never learned algebraic number theory. He was offended—actually, he was furious—when André Weil wrote that analytic number theory is good mathematics, but analysis, not number theory.[3] Paul's "tit-for-tat" response was that André Weil did good mathematics, but it was algebra, not number theory. I think Paul was a bit

[1] Christopher R. Browning, *Ordinary Men*, HarperCollins Publishers, New York, 1992.

[2] Irving Segal, "*Noncommutative Geometry* by Alain Connes (book review)," *Bull. Amer. Math. Soc.* 33 (1996), 459–465.

[3] Weil wrote, "…there is a subject in mathematics (it's a perfectly good and valid subject and it's perfectly good and valid mathematics) which is called Analytic Number Theory…. I would classify it under analysis…." (*Œuvres Scientifiques Collected Papers*, Springer-Verlag, New York, 1979, Volume III, p. 280).

shocked that a problem he did consider number theory, Fermat's Last Theorem, was solved using ideas and methods of Weil and other very sophisticated mathematicians.

It is idle to speculate about how great a mathematician Erdős was, as if one could put together a list of the top 10 or top 100 mathematicians of our century. His interests were broad, his conjectures, problems, and results profound, and his humanity extraordinary.

He was the "Bob Hope" of mathematics, a kind of vaudeville performer who told the same jokes and the same stories a thousand times. When he was scheduled to give yet another talk, no matter how tired he was, as soon as he was introduced to the audience, the adrenaline (or maybe amphetamine) would release into his system and he would bound onto the stage, full of energy, and do his routine for the 1001st time.

If he were here today, he would be sitting in the first row, half asleep, happy to be in the presence of so many colleagues, collaborators, and friends.

Yitgadal v'yitkadash sh'mei raba.

Y'hei zekronoh l'olam.

May his memory be with us forever.[4]

2 Reconsideration

My brief talk at the Erdős conference was not intended for publication. Someone asked me for a copy, and it subsequently spread via e-mail. Many people who heard me in Budapest or who later read my eulogy told me that it helped them remember Paul as a human being, but others clearly disliked what I said. I confess I still don't know what disturbed them so deeply. It has less to do with Erdős, I think, than with the status of "Hungarian mathematics" in the scientific world.[5]

Everyone understands that Erdős was an extraordinary human being and a great mathematician who made major contributions to many parts of mathematics. He was a central figure in the creation of new fields, such as probabilistic number theory and random graphs. This part of the story is trivial.

It is also true, understood by almost everyone, and not controversial, that Erdős did not work in and never learned the central core of twentieth-century mathematics. It is amazing to me how great were Erdos's contributions to mathematics, and how little he knew. He never learned, for example, the great discoveries in number theory that were made at the beginning of the twentieth century. These include, for example, Weil's work on Diophantine equations, Artin's class field theory, and Hecke's monumental contributions to modular forms and analytic number theory. Erdős apparently

[4] I ended my eulogy with a sentence in Aramaic and a sentence in Hebrew. The first is the first line of the Kaddish, the Jewish prayer for the dead. Immediately following the second sentence is its English translation.

[5] cf. L. Babai, "In and out of Hungary: Paul Erdős, his friends, and times," in: *Combinatorics, Paul Erdős is Eighty (Volume 2), Keszthely (Hungary) 1993*, Bolyai Society Mathematical Studies, Budapest, 1996, pp. 7–95.

knew nothing about Lie groups, Riemannian manifolds, algebraic geometry, algebraic topology, global analysis, or the deep ocean of mathematics connected with quantum mechanics and relativity theory. These subjects, already intensely investigated in the 1930s, were at the heart of twentieth-century mathematics. How could a great mathematician not want to study these things?[6] This is the first Erdős paradox.

In the case of the Indian mathematician Ramanujan, whose knowledge was also deep but narrow, there is a discussion in the literature about the possible sources of his mathematical education. The explanation of Hardy[7] and others is that the only serious book that was accessible to Ramanujan in India was Carr's *A Synopsis of Elementary Results in Pure and Applied Mathematics*, and that Ramanujan lacked a broad mathematical culture because he did not have access to books and journals in India. But Hungary was not India; there were libraries, books, and journals in Budapest, and in other places where Erdős lived in the 1930s and 1940s.

For the past half-century, "Hungarian mathematics" has been a term of art to describe the kind of mathematics that Erdős did.[8] It includes combinatorics, graph theory, combinatorial set theory, and elementary and combinatorial number theory. Not all Hungarians do this kind of mathematics, of course, and many non-Hungarians do Hungarian mathematics. It happens that combinatorial reasoning is central to theoretical computer science, and "Hungarian mathematics" commands vast respect in the computer science world. It is also true, however, that for many years combinatorics did not have the highest reputation among mathematicians in the ruling subset of the research community, exactly because combinatorics was concerned largely with questions that they believed (incorrectly) were not central to twentieth-century mathematics.[9]

In a volume in honor of Erdős's 70th birthday, Ernst Straus wrote, "In our century, in which mathematics is so strongly dominated by 'theory constructors' [Erdős] has remained the prince of problem solvers and the absolute monarch of problem posers."[10] I disagree. There is, as Gel'fand often said, only one mathematics. There is no separation of mathematics into "theory" and "problems." But there is an interesting lurking issue.

In his lifetime, did Erdős get the recognition he deserved? Even though Erdős received almost every honor that can be given to a mathematician, some of his friends believe that he was still insufficiently appreciated, and they are bitter on his behalf.

[6] This suggests the fundamental question: How much, or how little, must one know in order to do great mathematics?.

[7] "It was a book of a very different kind, Carr's *Synopsis*, which first aroused Ramanujan's full powers," according to G. H. Hardy, in his book *Ramanujan*, Chelsea Publishing, New York, 1959, p. 2

[8] For example, Joel Spencer, "I felt … I was working on 'Hungarian mathematics'," quoted in Babai, *op. cit.*

[9] For example, S. Mac Lane criticized "emphasizing too much of a Hungarian view of mathematics," in: "The health of mathematics," *Math.Intelligencer* 5 (1983), 53–55.

[10] E. G. Straus, "Paul Erdős at 70," *Combinatorica* 3 (1983), 245–246. Tim Gowers revisited this notion in his essay, "The two cultures of mathematics," published in *Mathematics: Frontiers and Perspectives*, American Mathematical Society, 2000.

He was awarded a Wolf Prize and a Cole Prize, but he did not get a Fields Medal or a permanent professorship at the Institute for Advanced Study. He traveled from one university to another across the USA and was never without an invitation to lecture somewhere, but his mathematics was not highly regarded by the power brokers of mathematics. To them, his methods were insufficiently abstruse and obscure; they did not require complicated machinery. Paul invented diabolically clever arguments from arithmetic, combinatorics, and probability to solve problems. But the technique was too simple, too elementary. It was suspicious. The work could not be "deep."

None of this seemed to matter to Erdős, who was content to prove and conjecture and publish more than 1,500 papers.

Not because of politicking, but because of computer science and because his mathematics was always beautiful, in the past decade the reputation of Erdős and the respect paid to discrete mathematics have increased exponentially. The *Annals of Mathematics* will now publish papers in combinatorics, and the most active seminar at the Institute for Advanced Study is in discrete mathematics and theoretical computer science. Fields Medals are awarded to mathematicians who solve Erdős-type problems. Science has changed.

In 1988, Alexander Grothendieck was awarded the Crafoord Prize of the Swedish Academy of Sciences. In the letter to the Swedish Academy in which he declined the prize, he wrote, "Je suis persuadé que la seule épreuve décisive pour la fécundité d'idées ou d'une vision nouvelles est celle du temps. La fécondité se reconnait par la progéniture, et non par les honneurs."[11]

Time has proved the fertility and richness of Erdős's work. The second Erdős paradox is that his methods and results, considered marginal in the twentieth century, have become central in twenty-first-century mathematics.

May his memory be with us forever.

[11] "I believe that time gives the only definite proof of the fertility of new ideas or a new vision. We recognize fertility by its offspring, and not by honors."

Limits and Decomposition of de Bruijn's Additive Systems

Melvyn B. Nathanson

Abstract An *additive system* for the nonnegative integers is a family $(A_i)_{i \in I}$ of sets of nonnegative integers with $0 \in A_i$ for all $i \in I$ such that every nonnegative integer can be written uniquely in the form $\sum_{i \in I} a_i$ with $a_i \in A_i$ for all i and $a_i \neq 0$ for only finitely many i. In 1956, de Bruijn proved that every additive system is constructed from an infinite sequence $(g_i)_{i \in \mathbf{N}}$ of integers with $g_i \geq 2$ for all i or is a contraction of such a system. This paper discusses limits and the stability of additive systems and also describes the "uncontractable" or "indecomposable" additive systems.

Keywords Additive system · Additive basis · British number system

2010 Mathematics Subject Classification: 11A05 · 11B75

1 Additive Systems and de Bruijn's Theorem

Let $\mathbf{N}_0$ and $\mathbf{N}$ denote the sets of nonnegative integers and positive integers, respectively. For real numbers a and b, we define the interval of integers $[a, b) = \{x \in \mathbf{Z} : a \leq x < b\}$ and $[a, b] = \{x \in \mathbf{Z} : a \leq x \leq b\}$.

Let I be a nonempty finite or infinite set, and let $\mathcal{A} = (A_i)_{i \in I}$ be a family of sets of integers with $0 \in A_i$ and $|A_i| \geq 2$ for all $i \in I$. Each set A_i can be finite or infinite. The *sumset* $S = \sum_{i \in I} A_i$ is the set of all integers n that can be represented in the form $n = \sum_{i \in I} a_i$, where $a_i \in A_i$ for all $i \in I$ and $a_i \neq 0$ for only finitely many $i \in I$. If every element of S has a *unique* representation in the form $n = \sum_{i \in I} a_i$, then we call $\mathcal{A}$ a *unique representation system for* S, and we write $S = \bigoplus_{i \in I} A_i$.

In a unique representation system $\mathcal{A}$ for S, we have $A_i \cap A_j = \{0\}$ for all $i \neq j$. The condition $|A_i| \geq 2$ for all $i \in I$ implies that $A_i = S$ for some $i \in I$ if and only if $|I| = 1$. Moreover, if $I^\flat \subseteq I$ and $S = \sum_{i \in I^\flat} A_i$, then $S = \bigoplus_{i \in I^\flat} A_i$ and $I = I^\flat$.

M. B. Nathanson (✉)
Department of Mathematics, Lehman College (CUNY),
Bronx, NY 10468, USA
e-mail: melvyn.nathanson@lehman.cuny.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_18

The family $\mathcal{A} = (A_i)_{i \in I}$ is an *additive system* if $\mathcal{A}$ is a unique representation system for the set of nonnegative integers, that is, if $\mathbf{N}_0 = \bigoplus_{i \in I} A_i$. The following lemma follows immediately from the definition of an additive system.

Lemma 1 *Let $\mathcal{B} = (B_j)_{j \in J}$ be an additive system. If $\{J_i\}_{i \in I}$ is a partition of J into pairwise disjoint nonempty sets, and if*

$$A_i = \sum_{j \in J_i} B_j$$

then $\mathcal{A} = (A_i)_{i \in I}$ is an additive system.

The additive system $\mathcal{A}$ obtained from the additive system $\mathcal{B}$ by the partition procedure described in Lemma 1 is called a *contraction* of $\mathcal{B}$. (In [1], de Bruijn called $\mathcal{A}$ a *degeneration* of $\mathcal{B}$.) If $I = J$ and if σ is a permutation of J such that $J_i = \{\sigma(i)\}$ for all $i \in J$, then $\mathcal{A}$ and $\mathcal{B}$ contain exactly the same sets. Thus, every additive system is a contraction of itself. An additive system $\mathcal{A}$ is a *proper contraction* of $\mathcal{B}$ if at least one set $A_i \in \mathcal{A}$ is the sum of at least two sets in $\mathcal{B}$.

Let X be a set of integers, and let g be an integer. The *dilation* of X by g is the set $g * X = \{gx : x \in X\}$.

Lemma 2 *Let $\mathcal{B} = (B_j)_{j \in J}$ be an additive system and let $I = \{i_0\} \cup J$, where $i_0 \notin J$. If*

$$A_{i_0} = [0, g)$$

and

$$A_j = g * B_j \qquad \text{for all } j \in J$$

then $\mathcal{A} = (A_i)_{i \in I}$ is an additive system.

The additive system $\mathcal{A}$ obtained from the additive system $\mathcal{B}$ by the procedure described in Lemma 2 is called the *dilation* of $\mathcal{B}$ by g.

There are certain additive systems that de Bruijn called *British number systems*. A British number system is an additive system constructed from an infinite sequence of integers according to the algorithm in Theorem 1 below. de Bruijn [1] proved that British number systems are essentially the only additive systems.

Theorem 1 *Let $(g_i)_{i \in \mathbf{N}}$ be an infinite sequence of integers such that $g_i \geq 2$ for all $i \geq 1$. Let $G_0 = 1$ and, for $i \in \mathbf{N}$, let $G_i = \prod_{j=1}^{i} g_j$ and*

$$A_i = \{0, G_{i-1}, 2G_{i-1}, \ldots, (g_i - 1)G_{i-1}\} = G_{i-1} * [0, g_i).$$

Then $\mathcal{A} = (A_i)_{i \in \mathbf{N}}$ is an additive system.

Theorem 2 *Every additive system is a British number system or a proper contraction of a British number system.*

The proof of Theorem 2 depends on the following fundamental lemma.

Lemma 3 *Let $\mathcal{A} = (A_i)_{i \in I}$ be an additive system with $|I| \geq 2$, and let i_1 be the unique element of I such that $1 \in I_{i_1}$. There exist an integer $g \geq 2$ and a family of sets $\mathcal{B} = (B_i)_{i \in I}$ such that*

$$A_{i_1} = [0, g) \oplus g * B_{i_1}$$

and, for all $i \in I \setminus \{i_1\}$,

$$A_i = g * B_i.$$

If $B_{i_1} = \{0\}$, then $\mathcal{B} = (B_i)_{i \in I \setminus \{i_1\}}$ is an additive system, and $\mathcal{A}$ is the dilation of the additive system $\mathcal{B}$ by the integer g. If $B_{i_1} \neq \{0\}$, then $\mathcal{B} = (B_i)_{i \in I}$ is an additive system and $\mathcal{A}$ is a contraction of the additive system $\mathcal{B}$ dilated by g.

For proofs of Lemmas 1, 2, and 3 and Theorems 1 and 2, see Nathanson [4].

This paper gives a refinement of de Bruijn's theorem. Every additive system is a contraction of a British number system, but even a British number system can be a proper contraction of another British number system. An additive system that is not a proper contraction of another number system will be called *indecomposable*. In Sect. 3, we describe all indecomposable British number systems. Unsurprisingly, there is a one-to-one correspondence between indecomposable British number systems and infinite sequences of prime numbers.

In Sect. 4, we define the limit of a sequence of additive systems and discuss the stability of British number systems.

Maltenfort [2] and Munagi [3] have also studied de Bruijn's additive systems.

2 Decomposable and Indecomposable Sets

The set A of integers is a *proper sumset* if there exist sets B and C of integers such that $|B| \geq 2$, $|C| \geq 2$, and $A = B + C$. For example, if u and v are integers and $v - u \geq 3$, then the interval $[u, v)$ is a proper sumset:

$$[u, v) = [0, i) + [u, v + 1 - i)$$

for every $i \in [2, v - u)$.

The set A of integers is *decomposable* if there exist sets B and C such that (B, C) is a unique representation system for A, that is, if $|B| \geq 2$, $|C| \geq 2$, and $A = B \oplus C$. A decomposition $A = B \oplus C$ is also called a *tiling* of A by B. For example,

$$[0, 12) = \{0, 3\} \oplus \{0, 1, 2, 6, 7, 8\}.$$

If $A = B \oplus C$ is a decomposition, then $|A| = |B|\ |C|$ and so the integer $|A|$ is composite.

Let $n \geq 2$ and consider the interval of integers $A = [0, n)$. A *proper divisor* of n is a divisor d of n such that $1 < d < n$. Associated to every proper divisor d of n is the decomposition

$$[0, n) = [0, d) \oplus d * [0, n/d). \tag{1}$$

This is simply the division algorithm for integers. The number of decompositions of type (1) is the number of proper divisors d of n. There is exactly one such decomposition if and only if the integer n has a unique proper divisor if and only if n is the square of a prime number.

Lemma 4 *Let $n \geq 2$. The interval $[0, n)$ is indecomposable if and only if n is prime.*

Proof If n is prime then $[0, n)$ is indecomposable, and if n is composite, then $[0, n)$ is decomposable.

If $A = B \oplus C$ and g is a nonzero integer, then $g * A = g * B \oplus g * C$, and so every dilation of a decomposable set is decomposable.

The *translate* of the set A by an integer t is the set

$$A + t = \{a + t : a \in A\}.$$

Let $t_1, t_2 \in \mathbf{Z}$ with $t = t_1 + t_2$. If $A = B + C$, then

$$A + t = (B + t_1) + (C + t_2).$$

In particular, $A + t = (B + t) + C$. If $A = B \oplus C$, then $A + t = (B + t) \oplus C$, and so every translate of a decomposable set is decomposable. Similarly, if $A = B \oplus C$, then $A = (B - t) \oplus (C + t)$ for every integer t.

Let A be a set of nonnegative integers with $0 \in A$, and let B and C be sets of integers with $A = B \oplus C$. Let $t = \min(B)$. Defining $B' = B - t$ and $C' = C + t$, we obtain $A = B' \oplus C'$. Because $\min(B') = 0$, we obtain

$$0 = \min(A) = \min(B') + \min(C') = 0 + \min(C') = \min(C')$$

and so B' and C' are sets of nonnegative integers with $0 \in B' \cap C'$.

Not every set with a composite number of elements is decomposable. For example, the n-element set $\{0, 1, 2, 2^2, \ldots, 2^{n-2}\}$ is indecomposable for every $n \geq 2$. This is a special case of the following result.

Lemma 5 *Let $m \geq 2$. Let A be a set of integers that contains integers a_0 and a_1 such that $a_0 \not\equiv a_1 \pmod{m}$, and $a \equiv a_0 \pmod{m}$ for all $a \in A \setminus \{a_1\}$. The set A is indecomposable.*

Proof The distinct congruence classes $a_0 \pmod{m}$ and $a_1 \pmod{m}$ contain elements of A. Let B and C be sets of integers such that $A = B + C$ with $|B|, |C| \geq 2$. If B is contained in the congruence class $r \pmod{m}$ and C is contained in the congruence class $s \pmod{m}$, then $B + C$ is contained in the congruence class $r + s$

$\pmod{m}$, and so $A \neq B + C$ (because A intersects two congruence classes). Therefore, at least one of the sets B and C must contain elements from distinct congruence classes modulo m. Let $b_1, b_2 \in B$ with $b_1 \not\equiv b_2 \pmod{m}$, and let $c_1, c_2 \in C$ with $c_1 \neq c_2$. We have $b_i + c_1 \in B + C$ for $i = 1, 2$ and $b_1 + c_1 \not\equiv b_2 + c_1 \pmod{m}$. Because A intersects only two congruence classes modulo m, and because the intersection with the congruence class $a_1 \pmod{m}$ contains only the integer a_1, we must have $b_i + c_1 = a_1$ for some $i \in \{1, 2\}$.

Similarly, $b_j + c_2 \in B + C$ for $j = 1, 2$ with $b_1 + c_2 \not\equiv b_2 + c_2 \pmod{m}$, and so $b_j + c_2 = a_1$ for some $j \in \{1, 2\}$. The equation $b_i + c_1 = b_j + c_2$ implies that $A \neq B \oplus C$. This completes the proof.

The following examples show that, in Lemma 5, the condition that the set A contains exactly one element of the congruence class $a_1 \pmod{m}$ is necessary.

Let $m \geq 2$, and let $R \subseteq [0, m)$ with $|R| \geq 2$. For every set J of integers with $|J| \geq 2$, we have

$$A = \{jm + r : j \in J \text{ and } r \in R\} = B \oplus C$$

where

$$B = \{jm : j \in J\} \qquad \text{and} \qquad C = R.$$

Let k, ℓ, and m be integers with $k \geq 2$, $\ell \geq 2$, and $m \geq 2$, and let u and v be integers such that $u \not\equiv v \pmod{m}$. Consider the set

$$A = \{im + u : i \in [0, \ell)\} \cup \{jm + v : j \in [0, k\ell)\}.$$

The sets

$$B = \{u\} \cup \{q\ell m + v : q \in [0, k)\}$$

and

$$C = \{im : i \in [0, \ell)\}$$

satisfy $|B| = 1 + k\ell \geq 2$, $|C| = \ell \geq 2$ and

$$A = B \oplus C.$$

3 Decomposition of Additive Systems

Contraction and dilation are two methods to construct new additive systems from old ones. Decomposition is a third method to produce new additive systems.

An additive system $\mathcal{A} = (A_i)_{i \in I}$ is called *decomposable* if the set A_{i_0} is decomposable for some $i_0 \in I$ and *indecomposable* if A_i is indecomposable for all $i \in I$.

Equivalently, an indecomposable additive system is an additive system that is not a proper contraction of another additive system.

Theorem 3 *Let $\mathcal{A} = (A_i)_{i \in I}$ be a decomposable additive system, and let A_{i_0} be a decomposable set in $\mathcal{A}$. Choose sets B and C of nonnegative integers such that $0 \in B \cap C$, $|B| \geq 2$, $|C| \geq 2$, and $A_{i_0} = B \oplus C$. Let*

$$I' = \{j_1, j_2\} \cup I \setminus \{i_0\}.$$

The family of sets $\mathcal{A}' = (A'_i)_{i \in I'}$ defined by

$$A'_i = \begin{cases} A_i & \text{if } i \in I \setminus \{i_0\} \\ B & \text{if } i = j_1 \\ C & \text{if } i = j_2 \end{cases}$$

is an additive system.

Proof This follows immediately from the definitions of additive system and indecomposable set.

We call $\mathcal{A}'$ a *decomposition* of the additive system $\mathcal{A}$.

Lemma 6 *Let a and b be positive integers, and let X be a set of integers. Then*

$$[0, ab) = [0, a) \oplus X \tag{2}$$

if and only if

$$X = a * [0, b).$$

Proof The division algorithm implies that $[0, ab) = [0, a) \oplus a * [0, b)$, and so $X = a * [0, b)$ is a solution of the additive set equation (2).

Conversely, let X be any solution of (2). Let $I = \{1, 2, 3\}$ and let $A_1 = [0, a)$, $A_2 = X$, and $A_3 = ab * N_0$. By the division algorithm, $\mathcal{A} = (A_i)_{i \in I}$ is an additive system. Applying Lemma 3 to $\mathcal{A}$, we obtain an integer $g \geq 2$ and sets B_1, B_2, and B_3 such that

$$\begin{aligned} [0, a) &= [0, g) \oplus g * B_1 \\ X &= g * B_2 \\ ab * \mathbf{N}_0 &= g * B_3. \end{aligned}$$

It follows that $g = a$, $B_1 = \{0\}$, $B_3 = b * \mathbf{N}_0$, and

$$\mathbf{N}_0 = B_2 \oplus B_3 = B_2 \oplus b * \mathbf{N}_0.$$

This implies that $B_2 = [0, b)$ and $X = a * [0, b)$.

There is also a nice polynomial proof of Lemma 6. Let

$$\begin{aligned} f(t) &= \sum_{i\in[0,ab)} t^i \\ g(t) &= \sum_{j\in[0,a)} t^j \\ h(t) &= \sum_{k\in[0,b)} t^{ak} \\ h_X(t) &= \sum_{x\in X} t^x. \end{aligned}$$

The set equation $[0, ab) = [0, a) \oplus a * [0, b)$ implies that

$$f(t) = g(t)h(t).$$

If $[0, ab) = [0, a) \oplus X$, then

$$f(t) = g(t)h_X(t)$$

and so

$$g(t)(h(t) - h_X(t)) = 0.$$

Because $g(t) \neq 0$, it follows that $h(t) = h_X(t)$ or, equivalently, $a * [0, b) = X$.

By Theorem 2, every additive system is a British number system or a proper contraction of a British number system. However, a British number system can also be a proper contraction of another British number system. Consider, for example, the British number systems $\mathcal{A}_2$ and $\mathcal{A}_4$ generated by the sequences $(2)_{i\in\mathbf{N}}$ and $(4)_{i\in\mathbf{N}}$, respectively:

$$\begin{aligned} \mathcal{A}_2 &= (\{0, 2^{i-1}\})_{i\in\mathbf{N}} = (2^{i-1} * [0, 2))_{i\in\mathbf{N}} \\ &= (\{0, 1\}, \{0, 2\}, \{0, 4\}, \{0, 8\}, \ldots) \end{aligned}$$

and

$$\begin{aligned} \mathcal{A}_4 &= (\{0, 4^{i-1}, 2 \cdot 4^{i-1}, 3 \cdot 4^{i-1}\})_{i\in\mathbf{N}} = (4^{i-1} * [0, 4))_{i\in\mathbf{N}} \\ &= (\{0, 1, 2, 3\}, \{0, 4, 8, 12\}, \{0, 16, 32, 48\}, \{0, 64, 128, 192, 256\}, \ldots) . \end{aligned}$$

Because

$$4^{i-1} * [0, 4) = \{0, 2^{2i-2}\} + \{0, 2^{2i-1}\} = 2^{2i-2} * [0, 2) + 2^{2i-1} * [0, 2)$$

we see that $\mathcal{A}_4$ is a contraction of $\mathcal{A}_2$.

de Bruijn [1] asserted the following necessary and sufficient condition for one British number system to be a contraction of another British number system.

Theorem 4 *Let $\mathcal{B} = (B_j)_{j\in\mathbf{N}}$ be the British number system constructed from the integer sequence $(h_j)_{j\in\mathbf{N}}$, and let $\mathcal{A} = (A_i)_{i\in\mathbf{N}}$ be the contraction of $\mathcal{B}$ constructed from a partition $(J_i)_{i\in\mathbf{N}}$ of* **N** *into nonempty finite sets. Then, $\mathcal{A}$ is a British number system if and only if J_i is a finite interval of integers for all $i \in \mathbf{N}$.*

Proof Let $(J_i)_{i\in\mathbf{N}}$ be a partition of **N** into nonempty finite intervals of integers. After re-indexing, there is a strictly increasing sequence $(u_i)_{i\in\mathbf{N}_0}$ of integers with $u_0 = 0$ such that $J_i = [u_{i-1} + 1, u_i]$ for all $i \in \mathbf{N}$.

If $\mathcal{B} = (B_j)_{j\in\mathbf{N}}$ is the British number system constructed from the integer sequence $(h_j)_{j\in\mathbf{N}}$, then $B_j = H_{j-1} * [0, h_j)$, where $H_0 = 1$ and $H_j = \prod_{k=1}^{j} h_k$. Let $G_0 = 1$. For $i \in \mathbf{N}$ we define

$$g_i = \frac{H_{u_i}}{H_{u_{i-1}}}$$

and

$$G_i = \prod_{j=1}^{i} g_j = \prod_{j=1}^{i} \frac{H_{u_j}}{H_{u_{j-1}}} = H_{u_i}.$$

We have

$$\begin{aligned}
A_i = \bigoplus_{j\in J_i} B_j &= \bigoplus_{j=u_{i-1}+1}^{u_i} H_{j-1} * [0, h_j) \\
&= H_{u_{i-1}} * \bigoplus_{j=u_{i-1}+1}^{u_i} \frac{H_{j-1}}{H_{u_{i-1}}} * [0, h_j) \\
&= H_{u_{i-1}} * \big([0, h_{u_{i-1}+1}) + h_{u_{i-1}+1} * [0, h_{u_{i-1}+2}) \\
&\quad + h_{u_{i-1}+1} h_{u_{i-1}+2} * [0, h_{u_{i-1}+3}) + \cdots \\
&\quad + h_{u_{i-1}+1} \cdots h_{u_i - 1} * [0, h_{u_i})\big) \\
&= H_{u_{i-1}} * \left[0, \frac{H_{u_i}}{H_{u_{i-1}}}\right) \\
&= G_{i-1} * [0, g_i)
\end{aligned}$$

and so $\mathcal{A} = (A_i)_{i\in\mathbf{N}}$ is the British number system constructed from the integer sequence $(g_i)_{i\in\mathbf{N}}$.

Conversely, let $\mathcal{A} = (A_i)_{i\in\mathbf{N}}$ be a contraction of $\mathcal{B}$ constructed from a partition $(J_i)_{i\in\mathbf{N}}$ of **N** in which some set J_{i_0} is a not a finite interval of integers. Let $u = \min\left(J_{i_0}\right)$ and $w = \max\left(J_{i_0}\right)$. Because J_{i_0} is not an interval, there is a smallest integer v such that

$$u < v < w$$

and $[u, v-1] \subseteq J_{i_0}$, but $v \notin I_{j_0}$. Because

$$[u, v-1] \cup \{w\} \subseteq J_{i_0} \subseteq [u, v-1] \cup [v+1, w]$$

and

$$A_{i_0} = \sum_{j \in J_{i_0}} H_{j-1} * [0, h_j)$$

we have

$$\begin{aligned} H_{u-1} * [0, h_u) \cup H_{w-1} * [0, h_w) &\subseteq A_{i_0} \\ &\subseteq \sum_{j \in [u, v-1]} H_{j-1} * [0, h_j) + \sum_{j \in [v+1, w]} H_{j-1} * [0, h_j) \\ &\subseteq H_{u-1} * \left[0, \frac{H_{v-1}}{H_{u-1}}\right) + H_v * \left[0, \frac{H_w}{H_v}\right) \end{aligned}$$

Because $h_u \geq 2$ and $h_v \geq 2$, it follows that

$$H_{u-1} \in A_{i_0}$$

and

$$H_{w-1} = H_{u-1}\left(\frac{H_{w-1}}{H_{u-1}}\right) \in A_{i_0}.$$

The largest multiple of H_{u-1} in $H_{u-1} * \left[0, H_{v-1}/H_{u-1}\right)$ is $H_{u-1}(H_{v-1}/H_{u-1} - 1)$. The smallest positive multiple of H_{u-1} in $H_v * [0, H_w/H_v)$ is $H_v = H_{u-1}(H_v/H_{u-1})$. The inequality

$$1 \leq \frac{H_{v-1}}{H_{u-1}} - 1 < \frac{H_{v-1}}{H_{u-1}} < \frac{H_v}{H_{u-1}} \leq \frac{H_{w-1}}{H_{u-1}}$$

implies that the set A_{i_0} does not contain the integer $H_{u-1}(H_{v-1}/H_{u-1})$. In a British number system, every set consists of consecutive multiples of its smallest positive element. Because the set A_{i_0} lacks this property, it follows that $\mathcal{A}$ is not a British number system. This completes the proof.

Theorem 5 *There is a one-to-one correspondence between sequences $(p_i)_{i \in \mathbf{N}}$ of prime numbers and indecomposable British number systems. Moreover, every additive system is either indecomposable or a contraction of an indecomposable system.*

Proof Let $\mathcal{A}$ be a British number system generated by the sequence $(g_i)_{i \in \mathbf{N}}$, so that

$$\mathcal{A} = (G_{i-1} * [0, g_i))_{i \in \mathbf{N}}.$$

Suppose that g_k is composite for some $k \in \mathbf{N}$. Then $g_k = rs$, where $r \geq 2$ and $s \geq 2$ are integers. Construct the sequence $(g'_i)_{i \in \mathbf{N}}$ as follows:

$$g'_i = \begin{cases} g_i & \text{if } i \leq k-1 \\ r & \text{if } i = k \\ s & \text{if } i = k+1 \\ g_{i-1} & \text{if } i \geq k+2. \end{cases}$$

Then,

$$G'_i = \prod_{j=1}^{i} g'_j = \begin{cases} G_i & \text{if } i \leq k-1 \\ rG_{k-1} & \text{if } i = k \\ G_k & \text{if } i = k+1 \\ G_{i-1} & \text{if } i \geq k+2 \end{cases}$$

and

$$\mathcal{A}' = \left(G'_{i-1} * [0, g'_i)\right)_{i \in \mathbf{N}}$$

is the British number system generated by the sequence $(g'_i)_{i \in \mathbf{N}}$. We have

$$G_{i-1} * [0, g_i) = \begin{cases} G'_{i-1} * [0, g'_i) & \text{if } i \leq k-1 \\ G'_i * [0, g'_{i+1}) & \text{if } i \geq k+1. \end{cases}$$

The identity

$$[0, g_k) = [0, rs) = [0, r) \oplus r * [0, s) = [0, g'_k) + \frac{G'_k}{G_{k-1}} * [0, g'_{k+1})$$

implies that

$$G_{k-1} * [0, g_k) = G'_{k-1} * [0, g'_k) + G'_k * [0, g'_{k+1}) = \sum_{i \in \{k, k+1\}} G'_{i-1} * [0, g'_i)$$

and so the British number system $\mathcal{A}$ is a contraction of the British number system $\mathcal{A}'$.

Conversely, if $\mathcal{A}$ is a contraction of a British number system $\mathcal{A}' = \left(G'_{i-1} * [0, g'_i)\right)_{i \in \mathbf{N}}$, then there are a positive integer k and a set I_k of positive integers with $|I_k| \geq 2$ such that

$$G_{k-1} * [0, g_k) = \sum_{i \in I_k} G'_{i-1} * [0, g'_i).$$

Therefore,

$$g_k = |G_{k-1} * [0, g_k)| = \left| \sum_{i \in I_k} G'_{i-1} * [0, g'_i) \right| = \prod_{i \in I_k} g'_i.$$

Because $|I_k| \geq 2$ and $|g'_i| \geq 2$ for all $i \in \mathbf{N}$, it follows that the integer g_k is composite. Thus, the British number system generated by $(g_i)_{i\in\mathbf{N}}$ is decomposable if and only if g_i is composite for at least one $i \in \mathbf{N}$. Equivalently, the British number system generated by $(g_i)_{i\in\mathbf{N}}$ is indecomposable if and only if $(g_i)_{i\in\mathbf{N}}$ is a sequence of prime numbers. This completes the proof.

Theorem 5 has also been observed by Munagi [3].

4 Limits of Additive Systems

Let $\mathcal{A} = (A_i)_{i\in\mathbf{N}_0}$ be an additive system, and let $(g_i)_{i\in[1,n]}$ be a finite sequence of integers with $g_i \geq 2$ for all $i \in [1, n]$. The *dilation* of $\mathcal{A}$ by the sequence $(g_i)_{i\in[1,n]}$ is the additive system defined inductively by

$$(g_i)_{i\in[1,n]} * \mathcal{A} = g_1 * \left((g_i)_{i\in[2,n]} * \mathcal{A}\right).$$

For $n = 1$, we have

$$\begin{aligned}\mathcal{A}^{(1)} &= (g_i)_{i\in[1,1]} * \mathcal{A} = g_1 * \mathcal{A}\\ &= [0, g_1) \cup (g_1 * A_i)_{i\in\mathbf{N}_0}\\ &= \left(A_i^{(1)}\right)_{i\in\mathbf{N}_0}\end{aligned}$$

where

$$A_1^{(1)} = [0, g_1)$$

and

$$A_i^{(1)} = g_1 * A_{i-1} \text{ for} i \geq 2.$$

For $n = 2$, we have

$$\begin{aligned}\mathcal{A}^{(2)} &= (g_i)_{i\in[1,2]} * \mathcal{A} = g_1 * (g_2 * \mathcal{A})\\ &= g_1 * \left([0, g_2) \cup (g_2 * A_i)_{i\in\mathbf{N}_0}\right)\\ &= [0, g_1) \cup (g_1 * [0, g_2)) \cup (g_1 g_2 * A_i)_{i\in\mathbf{N}_0}\\ &= \left(A_i^{(2)}\right)_{i\in\mathbf{N}_0}\end{aligned}$$

where

$$\begin{aligned}A_1^{(2)} &= [0, g_1)\\ A_2^{(2)} &= g_1 * [0, g_2)\end{aligned}$$

and

$$A_i^{(2)} = g_1 g_2 * A_{i-2} \text{ for } i \geq 3.$$

For $n = 3$, we have

$$\begin{aligned} g_3 * \mathcal{A} &= [0, g_3) \cup (g_3 * A_i)_{i \in \mathbf{N}_0} \\ g_2 * (g_3 * \mathcal{A}) &= [0, g_2) \cup g_2 * [0, g_3) \cup (g_2 g_3 * A_i)_{i \in \mathbf{N}_0} \end{aligned}$$

and

$$\begin{aligned} \mathcal{A}^{(3)} &= (g_i)_{i \in [1,3]} * \mathcal{A} = g_1 * (g_2 * (g_3 * \mathcal{A})) \\ &= [0, g_1) \cup (g_1 * [0, g_2)) \cup (g_1 g_2 * [0, g_3)) \cup (g_1 g_2 g_3 * A_i)_{i \in \mathbf{N}_0} \\ &= \left(A_i^{(3)} \right)_{i \in \mathbf{N}_0} \end{aligned}$$

where

$$\begin{aligned} A_1^{(3)} &= [0, g_1) \\ A_2^{(3)} &= g_1 * [0, g_2) \\ A_3^{(3)} &= g_1 g_2 * [0, g_3) \\ A_i^{(3)} &= g_1 g_2 g_3 A_{i-3} \qquad \text{for } i \geq 4. \end{aligned}$$

Lemma 7 *Let $(g_i)_{i=1}^n$ be a sequence of integers such that $g_i \geq 2$ for all i. For every additive system $\mathcal{A} = (A_i)_{i \in \mathbf{N}}$,*

$$\mathcal{A}^{(n)} = (g_i)_{i=1}^n * \mathcal{A} = \left(A_i^{(n)} \right)_{i \in \mathbf{N}}$$

where

$$A_i^{(n)} = g_1 g_2 \cdots g_{i-1} * [0, g_i) \qquad \textit{for } i = 1, \ldots, n$$

and

$$A_i^{(n)} = g_1 g_2 \cdots g_n * A_{i-n-1} \qquad \textit{for } i \geq n + 1.$$

Proof Induction on n.

Let $(\mathcal{A}^{(n)})_{n \in \mathbf{N}}$ be a sequence of additive systems. The additive system $\mathcal{A}$ is the *limit* of the sequence $(\mathcal{A}^{(n)})_{n \in \mathbf{N}}$ if it satisfies the following condition: The set S belongs to $\mathcal{A}$ if and only if S belongs to $\mathcal{A}^{(n)}$ for all sufficiently large n. We write

$$\lim_{n \to \infty} \mathcal{A}^{(n)} = \mathcal{A}$$

if $\mathcal{A}$ is the limit of the sequence $(\mathcal{A}^{(n)})_{n \in \mathbf{N}}$. The following result indicates the remarkable stability of a British number system.

Theorem 6 *Let $(g_i)_{i \in \mathbf{N}}$ be a sequence of integers such that $g_i \geq 2$ for all $i \in \mathbf{N}$, and let $\mathcal{G}$ be the British number system generated by $(g_i)_{i \in \mathbf{N}}$. Let $\mathcal{A}$ be an additive system and let $\mathcal{A}^{(n)} = (g_i)_{i \in [1,n]} * \mathcal{A}$. Then,*

$$\lim_{n \to \infty} \mathcal{A}^{(n)} = \mathcal{G}.$$

Proof If S is a set in $\mathcal{G}$, then $S = g_1 g_2 \cdots g_{i-1} * [0, g_i)$ for some $i \in \mathbf{N}$. By Lemma 7, S is a set in $\mathcal{A}^{(n)}$ for all $n \geq i$, and so $S \in \lim_{n \to \infty} \mathcal{A}^{(n)}$.

Conversely, let S be a set that is in $\mathcal{A}^{(n)}$ for all sufficiently large n. If S is finite, then $\max(S) < g_1 g_2 \cdots g_k$ for some integer k. If $n \geq k$ and $i \geq n + 1$, then

$$\max\left(A_i^{(n)}\right) \geq g_1 \ldots g_n \geq g_1 \ldots g_k$$

and so $S \neq A_i^{(n)}$. Therefore, $S = A_i^{(n)}$ for some $i \leq n$, and so $S = g_1 g_2 \cdots g_{i-1} * [0, g_i)$ for some $i \leq n$.

If T is an infinite set in $\mathcal{A}^{(n)}$, then $T = g_1 g_2 \cdots g_n * A_{i-n-1}$ for some $i \geq n + 1$, and so $\min(T \setminus \{0\}) \geq g_1 g_2 \cdots g_n \geq 2^n$. If $T \in \mathcal{A}^{(n)}$ for all $n \geq N$, then $\min(T \setminus \{0\}) \geq 2^n$ for all $n \geq N$, which is absurd. It follows that the set S is in $\mathcal{A}^{(n)}$ for all sufficiently large n if and only if S is finite and S is a set in the British number system generated by $(g_i)_{i \in \mathbf{N}}$. This completes the proof.

Corollary 8 *Let $(g_i)_{i \in \mathbf{N}}$ be a sequence of integers such that $g_i \geq 2$ for all $i \in \mathbf{N}$, and let $\mathcal{G}$ be the British number system generated by $(g_i)_{i \in \mathbf{N}}$. If $\mathcal{G}_n = (g_i)_{i \in [1,n]} * \mathbf{N}_0$, then*

$$\lim_{n \to \infty} \mathcal{G}_n = \mathcal{G}.$$

Acknowledgements Supported in part by a grant from the PSC-CUNY Research Award Program.

References

1. N.G. de Bruijn, On number systems. Nieuw Arch. Wisk. **4**(3), 15–17 (1956)
2. M. Maltenfort, Characterizing additive systems. Am. Math. Monthly **124**, 132–148 (2017)
3. A.O. Munagi, *k*-complementing subsets of nonnegative integers. Int. J. Math. Math. Sci. 215–224 (2005)
4. M.B. Nathanson, Additive systems and a theorem of de Bruijn. arXiv:1301.6208 (2013)

Extending Babbage's (Non-)Primality Tests

Jonathan Sondow

Abstract We recall Charles Babbage's 1819 criterion for primality, based on simultaneous congruences for binomial coefficients, and extend it to a least-prime-factor test. We also prove a partial converse of his non-primality test, based on a single congruence. Along the way we encounter Bachet, Bernoulli, Bézout, Euler, Fermat, Kummer, Lagrange, Lucas, Vandermonde, Waring, Wilson, Wolstenholme, and several contemporary mathematicians.

Keywords Charles Babbage · Primality test · Binomial coefficient · Congruence Wolstenholme prime · Lucas's theorem

1 Introduction

Charles Babbage was an English mathematician, philosopher, inventor, mechanical engineer, and "irascible genius" who pioneered computing machines [2, 4, 10, 21–23]. Although he held the Lucasian Chair of Mathematics at Cambridge University from 1828 to 1839, during that period he never resided in Cambridge or delivered a lecture [5, 7, p. 7].

In 1819, he published his only work on number theory, a short paper [1] that begins:

> The singular theorem of Wilson respecting Prime Numbers, which was first published by Waring in his *Meditationes Analyticae* [31, p. 218], and to which neither himself nor its author could supply the demonstration, excited the attention of the most celebrated analysts of the continent, and to the labors of Lagrange [14] and Euler we are indebted for several modes of proof . . . .

Babbage formulated **Wilson's theorem** as a criterion for primality: *an integer* $p > 1$ *is a prime if and only if* $(p-1)! \equiv -1 \pmod{p}$. (For a modern proof, see Moll [20, p. 66].) He then introduced several such criteria, involving congruences for

J. Sondow (✉)
209 West 97th Street, 10025 New York, NY, USA
e-mail: jsondow@alumni.princeton.edu

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_19

binomial coefficients (see Granville [11, Sections 1 and 4]). However, some of his claims were unproven or even wrong (as Dubbey points out in [7, pp. 139–141]). One of his valid results is a necessary and sufficient condition for primality, based on a number of simultaneous congruences. Henceforth, let n denote an integer.

Theorem 1 (Babbage's Primality Test) *An integer $p > 1$ is a prime if and only if*

$$\binom{p+n}{n} \equiv 1 \pmod{p} \tag{1}$$

for all n satisfying $0 \leq n \leq p-1$.

This is of only theoretical interest, the test being slower than trial division.

The "only if" part is an immediate consequence of the beautiful **theorem of Lucas** [15] (see [8, 11, 17, 19] and [20, p. 70]), which asserts that *if p is a prime and the non-negative integers $a = \alpha_0 + \alpha_1 p + \cdots + \alpha_r p^r$ and $b = \beta_0 + \beta_1 p + \cdots + \beta_r p^r$ are written in base p (so that* $0 \leq \alpha_i, \beta_i \leq p-1$ for all i), then

$$\binom{a}{b} \equiv \prod_{i=0}^{r} \binom{\alpha_i}{\beta_i} \pmod{p}. \tag{2}$$

(Here the convention is that $\binom{\alpha}{\beta} = 0$ if $\alpha < \beta$.) The congruence (1) follows if $0 \leq n \leq p-1$, for then all the binomial coefficients formed on the right-hand side of (2) are of the form $\binom{\alpha}{\alpha} = 1$, except the last one, which is $\binom{1}{0} = 1$.

However, the theorem was not available to Babbage because when it was published in 1878 he had been dead for seven years.

Lucas's theorem implies more generally that *for p a prime and m a power of p*, the congruences

$$\binom{m+n}{n} \equiv 1 \pmod{p} \qquad (0 \leq n \leq m-1) \tag{3}$$

hold. A converse was proven in 2013: **Meštrović's theorem** [19] states that *if $m > 1$ and $p > 1$ are integers such that* (3) *holds, then p is a prime and m is a power of p*. To begin the proof, Meštrović noted that for $n = 1$, the hypothesis gives

$$\binom{m+1}{1} = m + 1 \equiv 1 \pmod{p} \quad \Longrightarrow \quad p \mid m.$$

The rest of the proof involves combinatorial congruences modulo prime powers.

As Meštrović pointed out, "the 'if' part of Theorem 1 is an immediate consequence of [his theorem] (supposing a priori [that $m = p$]). Accordingly, [his theorem] may be considered as a generalization of Babbage's criterion for primality."

Here we offer another generalization of Babbage's primality test.

Theorem 2 (Least-Prime-Factor Test) *The least prime factor of an integer $m > 1$ is the smallest natural number ℓ satisfying*

$$\binom{m+\ell}{\ell} \not\equiv 1 \pmod{m}. \tag{4}$$

For that value of ℓ, *the least non-negative residue of* $\binom{m+\ell}{\ell}$ *modulo* m *is* $\frac{m}{\ell}+1$.

The proof is given in Sect. 2.

Babbage's primality test is an easy corollary of the least-prime-factor test. Indeed, Theorem 2 implies a sharp version of Theorem 1 noticed by Granville [11] in 1995.

Corollary 1 (Sharp Babbage Primality Test) *Theorem 1 remains true if the range for n is shortened to* $0 \le n \le \sqrt{p}$.

Proof An integer $m > 1$ is a prime if and only if its least prime factor ℓ exceeds $\sqrt{m}$. The corollary follows by setting $m = p$ in Theorem 2. □

To see that *Corollary 1 is sharp in that the range for n cannot be further shortened to* $0 \le n \le \sqrt{p} - 1$, let q be any prime and set $p = q^2$. Then p is not a prime, but the least-prime-factor test with $m = p$ and $\ell = q$ implies (1) when $0 \le n \le q - 1$.

Problem 1 Since the "if" part of Babbage's primality test is a consequence both of Meštrović's theorem and of the least-prime-factor test, one may ask, *Is there a common generalization of Meštrović's theorem and Theorem 2*? (Note, though, that the modulus in the former is p, while that in the latter is m.)

Actually, the incongruence (4) holds more generally if the *least* prime factor $\ell \mid m$ is replaced with *any* prime factor $p \mid m$. The following extension of the least-prime-factor test is proven in Sect. 2.

Theorem 3 (i) *Given a positive integer m and a prime factor* $p \mid m$, *we have*

$$\binom{m+p}{p} \not\equiv 1 \pmod{m}. \tag{5}$$

(ii) *If in addition* $p^r \mid m$ *but* $p^{r+1} \nmid m$, *where* $r \ge 1$, *then*

$$\binom{m+p}{p} \equiv \frac{m}{p} + 1 \not\equiv 1 \pmod{p^r}. \tag{6}$$

Part (i) is clearly equivalent to the statement that *if* $d > 1$ *divides m and* $\binom{m+d}{d} \equiv 1 \pmod{m}$, *then d is composite.* As an example, for $m = 260$ and $d = 10$, we have

$$\binom{m+d}{d} = \binom{270}{10} = 479322759878148681 \equiv 1 \pmod{260}.$$

The sequence of integers $m > 1$, for which some integer d (necessarily composite) satisfies

$$d > 1, \qquad d \mid m, \qquad \binom{m+d}{d} \equiv 1 \pmod{m},$$

begins [28, Seq. A290040]

$$m = 260, 1056, 1060, 3460, 3905, 4428, 5000, 5060, 5512, 5860, 6372, 6596, \ldots$$

and the sequence of smallest such divisors d is, respectively, [28, Seq. A290041]

$$d = 10, 264, 10, 10, 55, 18, 20, 10, 52, 10, 18, 34, \ldots . \tag{7}$$

Problem 2 Does Theorem 3 extend to prime power factors, i.e., does (5) also hold when p is replaced with p^k, where $p^k \mid m$ and $k > 1$? In particular, in the sequence (7), is any term d a prime power?

Babbage also claimed a necessary and sufficient condition for primality based on a *single* congruence. But he proved only necessity, so we call it a test for non-primality.

Theorem 4 (Babbage's Non-Primality Test) *An integer $m \geq 3$ is composite if*

$$\binom{2m-1}{m-1} \not\equiv 1 \pmod{m^2}. \tag{8}$$

Our version of his proof is given in Sect. 3.

Not only did Babbage not prove the claimed converse, but in fact it is false. Indeed, *the numbers $m_1 = p_1^2 = 283686649$ and $m_2 = p_2^2 = 4514260853041$ are composite but do not satisfy* (8), where $p_1 = 16843$ and $p_2 = 2124679$ are primes.

Here p_1 (indicated by Selfridge and Pollack in 1964) and p_2 (discovered by Crandall, Ernvall, and Metsänkylä in 1993) are *Wolstenholme primes*, so called by Mcintosh [16] because, while **Wolstenholme's theorem** [32] (see [11, 18, 29] and [20, p. 73]) of 1862 guarantees that *every prime $p \geq 5$ satisfies*

$$\binom{2p-1}{p-1} \equiv 1 \pmod{p^3}, \tag{9}$$

in fact p_1 and p_2 satisfy the congruence in (9) modulo p^4, not just p^3 (see Guy [12, p. 131] and Ribenboim [25, p. 23]).

Note that (9) strengthens Babbage's non-primality test, as Theorem 4 is equivalent to the statement that *the congruence in* (9) *holds modulo p^2 for any prime $p \geq 3$.*

In their solutions to a problem by Segal in the *Monthly*, Brinkmann [26] and Johnson [27] made Babbage's and Wolstenholme's theorems more precise by showing that *every prime $p \geq 5$ satisfies the congruences*

$$\binom{2p-1}{p-1} \equiv 1 - \frac{2}{3} p^3 B_{p-3} \equiv \binom{2p^2-1}{p^2-1} \pmod{p^4},$$

where B_k denotes the kth *Bernoulli number*, a rational number. (See also Gardiner [9] and Mcintosh [16].) Thus, *a prime* $p \geq 5$ *is a Wolstenholme prime if and only if* $B_{p-3} \equiv 0 \pmod{p}$. (The congruence means that p divides the numerator of B_{p-3}.) In that case, the square of that prime, say $m = p^2$, is composite but must satisfy

$$\binom{2m-1}{m-1} \equiv 1 \pmod{m^2},$$

thereby providing a counterexample to the converse of Babbage's non-primality test.

Johnson [27] commented that "interest in [Wolstenholme primes] arises from the fact that in 1857, Kummer proved that the first case of [Fermat's Last Theorem] is true for all prime exponents p such that $p \nmid B_{p-3}$."

We have seen that the converse of Babbage's non-primality test is false. The converse of Wolstenholme's theorem is the statement that *if* $p \geq 5$ *is composite, then* (9) *does not hold.* It is not known whether this is generally true. A proof that it is true for *even* positive integers was outlined by Trevisan and Weber [29] in 2001. In Sect. 3, we fill in some details omitted from their argument and extend it to prove the following stronger result.

Theorem 5 (Converse of Babbage's Non-Primality Test for Even Numbers) *If a positive integer m is even, then*

$$\binom{2m-1}{m-1} \not\equiv 1 \pmod{m^2}. \tag{10}$$

2 Proofs of the Least-Prime-Factor Test and Its Extension

We prove Theorems 2 and 3. The arguments use only mathematics available in Babbage's time.

Proof (Theorem 2) As ℓ is the smallest prime factor of m, if $0 < k < \ell$ then $k!$ and m are coprime. In that case, **Bézout's identity** (proven in 1624 by Bachet in a book with the charming title *Pleasant and Delectable Problems* [3, p. 18, Proposition XVIII]—see [6, Section 4.3]) gives integers a and b with $ak! + bm = 1$. Multiplying Bézout's equation by the number $\binom{m}{k} = m(m-1)\cdots(m-k+1)/k!$ yields

$$am(m-1)\cdots(m-k+1) + bm\binom{m}{k} = \binom{m}{k},$$

so $\binom{m}{k} \equiv 0 \pmod{m}$ if $1 \leq k \leq \ell - 1$. Now, for $n = 0, 1, \ldots, \ell - 1$, **Vandermonde's convolution** [30] (see [20, p. 164]) of 1772 gives

$$\binom{m+n}{n} = \sum_{k=0}^{n} \binom{m}{k}\binom{n}{n-k} \equiv \binom{m}{0}\binom{n}{n} \pmod{m}.$$

(To see the equality, equate the coefficients of x^n in the expansions of $(1+x)^{m+n}$ and $(1+x)^m(1+x)^n$). Thus, we arrive at the congruences

$$\binom{m+n}{n} \equiv 1 \pmod{m} \qquad (0 \le n \le \ell - 1). \tag{11}$$

On the other hand, from the identity

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1} \tag{12}$$

(to prove it, use factorials), the congruence (11) for $n = \ell - 1$, the integrality of $\frac{m+\ell}{\ell} = \frac{m}{\ell} + 1$, and the inequality $\ell > 1$ (as ℓ is a prime), we deduce that

$$\binom{m+\ell}{\ell} = \frac{m+\ell}{\ell}\binom{m+\ell-1}{\ell-1} \equiv \frac{m}{\ell} + 1 \not\equiv 1 \pmod{m}.$$

Together with (11), this implies the least-prime-factor test. □

Proof (Theorem 3) It suffices to prove (ii). Set

$$g \overset{\text{def}}{=} \gcd((p-1)!, m) \qquad \text{and} \qquad m_p \overset{\text{def}}{=} \frac{m}{g}.$$

Note that

$$p \text{ prime} \implies p \nmid g \implies p^r \mid m_p, \tag{13}$$

since $p^r \mid m$. Bézout's identity gives integers a and b with $a(p-1)! + bm = g$. When $0 < k < p$, multiplying Bézout's equation by $\binom{m}{k}$ yields

$$am(m-1)\cdots(m-k+1)\frac{(p-1)!}{k!} + bm\binom{m}{k} = g\binom{m}{k}$$

with $(p-1)!/k!$ an integer, so $g\binom{m}{k} \equiv 0 \pmod{m}$. Dividing by g gives

$$\binom{m}{k} \equiv 0 \pmod{m_p} \quad (1 \le k \le p-1).$$

Combining this with (12) and Vandermonde's convolution, we get

$$\begin{aligned}\binom{m+p}{p} = \frac{m+p}{p}\binom{m+p-1}{p-1} &= \frac{m+p}{p}\sum_{k=0}^{p-1}\binom{m}{k}\binom{p-1}{p-1-k} \\ &\equiv \frac{m}{p} + 1 \pmod{m_p}.\end{aligned} \tag{14}$$

As $p^{r+1} \nmid m$, we have $p^r \nmid \frac{m}{p}$. Now, (13) and (14) imply (6), as required. □

3 Proofs of Babbage's Non-primality Test and Its Converse for Even Numbers

The following proof is close to the one Babbage gave.

Proof (Theorem 4) Suppose on the contrary that m is prime. If we have $1 \leq n \leq m-1$, then m divides the numerator of $\binom{m}{n} = m!/n!(m-n)!$ but not the denominator, so $\binom{m}{n} \equiv 0 \pmod{m}$. Thus, by (12) and a famous case of Vandermonde's convolution,

$$2\binom{2m-1}{m-1} = \binom{2m}{m} = \sum_{n=0}^{m}\binom{m}{n}^2 \equiv 1^2 + 1^2 \equiv 2 \pmod{m^2}.$$

But as $m \geq 3$ is odd, (3) contradicts (8). Therefore, m is composite. □

Before giving the proof of Theorem 5, we establish two lemmas. For any positive integer k, let $2^{v(k)}$ denote the highest power of 2 that divides k.

Lemma 1 *If $m \geq n \geq 1$ are integers satisfying $n \leq 2^{v(m)}$, then the formula $v(\binom{m}{n}) = v(m) - v(n)$ holds.*

Proof Let $m = 2^r m'$ with m' odd. Note that $v(2^r m' - k) = v(k)$ if $0 < k < 2^r$. (*Proof.* Write $k = 2^t k'$, where $0 \leq t = v(k) \leq r-1$ and k' is odd. Then $2^{r-t}m' - k'$ is also odd, so $v(2^r m' - k) = v(2^t(2^{r-t}m' - k')) = t = v(k)$.) The logarithmic formula $v(ab) = v(a) + v(b)$ then implies that when $1 \leq n \leq 2^r$, the exponent of the highest power of 2 that divides the product

$$n!\binom{m}{n} = 2^r m'(2^r m' - 1)(2^r m' - 2)\cdots(2^r m' - (n-1))$$

is $v(n!) + v(\binom{m}{n}) = r + v(1 \cdot 2 \cdots (n-1))$, so $v(\binom{m}{n}) = r - v(n)$. As $r = v(m)$, this proves the desired formula. □

Lemma 1 is sharp in that the hypothesis $n \leq 2^{v(m)}$ cannot be replaced with the weaker hypothesis $v(n) \leq v(m)$. For example, $v(\binom{10}{6}) = v(210) = 1$, but $v(10) - v(6) = 0$.

Lemma 2 *A binomial coefficient $\binom{2m-1}{m-1}$ is odd if and only if $m = 2^r$ for some $r \geq 0$.*

Proof **Kummer's theorem** [13] (see [20, p. 78] or [24]) for the prime 2 states that $v(\binom{a+b}{a})$ equals the number of carries when adding a and b in base 2 arithmetic. Hence, $v(\binom{m+m}{m})$ is the number of ones in the binary expansion of m, and so $v(\binom{2m}{m}) = 1$ if and only if $m = 2^r$ for some $r \geq 0$. As $\binom{2m}{m} = 2\binom{2m-1}{m-1}$ by (12), we are done. □

We can now prove the converse of Babbage's non-primality test for even numbers.

Proof (Theorem 5) For $m \geq 2$ not a power of 2, Lemma 2 implies that $\binom{2m-1}{m-1}$ is even, so $\binom{2m-1}{m-1}$ is congruent modulo 4 to either 0 or 2. For $m \geq 2$ a power of 2, say $m = 2^r$, the equalities in (3) and the symmetry $\binom{m}{n} = \binom{m}{m-n}$ yield

$$\binom{2m-1}{m-1} = 1 + \frac{1}{2}\binom{2^r}{2^{r-1}}^2 + \sum_{k=1}^{2^{r-1}-1} \binom{2^r}{k}^2,$$

and Lemma 1 implies that $\frac{1}{2}\binom{2^r}{2^{r-1}}^2 \equiv 2 \pmod 4$ and that $\binom{2^r}{k}^2 \equiv 0 \pmod 4$ when $0 < k < 2^{r-1}$; thus, by addition $\binom{2m-1}{m-1} \equiv 3 \pmod 4$. Hence for all $m \geq 2$, we have $\binom{2m-1}{m-1} \not\equiv 1 \pmod 4$. Now as 4 divides m^2 when m is even, (10) holds a fortiori. This completes the proof. □

References

1. C. Babbage, Demonstration of a theorem relating to prime numbers, Edinburgh Phil. J. **1**, 46–49 (1819), http://books.google.com/books?id=KrA-AAAAYAAJ&pg=PA46
2. C. Babbage, *Passages from the Life of a Philosopher*, (Longman, Green, Longman, Roberts, & Green, London, 1864), http://djm.cc/library/Passages_Life_of_a_Philosopher_Babbage_edited.pdf
3. C.G. Bachet, *Problèmes plaisants et délectables, qui se font par les nombres*, 2nd edn. (Rigaud, Lyon, 1624), http://bsb3.bsb.lrz.de/~db/1008/bsb10081407/images/bsb10081407_00036
4. W.A. Beyer, Review of [7]. Am. Math. Mon. **86**, 66–67 (1979)
5. B.D. Blackwood, Charles Babbage. In: ed. by D.R. Franceschetti *Biographical Encyclopedia of Mathematicians*. (Cavendish, New York, 1998), pp. 33–36, http://www.blackwood.org/Babbage.htm
6. É. Barbin, J. Borowczyk, J.-L. Chabert, A. Djebbar, M. Guillemot, J.-C. Martzloff, A. Michel-Pajus, *A History of Algorithms: From the Pebble to the Microchip*. ed. by J.-L. Chabert. Trans. by C. Weeks (Springer, Berlin and Heidelberg, 2012)
7. J.M. Dubbey, *The Mathematical Work of Charles Babbage* (Cambridge University Press, Cambridge, 1978)
8. N.J. Fine, Binomial coefficients modulo a prime. Am. Math. Mon. **54**, 589–592 (1947)
9. A. Gardiner, Four problems on prime power divisibility. Am. Math. Mon. **95**, 926–931 (1988)
10. J. Grabiner, Review of From Newton to Hawking: A History of Cambridge University's Lucasian Professors of Mathematics by K.C. Knox, R. Noakes. Am. Math. Mon. **112**, 757–762 (2005)
11. A. Granville, Arithmetic properties of binomial coefficients I: binomial coefficients modulo prime powers. In: J. Borwein (ed), *Organic mathematics (Burnaby, BC, 1995)*. CMS Conference Proceeding Vol. 20 (American Mathematical Society, Providence, RI, 1997), pp. 253–275, http://www.dms.umontreal.ca/~andrew/Binomial/
12. R.K. Guy, *Unsolved Problems in Number Theory*, 3rd edn. (Springer, New York, 2004)
13. E. Kummer, Über die Ergänzungssätze zu den allgemeinen Reciprocitätsgesetzen. J. Reine Angew. Math. **44**, 93–146 (1852)
14. J.L. Lagrange, Démonstration d'un théorème nouveau concernant les nombres premiers, Nouv. Mém. Acad. Roy. Sci. Belles-Letters, Berlin **2**, 125–137 (1771); available at https://books.google.com/books?id=_-U_AAAAYAAJ&pg=PA125

15. É. Lucas, Sur les congruences des nombres eulériens et des coefficients différentiels des fonctions trigonométriques, suivant un module premier, Bull. Soc. Math. France **6**, 49–54 (1878), http://archive.numdam.org/ARCHIVE/BSMF/BSMF_1878__6_/BSMF_1878__6__49_1/BSMF_1878__6__49_1.pdf
16. R.J. McIntosh, On the converse of Wolstenholme's theorem. Acta Arith. **71**, 381–389 (1995)
17. R. Meštrović, A note on the congruence $\binom{nd}{md} \equiv \binom{n}{m} \pmod{q}$. Am. Math. Mon. **116**, 75–77 (2009)
18. R. Meštrović, Wolstenholme's theorem: its generalizations and extensions in the last hundred and fifty years (1862–2011), arXiv:1111.3057 [math.NT] (2011)
19. R. Meštrović, An extension of Babbage's criterion for primality, Math. Slovaca **63**, 1179–1182 (2013). http://dx.doi.org/10.2478/s12175-013-0164-8
20. V.H. Moll, *Numbers and Functions: From a Classical-Experimental Mathematician's Point of View*. Student Mathematical Library, Vol. 65 (American Mathematical Society, Providence, RI, 2012)
21. M. Moseley, *Irascible Genius: A Life of Charles Babbage, Inventor* (Hutchinson, London, 1964)
22. J.J. O'Connor, E.F. Robertson, *Charles Babbage, MacTutor History of Mathematics*, http://www-groups.dcs.st-and.ac.uk/history/Biographies/Babbage.html
23. J.T. O'Donnell, Review of Charles Babbage: Pioneer of the Computer by A. Hymanl. Am. Math. Mon. **92**, 522–525 (1985)
24. C. Pomerance, Divisors of the middle binomial coefficient. Am. Math. Mon. **122**, 636–644 (2015)
25. P. Ribenboim, *The Little Book of Bigger Primes* (Springer, New York, 2004)
26. D. Segal, H.W. Brinkmann, E435, Am. Math. Mon. **48**, 269–271 (1941)
27. D. Segal, W. Johnson, E435. Am. Math. Mon. **83**, 813 (1976)
28. Sloane, N.J.A. *The On-Line Encyclopedia of Integer Sequences*. Published electronically at http://oeis.org/ (2017)
29. V. Trevisan, K. Weber, Testing the converse of Wolstenholme's theorem. Mat. Contemp. **21**, 275–286 (2001)
30. A.-T. Vandermonde, Mémoire sur des irrationnelles de différens ordres, avec une application au cercle, Mém. Acad. Roy. Sci. Paris (1772), 489–498, http://gallica.bnf.fr/ark:/12148/bpt6k3570q/f79
31. E. Waring, *Meditationes Algebraicae* (Cambridge University Press, Cambridge, 1770)
32. J. Wolstenholme, On certain properties of prime numbers, Q. J. Pure Appl. Math. **5**, 35–39 (1862), http://books.google.com/books?id=vL0KAAAAIAAJ&pg=PA35

Conjectures on Representations Involving Primes

Zhi-Wei Sun

Abstract We pose 100 new conjectures on representations involving primes or related things, which might interest number theorists and stimulate further research. Below are five typical examples: (i) For any positive integer n, there exists $k \in \{0, \ldots, n\}$ such that $n + k$ and $n + k^2$ are both prime. (ii) Each integer $n > 1$ can be written as $x + y$ with $x, y \in \{1, 2, 3, \ldots\}$ such that $x + ny$ and $x^2 + ny^2$ are both prime. (iii) For any rational number $r > 0$, there are distinct primes $q_1, \ldots, q_k$ with $r = \sum_{j=1}^{k} 1/(q_j - 1)$. (iv) Every $n = 4, 5, \ldots$ can be written as $p + q$, where p is a prime with $p - 1$ and $p + 1$ both practical, and q is either prime or practical. (v) Any positive rational number can be written as m/n, where m and n are positive integers with $p_m + p_n$ a square (or $\pi(m)\pi(n)$ a positive square), p_k is the kth prime and $\pi(x)$ is the prime-counting function.

Keywords Conjectures · Primes · Practical numbers · Representations

2010 Mathematics Subject Classification. Primary 11A41, 11P32 · Secondary 11B13, 11D68

1 Introduction

Primes have been investigated for over 2000 years. Nevertheless, there are many problems on primes that remain open. The famous Goldbach's conjecture (cf. [2, 14]) states that any even integer $n > 2$ can be represented as a sum of two primes. Lemoine's conjecture (see [10]) asserts that any odd integer $n > 6$ can be written as $p + 2q$ with p and q both prime; this is a refinement of the weak Goldbach's

The initial version of this paper was posted to `arXiv` in November 2012 as a preprint with the ID arXiv:1211.1588.

Z.-W. Sun (✉)
Department of Mathematics, Nanjing University, Nanjing 210093,
People's Republic of China
e-mail: zwsun@nju.edu.cn

M. B. Nathanson (ed.), *Combinatorial and Additive Number Theory II*,
Springer Proceedings in Mathematics & Statistics 220,
https://doi.org/10.1007/978-3-319-68032-3_20

conjecture (involving sums of three primes) proved by Vinogradov [24] for large odd numbers and confirmed by Helfgott [9] completely. Legendre's conjecture states that for any positive integer n, there is a prime between n^2 and $(n+1)^2$. Another well-known conjecture of A. de Polignac asserts that for any positive even number d, there are infinitely many positive integers n with $p_{n+1} - p_n = d$, where p_k denotes the kth prime. (This conjecture in the case $d = 2$ is the famous twin prime conjecture; recently Zhang [26] made an important breakthrough along this line.) Polignac's conjecture follows from the following well-known hypothesis due to A. Schinzel.

Schinzel's Hypothesis. *If $f_1(x), \ldots, f_k(x)$ are irreducible polynomials with integer coefficients and positive leading coefficients such that there is no prime dividing the product $f_1(q)f_2(q)\ldots f_k(q)$ for all $q \in \mathbb{Z}$, then there are infinitely many positive integers n such that $f_1(n), f_2(n), \ldots, f_k(n)$ are all primes.*

A positive integer n is said to be *practical* if every $m = 1, \ldots, n$ can be written as the sum of some distinct (positive) divisors of n. In 1954, Stewart [15] showed that if $q_1 < \cdots < q_r$ are distinct primes and $a_1, \ldots, a_r$ are positive integers, then $m = q_1^{a_1} \cdots q_r^{a_r}$ is practical if and only if $q_1 = 2$ and

$$q_{s+1} - 1 \leqslant \sigma(q_1^{a_1} \cdots q_s^{a_s}) \quad \text{for all } \ 0 < s < r,$$

where $\sigma(n)$ stands for the sum of all divisors of n. The behavior of practical numbers is quite similar to that of primes. For example, Melfi [12] proved the following Goldbach-type conjecture of Margenstern [11]: Each positive even integer is a sum of two practical numbers, and there are infinitely many practical numbers m with $m-2$ and $m+2$ also practical. Recently, Weingartner [25] proved that the number of practical numbers not exceeding $x \geqslant 2$ is asymptotically equivalent to $cx/\log x$, where c is a positive constant close to 1; this analog of the Prime Number Theorem for practical numbers was first conjectured by Margenstern [11] in 1991.

In the published papers [19, 20, 22, 23], the author posed many conjectures on primes. For example, [22] contains 60 problems on combinatorial properties of primes many of which depend on some exact values of the prime-counting function $\pi(x)$ ($\pi(x)$ with $x \geqslant 0$ denotes the number of primes not exceeding x).

In this paper, we present 100 new conjectures on representations involving primes or related things. In particular, we find some surprising refinements of Goldbach's conjecture, Lemoine's conjecture, Legendre's conjecture, and the twin prime conjecture. The next section contains 25 conjectures, the first of which is a general hypothesis (similar to Schinzel's Hypothesis) on representations of integers involving primes, and the other 24 conjectures are closely related to this general hypothesis. In Sect. 3, we include 45 conjectures on various other representation problems for integers. In Sect. 4, we pose 30 conjectures on representations of positive rational numbers and related things. For numbers of representations related to some conjectures in Sects. 2–4, the reader may consult [16] for certain sequences in the OEIS.

We hope that the 100 conjectures collected here might interest some number theorists and stimulate further research.

Throughout this paper, we set $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$. For a real number x, the fractional part of $\{x\}$ is given by $x - \lfloor x \rfloor$. For $a \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$, by $\{a\}_n$ we mean the least nonnegative residue of a modulo n, i.e., $\{a\}_n = n\{a/n\}$. For $a \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$ with $2 \nmid n$, $(\frac{a}{n})$ denotes the Jacobi symbol. As usual, φ stands for Euler's totient function.

2 A General Hypothesis and Related Conjectures

Note that Schinzel's Hypothesis does not imply Goldbach's conjecture. Here, we pose a general hypothesis on representations of integers.

Conjecture 2.1 (General Hypothesis, 2012-12-28) *Let*

$$f_1(x, y), \ \ldots, \ f_m(x, y)$$

be non-constant polynomials with integer coefficients. Suppose that for all large $n \in \mathbb{Z}^+$, those $f_1(x, n-x), \ldots, f_m(x, n-x)$ are irreducible, and there is no prime dividing all the products $\prod_{k=1}^m f_k(x, n-x)$ with $x \in \mathbb{Z}$. If $n \in \mathbb{Z}^+$ is large enough, then we can write $n = x + y$ $(x, y \in \mathbb{Z}^+)$ such that $|f_1(x, y)|, \ldots, |f_m(x, y)|$ are all prime.

Remark 2.1 In view of this general hypothesis, almost all of the other conjectures in this section are essentially reasonable.

Conjecture 2.2 (Symmetric Conjecture, 2015-08-27) *For any integer $n > 6$, there is a prime $p < n/n'$ such that $n - (pn' - 1)$ and $n + (pn' - 1)$ are both prime, where $n' = 2 - \{n\}_2$ is 1 or 2 according as n is odd or even.*

Remark 2.2 Conjecture 2.2 is stronger than Goldbach's conjecture and Lemoine's conjecture. We have verified Conjecture 2.2 for all $n = 7, \ldots, 10^8$; see [16, A261627 and A261628] for related data. Conjecture 2.1 implies that Conjecture 2.2 holds for all sufficiently large integers n. In fact, if we apply Conjecture 2.1 with $f_1(x, y) = x$, $f_2(x, y) = 2y + 1$ and $f_3(x, y) = 4x + 2y - 1$, then for sufficiently large $n \in \mathbb{Z}^+$ there are primes p and q with $n = p + (q-1)/2$ (i.e., $2n - (2p - 1) = q$) such that $2n + 2p - 1 = 4p + q - 2$ is prime; if we apply Conjecture 2.1 with $f_1(x, y) = 2x + 1$, $f_2(x, y) = 2y - 1$ and $f_3(x, y) = 4x + 2y - 1$, then for sufficiently large $n \in \mathbb{Z}^+$ there are primes p and q with $n = (p-1)/2 + (q+1)/2$ (i.e., $2n - 1 - (p-1) = q$) such that $2n - 1 + (p - 1) = 2p + q - 2$ is prime.

Conjecture 2.3 *For each $n = 6, 7, \ldots$ there is a prime $p < n$ such that both $6n - p$ and $6n + p$ are prime.*

Remark 2.3 We also have some conjectures involving practical numbers similar to Conjectures 2.2 and 2.3; see [16, A261641] and Conjectures 3.43 and 3.44. Conjecture 2.1 with $f_1(x, y) = x$, $f_2(x, y) = 5x + 6y$ and $f_3(x, y) = 7x + 6y$ implies that Conjecture 2.3 holds for sufficiently large integers n.

Conjecture 2.4 (2012-12-22) *Any integer $n \geqslant 12$ can be written as $p + q$ ($q \in \mathbb{Z}^+$) with $p,\ p + 6,\ 6q - 1$, and $6q + 1$ all prime.*

Remark 2.4 Conjecture 2.1 implies that Conjecture 2.4 holds for all sufficiently large integers n. We have verified Conjecture 2.4 for n up to 10^9; see [16, A199920] for numbers of such representations. Conjecture 2.4 implies that there are infinitely many twin primes and also infinitely many sexy primes, because for any $m = 2, 3, \ldots$ the interval $[m! + 2, m! + m]$ of length $m - 2$ contains no prime.

Conjecture 2.5 (2013-10-09) *Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $6k - 1$ a Sophie Germain prime and $\{6m - 1, 6m + 1\}$ a twin prime pair.*

Remark 2.5 Recall that a Sophie Germain prime is a prime p with $2p + 1$ also prime. Conjecture 2.1 implies that Conjecture 2.5 holds for all sufficiently large integers n. We have verified Conjecture 2.5 for all $n = 2, \ldots, 10^8$; see [16, A227923] for numbers of such representations. Conjecture 2.5 implies that there are infinitely many twin primes and also infinitely many Sophie Germain primes. For example, if all twin primes do not exceed an integer $N > 2$ and $(N + 1)!/6 = k + m$ ($k, m \in \mathbb{Z}^+$) with $6k - 1$ a Sophie Germain prime and $\{6m - 1, 6m + 1\}$ a twin prime pair, then $6k - 1 = (N + 1)! - (6m + 1)$ with $2 \leqslant 6m + 1 \leqslant N$ which contradicts that $6k - 1$ is prime.

Recall that for two subsets X and Y of $\mathbb{Z}$, the sumset $X + Y$ is defined as $\{x + y : x \in X \ \text{and}\ y \in Y\}$.

Conjecture 2.6 (2013-01-03) *Let*

$$\begin{aligned} A &= \{x \in \mathbb{Z}^+ :\ 6x - 1 \textit{ and } 6x + 1 \textit{ are both prime}\},\\ B &= \{x \in \mathbb{Z}^+ :\ 6x + 1 \textit{ and } 6x + 5 \textit{ are both prime}\},\\ C &= \{x \in \mathbb{Z}^+ :\ 2x - 3 \textit{ and } 2x + 3 \textit{ are both prime}\}. \end{aligned}$$

Then

$$A + B = \{2, 3, \ldots\},\ B + C = \{5, 6, \ldots\},\ \ A + C = \{5, 6, \ldots\} \setminus \{161\}.$$

Also, if we set $2X := X + X$ for $X \subseteq \mathbb{Z}$, then

$$2A \supseteq \{702, 703, \ldots\},\ 2B \supseteq \{492, 493, \ldots\},\ 2C \supseteq \{4006, 4007, \ldots\}.$$

Remark 2.6 Conjecture 2.1 implies that each of the sumsets $A + B, B + C, A + C, 2A, 2B, 2C$ in Conjecture 2.6 contains all sufficiently large integers.

Conjecture 2.7 (2013-10-12)

(i) *For any integer $n > 3$, we can write $2n$ as $p + q$ with $p, q, 3p - 10, 3q + 10$ all prime.*

(ii) *For any integer $n > 4$ not equal to 76, we can write $2n$ as $p+q$ with $p, 3p-10, q, 3q-10$ all prime.*

Remark 2.7 Note that if $2n = p+q$, then $6n = (3p-10)+(3q+10)$. We have verified Conjecture 2.7 for n up to 10^8. See [16, A230230] for related data. Conjecture 2.1 implies that Conjecture 2.7 holds for all sufficiently large integers n.

Conjecture 2.8 (2012-11-07) *For any integer $n > 8$, we can write $2n-1$ as $p+2q$ with p, q, and p^2+60q^2 all prime.*

Remark 2.8 This is stronger than Lemoine's conjecture. We have verified Conjecture 2.8 for n up to 10^8. See [16, A218825] for related data.

Conjecture 2.9 (2013-10-16) *Any integer $n > 3$ can be written as $p+q$ ($q \in \mathbb{Z}^+$) with p, $2p^2-1$, and $2q^2-1$ all prime.*

Remark 2.9 See [16, A230351] for related data. Note that each of 7, 12, 68, 330 has a unique required representation:

$$\begin{gathered}7 = 3+4,\ 2\cdot 3^2-1 = 17,\ 2\cdot 4^2-1 = 31;\\ 12 = 2+10,\ 2\cdot 2^2-1 = 7,\ 2\cdot 10^2-1 = 199;\\ 68 = 43+25,\ 2\cdot 43^2-1 = 3697,\ 2\cdot 25^2-1 = 1249;\\ 330 = 7+323,\ 2\cdot 7^2-1 = 97,\ 2\cdot 323^2-1 = 208657.\end{gathered}$$

In 2001, A. Murthy (cf. [13]) conjectured that for any integer $n > 1$, there is an integer $0 < k < n$ such that $kn+1$ is prime. In 2005, he [13] conjectured that any integer $n > 3$ can be written as $x+y$ ($x, y \in \mathbb{Z}^+$) with $xy-1$ prime. In the 1990s, Ming-Zhi Zhang (cf. [6, p. 161]) asked whether any odd integer $n > 1$ can be written as $a+b$ with $a, b \in \mathbb{Z}^+$ and a^2+b^2 prime.

Conjecture 2.10 (2012-12-20)

(i) *For any integer $n > 3$, there is an integer $k \in \{1, \dots, n-1\}$ such that $kn+1$ and $k(n-k)-1$ are both prime.*
(ii) *For any odd integer $n > 1$, there is an integer $k \in \{1, \dots, n-1\}$ such that $kn+1$ and $k^2+(n-k)^2$ are both prime.*

Remark 2.10 This combines Murthy's conjectures and Zhang's conjecture. We also conjecture that any integer $n > 3$ can be written as $x+y$ with $x, y \in \mathbb{Z}^+$ such that $3x \pm 1$ and $xy-1$ are all prime (cf. [16, A220431]).

Conjecture 2.11 (2013-11-12)

(i) *Any integer $n > 2$ can be written as $k+m$ ($k, m \in \mathbb{Z}^+$) with k^2m-1 prime. Also, each integer $n > 4$ can be written as $k+m$ ($k, m \in \mathbb{Z}^+$) with k^2m+1 prime.*
(ii) *Any integer $n > 1$ can be written as $k+m$ ($k, m \in \mathbb{Z}^+$) with $(km)^2+km+1$ prime. Also, each integer $n > 2$ can be written as $k+m$ ($k, m \in \mathbb{Z}^+$) with $(km)^2+km-1$ (or $2k^2m^2-1$) prime.*

Remark 2.11 See [16, A231633] for related data.

Conjecture 2.12 (2013-10-13)

(i) *For any integer $n > 1$, there is a prime $p \leqslant n$ such that $(p-1)n+1$ is prime. Moreover, for any integer $n > 4$, there is a prime $p < n$ such that $3p+8$ and $(p-1)n+1$ are both prime.*

(ii) *Any integer $n > 5$ can be written as $p+q$ $(q \in \mathbb{Z}^+)$ with p, $3p-10$, and $(p-1)q-1$ all prime.*

Remark 2.12 See [16, A230243 and A230241] for related data.

Conjecture 2.13 (2012-12-16) *For any integer $n > 1$, we can write $2n$ as $p+q$, where p is a Sophie Germain prime, q is a positive integer, and $(p-1)^2+q^2$ is prime.*

Remark 2.13 This is stronger than Zhang's conjecture. Conjecture 2.1 implies that any sufficiently large n can be written as $x+y$ $(x, y \in \mathbb{Z}^+)$ with $p = 2x+1$, $2p+1 = 4x+3$ and

$$(p-1)^2+(2n-p)^2 = (2x)^2+(2y-1)^2$$

all prime. See [16, A220554] for related data. For example, $32 = 11+21$ with 11 a Sophie Germain prime and $(11-1)^2+21^2 = 541$ a prime.

Conjecture 2.14 *(i)* *(2011-11-04) Any odd integer $n > 1$ can be written as $x+y$ with $x, y \in \mathbb{Z}^+$ such that x^4+y^4 is prime.*

(ii) *(2012-12-01) Any integer $n > 10$ can be written as $p+q$ $(q \in \mathbb{Z}^+)$ with p, $p+6$, and $p^2+3pq+q^2 = n^2+pq$ all prime.*

(iii) *(2013-11-21) Let $n > 1$ be an odd integer. We can write $n = k+m$ with $k, m \in \mathbb{Z}^+$ such that both k^2+m^2 and k^3+m^2 are prime.*

Remark 2.14 See [16, A218656, A218654, A218754 and A232269] for related data.

Conjecture 2.15 (Olivier Gerard and Zhi-Wei Sun, 2013-10-13). *For any integer $n > 1$, we can write $2n$ as $p+q$ with p, q, and $(p-1)(q+1)-1$ all prime.*

Remark 2.15 This is stronger than Goldbach's conjecture. Note also that $(p-1)+(q+1) = p+q$. We have verified Conjecture 2.15 for all $n = 2, \ldots, 10^8$. See [16, A227909] for related data.

Conjecture 2.16 (2012-11-30) *Any integer $n > 7$ can be written as $p+q$ $(q \in \mathbb{Z}^+)$ with p and $2pq+1$ both prime. In general, for each $m \in \mathbb{N}$, any sufficiently large integer n can be written as $x+y$ $(x, y \in \mathbb{Z}^+)$ with $x-m$, $x+m$, and $2xy+1$ all prime.*

Remark 2.16 We have verified the first assertion in Conjecture 2.16 for all $n = 8, 9, \ldots, 10^9$. See [16, A219864] for related data. Concerning the general statement in Conjecture 2.16, for $m = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, it suffices to require that n is greater than

$$623,\ 28,\ 151,\ 357,\ 199,\ 307,\ 357,\ 278,\ 697,\ 263$$

respectively.

Conjecture 2.17 (2013-10-14) *Any integer $n > 3$ can be written as $p + q$ $(q \in \mathbb{Z}^+)$ with p and $(p+1)q/2 + 1$ both prime.*

Remark 2.17 We have verified this conjecture for all $n = 4, \ldots, 10^8$. See [16, A230254] for related data. For example, 30 has a unique representation $2 + 28$ with $(2+1)28/2 + 1 = 43$ prime.

Bertrand's Postulate proved by Chebyshev in 1852 states that for any positive integer n, the interval $[n, 2n]$ contains at least a prime. Goldbach's conjecture essentially asserts that for any integer $n > 1$, there is an integer $k \in \{0, \ldots, n\}$ such that $n - k$ and $n + k$ are both prime. The following conjecture is of a similar flavor.

Conjecture 2.18 (2012-12-18) *For each positive integer n, there is an integer $k \in \{0, \ldots, n\}$ such that $n + k$ and $n + k^2$ are both prime.*

Remark 2.18 We have verified this for n up to 10^8. See [16, A185636 and A204065] for related data. The author would like to offer 100 US dollars as the prize for the first solution of Conjecture 2.18.

Conjecture 2.19 (2013-04-15) *For any positive integer n, there is a positive integer $k \leqslant 4\sqrt{n+1}$ such that $n^2 + k^2$ is prime.*

Remark 2.19 Note that the least $k \in \mathbb{Z}^+$ with $63^2 + k^2$ prime is $32 = 4\sqrt{63+1}$.

Conjecture 2.20 (2013-10-15)

(i) For any integer $n > 5$, there is a prime $p < n$ with $p + 6$ and $n + (n-p)^2$ both prime.
(ii) For any integer $n > 3$, there is a prime $p < n$ with $3p - 4$ and $n^2 + (n-p)^2$ both prime.

Remark 2.20 See [16, A227898 and A227899] for related data.

Conjecture 2.21 (2013-11-20)

(i) Any integer $n > 1$ can be written as $x + y$ with $x, y \in \mathbb{Z}^+$ such that $x + ny$ and $x^2 + ny^2$ are both prime.
(ii) Any integer $n > 2$ can be written as $x + y$ with $x, y \in \mathbb{Z}^+$ such that $nx + y$ and $nx - y$ are both prime. Also, any integer $n > 2$ can be written as $x + y$ with $x, y \in \mathbb{Z}^+$ such that $x^2 + (n-2)y^2$ is prime.
(iii) Any integer $n > 2$ can be written as $p + q$ with $q \in \mathbb{Z}^+$ such that p and $p^3 + nq^2$ (or $p + nq$) are both prime.

Remark 2.21 See [16, A232174, A231883 and A232186] for related data. For example, $20 = 11 + 9$ with $11 + 20 \cdot 9 = 191$ and $11^2 + 20 \cdot 9^2 = 121 + 20 \times 81 = 1741$ both prime. The author would like to offer 200 US dollars as the prize for the first solution to part (i) of Conjecture 2.21. We also conjecture that there are infinitely many $n \in \mathbb{Z}^+$ such that $p_n = x^2 + ny^2$ for some $x, y \in \mathbb{Z}^+$ (where p_n is the nth prime).

Conjecture 2.22 (2013-10-14) *Any integer $n > 1$ can be written as $x + y$ with $x, y \in \mathbb{Z}^+$ such that $2x + 1$, $x^2 + x + 1$ and $y^2 + y + 1$ are all prime. Also, each integer $n > 1$ can be written as $x + y$ with $x, y \in \mathbb{Z}^+$ such that $x^2 + 1$ (or $4x^2 + 1$) and $4y^2 + 1$ are both prime.*

Remark 2.22 See [16, A230252] for related data. For example, $31 = 14 + 17$ with $2 \cdot 14 + 1 = 29$, $14^2 + 14 + 1 = 211$, and $17^2 + 17 + 1 = 307$ all prime.

In 2001, Heath-Brown [8] proved that there are infinitely many primes of the form $x^3 + 2y^3$ where x and y are positive integers.

Conjecture 2.23 (2012-12-14) *Any positive integer n can be written as $x + y$ $(x, y \in \mathbb{N})$ with $x^3 + 2y^3$ prime. In general, for each positive odd integer m, any sufficiently large integer can be written as $x + y$ $(x, y \in \mathbb{N})$ with $x^m + 2y^m$ prime.*

Remark 2.23 See [16, A220413] for related data. For any integer $d > 2$, not every sufficiently large integer n can be written as $x + y$ $(x, y \in \mathbb{N})$ with $x^3 + dy^3$ prime. For, if n is a multiple of a prime divisor p of $d - 1$, then $x^3 + d(n - x)^3 \equiv (1 - d)x^3 \equiv 0 \pmod{p}$ for any integer x.

Conjecture 2.24 (2013-04-15) *For any integer $n > 4$, there is a positive integer $k < n$ such that $p = 2n + k$ and $2n^3 + k^3 = 2n^3 + (p - 2n)^3$ are both prime.*

Remark 2.24 See [16, A224030] for related data.

Conjecture 2.25 (2012-12-16) *Let m be a positive integer. Then, any sufficiently large odd integer n can be written as $x + y$ $(x, y \in \mathbb{Z}^+)$ with $x^m + 3y^m$ prime (and any sufficiently large even integer n can be written as $x + y$ $(x, y \in \mathbb{Z}^+)$ with $x^m + 3y^m + 1$ prime). In particular, if $m \leqslant 6$ or $m = 18$, then each positive odd integer can be written as $x + y$ $(x, y \in \mathbb{N})$ with $x^m + 3y^m$ prime.*

Remark 2.25 See [16, A220572] for related data and comments. For example, 5 can be written as $1 + 4$ with

$$1^{18} + 3 \cdot 4^{18} = 206158430209$$

prime.

3 Other Representation Problems for Positive Integers

Conjecture 3.1 (2013-11-12) *Any integer $n > 1$ can be written as $x + y$ ($x, y \in \mathbb{Z}^+$) with $[x, y] + 1$ prime, where $[x, y]$ is the least common multiple of x and y. Also, each integer $n > 3$ can be written as $x + y$ ($x, y \in \mathbb{Z}^+$) with $[x, y] - 1$ prime.*

Remark 3.1 See [16, A231635] for related data. For example, $10 = 4 + 6$ with $[4, 6] + 1 = 13$ and $[4, 6] - 1 = 11$ both prime.

As usual, for $x \in \mathbb{Z}$, we let T_x denote the triangular number $x(x + 1)/2$.

Conjecture 3.2 *(i) (2013-11-10) Any integer $n > 1$ can be written as $x + y$ ($x, y \in \mathbb{Z}^+$) with $T_x + y^2$ prime. Also, any integer $n > 6$ can be written as $x + y$ ($x, y \in \mathbb{Z}^+$) with $T_x + y^4$ prime.*

(ii) (2013-11-18) Any integer $n > 1$ can be written as $x + y$ ($x, y \in \mathbb{Z}^+$) with $p = 2x + 1$ and $T_x + y = n + (p - 1)(p - 3)/8$ both prime.

Remark 3.2 See [16, A228425 and A232109] for related data. For example, $18 = 7 + 11$ with $T_7 + 11^2 = 149$ prime, $27 = 5 + 22$ with $T_5 + 22^4 = 234271$ prime, and $18 = 11 + 7$ with $2 \cdot 11 + 1 = 23$ and $T_{11} + 7 = 73$ both prime.

Conjecture 3.3 (2012-10-15) *Each $n = 1, 2, 3. \ldots$ can be written as $T_x + y$ with $x, y \in \mathbb{N}$ such that $T_y + 1$ is prime.*

Remark 3.3 See [16, A229166] for related data. For example, 34 has a unique required representation: $34 = T_5 + 19$ with $T_{19} + 1 = 191$ prime.

Conjecture 3.4 (2012-12-09) *Any integer $n > 2$ can be written as $x^2 + y$ ($x, y \in \mathbb{Z}^+$) with $2xy - 1$ prime. In other words, for each $n = 3, 4, \ldots,$ there is a prime of the form $2k(n - k^2) - 1$ with $k \in \mathbb{Z}^+$.*

Remark 3.4 We have verified Conjecture 3.4 for all $n = 3, 4, \ldots, 3 \cdot 10^9$. See [16, A220272] for related data. For example, $18 = 3^2 + 9$ with $2 \times 3 \times 9 - 1 = 53$ prime.

Conjecture 3.5 (2013-10-21) *Any integer $n > 1$ can be written as $x^2 + y$ with $2y^2 - 1$ prime, where $x, y \in \mathbb{N}$. In other words, for each $n = 2, 3, 4, \ldots$ there is an integer $0 \leqslant k \leqslant \sqrt{n}$ such that $2(n - k^2)^2 - 1$ is prime.*

Remark 3.5 We have verified this conjecture for all $n = 2, 3, \ldots, 10^8$. See [16, A230494] for related data. For example, $9 = 1^2 + 8$ with $2 \cdot 8^2 - 1 = 127$ prime.

Conjecture 3.6 (2013-11-11)

(i) Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^k + m$ prime. In other words, for each $n = 2, 3, \ldots,$ there is a positive integer $k < n$ with $n + 2^k - k$ prime.

(ii) For any integer $n > 3$, there is a positive integer $k < n$ such that $n + 2^k + k$ is prime.

Remark 3.6 We have verified parts (i) and (ii) of this conjecture for n up to 10^7 and 3.8×10^6, respectively, see [16, A231201, A231557 and A231725] for related data and other similar conjectures. For example, $9302003 = 311468 + 8990535$ with $2^{311468} + 8990535$ a prime of 93762 decimal digits. In [21], the author proved that the set $\{2^k - k : k = 1, 2, 3, \ldots\}$ contains a complete system of residues modulo any positive integer. The author would like to offer 1000 US dollars as the prize for the first solution to part (i) of Conjecture 3.6.

Conjecture 3.7 (2013-11-23) *Any integer $n > 3$ can be written as $p + (2^k - k) + (2^m - m)$ with p prime and $k, m \in \mathbb{Z}^+$.*

Remark 3.7 For example, 94 has a unique required representation $31 + (2^3 - 3) + (2^6 - 6)$. See [16, A232398] for related data. After the author verified this conjecture for n up to 2×10^8, Qing-Hu Hou extended the verification to 10^{10} in December 2013. In contrast with Conjecture 3.7, Crocker [3] proved in 1971 that there are infinitely many positive odd numbers not of the form $p + 2^k + 2^m$ with p prime and $k, m \in \mathbb{Z}^+$.

Conjecture 3.8 (2013-11-11) *Let $r \in \{1, 2\}$. Then, any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^k m^r + 1$ prime. Also, any integer $n > 2$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^k m^r - 1$ prime.*

Remark 3.8 See [16, A231561] for related data.

Conjecture 3.9 *(i) (2013-11-10) Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $F_k + m$ (or $F_k + 2m$, or $F_k + m(m + 1)$) prime, where the Fibonacci sequence $(F_j)_{j \geqslant 0}$ is given by $F_0 = 0$, $F_1 = 1$, and $F_{j+1} = F_j + F_{j-1}$ for $j \in \mathbb{Z}^+$.*

(ii) (2014-04-27) Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $L_k + m$ prime, where the Lucas sequence $(L_j)_{j \geqslant 0}$ is given by $L_0 = 2$, $L_1 = 1$, and $L_{j+1} = L_j + L_{j-1}$ for $j \in \mathbb{Z}^+$.

Remark 3.9 See [16, A231555 and A241844] for related data. We have verified parts (i) and (ii) of Conjecture 3.9 for n up to 3.7×10^6 and 7×10^6, respectively.

Conjecture 3.10 *(i) (2013-11-11) Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $k!m + 1$ prime. Also, any integer $n > 3$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $k!m - 1$ prime.*

(ii) (2014-03-19) Let $r \in \{1, -1\}$. For each integer $n > 1$, there is a number $k \in \{1, \ldots, n\}$ with $k!n + r$ prime.

Remark 3.10 See [16, A231516 and A231631] for related data. We have verified part (i) of Conjecture 3.10 for n up to 10^6. We also conjecture that for any integer $n > 2$, there is a positive integer $k < \sqrt{n} \log n$ with $k!(n - k) + 1$ prime.

Conjecture 3.11 *(i) (2015-04-01) Let $k, m \in \mathbb{Z}^+$ with $k + m > 2$. Then, any integer $n > 2$ can be written as $\lfloor p/k \rfloor + \lfloor q/m \rfloor$ with p and q both prime.*

(ii) (2015-04-24) Let

$$T := \left\{ \left\lfloor \frac{x}{9} \right\rfloor : \ x - 1 \textit{ and } x + 1 \textit{ are twin prime} \right\}$$
$$= \left\{ \left\lfloor \frac{x}{3} \right\rfloor : \ 3x - 1 \textit{ and } 3x + 1 \textit{ are twin prime} \right\}.$$

Then, any positive integer can be written as the sum of two distinct elements of T one of which is even.

Remark 3.11 See [16, A256555 and A256707] for related data. Part (i) of Conjecture 3.11 in the case $k = m = 2$ reduces to Goldbach's conjecture, and it reduces to Lemoine's conjecture when $\{k, m\} = \{1, 2\}$. Part (ii) of Conjecture 3.11 implies the twin prime conjecture.

Conjecture 3.12 (2014-03-03)

(i) Let $1 < m < n$ be integers with $m \nmid n$. Then, $\lfloor kn/m \rfloor$ is prime for some $k = 1, \ldots, n - 1$.
(ii) Let $m > 2$ and $n > 2$ be integers. Then, there is a prime $p < n$ with $\lfloor (n - p)/m \rfloor$ a square. Also, there is a prime $p < n$ such that $\lfloor (n - p)/m \rfloor$ is a triangular number of the form $T_{(q-3)/2} = (q - 1)(q - 3)/8$ with q an odd prime.
(iii) For each $n = 3, 4, \ldots$, there is a prime $p < n$ with $\lfloor (n - p)/5 \rfloor$ a cube.

Remark 3.12 See [16, A238703, A238732 and A238733] for related data.

Conjecture 3.13 *(i) (2013-10-21) Let*

$$S = \{n \in \mathbb{Z}^+ : \ 2n + 1 \textit{ and } 2n^3 + 1 \ \textit{ are both prime}\}.$$

Then, any integer $n > 2$ is a sum of three elements of S.
(ii) (2013-10-22) Any integer $n > 5$ can be written as $a + b + c$ with $a, b, c \in \mathbb{Z}^+$ such that

$$\{a^2 + a \pm 1\}, \ \{b^2 + b \pm 1\}, \ \{c^2 + c \pm 1\}$$

are all twin prime pairs!

Remark 3.13 See [16, A230507 and A230516] for related data and comments.

Conjecture 3.14 (2013-10-11) *Let*

$$P = \{p : \ p, \ p + 6 \textit{ and } 3p + 8 \ \textit{ are all prime}\}.$$

Then, for any integer $n > 6$, we can write $2n + 1 = p + q + r$ with $p, q, r \in P$ such that $p + q + 9$ is also prime.

Remark 3.14 This implies not only Goldbach's weak conjecture but also Goldbach's conjecture for even numbers. See [16, A230217 and A230219] for related data. Note that 37 has a unique required representation 7 + 13 + 17; in fact,

$$\begin{aligned}&7,\ 7+6=13,\ 3\times 7+8=29,\\ &13,\ 13+6=19,\ 3\cdot 13+8=47,\\ &17,\ 17+6=23,\ 3\cdot 17+8=59,\end{aligned}$$

and $7+13+9=29$ are all prime.

Conjecture 3.15 (2013-10-12) *Let*

$$P'=\{p:\ p,\ 3p-4,\ 3p-10 \textit{ and } 3p-14 \ \textit{ are all prime}\}.$$

Then, for any integer $n>17$, we can write $2n=p+q+r+s$ with $p,q,r,s\in P'$.

Remark 3.15 See [16, A230223 and A230224] for related data. Note that such a representation involves 16 primes! For example, 54 has a unique required representation $7+11+17+19$; in fact,

$$\begin{aligned}&7,\ 3\cdot 7-4=17,\ 3\cdot 7-10=11,\ 3\cdot 7-14=7,\\ &11,\ 3\cdot 11-4=29,\ 3\cdot 11-10=23,\ 3\cdot 11-14=19,\\ &17,\ 3\cdot 17-4=47,\ 3\cdot 17-10=41,\ 3\cdot 17-14=37,\\ &19,\ 3\cdot 19-4=53,\ 3\cdot 19-10=47,\ 3\cdot 19-14=43\end{aligned}$$

are all prime.

Conjecture 3.16 *(i) (2015-10-01) Any integer $n>1$ can be written as $x^2+y^2+\varphi(z^2)$ with $x,y\in\mathbb{N}$, $x\leqslant y$ and $z\in\mathbb{Z}^+$ such that y or z is prime.*

(ii) (2015-10-02) Each positive integer can be written as $x^2+y^2+p(p+\varepsilon)/2$, where $x,y\in\mathbb{Z}$, $\varepsilon\in\{\pm 1\}$, and p is a prime.

Remark 3.16 See [16, A262311, A262785, A262982, A262985, A263992, A263998, A264010 and A264025] for related data and similar conjectures. For example, $13=1^2+2^2+\varphi(4^2)$ with 2 prime, $94415=115^2+178^2+\varphi(223^2)$ with 223 prime, $97=1^2+9^2+5(5+1)/2$ with 5 prime, and $538=3^2+8^2+31(31-1)/2$ with 31 prime. It is known that each $n\in\mathbb{N}$ can be expressed as the sum of two squares and a triangular number (cf. [17]).

Conjecture 3.17 (2014-02-26)

(i) Any integer $n>6$ can be written as $k+m$ $(k,m\in\mathbb{Z}^+)$ such that $p_k+\pi(m)$ is a triangular number.

(ii) Any integer $n>10$ can be written as $k+m$ $(k,m\in\mathbb{Z}^+)$ such that $p=p_k+\pi(m)$ and $p+2$ are both prime.

Remark 3.17 See [16, A238405 and A238386] for related data. For example, $72=41+31$ with $p_{72}+\pi(31)=179+11=19\cdot 20/2$ a triangular number, and $108=15+93$ with $p_{15}+\pi(93)=47+24=71$ and $71+2=73$ twin prime.

Conjecture 3.18 (2014-03-05) *Any integer $n > 2$ can be written as $q + m$ with $m \in \mathbb{Z}^+$ such that q, $p_q - q + 1$, and $p_{p_m} - p_m + 1$ are all prime.*

Remark 3.18 See [16, A237715] for related data.

Conjecture 3.19 (2014-01-04) *For any integer $n > 6$, there is a prime $q < n/2$ with $p_q - q + 1$ prime such that $n - (1 + \{n\}_2)q$ is prime.*

Remark 3.19 This conjecture is stronger than Goldbach's conjecture and Lemoine's conjecture, and it also implies that there are infinitely many primes q with $p_q - q + 1$ prime. See [16, A235189] for related data. For example, 7, $p_7 - 7 + 1 = 17 - 6 = 11$, and $61 - 2 \cdot 7 = 47$ are all prime, and 31, $p_{31} - 31 + 1 = 97$ and $98 - 31 = 67$ are all prime.

Conjecture 3.20 (2014-02-04)

(i) For any integer $n > 2$, we can write $2n = p + q$ with p, q, and $\varphi(p + 2) \pm 1$ all prime. Also, for any integer $n > 12$ we can write $2n - 1 = 2p + q$ with p, q and $\varphi(p + 1) \pm 1$ all prime.

(ii) Any integer $n \geqslant 24$ can be written as $(1 + \{n\}_2)p + q$ with $p, q, \varphi(p + 1) - 1$, and $\varphi(q - 1) + 1$ all prime.

Remark 3.20 See [16, A237168, A237183 and A237184] for related data. Note that either of the two parts is stronger than Goldbach's conjecture and Lemoine's conjecture. Also, part (i) of Conjecture 3.20 implies the twin prime conjecture.

Conjecture 3.21 (2014-02-04)

(i) Any integer $n \geqslant 12$ can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ and $k \neq m$ such that $\varphi(k) \pm 1$ and $\varphi(m) \pm 1$ are all prime.

(ii) Any integer $n > 6$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ such that both $\{p_k, p_k + 2\}$ and $\{\varphi(m) - 1, \varphi(m) + 1\}$ are twin prime pairs.

(iii) Any integer $n \geqslant 6$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ with $p_{p_k} + 2$ and $\varphi(m) \pm 1$ all prime. Also, each $n = 2, 3, 4, \ldots$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ with $p_{p_k} + 2$ and $6m \pm 1$ all prime.

(iv) Any integer $n \geqslant 8$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ with $p_{p_{p_k}} - 2$ and $\varphi(m) \pm 1$ all prime.

(v) Any integer $n > 8$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ with $3k \pm 1$ and $\varphi(m) \pm 1$ all prime.

(vi) Any integer $n \geqslant 12$ can be written as $p + q$ $(q \in \mathbb{Z}^+)$ with p, $p + 6$, and $\varphi(q) \pm 1$ all prime.

Remark 3.21 See [16, A237127, A237130, A218829 and A237253] for related data and comments. Clearly each part of Conjecture 3.21 implies the twin prime conjecture.

Conjecture 3.22 (2013-12-31)

(i) *Any integer* $n > 1$ *with* $n \neq 8$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $p = k + \varphi(m)$ *and* $2n - p$ *are both prime.*

(ii) *Each integer* $n > 2$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $p = k + \varphi(m)$ *and* $2n + 1 - 2p$ *are both prime.*

Remark 3.22 Clearly parts (i) and (ii) are stronger than Goldbach's conjecture and Lemoine's conjecture, respectively. See [16, A234808 and A234809] for related data. For example, $24 = 9 + 15$ with $9 + \varphi(15) = 17$ and $2 \cdot 24 - 17 = 31$ both prime, and $41 = 7 + 34$ with $7 + \varphi(34) = 23$ and $2 \cdot 41 + 1 - 2 \cdot 23 = 37$ both prime.

Conjecture 3.23 (2014-02-02)

(i) *Any integer* $n > 1$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $6k \pm 1$ *and* $k + \varphi(m)$ *are all prime.*

(ii) *Any integer* $n > 3$ *with* $n \neq 12$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $6k \pm 1$ *and* $k + \varphi(m)/2$ *are all prime.*

(iii) *Each integer* $n > 5$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *with* $k + \varphi(m)/2$ *a square.*

Remark 3.23 See [16, A236968 and A236567] for related data.

Conjecture 3.24 (2014-01-13) *Define*

$$K := \{k \in \mathbb{Z}^+ : \ k(k+1) - p_k \ \textit{ is prime}\}.$$

(i) *Any integer* $n > 3$ *can be written as* $a + b$ *with* $a, b \in K$.

(ii) *Any integer* $n > 2$ *can be expressed as the sum of an element of* K *and a positive triangular number.*

(iii) *Any integer* $n > 3$ *can be written as the sum of an element of* K *and a prime* q *with* $p_q - q + 1$ *also prime.*

(iv) *Any integer* $n > 7$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $q = p_k + \varphi(m)$ *and* $q(q+1) - p_q$ *are both prime.*

Remark 3.24 See [16, A235592, A235613, A235614, A235661, A235703, A232353] for related data.

Conjecture 3.25 (2014-01-18)

(i) *Any integer* $n > 7$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $p = \varphi(k) + \varphi(m)/2 - 1$ *is a prime and also* 2 *is a primitive root modulo* p.

(ii) *Any integer* $n \geqslant 38$ *can be written as* $k + m$ $(k, m \in \mathbb{Z}^+)$ *such that* $p = p_k + \varphi(m)$ *is a Sophie Germain prime and also* 2 *is a primitive root modulo* p.

Remark 3.25 See [16, A235987] for related data and comments. For example, $79 = 19 + 60$, $p_{19} + \varphi(60) = 67 + 16 = 83$ is a Sophie Germain prime and 2 is a primitive root modulo 83.

Conjecture 3.26 *(i) (2012-12-23) Any integer $n > 5$ can be written as $k + m$ ($k, m \in \{3, 4, \ldots\}$) with $2^{\varphi(k)} + 2^{\varphi(m)} - 1$ prime.*

(ii) (2012-12-24) For any integer $a > 1$, there is a positive integer $N(a)$ such that any integer $n > N(a)$ can be written as $k + m$ with $k, m \in \{3, 4, \ldots\}$ such that $a^{\varphi(k)} + a^{\varphi(m)/2} - 1$ is prime. Moreover, we may take $N(2) = N(3) = \ldots = N(6) = N(8) = 5$ and $N(7) = 17$.

Remark 3.26 See [16, A234309, A234347 and A234359] for related data and comments. Clearly, part (ii) of Conjecture 3.26 implies that for each $a = 2, 3, \ldots$, there are infinitely many primes of the form $a^{2k} + a^m - 1$ with $k, m \in \mathbb{Z}^+$.

Conjecture 3.27 (2013-12-26)

(i) Any integer $n \geqslant 10$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^{\varphi(k)/2+\varphi(m)/6} + 3$ prime. Also, any integer $n > 13$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^{\varphi(k)/2+\varphi(m)/6} - 3$ prime.

(ii) Any integer $n > 25$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $3 \cdot 2^{\varphi(k)/2+\varphi(m)/8} + 1$ prime. Also, any integer $n \geqslant 15$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $3 \cdot 2^{\varphi(k)/2+\varphi(m)/12} - 1$ prime.

(iii) Any integer $n \geqslant 27$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2 \cdot 3^{\varphi(k)/2+\varphi(m)/12} +1$ prime. Also, any integer $n > 37$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2 \cdot 3^{\varphi(k)/2+\varphi(m)/12} - 1$ prime.

(iv) Any integer $n > 10$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $2^{\varphi(k)+\varphi(m)/4} - 5$ prime.

Remark 3.27 This implies that there are infinitely many primes in any of the following seven forms:

$$2^n + 3,\ 2^n - 3,\ 3 \cdot 2^n + 1,\ 3 \cdot 2^n - 1,\ 2 \cdot 3^n + 1,\ 2 \cdot 3^n - 1,\ 2^n - 5.$$

We have verified Conjecture 3.27 for n up to 50,000. See [16, A234451, A236358 and A234504] for related data.

Conjecture 3.28 (2012-12-24)

(i) Any integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $(k + 1)^{\varphi(m)} + k$ prime. Also, each integer $n > 1$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $k(k + 1)^{\varphi(m)} + 1$ prime.

(ii) Any integer $n > 5$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $(k + 1)^{\varphi(m)/2} - k$ prime. Also, each integer $n > 3$ can be written as $k + m$ ($k, m \in \mathbb{Z}^+$) with $k(k + 1)^{\varphi(m)/2} - 1$ prime.

Remark 3.28 This conjecture is somewhat curious. See [16, A234360] for related data.

Conjecture 3.29 *(i) (2014-02-02) Any integer $n > 8$ can be written as $i + j$ with $i, j \in \mathbb{Z}^+$ and $i < j$ such that $\varphi(i)\varphi(j)$ is a square. Also, for each $k = 3, 4, \ldots$, any integer $n \geqslant 3k$ can be written as $i_1 + i_2 + \ldots + i_k$ with $i_1, i_2, \ldots, i_k \in \mathbb{Z}^+$ not all equal such that $\varphi(i_1)\varphi(i_2) \cdots \varphi(i_k)$ is a kth power.*

(ii) *(2014-02-09) Any integer $n \geqslant 8$ can be written as $i + j$ with $i, j \in \mathbb{Z}^+$ and $i < j$ such that $\varphi(ij) + 1$ is a square. Also, for each $k = 3, 4, \ldots$, any integer $n > 2k + 1$ can be written as $\sum_{j=1}^{k} i_j$ with $i_1, i_2, \ldots, i_k \in \mathbb{Z}^+$ such that $\varphi(i_1 i_2 \ldots i_k)$ is a kth power.*

(iii) *(2014-02-04) Let $k > 1$ be an integer. Any sufficiently large integer n can be written as $\sum_{j=1}^{k} i_j$ with $i_1, \ldots, i_k \in \mathbb{Z}^+$ and $i_1 < \ldots < i_k$ such that all those $\varphi(i_j)$ $(j = 1, \ldots, k)$ are kth powers.*

(iv) *(2014-02-02) For each $k = 3, 4, \ldots$ any sufficiently large integer n can be written as $i_1 + i_2 + \ldots + i_k$ with $i_1, i_2, \ldots, i_k$ not all equal such that the product $i_1 i_2 \ldots i_k$ is a kth power.*

Remark 3.29 See [16, A236998, A233386, A237523, A237524, A237123, A237050] for related data. For any integer $k > 1$, we clearly have $2k + 2 = 4 + (k - 1)2$ with $\varphi(4 \cdot 2^{k-1}) = 2^k$ a kth power. In contrast with part (i) of Conjecture 3.29, we also conjecture that (cf. [16, A237049]) for each $k = 2, 3, 4, \ldots$, any sufficiently large integer n can be written as $\sum_{j=1}^{k} i_j$ with $i_1, i_2, \ldots, i_k \in \mathbb{Z}^+$ not all equal such that $\prod_{j=1}^{k} \sigma(i_j)$ is a kth power, where $\sigma(m)$ denotes the sum of all positive divisors of $m \in \mathbb{Z}^+$.

Conjecture 3.30 *(i)* *(2013-12-21) Any integer $n > 5$ can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ such that $(\varphi(k) + \varphi(m))/2$ is prime.*

(ii) *(2013-12-22) Any positive integer n not dividing 6 can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ such that $k\varphi(m) + 1$ is a square. Also, any integer $n > 4$ can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ and $k < m$ such that $k\varphi(m) - 1$ and $k\varphi(m) + 1$ are both prime.*

(iii) *(2013-12-12) Any integer $n > 5$ can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ such that $\varphi(k)\varphi(m) - 1$ and $\varphi(k)\varphi(m) + 1$ are both prime.*

(iv) *(2013-12-23) Any integer $n > 4$ can be written as $k + m$ $(k, m \in \mathbb{Z}^+)$ with $\varphi(k^2)\varphi(m) - 1$ a Sophie Germain prime.*

Remark 3.30 See [16, A233918, A234200, A234246, A233547, A234308] for related data. For example, $13 = 3 + 10$ with $(\varphi(3) + \varphi(10))/2 = 3$ prime, $13 = 4 + 9$ with $4\varphi(9) + 1 = 25$ a square, $18 = 5 + 13$ with $\{5\varphi(13) \pm 1\} = \{59, 61\}$ a twin prime pair, $26 = 7 + 19$ with $\{\varphi(7)\varphi(19) \pm 1\} = \{107, 109\}$ a twin prime pair, and $30 = 2 + 28$ with $\varphi(2^2)\varphi(28) - 1 = 23$ a Sophie Germain prime.

Conjecture 3.31 (2013-12-12)

(i) *Any integer $n > 1$ can be written as $k^2 + m$ with $\sigma(k^2) + \varphi(m)$ prime, where k and m are positive integers with $k^2 \leqslant m$.*

(ii) *Any integer $n > 1$ can be written as $k + m$ with $k, m \in \mathbb{Z}^+$ such that $\sigma(k)^2 + \varphi(m)$ (or $\sigma(k) + \varphi(m)^2$) is prime.*

Remark 3.31 See [16, A233544] for related data and comments. We have verified part (i) of Conjecture 3.31 for all $n = 2, \ldots, 10^8$; for example, $25 = 2^2 + 21$ with $\sigma(2^2) + \varphi(21) = 7 + 12 = 19$ prime.

Conjecture 3.32 *Let $n > 2$ be an integer.*

(i) *(2013-12-14) If n is even, then n can be written as $p + \sigma(k)$, where p is an odd prime and $k \in \{1, \ldots, n-1\}$.*

(ii) *(2013-12-17) If n is odd, then n can be written as $p + \varphi(k^2)$, where p is a prime and k is a positive integer with $k^2 < n$.*

Remark 3.32 See [16, A233654, A233793 and A233867] for related data. For example, $28 = 13 + \sigma(8)$ with 13 prime, and $29 = 23 + \varphi(3^2)$ with 23 prime. Note that if $n = p + q$ with p and q both prime, then $n + 1 = p + (q + 1) = p + \sigma(q)$ and $n - 1 = p + (q - 1) = p + \varphi(q)$.

Conjecture 3.33 (2012-12-29)

(i) *For each integer $n > 8$ with $n \neq 14$, there is a prime p between n and $2n$ with $(\frac{n}{p}) = 1$. If $n \in \mathbb{Z}^+$ is not a square, then there is a prime p between n and $2n$ with $(\frac{n}{p}) = -1$.*

(ii) *For any integer $n > 5$, there is a prime $p \in (n, 2n)$ with $(\frac{2n}{p}) = 1$. For any integer $n > 6$, there is a prime $p \in (n, 2n)$ with $(\frac{-n}{p}) = -1$.*

Remark 3.33 We have verified this refinement of Bertrand's Postulate for n up to 5×10^8.

Conjecture 3.34 (2012-12-29) *For any positive integer n, there is a prime p between n^2 and $(n+1)^2$ with $(\frac{n}{p}) = 1$. Also, for any integer $n > 1$, we have $(\frac{n(n+1)}{p}) = 1$ for some prime $p \in (n^2, (n+1)^2)$.*

Remark 3.34 We have verified this refinement of Legendre's conjecture for n up to 10^9.

Conjecture 3.35 (Olivier Gerard and Zhi-Wei Sun, 2012-11-19) *For any integer $n \geqslant 400$ with $n \neq 757,\ 1069,\ 1238$, there are odd primes p and q with $(\frac{p}{q}) = (\frac{q}{p}) = 1$ such that $p + (1 + \{n\}_2)q = n$.*

Remark 3.35 We have verified Conjecture 3.35 for n up to 10^8. See [5] for the announcement of this conjecture.

Conjecture 3.36 (2012-11-22) *Let m be any integer. Then, for every sufficiently large integer n, there are primes $p > q > 2$ with $(\frac{p-(1+\{n\}_2)m}{q}) = (\frac{q+m}{p}) = 1$ and $p + (1 + \{n\}_2)q = n$.*

Remark 3.36 Conjecture 3.36 in the case $m = 0$ corresponds to Conjecture 3.35.

Conjecture 3.37 (2012-12-30) *Any integer $n > 5$ can be written as $p + (1 + \{n\}_2)q$, where p is an odd prime and q is a prime not exceeding $n/2$ such that $(\frac{q}{n}) = 1$ if $2 \nmid n$, and $(\frac{(q+1)/2}{n+1}) = 1$ if $2 \mid n$.*

Remark 3.37 We have verified this refinement of Goldbach's and Lemoine's conjectures for n up to 10^9.

Conjecture 3.38 (2013-01-19)

(i) *Any even integer* $2n > 4$ *can be written as* $p + q = (p + 1) + (q - 1)$*, where* p *and* q *are primes with* $p + 1$ *and* $q - 1$ *both practical.*

(ii) *For each integer* $n > 8$*, we can write* $2n - 1 = p + q = 2p + (q - p)$*, where* p *and* $q - p$ *are both prime, and* q *is practical.*

Remark 3.38 We have verified both parts of Conjecture 3.38 for n up to 10^8. See [16, A209320 and A209315] for related data.

If one of n and $n + 1$ is prime and the other is practical, then we call $\{n, n + 1\}$ a *couple*. As powers of two are practical numbers, $\{2^p - 1, 2^p\}$ is a couple if $2^p - 1$ is a Mersenne prime, and $\{2^{2^n}, 2^{2^n} + 1\}$ is a couple if $2^{2^n} + 1$ is a Fermat prime. If p is a prime and $p - 1$ and $p + 1$ are both practical, then we call $\{p - 1, p, p + 1\}$ a *sandwich of the first kind*. If $\{p, p + 2\}$ is a twin prime pair and $p + 1$ is practical, then we call $\{p, p + 1, p + 2\}$ a *sandwich of the second kind*. For example, $\{88, 89, 90\}$ is a sandwich of the first kind, while $\{59, 60, 61\}$ is a sandwich of the second kind. See [16, A210479] for the list of the first 10,000 sandwiches of the first kind, and [16, A258838] for the list of the first 10,000 sandwiches of the second kind.

Conjecture 3.39 (2013-01-12)

(i) *For any integer* $n > 8$*, the interval* $[n, 2n]$ *contains a sandwich of the first kind.*

(ii) *For each* $n = 7, 8, \ldots$*, the interval* $[n, 2n]$ *contains a sandwich of the second kind.*

(iii) *For any integer* $n > 231$*, the interval* $[n, 2n]$ *contains four consecutive integers* $p - 1, p, p + 1, p + 2$ *with* $\{p, p + 2\}$ *a twin prime pair and* $\{p - 1, p + 1\}$ *a twin practical pair.*

(iv) *There are infinitely many quintuples* $\{m - 2, m - 1, m, m + 1, m + 2\}$ *with* $\{m - 1, m + 1\}$ *a twin prime pair and* $m, m \pm 2$ *all practical.*

Remark 3.39 For those middle terms m described in part (iv) of Conjecture 3.39, the reader may consult [16, A209236]. It is known that (cf. [12]) there are infinitely many practical numbers m with $m \pm 2$ also practical.

Conjecture 3.40 *(i)* *(2013-01-23) Each* $n = 4, 5, \ldots$ *can be written as* $p + q$*, where* $\{p - 1, p, p + 1\}$ *is a sandwich of the first kind, and* q *is either prime or practical.*

(ii) *(2013-01-29) Any integer* $n > 11$ *can be written as* $(1 + \{n\}_2)p + q + r$*, where* $\{p - 1, p, p + 1\}$ *and* $\{q - 1, q, q + 1\}$ *are sandwiches of the first kind, and* $\{r - 1, r, r + 1\}$ *is a sandwich of the second kind.*

Remark 3.40 We have verified parts (i) and (ii) of Conjecture 3.40 for n up to 10^8 and 10^7, respectively. For numbers of representations related to parts (i) and (ii), see [16, A210480 and A210681].

Conjecture 3.41 (2013-01-29)

(i) *Any integer $n > 6$ can be written as $p + q + r$ such that $\{p - 1, p, p + 1\}$ and $\{q - 1, q, q + 1\}$ are sandwiches of the first kind, and $\{6r - 1, 6r, 6r + 1\}$ is a sandwich of the second kind.*

(ii) *Every $n = 3, 4, \dots$ can be expressed as $x + y + z$ with $x, y, z \in \mathbb{Z}^+$ such that $\{6x - 1, 6x, 6x + 1\}$, $\{6y - 1, 6y, 6y + 1\}$, and $\{6z - 1, 6z, 6z + 1\}$ are all sandwiches of the second kind.*

(iiii) *Each integer $n > 7$ can be written as $p + q + x^2$ with $x \in \mathbb{Z}$ such that $\{p - 1, p, p + 1\}$ is a sandwich of the first kind and $\{q - 1, q, q + 1\}$ is a sandwich of the second kind.*

Remark 3.41 We also conjecture that each $n = 3, 4, \dots$ can be written as the sum of two triangular numbers and a prime p with $\{p - 1, p, p + 1\}$ a sandwich of the first kind. See [16, A210681] for related comments.

Conjecture 3.42 (2013-01-30)

(i) *For any integer $n > 8$, we can write $2n = p + 2q + 3r$, where $\{p - 1, p, p + 1\}$, $\{q - 1, q, q + 1\}$, and $\{r - 1, r, r + 1\}$ are all sandwiches of the first kind.*

(ii) *Each integer $n > 5$ can be written as the sum of a prime p with $\{p - 1, p, p + 1\}$ a sandwich of the first kind, a prime q with $q + 2$ also prime, and a Fibonacci number.*

Remark 3.42 See [16, A211190 and A211165] for related data. We have verified part (ii) of Conjecture 3.42 for n up to 2, 000, 000.

Conjecture 3.43 *(i)* *(2013-01-14) Any odd number $n > 1$ can be expressed as $p + q$, where p is a Sophie Germain prime and q is a practical number.*

(ii) *(2013-01-19) For any integer $n > 2$, there is a practical number $q < n$ such that $n - q$ and $n + q$ are both prime or both practical.*

Remark 3.43 We have verified this conjecture for n up to 10^8. See [16, A209253 and A209312] for related data. We also conjecture that each positive integer can be represented as the sum of a practical number and a triangular number (cf. [16, A208244]), which is an analog of the author's conjecture on sums of primes and triangular numbers (cf. [18]).

Conjecture 3.44 (2015-08-28)

(i) *For any integer $n > 6$, there is a prime $p < n$ such that $n - (p + 1)$ and $n + (p + 1)$ are both prime or both practical.*

(ii) *For any integer $n > 2$, there is a prime $p < n$ such that $n - (p - 1)$ and $n + (p - 1)$ are both prime or both practical.*

Remark 3.44 See [16, A261653] for related data, and compare this conjecture with Conjectures 2.2, 2.3 and 3.43.

Conjecture 3.45 (2015-07-12)

(i) There are infinitely many sandwiches $\{n-1, n, n+1\}$ *of the first kind such that* $\{p_n-1, p_n, p_n+1\}$ *is also a sandwich of the first kind.*

(ii) There are infinitely many sandwiches $\{n-1, n, n+1\}$ *of the second kind such that* $\{p_n-1, p_n, p_n+1\}$ *is a sandwich of the first kind.*

Remark 3.45 See [16, A257924 and A257922] for related data.

4 On Representations of Positive Rational Numbers

It is well known that any positive rational number can be written as finitely many distinct unit fractions. It is also known that the series $\sum_{n=1}^{\infty} 1/p_n$ diverges as proved by Euler.

Conjecture 4.1 *(i) (2015-09-09) For any positive rational number r, there are finitely many distinct primes $q_1, \ldots, q_k$ such that*

$$r=\sum_{j=1}^{k}\frac{1}{q_j-1}.$$

(ii) (2015-09-12) For any positive rational number r, there are finitely many distinct primes $q_1, \ldots, q_k$ such that

$$r=\sum_{j=1}^{k}\frac{1}{q_j+1}.$$

(iii) (2015-09-12) For any positive rational number r, there are finitely many distinct practical numbers $q_1, \ldots, q_k$ with $r=\sum_{j=1}^{k} 1/q_j$.

Remark 4.1 For example,

$$2=\frac{1}{2-1}+\frac{1}{3-1}+\frac{1}{5-1}+\frac{1}{7-1}+\frac{1}{13-1}=\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{12}$$

with 2, 3, 5, 7 all prime and 1, 2, 4, 6, 12 all practical, and

$$1=\frac{1}{2+1}+\frac{1}{3+1}+\frac{1}{5+1}+\frac{1}{7+1}+\frac{1}{11+1}+\frac{1}{23+1}$$

with 2, 3, 5, 7, 11, 23 all prime. Also,

$$\begin{aligned}\frac{10}{11}&=\frac{1}{3-1}+\frac{1}{5-1}+\frac{1}{13-1}+\frac{1}{19-1}+\frac{1}{67-1}+\frac{1}{199-1}\\&=\frac{1}{2+1}+\frac{1}{3+1}+\frac{1}{5+1}+\frac{1}{7+1}+\frac{1}{43+1}+\frac{1}{131+1}+\frac{1}{263+1}\\&=\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{48}+\frac{1}{132}+\frac{1}{176}\end{aligned}$$

with $2, 3, 5, 7, 13, 19, 43, 67, 131, 199, 263$ all prime and $2, 4, 8, 48, 132, 176$ all practical. After learning Conjecture 4.1 from the author, Qing-Hu Hou verified parts (i) and (ii) in November 2015 for all rational numbers $r \in (0, 1)$ with denominators not exceeding 100. The author would like to offer 500 US dollars as the prize for the first solution to parts (i) and (ii) of Conjecture 4.1.

Conjecture 4.2 (2015-09-09) *Let m be any positive integer.*

(i) All the rational numbers

$$\sum_{i=j}^{k}\frac{1}{(p_i-1)^m}\ \text{with } 1 \leqslant j \leqslant k$$

are pairwise distinct! If

$$\sum_{i=j}^{k}\frac{1}{(p_i-1)^m}\quad\text{and}\quad\sum_{r=s}^{t}\frac{1}{(p_r-1)^m}$$

have the same fractional part with

$$0<\min\{2,k\}\leqslant j\leqslant k,\ 0<\min\{2,t\}\leqslant s\leqslant t \text{ and } j\leqslant s,$$

but the ordered pairs (j, k) and (s, t) are different, then we must have $m = 1$ and

$$\sum_{i=j}^{k}\frac{1}{p_i-1}=1+\sum_{r=s}^{t}\frac{1}{p_r-1};$$

moreover, either $(j, k) = (2, 6)$ and $(s, t) = (5, 5)$, or $(j, k) = (2, 5)$ and $(s, t) = (18, 18)$, or $(j, k) = (2, 17)$ and $(s, t) = (6, 18)$.

(ii) If

$$\sum_{i=j}^{k}\frac{1}{(p_i+1)^m}\quad\text{and}\quad\sum_{r=s}^{t}\frac{1}{(p_r+1)^m}$$

have the same fractional part with

$$1\leqslant j\leqslant k,\ 1\leqslant s\leqslant t \text{ and } j\leqslant s,$$

but the ordered pairs (j,k) *and* (s,t) *are different, then we must have* $m=1$ *and*

$$\sum_{i=j}^{k}\frac{1}{p_i+1}-\sum_{r=s}^{t}\frac{1}{p_r+1}\in\{0,1\};$$

moreover, $(j,k)=(1,9)$ *and* $(s,t)=(6,8)$*, or* $(j,k)=(4,4)$ *and* $(s,t)=(8,10)$*, or* $(j,k)=(4,7)$ *and* $(s,t)=(5,10)$*, or* $(j,k)=(1,10)$ *and* $(s,t)=(5,7)$.

(iii) For any integer $d>1$*, the rational numbers*

$$\sum_{i=j}^{k}\frac{1}{(p_i+d)^m}\ \ \textit{with}\ 1\leqslant j\leqslant k$$

have pairwise distinct fractional parts.

Remark 4.2 Recall that $\sum_{j=1}^{\infty}1/p_j$ diverges.

Actually, Conjecture 4.2 was motivated by our following conjecture whose proofs might involve primes.

Conjecture 4.3 *(i) (2015-09-09) If* $1/j+\cdots+1/k$ *and* $1/s+\cdots+1/t$ *have the same fractional part with*

$$0<\min\{2,k\}\leqslant j\leqslant k,\ 0<\min\{2,t\}\leqslant s\leqslant t\ \textit{and}\ j\leqslant s,$$

but the ordered pairs (j,k) *and* (s,t) *are different, then we have*

$$\frac{1}{j}+\ldots+\frac{1}{k}=1+\frac{1}{s}+\ldots+\frac{1}{t};$$

moreover, one of the following (a)-(d) *holds.*

(a) $(j,k)=(2,6)$ *and* $(s,t)=(4,5)$*,*
(b) $(j,k)=(2,4)$ *and* $(s,t)=(12,12)$*,*
(c) $(j,k)=(2,11)$ *and* $(s,t)=(5,12)$*,*
(d) $(j,k)=(3,20)$ *and* $(s,t)=(7,19)$.

(ii) (2015-09-11) Let $a>b\geqslant 0$ *and* $m>0$ *be integers with* $\gcd(a,b)=1<\max\{a,m\}$*. Then, the numbers*

$$\sum_{i=j}^{k}\frac{1}{(ai-b)^m}\ \ \textit{with}\ 1\leqslant j\leqslant k\ \textit{and}\ (j>1\ \textit{if}\ k>a-b=1)$$

have pairwise distinct fractional parts. Also, for each $r=0,1$*, the numbers*

$$\sum_{i=j}^{k}\frac{(-1)^{i-jr}}{(ai-b)^m}\ \text{with } 1 \leqslant j \leqslant k \text{ and } (j > 1 \text{ if } k > a-b=1)$$

have pairwise distinct fractional parts.

Remark 4.3 In 1918, J. Kürschak proved that for any integers $k \geqslant j > 1$, the number $1/j + \cdots + 1/k$ is not an integer. In 1946, Erdős and Niven [4] used Sylvester's theorem (which states that the product of n consecutive integers greater than n is divisible by a prime greater than n) to show that all the numbers $1/j + \cdots + 1/k$ with $1 \leqslant j \leqslant k$ are pairwise distinct.

If $d \in \mathbb{Z}^+$ is not a square, then the Pell equation $x^2 - dy^2 = 1$ has infinitely many integral solutions. Thus, for $r = a/b$ with $a, b \in \mathbb{Z}^+$ and $\gcd(a, b) = 1$, if r is not a square of rational numbers, then there is a positive integer k such that $(ka)(kb) + 1$ is a square, i.e., we can write $a/b = m/n$ with $m, n \in \mathbb{Z}^+$ such that $mn + 1$ is a square. Motivated by this, below we consider various representations of positive rational numbers.

Conjecture 4.4 *(i) (2015-07-03) The set*

$$\left\{\frac{m}{n} :\ m, n \in \mathbb{Z}^+ \text{ and } p_m + p_n \text{ is a square}\right\}$$

contains any positive rational number r. Also, any rational number $r > 1$ can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $p_m - p_n$ is a square.

(ii) (2015-08-20) Any positive rational number $r \neq 1$ can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $p_{p_m} + p_{p_n}$ is a square.

Remark 4.4 We have verified part (i) of Conjecture 4.4 for all those rational numbers $r = a/b$ with $a, b \in \{1, \ldots, 200\}$ (cf. [16, A259712 and A257856]) and part (ii) of Conjecture 4.4 for all those rational numbers $r = a/b \neq 1$ with $a, b \in \{1, \ldots, 60\}$. For example, $2 = 20/10$ with $p_{20} + p_{10} = 71 + 29 = 10^2$, and $2 = 92/46$ with $p_{p_{92}} + p_{p_{46}} = p_{479} + p_{199} = 3407 + 1217 = 68^2$.

Conjecture 4.5 (2015-07-08) *The set*

$$\left\{\frac{m}{n} :\ m, n \in \mathbb{Z}^+, \text{ and } \varphi(m) \text{ and } \sigma(n) \text{ are both squares}\right\}$$

contains any positive rational number r.

Remark 4.5 We have verified Conjecture 4.5 for all those $r = a/b$ with $a, b \in \{1, \ldots, 150\}$ (cf. [16, A259915 and A259916]). For example, $4/5 = 136/170$ with $\varphi(136) = 8^2$ and $\sigma(170) = 18^2$, and $5/4 = 1365/1092$ with $\varphi(1365) = 24^2$ and $\sigma(1092) = 56^2$.

Conjecture 4.6 *(i) (2015-07-05) Any positive rational number r can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $\pi(m)\pi(n)$ is a positive square.*

(ii) *(2015-07-06) Any positive rational number r can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $\pi(m)$ and $\pi(\pi(n))$ are positive squares.*

Remark 4.6 We have verified part (i) of this conjecture for all those rational numbers $r = a/b$ with $a, b \in \{1, \ldots, 60\}$. See [16, A259789] for related data. For example, $49/58 = 1076068567/1273713814$ with

$$\pi(1076068567)\pi(1273713814) = 54511776 \cdot 63975626 = 59054424^2.$$

Conjecture 4.7 (2015-07-10) *Each positive rational number $r < 1$ can be written as m/n with $1 < m < n$ such that $\pi(m)^2 + \pi(n)^2$ is a square. Also, any rational number $r > 1$ can be written as m/n with $m > n > 1$ such that $\pi(m)^2 - \pi(n)^2$ is a square.*

Remark 4.7 We have verified this conjecture for all those rational numbers $r = a/b$ with $a, b \in \{1, \ldots, 50\}$. See [16, A255677] for related data. For example, $23/24 = 19947716/20815008$ with

$$\pi(19947716)^2 + \pi(20815008)^2 = 1267497^2 + 1319004^2 = 1829295^2,$$

and $7/3 = 26964/11556$ with

$$\pi(26964)^2 - \pi(11556)^2 = 2958^2 - 1392^2 = 2610^2.$$

Motivated by Conjecture 4.7, we raise the following conjecture which sounds interesting and challenging.

Conjecture 4.8 *(i) (2015-07-11) For any $n \in \mathbb{Z}^+$, there are distinct primes p, q, r such that $\pi(pn)^2 = \pi(qn)^2 + \pi(rn)^2$.*

(ii) (2015-07-13) For any $n \in \mathbb{Z}^+$, there are distinct primes p, q, r with $\pi(pn) = \pi(qn)\pi(rn)$ (or $\pi(pn) = \pi(qn) + \pi(rn)$).

Remark 4.8 See [16, A257364 and A257928] for related data.

Conjecture 4.9 *(i) (2015-07-02) Any positive rational number r can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $p(m)^2 + p(n)^2$ is prime, where $p(\cdot)$ is the partition function.*

(ii) (2015-08-20) Any positive rational number $r \neq 1$ can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $p(p_m) + p(p_n)$ is prime.

Remark 4.9 Conjecture 4.9 implies that there are infinitely many primes of the form $p(m)^2 + p(n)^2$ with $m, n \in \mathbb{Z}^+$ as well as primes of the form $p(q) + p(r)$ with q and r both prime. We have verified part (i) for all those rational numbers $r = a/b$ with $a, b \in \{1, \ldots, 100\}$, and part (ii) for all those rational numbers $r = a/b \neq 1$ with $a, b \in \{1, \ldots, 37\}$. See [16, A259531, A259678, A261513 and A261515] for related data. For example, $4/5 = 124/155$ with

$$\begin{aligned} p(124)^2 + p(155)^2 &= 2841940500^2 + 66493182097^2 \\ &= 4429419891190341567409 \end{aligned}$$

prime, and $3 = 138/46$ with

$$\begin{aligned} p(p_{138}) + p(p_{46}) &= p(787) + p(199) \\ &= 3223934948277725160271634798 + 3646072432125 \\ &= 3223934948277728806344066923 \end{aligned}$$

prime.

Conjecture 4.10 (2015-08-17) *Any positive rational number r can be written as m/n, where m and n are positive integers with $(m \pm 1)^2 + n^2$ and $m^2 + (n \pm 1)^2$ all prime.*

Remark 4.10 We have verified this for all those $r = a/b$ with $a, b \in \{1, \ldots, 60\}$. See [16, A261382] for related data. It is easy to prove that if m and n are positive integers with $(m \pm 1)^2 + n^2$ and $m^2 + (n \pm 1)^2$ all prime, then either $m = n = 2$ or $m \equiv n \equiv 0 \pmod 5$.

Conjecture 4.11 *(i) (2015-06-28) Each rational number $r > 0$ can be written as m/n, where m and n are positive integers with*

$$p_m \pm m,\ \ p_n \pm n,\ \ p_m + n \text{ and } p_n + m$$

all prime.

(ii) (2015-07-02) Any rational number $r > 0$ can be written as m/n, where m and n are positive integers with

$$m^2 + p_m^2,\ \ n^2 + p_n^2,\ \ m^2 + p_n^2 \text{ and } n^2 + p_m^2$$

all prime.

(iii) (2015-08-15) Any rational number $r > 0$ can be written as m/n with m and n in the set

$$\begin{aligned} &\{k \in \mathbb{Z}^+ :\ k+1,\ k^2+1 \text{ and } k^2 + p_k^2 \text{ are all prime}\} \\ =&\{q-1 :\ q,\ (q-1)^2+1 \text{ and } (q-1)^2 + p_{(q-1)}^2 \text{ are all prime}\}. \end{aligned}$$

Remark 4.11 We have verified parts (i)–(ii) for those $r = a/b$ with $a, b \in \{1, \ldots, 150\}$ and part (iii) for those $r = a/b$ with $a, b \in \{1, \ldots, 60\}$. See [16, A259492 and A261339] for related data.

Conjecture 4.12 *(i) (2015-06-30) Let*

$$U := \{n \in \mathbb{Z}^+ :\ n \pm 1 \text{ and } p_n + 2 \text{ are all prime}\}.$$

Then, any positive rational number r can be written as m/n with $m, n \in U$.

(ii) (2015-06-28) Let

$$V := \{n \in \mathbb{Z}^+ : \ p_n + 2 \text{ and } p_{p_n} + 2 \text{ are both prime}\}.$$

Then, any positive rational number r can be written as m/n with $m, n \in V$.

(iii) (2015-06-12) Let

$$Q := \{q \in \mathbb{Z}^+ : \ q \text{ is practical with } q \pm 1 \text{ twin prime}\}.$$

Then, any positive rational number r can be written as q/q' with $q, q' \in Q$.

Remark 4.12 We have verified part (i) for all those $r = a/b$ with $a, b \in \{1, \ldots, 100\}$, part (ii) for all those $r = a/b$ with $a, b \in \{1, \ldots, 400\}$, and part (iii) for all those $r = a/b$ with $a, b \in \{1, \ldots, 1000\}$. See [16, A259539, A259540, A259487, A259488 and A258836] for related data. For example, $4/5 = 11673840/14592300$ with 11673840 and 14592300 in the set U.

Motivated by part (i) of Conjecture 4.12 and [22, Conjecture 3.7(i)], we pose the following conjecture.

Conjecture 4.13 (2015-07-01) *There are infinitely many positive integers n such that the seven numbers*

$$n \pm 1, \ \ p_n + 2, \ \ p_n \pm n, \ \ np_n \pm 1$$

are all prime.

Remark 4.13 We have listed the first 160 such positive integers n the least of which is 2523708 (cf. [16, A259628]).

Conjecture 4.14 (2015-08-24) *Any positive rational number r can be written as m/n, where m and n belong to the set*

$$\{k \in \mathbb{Z}^+ : \ p_k + 2, \ \ p_k + 6 \text{ and } p_k + 8 \ \text{ are all prime}\}.$$

Also, each positive rational number r can be written as m/n, where m and n belong to the set

$$\{k \in \mathbb{Z}^+ : \ p_k + 4, \ \ p_k + 6 \text{ and } p_k + 10 \ \text{ are all prime}\}.$$

Remark 4.14 This conjecture implies that there are infinitely many prime quadruples $(p, p + 2, p + 6, p + 8)$ as well as $(p, p + 4, p + 6, p + 10)$, which is a special case of Schinzel's Hypothesis. See [16, A261541] for related data. For example, $3/4 = m/n$ with $m = 20723892$ and $n = 27631856$, and

$$p_m+2=387875563,\ p_m+6=387875567,\ p_m+8=387875569,$$
$$p_n+2=525608593,\ p_n+6=525608597,\ p_n+8=525608599$$

are all prime.

Conjecture 4.15 (2015-08-23) *Any positive rational number r can be written as m/n, where m and n belong to the set*

$$W=\{k\in\mathbb{Z}^+:\ p_k+2 \text{ is prime and } p_{p_k+2}-p_{p_k}=6\}.$$

Remark 4.15 See [16, A261528 and A261533] for related data. For example, $2=1782/891$ with 891 and 1782 in the set W. Conjecture 4.15 implies that there are infinitely many twin prime pairs $\{q,q+2\}$ with $p_{q+2}-p_q=6$.

Conjecture 4.16 (2015-08-14) *Each positive rational number r can be written as m/n with m and n in the set*

$$\{k\in\mathbb{Z}^+:\ p_k^2-2 \text{ and } p_{p_k}^2-2 \text{ are both prime}\}.$$

Remark 4.16 We have verified this for all those $r=a/b$ with $a,b\in\{1,\ldots,300\}$. See [16, A261281] for related data.

Conjecture 4.17 *(i) (2014-05-14) For any prime $p>5$, there is a positive square $k^2<p$ such that the inverse of k^2 modulo p is prime, where the inverse of $a\in\{1,\ldots,p-1\}$ modulo p denotes the unique $x\in\{1,\ldots,p-1\}$ with $ax\equiv 1 \pmod p$.*

(ii) (2015-08-18) Any positive rational number $r\leqslant 1$ can be written as m/n with $m,n\in\mathbb{Z}^+$ such that the inverse of m modulo p_n is a square.

Remark 4.17 We have checked part (i) of Conjecture 4.17 for those primes $p<1.8\times 10^8$. See [16, A242425 and A242441] for related data. For example, the inverse of 4^2 modulo 23 is the prime 13.

Conjecture 4.18 (2014-08-26)

(i) Any integer $n>2$ with $n\neq 8$ can be written as $k+m$ with $k,m\in\mathbb{Z}^+$ and $k\neq m$ such that p_k is a primitive root modulo p_m and p_m is also a primitive root modulo p_k.

(ii) Any positive rational number $r\neq 1$ can be written as m/n with $m,n\in\mathbb{Z}^+$ such that p_m is a primitive root modulo p_n and also p_n is a primitive root modulo p_m.

Remark 4.18 See [16, A261387] for related data and comments.

Conjecture 4.19 (2015-07-20) *Let $n\in\mathbb{Z}^+$ and $s,t\in\{1,-1\}$. Then, any positive rational number r_0 can be written as $(p_{qn}+s)/(p_{rn}+t)$ with q and r both prime, unless $n>r_0=1$ and $\{s,t\}=\{1,-1\}$.*

Remark 4.19 We have verified this conjecture in the case $n = 1$ for all those $r_0 = a/b$ with $a, b \in \{1, \ldots, 500\}$ (cf. [16, A258803]). For $n = 2, \ldots, 10$ we have verified Conjecture 4.19 for all those $r_0 = a/b$ with $a, b \in \{1, \ldots, 30\}$ (cf. [16, A260252]). For example, $23 = (p_{17209} - 1)/(p_{1039} - 1) = (190579 - 1)/(8287 - 1)$ with 1039 and 17209 both prime.

Conjecture 4.20 (2015-08-02)

(i) If a, b, c are positive integers with $\gcd(a, b) = \gcd(a, c) = \gcd(b, c) = 1$, and $a \neq b$ and $a + b \equiv c \pmod 2$, then for any $n \in \mathbb{Z}^+$ the linear equation $ax - by = c$ has solutions with x and y in the set $\{p_{qn} : q \text{ is prime}\}$.

(ii) Let a and b be relatively prime positive integers, and let c be any integer. For any $n \in \mathbb{Z}^+$, the linear equation $ax - by = c$ has solutions with x and y in the set $\{\pi(pn) : p \text{ is prime}\}$.

Remark 4.20 Note that part (i) of Conjecture 4.20 is an extension of Conjecture 4.19. In the $a = c = 1$ and $b = 2$, it asserts that for any $n \in \mathbb{Z}^+$, there are primes q and r such that $2p_{qn} + 1 = p_{rn}$. This implies that there are infinitely many Sophie Germain primes. Also, part (ii) of Conjecture 4.20 with $c = 0$ asserts that for any $n \in \mathbb{Z}^+$, the set

$$\left\{\frac{\pi(pn)}{\pi(qn)} :\ p \text{ and } q \text{ are primes}\right\}$$

contains all positive rational numbers (cf. [16, A260232]). We have checked both parts of the conjecture for $a, b, c = 1, \ldots, 20$ and $n = 1, \ldots, 30$. For related data, see [16, A260886 and A260888].

Recall that a prime p is called a Chen prime if $p + 2$ is a product of at most two primes. In 1973, Chen [1] proved that there are infinitely many Chen primes.

Conjecture 4.21 *(i) (2015-07-14) For any positive integer n, there are $i, j, k \in \mathbb{Z}^+$ with $i \neq j$ such that $p_{kn} + 2 = p_{in}p_{jn}$.*

(ii) (2015-07-15) For any positive integer n, there are $i, j, k \in \mathbb{Z}^+$ with $i \neq j$ such that $p_{kn}^2 - 2 = p_{in}p_{jn}$.

Remark 4.21 See [16, A257926 and A260080] for related data. Clearly, part (i) of Conjecture 4.21 implies that there are infinitely many Chen primes.

Conjecture 4.22 (2015-07-15) *Let d be a nonzero integer and let $n \in \mathbb{Z}^+$. Set*

$$D := \{p_{kn} + d :\ k = 1, 2, 3, \ldots\}.$$

(i) If $\gcd(6, d) = 1$, then there are two distinct elements x and y of D with $x + y \in D$ and $x - y \in D$.

(ii) For each $k = 1, 2$, we have $xy = z^k$ for some distinct elements x, y, z of D.

Remark 4.22 See [16, A260078, A257938 and A260082] for related data.

Conjecture 4.23 (2015-07-17)

(i) *Let $a, n \in \mathbb{Z}^+$ and $b, c \in \mathbb{Z}$ with $\gcd(a, b, c) = 1$, $2 \nmid (a + b + c)$ and $3 \nmid \gcd(b, a + c)$. If $b^2 - 4ac$ is not a square, then there are $x, y \in \{p_{kn} : k = 1, 2, 3, \ldots\}$ such that $y = ax^2 + bx + c$.*

(ii) *For any $a, n \in \mathbb{Z}^+$ and $b, c \in \mathbb{Z}$, there are $x, y \in \{\pi(pn) : p$ is prime$\}$ such that $y = ax^2 + bx + c$.*

Remark 4.23 See [16, A260120 and A260140] for related data. Part (i) of Conjecture 4.23 implies that for any $n \in \mathbb{Z}^+$, there are $j, k \in \mathbb{Z}^+$ with $p_{kn}^2 - 2 = p_{jn}$ (or $(p_{kn} - 1)^2 = p_{jn} - 1$). Part (ii) of Conjecture 4.23 implies that for any $n \in \mathbb{Z}^+$, there are primes p and q with $\pi(pn) = \pi(qn)^2$.

Conjecture 4.24 (2015-08-14) *Let*

$$S_1 := \{q + 1 : \ q \text{ and } p_q + 2 \ \text{ are both prime}\}$$

and

$$S_2 := \{q - 1 : \ q \text{ and } p_q - 2 \ \text{ are both prime}\}.$$

For any $i, j \in \{1, 2\}$, each positive rational number r can be written as m/n with $m \in S_i$ and $n \in S_j$, unless $i \neq j$ and $r = 1$.

Remark 4.24 See [16, A261295] for related data. For example, $4/5 = 15648/19560$ with 15647, $p_{15647} + 2 = 171763$, 19559, and $p_{19559} + 2 = 219409$ all prime. A twin prime pair $\{p, p + 2\}$ with $\pi(p)$ also prime is called a super twin prime pair (cf. [22, Conjecture 3.2 and Remark 3.2]).

Conjecture 4.25 (2015-08-18) *Let $s, t \in \{1, -1\}$. Then, any positive rational number r can be written as m/n with m and n in the set*

$$K_{s,t} := \{k \in \mathbb{Z}^+ : \ p_{p_k} + sp_k + t = p_q \quad \text{for some prime } q\}.$$

Remark 4.25 This implies that for any $s, t \in \{\pm 1\}$, there are infinitely many primes q with $p = p_q + sq + t$ and $\pi(p)$ both prime. See [16, A260753 and A261136] for related data. For example, $3 = 6837/2279$, and

$$p_{p_{6837}} - p_{6837} + 1 = p_{68777} - 68777 + 1 = 865757 - 68776 = 796981 = p_{63737}$$

with 63737 prime, and

$$p_{p_{2279}} - p_{2279} + 1 = p_{20147} - 20147 + 1 = 226553 - 20146 = 206407 = p_{18503}$$

with 18503 prime.

Conjecture 4.26 (2015-08-16) *Any positive rational number can be written as m/n, where m and n are positive integers with $p_{p_m} p_{p_n} = p_q + 2$ for some prime q.*

Remark 4.26 See [16, A261352 and A261353] for related data. For example, $4 = 2424/606$ and

$$p_{p_{2424}} p_{p_{606}} = p_{21589} p_{4457} = 244471 \cdot 42643 = 10424976853 = p_{473490161} + 2$$

with 473490161 prime. Conjecture 4.26 implies that there are infinitely many prime triples (q, r, s) with $p_q + 2 = p_r p_s$.

Conjecture 4.27 (2014-08-17)

(i) Let d be any nonzero integer. Then any positive rational number r can be written as m/n with $m, n \in \mathbb{Z}^+$ such that $(p_{p_m} + d)(p_{p_n} + d) = p_q + d$ for some prime q.

(ii) For any nonzero integer d, there are infinitely many prime triples (q, r, s) with q, r, s distinct such that $(p_q + d)^2 = (p_r + d)(p_s + d)$.

Remark 4.27 See [16, A261385 and A261395] for related data and comments. Clearly, for each $d \in \mathbb{Z} \setminus \{0\}$, part (i) of Conjecture 4.27 implies that the equation $xy = z$ has infinitely many solutions with $x, y, z \in \{p_q + d : q \text{ is prime}\}$, and part (ii) of Conjecture 4.27 implies that the set $\{p_q + d : q \text{ is prime}\}$ contains infinitely many nontrivial three-term geometric progressions.

Conjecture 4.28 (2015-08-16)

(i) Let $a, b, c \in \mathbb{Z}^+$ with $a \neq b$, $a + b \equiv c \pmod 2$ and $\gcd(a, b) = \gcd(a, c) = \gcd(b, c) = 1$. Then, any positive rational number r can be written as m/n with m and n in the set

$$\{k \in \mathbb{Z}^+ : \ ap_q - bp_{p_k} = c \ \text{ for some prime } q\},$$

and thus there are infinitely many pairs of primes q and r such that $ap_q - bp_r = c$.

(ii) Let $a \in \mathbb{Z}^+$ and $b, c \in \mathbb{Z}$ with $\gcd(a, b, c) = 1$. If $2 \nmid (a + b + c)$, $3 \nmid \gcd(b, a + c)$, and $b^2 - 4ac$ is not a square, then the equation $y = ax^2 + bx + c$ has infinitely many solutions with $x, y \in \{p_q : q \text{ is prime}\}$.

Remark 4.28 See [16, A261361, A261362 and A261354] for related data and comments. Clearly, part (i) of Conjecture 4.28 implies that there are infinitely many prime pairs q and r with $2p_q + 1 = p_r$, and part (ii) of Conjecture 4.28 implies that there are infinitely many prime pairs q and r with $p_q^2 - 2 = p_r$.

Conjecture 4.29 *(i) (2015-08-18) For any $j = \pm 1$ and $n \in \mathbb{Z}^+$, there is a positive integer k such that $kn + j = p_q$ and $k^2 n + 1 = p_r$ for some pair of primes q and r.*

(ii) (2015-08-20) Each positive rational number $r \leqslant 1$ can be written as m/n, where m and n are positive integers such that $p_{p_m}, p_{p_n}, p_{p_k}, p_{p_l}$ form a four-term arithmetic progression for some $k, l \in \mathbb{Z}^+$.

(iii) (2015-08-25) Any positive rational number r can be written as m/n, where m and n are positive integers with $(p_{p_{p_m}} + p_{p_{p_n}})/2 = p_{p_q}$ for some prime q.

Remark 4.29 See [16, A261437, A261462 and A261583] for related data.

Motivated by Conjecture 4.29, we define

$$p_n^{(1)} = p_n, \text{ and } p_n^{(m+1)} = p_{p_n}^{(m)} \quad \text{for } m, n = 1, 2, 3, \ldots,$$

and pose the following conjecture.

Conjecture 4.30 (2015-08-25)

(i) If $q \in \mathbb{Z}^+$ and $a \in \mathbb{Z}$ are relatively prime, then for any $m \in \mathbb{Z}^+$ there are infinitely many $n \in \mathbb{Z}^+$ with $p_n^{(m)} \equiv a \pmod{q}$.

(ii) For any integer $k > 2$ and $m > 0$, the set $P_m := \{p_n^{(m)} : n \in \mathbb{Z}^+\}$ contains infinitely many nontrivial k-term arithmetic progressions.

(iii) For any $m, n \in \mathbb{Z}^+$, we have

$$\frac{\sqrt[n+1]{p_{n+1}^{(m+1)}}}{\sqrt[n]{p_n^{(m+1)}}} < \frac{\sqrt[n+1]{p_{n+1}^{(m)}}}{\sqrt[n]{p_n^{(m)}}} < 1.$$

Remark 4.30 Part (i) of Conjecture 4.30 is an extension of Dirichlet's theorem on primes in arithmetic progressions, and part (ii) of Conjecture 4.30 is an extension of the Green–Tao theorem [7]. Part (iii) is an analog of Firoobakht's conjecture (cf. [19]), and we also conjecture that the sequence $(\sqrt[n]{q_n})_{n \geqslant 3}$ is strictly decreasing if q_n denotes the nth practical number.

Acknowledgements The work was supported by the National Natural Science Foundation of China (Grant No. 11571162).

References

1. J.R. Chen, On the representation of a larger even integer as the sum of a prime and the product of at most two primes. Sci. Sinica **16**, 157–176 (1973)
2. R. Crandall, C. Pomerance, *Prime Numbers: A Computational Perspective*, 2nd edn. (Springer, New York, 2005)
3. R. Crocker, On the sum of a prime and two powers of two. Pac. J. Math. **36**, 103–107 (1971)
4. P. Erdős, I. Niven, Some properties of partial sums of the harmonic series. Bull. Am. Math. Soc. **52**(4), 248–251 (1946)
5. O. Gerard, Z.-W. Sun, Refining Goldbach's conjecture by using quadratic residues. A Message to Number Theory List (November 19, 2012), http://listserv.nodak.edu/cgi-bin/wa.exe?A2=NMBRTHRY;c08d598.1211
6. R.K. Guy, *Unsolved Problems in Number Theory*, 3rd edn. (Springer, New York, 2004)
7. B. Green, T. Tao, The primes contain arbitrary long arithmetic progressions. Ann. Math. **167**, 481–547 (2008)

8. D.R. Heath-Brown, Primes represented by $x^3 + 2y^3$. Acta Math. **186**, 1–84 (2001)
9. H.A. Helfgott, *The ternary Goldbach problem* (2015), arXiv:1501.05438
10. E. Lemoine, L'intermédiare des mathématiciens **1**, 179 (1894); ibid **3**, 151 (1896)
11. M. Margenstern, Les nombres pratiques: théorie, observations et conjectures. J. Number Theory **37**, 1–36 (1991)
12. G. Melfi, On two conjectures about practical numbers. J. Number Theory **56**, 205–210 (1996)
13. A. Murthy, Sequence A109909 in OEIS (On-Line Encyclopedia of Integer Sequences) and comments for A034693 in OEIS, http://www.oeis.org
14. M.B. Nathanson, in *Additive Number Theory: the Classical Bases*. Graduate Texts in Mathematics, vol. 164 (Springer, 1996)
15. B.M. Stewart, Sums of distinct divisors. Am. J. Math. **76**, 779–785 (1954)
16. Z.-W. Sun, Sequences A185636, A199920, A204065, A208244, A209236, A209253, A209312, A209315, A209320, A210479, A210480, A210681, A218654, A218656, A218754, A218825, A218829, A219864, A220272, A220413, A220431, A220554, A220572, A224030, A227898, A227899, A227909 (joint with O. Gerard), A227923, A228425, A229166, A230217, A230219, A230223, A230224, A230230, A230241, A230243, A230252, A230254, A230351, A230494, A230507, A230516, A231201, A231516, A231555, A231557, A231561, A231631, A231633, A231635, A231725, A231883, A232109, A232174, A232186, A232269, A232353, A233386, A232398, A233544, A233547, A233654, A233793, A233867, A233918, A234200, A234246, A234308, A234309, A234347, A234359, A234360, A234451, A234504, A234808, A234809, A235189, A235592, A235613, A235614, A235661, A235703, A235987, A236358, A236567, A236968, A236998, A237049, A237050, A237123, A237127, A237130, A237168, A237183, A237184, A237253, A237523, A237524, A237715, A238386, A238405, A238703, A238732, A238733, A241844, A242425, A242441, A255677, A256555, A256707, A257364, A257856, A257922, A257924, A257926, A257928, A257938, A258803, A258836, A258838, A259487, A259488, A259492, A259531, A259539, A259540, A259628, A259678, A259712, A259789, A259915, A259916, A260078, A260080, A260082, A260120, A260140, A260232, A260252, A260753, A260886, A260888, A261136, A261281, A261295, A261339, A261354, A261352, A261353, A261361, A261362, A261382, A261385, A261387, A261395, A261437, A261462, A261513, A261515, A261528, A261533, A261541, A261583, A261627, A261628, A261641, A261653, A262311, A262785, A262982, A262985, A263992, A263998, A264010, A264025 in OEIS, http://www.oeis.org
17. Z.-W. Sun, Mixed sums of squares and triangular numbers. Acta Arith. **127**, 103–113 (2007)
18. Z.-W. Sun, On sums of primes and triangular numbers. J. Comb. Number Theory **1**, 65–76 (2009)
19. Z.-W. Sun, Conjectures involving arithmetical sequences. In: Number Theory: Arithmetic in Shangri-La, ed. by S. Kanemitsu, H. Li, J. Liu, *Proceedings of 6th China-Japan Seminar*, Shanghai, August 15–17, 2011 (World Sci., Singapore, 2013), pp. 244–258
20. Z.-W. Sun, On functions taking only prime values. J. Number Theory **133**, 2794–2812 (2013)
21. Z.-W. Sun, *On $a^n + bn$ modulo m*, arXiv:1312.1166
22. Z.-W. Sun, Problems on combinatorial properties of primes. In: Plowing and Starring through High Wave Forms, ed. by M. Kaneko, S. Kanemitsu, J. Liu, Proceedings of 7th China-Japan Seminar on Number Theory, Fukuoka, Oct. 28–Nov. 1, 2013, Ser. Number Theory Applications, vol. 11 (World Sci., Singapore, 2015), pp. 169–187
23. Z.-W. Sun, A new theorem on the prime-counting function. Ramanujan J. **42**, 59–67 (2017)
24. I.M. Vinogradov, *The Method of Trigonometrical Sums in the Theory of Numbers* (Dover, New York, 2004)
25. A. Weingartner, Practical numbers and the distribution of divisors. Q. J. Math. **66**, 743–758 (2015)
26. Y. Zhang, Bounded gaps between primes. Ann. Math. **179**, 1121–1174 (2014)