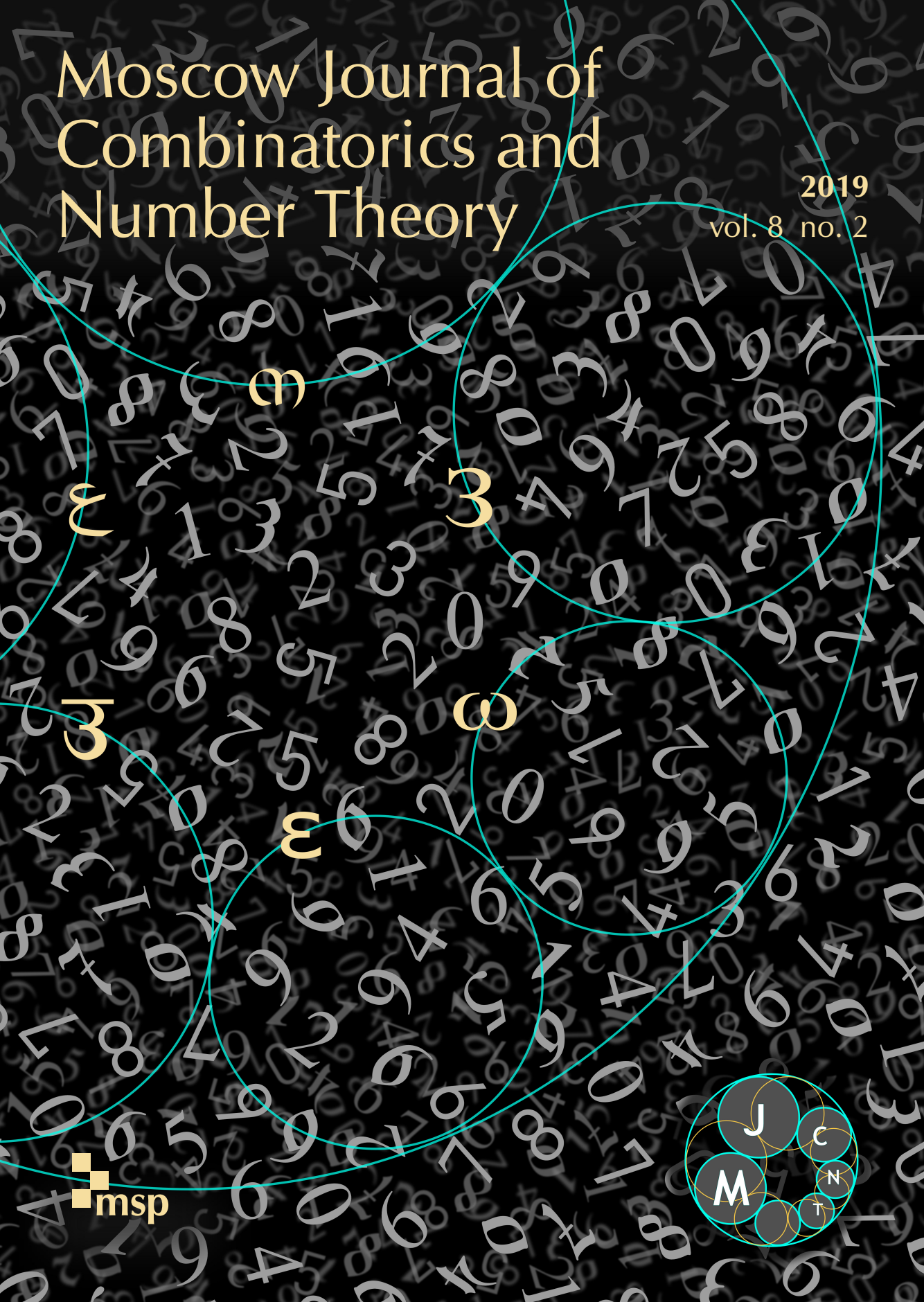
Moscow Journal of Combinatorics and Number Theory

2019
vol. 8 no. 2



msp

Moscow Journal of Combinatorics and Number Theory

msp.org/moscow

See inside back cover or msp.org/moscow for submission instructions.

The subscription price for 2019 is US $310/year for the electronic version, and $365/year (+$20, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Moscow Journal of Combinatorics and Number Theory (ISSN 2640-7361 electronic, 2220-5438 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

MJCNT peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.97

A simple proof of the Hilton–Milner theorem

Peter Frankl

Let $n \geq 2k \geq 4$ be integers and $\mathcal{F}$ a family of k-subsets of $\{1, 2, \ldots, n\}$. We call $\mathcal{F}$ *intersecting* if $F \cap F' \neq \varnothing$ for all $F, F' \in \mathcal{F}$, and we call $\mathcal{F}$ *nontrivial* if $\bigcap_{F\in\mathcal{F}} F = \varnothing$. Strengthening the famous Erdős–Ko–Rado theorem, Hilton and Milner proved that $|\mathcal{F}| \leq \binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1$ if $\mathcal{F}$ is nontrivial and intersecting. We provide a proof by injection of this result.

1. Introduction

The proof of the Hilton–Milner theorem that we are going to present is very short but it is based on the very useful operation of *shifting* and two old results of the author. We are going to review these in this section.

Let $[n] = \{1, \ldots, n\}$ be the standard n-element set and $2^{[n]}$ its power set. Subsets $\mathcal{F} \subset 2^{[n]}$ are called families. For $i \in [n]$ we use the standard notation $\mathcal{F}(i) = \{F \setminus \{i\} : i \in F \in \mathcal{F}\}$ and $\mathcal{F}(\bar{i}) = \{F : i \notin F \in \mathcal{F}\}$. Note that

$$|\mathcal{F}| = |\mathcal{F}(i)| + |\mathcal{F}(\bar{i})|.$$

For a positive integer t the family $\mathcal{F}$ is said to be *t-intersecting* if $|F \cap F'| \geq t$ for all $F, F' \in \mathcal{F}$. For $t = 1$ we use the term *intersecting*.

Let us recall the definition of the $S_{i,j}$ shift, an important operation on families, discovered by Erdős, Ko and Rado [Erdős et al. 1961].

Definition 1.1. For $1 \leq i < j \leq n$ and a family $\mathcal{F} \subset 2^{[n]}$, one defines $S_{i,j}(\mathcal{F}) = \{S_{i,j}(F) : F \in \mathcal{F}\}$, where

$$S_{i,j}(F) = \begin{cases} F' := (F \setminus \{j\}) \cup \{i\} & \text{if } j \in F,\ i \notin F \text{ and } F' \notin \mathcal{F}, \\ F & \text{otherwise.} \end{cases}$$

From the definition, $|S_{i,j}(\mathcal{F})| = |\mathcal{F}|$ and $|S_{i,j}(F)| = |F|$ should be obvious. More importantly, if $\mathcal{F}$ is t-intersecting then $S_{i,j}(\mathcal{F})$ is t-intersecting as well.

If $S_{i,j}(\mathcal{F}) = \mathcal{F}$ for all $1 \leq i < j \leq n$ then $\mathcal{F}$ is called *shifted*.

Let us use the notation $(a_1, a_2, \ldots, a_r)$ to denote the set $\{a_1, a_2, \ldots, a_r\}$, where $a_1 < a_2 < \cdots < a_r$. For two subsets $F = (a_1, \ldots, a_r)$ and $G = (b_1, \ldots, b_r)$ we say that F is smaller than G if $a_i \leq b_i$ for all $1 \leq i \leq r$. We denote this by $F \prec G$.

It is not hard to see that $\mathcal{F}$ is shifted if and only if for all pairs of F, G with $F \prec G$, we have $G \in \mathcal{F}$ implies $F \in \mathcal{F}$. For the proof of this and many other useful properties of shifting see [Frankl 1987b].

We shall need the following simple result.

MSC2010: 05D05.
Keywords: finite sets, intersection, hypergraphs.

Proposition 1.2 [Frankl 1978]. *Let $\mathcal{F} \subset 2^{[n]}$ be a shifted t-intersecting family. Then the following hold*:

(i) *For every $F \in \mathcal{F}$ there exists an integer $\ell \geq t$ such that*

$$|F \cap [2\ell - t]| \geq \ell.$$

(ii) *For all $F, G \in \mathcal{F}$ there exists an integer $h \geq t$ such that*

$$|F \cap [h]| + |G \cap [h]| \geq h + t. \tag{1-1}$$

Note that (1-1) implies $|F \cap G \cap [h]| \geq t$.

For $F \in \mathcal{F}$ define $\ell(F) = \left\{\max \ell, t \leq \ell \leq \frac{n}{2} : |F \cap [2\ell]| \geq \ell\right\}$. Note that if $2|F| \leq n$ then the maximality of $\ell(F)$ implies

$$|F \cap [2\ell(F)]| = \ell(F). \tag{1-2}$$

Let $k \geq s \geq 2$ be integers. Let $\binom{[n]}{k}$ denote the collection of all k-subsets of $[n]$.

Example 1.3. Define

$$\mathcal{E}(n,k,s) = \left\{E \in \binom{[n]}{k} : 1 \in E,\ E \cap [2, s+1] \neq \varnothing\right\} \cup \left\{F \subset \binom{[2n]}{k} : [2, s+1] \subset F\right\}.$$

Note that $\mathcal{E}(n,k,s)$ is intersecting, $E \cap [2, s+1] \neq \varnothing$ for all $E \in \mathcal{E}(n,k,s)$ and

$$|\mathcal{E}(n,k,s)| = \binom{n-1}{k-1} - \binom{n-s-1}{k-1} + \binom{n-s-1}{k-s}.$$

Theorem 1.4. *Let $n \geq 2k \geq 2s \geq 4$. Suppose that $\mathcal{F} \subset \binom{[n]}{k}$ is a shifted intersecting family satisfying $F \cap [2, s+1] \neq \varnothing$ for all $F \in \mathcal{F}$. Then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1} - \binom{n-s-1}{k-1} + \binom{n-s-1}{k-s}. \tag{1-3}$$

This result is somewhat technical but its proof is rather special. We are going to prove it through an explicit injection from $\mathcal{F}$ into $\mathcal{E}(n,k,s)$.

For sets A, B let $A \triangle B$ denote their symmetric difference. Let us define the map $\alpha : \mathcal{F} \to \mathcal{E}(n,k,s)$ by

$$\alpha(F) = \begin{cases} F & \text{if } 1 \in F \text{ or if } [2, s+1] \subset F, \\ F \triangle [2\ell(F)] & \text{otherwise.} \end{cases}$$

To prove (1-3) it is sufficient to prove the following.

Proposition 1.5. *The map α is an injection into $\mathcal{E}(n,k,s)$.*

Let us recall two important results concerning intersecting families of k-sets.

Erdős–Ko–Rado theorem [Erdős et al. 1961]. *Suppose that $n \geq 2k > 0$ and $\mathcal{F} \subset \binom{[n]}{k}$ is an intersecting family. Then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1}. \tag{1-4}$$

Taking all k-sets containing a fixed element shows that (1-4) is the best possible bound.

An intersecting family is called *nontrivial* if there is no element common to all its members. For $k = 1$ there is no nontrivial k-intersecting family. For $k = 2$ the only such family is the triangle: $\binom{[3]}{2}$.

Hilton–Milnor theorem [1967]. *Suppose that $n \geq 2k \geq 4$ and $\mathcal{F} \subset \binom{[n]}{k}$ is a nontrivial intersecting family. Then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1. \tag{1-5}$$

Recently Hurlbert and Kamat [2018] gave an injective proof for (1-4). We extend their work by providing an injective proof for (1-5). For this we need the following proposition.

Proposition 1.6 [Frankl 1987b]. *Suppose that $n \geq 2k \geq 4$ and $\mathcal{F} \subset \binom{[n]}{k}$ is a nontrivial intersecting family of maximal size. Then there exists a nontrivial intersecting family $\widetilde{\mathcal{F}} \subset \binom{[n]}{k}$ such that $|\widetilde{\mathcal{F}}| = |\mathcal{F}|$ and $\widetilde{\mathcal{F}}$ is shifted.*

Once one has Proposition 1.6, to establish (1-5) is easy. One only needs to apply the case $s = k$ of Theorem 1.4 to the family $\widetilde{\mathcal{F}}$. Indeed, since $\widetilde{\mathcal{F}}$ is nontrivial and shifted, $[2, k+1] \in \widetilde{\mathcal{F}}$ and $\widetilde{\mathcal{F}}$ being intersecting imply that $F \cap [2, k+1] \neq \varnothing$ holds for all $F \in \widetilde{\mathcal{F}}$.

Since the proof of Proposition 1.6 is quite short and somewhat hidden in [Frankl 1987b], we reproduce it in Section 2.

Let us mention that there are several other, known proofs of the Hilton–Milner theorem: [Frankl and Füredi 1986; Frankl and Tokushige 1992; Mörs 1985; Kupavskii and Zakharov 2018].

We should also mention that in [Hilton and Milner 1967] the essentially unique families attaining equality are determined as well. This can be done via the present proof as well. However, it is rather technical and very similar to the corresponding part of previous proofs. Therefore we prefer to omit it.

2. The proofs of Propositions 1.5 and 1.6

We divide the proof of Proposition 1.5 into two lemmas. The first shows that for $F \in \mathcal{F} \setminus \mathcal{E}(n, k, s)$ the image $\alpha(F)$ is in $\mathcal{E}(n, k, s) \setminus \mathcal{F}$.

The second shows that α is an injection.

Lemma 2.1. *Suppose that $F \in \mathcal{F}(\bar{1})$ and $[2, s+1] \not\subset F$. Then the following hold*:

(i) $1 \in \alpha(F)$.

(ii) $\alpha(F) \notin \mathcal{F}$.

(iii) $\alpha(F) \cap [2, s+1] \neq \varnothing$.

Proof. (i) Recall that $\alpha(F) = F \triangle [2\ell(F)]$. As $1 \notin F$ implies $1 \in \alpha(F)$, (i) is true.

(ii) Suppose for contradiction that $\alpha(F) \in \mathcal{F}$. Apply Proposition 1.2 to F and $\alpha(F)$. By (1-2), $F \cap [2\ell(F)]$ and $\alpha(F) \cap [2\ell(F)]$ are complementary ℓ-element subsets of $[2\ell(F)]$. Consequently $h > 2\ell(F)$.

However, for $h \geq 2\ell$, we have $|F \cap [h]| = |\alpha(F) \cap [h]|$. Thus $2|F \cap [h]| \geq h+1$ implies

$$|F \cap [h]| \geq \tfrac{1}{2}(h+1). \tag{2-1}$$

Thus

$$|F \cap [h+1]| \geq \tfrac{1}{2}(h+1)$$

as well, and we get a contradiction with the maximality of $\ell(F)$.

(iii) Define $i(F) = \min\{i : 2 \le i \le n,\ i \notin F\}$. As $\ell(F) \ge 2$, (1-2) implies $i(F) \le 2\ell(F)$. Also, $[2, s+1] \not\subset F$ implies $i(F) \le s+1$. Consequently $i(F) \in [2\ell(F)]$ and $i(F) \in [2, s+1]$ hold. Therefore $i(F) \in \alpha(F) \cap [2, s+1]$. □

Lemma 2.2. *For distinct $F, F' \in \mathcal{F} \setminus \mathcal{E}(n,k,s)$, it holds that $\alpha(F) \ne \alpha(F')$.*

Proof. Since $F, F' \notin \mathcal{E}(n,k,s)$, we have $\alpha(F) = F \triangle [2\ell(F)]$ and $\alpha(F') = F' \triangle [2\ell(F')]$. If $\ell(F) = \ell(F')$ then $\alpha(F) \ne \alpha(F')$ is evident from $F \ne F'$.

By symmetry suppose $\ell(F) < \ell(F')$. The maximality of $\ell(F)$ implies $|F \cap [2\ell(F')]| < \ell(F')$. Using $|F \cap [2\ell(F)]| = \ell(F) = |\alpha(F) \cap [2\ell(F)]|$, it follows that $|\alpha(F) \cap [2\ell(F')]| < \ell(F') = |\alpha(F') \cap [2\ell(F')]|$. This proves $\alpha(F) \ne \alpha(F')$. □

Since $\alpha(F) = F$ for $F \in \mathcal{F} \cap \mathcal{E}(n,k,s)$, Lemmas 2.1 and 2.2 prove that α is an injection into $\mathcal{E}(n,k,s)$.

The proof of Proposition 1.6. Starting with a nontrivial intersecting family $\mathcal{F} \subset \binom{[n]}{k}$ of maximal size, we can keep on applying the S_{ij} shift for various pairs until we run into trouble. The possible trouble is that $S_{ij}(\mathcal{F})$ ceases to be nontrivial, i.e., all its members contain the element i. Then $\{i, j\} \cap F \ne \varnothing$ must hold for all $F \in \mathcal{F}$. By symmetry let $i = 1$, $j = 2$.

The maximality of $|\mathcal{F}|$ implies that *all* k-sets G with $\{1,2\} \subset G \subset [n]$ are in $\mathcal{F}$. Therefore continuing with the $S_{a,b}$ shift for $3 \le a < b \le n$ will never produce a trivial intersecting family. Eventually we obtain a nontrivial intersecting family $\mathcal{H}$, with $|\mathcal{H}| = |\mathcal{F}|$, such that $S_{a,b}(\mathcal{H}) = \mathcal{H}$ for all $3 \le a < b \le n$.

Consequently, both $\{1, 3, 4, \dots, k+1\}$ and $\{2, 3, 4, \dots, k+1\}$ are in $\mathcal{H}$. Since all $G \in \binom{[n]}{k}$ with $\{1,2\} \subset G \subset [n]$ are unchanged under the shift $S_{a,b}$ for $3 \le a < b \le n$, we infer that $\binom{[k+1]}{k} \subset \mathcal{H}$.

Noting that $\binom{[k+1]}{k}$ is not affected by $S_{i,j}$ for $1 \le i < j \le n$, we can continue shifting and eventually obtain a shifted, nontrivial intersecting family of the same size. □

3. Concluding remarks

For a family $\mathcal{F} \subset 2^{[n]}$, let $\triangle(\mathcal{F})$ be its *maximum degree*, that is, $\max_i |\mathcal{F}(i)|$. Then $\gamma(\mathcal{F}) = |\mathcal{F}| - \triangle(\mathcal{F})$ is called the *diversity* of $\mathcal{F}$. With this terminology, for intersecting families $\mathcal{F}$, with $\mathcal{F} \subset \binom{[n]}{k}$, $n \ge 2k$, the Hilton–Milner theorem shows that $\gamma(\mathcal{F}) \ge 1$ implies

$$|\mathcal{F}| \le |\mathcal{E}(n,k,k)| = \binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1.$$

In [Frankl 1987a] the author proved that $\gamma(\mathcal{F}) \ge \binom{n-s-1}{k-s}$, $3 \le s \le k$, implies $|\mathcal{F}| \le |\mathcal{E}(n,k,s)|$. Kupavskii and Zakharov [2018] gave a new proof for a stronger version of this result. It would be desirable to have a proof by injection. Let us note that for $\mathcal{F} \subset \mathcal{G}$ necessarily $\gamma(\mathcal{F}) \le \gamma(\mathcal{G})$ holds.

In the case of Theorem 1.4, we may replace $\mathcal{F}$ by another family $\mathcal{G}$, with $\mathcal{F} \subset \mathcal{G} \subset \binom{[n]}{k}$ where $\mathcal{G}$ is shifted, intersecting and all $G \in \binom{[n]}{k}$ with $[2, s+1] \subset G$ are members of $\mathcal{G}$. For such a special case Theorem 1.4 provides an injective proof. However the general case seems to be harder.

The proofs in [Frankl 1987a; Kupavskii and Zakharov 2018] rely heavily on the Kruskal–Katona theorem; see [Kruskal 1963; Katona 1968]. Therefore we feel that it would be desirable to have a proof by injection for this important result as well.

Note in proof

Hurlbert and Kamat [2018] independently gave a very similar proof in the new version of their paper.

References

[Erdős et al. 1961] P. Erdős, C. Ko, and R. Rado, "Intersection theorems for systems of finite sets", *Quart. J. Math. Oxford Ser.* (2) **12** (1961), 313–320. MR Zbl

[Frankl 1978] P. Frankl, "The Erdős–Ko–Rado theorem is true for $n = ckt$", pp. 365–375 in *Combinatorics, I: Proc. Fifth Hungarian Colloq.* (Keszthely, 1976), edited by A. Hajnal and V. T. Sós, Colloq. Math. Soc. János Bolyai **18**, North-Holland, Amsterdam, 1978. MR Zbl

[Frankl 1987a] P. Frankl, "Erdős–Ko–Rado theorem with conditions on the maximal degree", *J. Combin. Theory Ser. A* **46**:2 (1987), 252–263. MR Zbl

[Frankl 1987b] P. Frankl, "The shifting technique in extremal set theory", pp. 81–110 in *Surveys in combinatorics 1987* (New Cross, 1987), edited by C. Whitehead, London Math. Soc. Lecture Note Ser. **123**, Cambridge Univ. Press, 1987. MR Zbl

[Frankl and Füredi 1986] P. Frankl and Z. Füredi, "Nontrivial intersecting families", *J. Combin. Theory Ser. A* **41**:1 (1986), 150–153. MR Zbl

[Frankl and Tokushige 1992] P. Frankl and N. Tokushige, "Some best possible inequalities concerning cross-intersecting families", *J. Combin. Theory Ser. A* **61**:1 (1992), 87–97. MR Zbl

[Hilton and Milner 1967] A. J. W. Hilton and E. C. Milner, "Some intersection theorems for systems of finite sets", *Quart. J. Math. Oxford Ser.* (2) **18** (1967), 369–384. MR Zbl

[Hurlbert and Kamat 2018] G. Hurlbert and V. Kamat, "New injective proofs of the Erdős–Ko–Rado and Hilton–Milner theorems", *Discrete Math.* **341**:6 (2018), 1749–1754. MR Zbl

[Katona 1968] G. Katona, "A theorem of finite sets", pp. 187–207 in *Theory of graphs* (Tihany, 1966), edited by P. Erős and G. Katona, Academic, New York, 1968. MR Zbl

[Kruskal 1963] J. B. Kruskal, "The number of simplices in a complex", pp. 251–278 in *Mathematical optimization techniques*, edited by R. Bellman, Univ. of California Press, Berkeley, CA, 1963. MR Zbl

[Kupavskii and Zakharov 2018] A. Kupavskii and D. Zakharov, "Regular bipartite graphs and intersecting families", *J. Combin. Theory Ser. A* **155** (2018), 180–189. MR Zbl

[Mörs 1985] M. Mörs, "A generalization of a theorem of Kruskal", *Graphs Combin.* **1**:2 (1985), 167–183. MR Zbl

Received 9 Oct 2017.

PETER FRANKL:

peter.frankl@gmail.com
Rényi Institute, Budapest, Hungary

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.103

On the quotient set of the distance set

Alex Iosevich, Doowon Koh and Hans Parshall

Let $\mathbb{F}_q$ be a finite field of order q. We prove that if $d \geq 2$ is even and $E \subset \mathbb{F}_q^d$ with $|E| \geq 9q^{d/2}$ then

$$\mathbb{F}_q = \frac{\Delta(E)}{\Delta(E)} = \left\{ \frac{a}{b} : a \in \Delta(E), b \in \Delta(E) \backslash \{0\} \right\},$$

where

$$\Delta(E) = \{\|x-y\| : x, y \in E\}, \quad \|x\| = x_1^2 + x_2^2 + \cdots + x_d^2.$$

If the dimension d is odd and $E \subset \mathbb{F}_q^d$ with $|E| \geq 6q^{d/2}$, then

$$\{0\} \cup \mathbb{F}_q^+ \subset \frac{\Delta(E)}{\Delta(E)},$$

where $\mathbb{F}_q^+$ denotes the set of nonzero quadratic residues in $\mathbb{F}_q$. Both results are, in general, best possible, including the conclusion about the nonzero quadratic residues in odd dimensions.

1. Introduction

The Erdős–Falconer distance problem in vector spaces over finite fields asks for the smallest possible size of

$$\Delta(E) = \{\|x-y\| : x, y \in E\}, \quad \|x\| = x_1^2 + \cdots + x_d^2,$$

given $E \subset \mathbb{F}_q^d$, $d \geq 2$. This problem was introduced by Bourgain, Katz and Tao [Bourgain et al. 2004]. Here $\mathbb{F}_q$ denotes the finite field with q elements and $\mathbb{F}_q^d$ is the d-dimensional vector space over this field.

In [Iosevich and Rudnev 2007], one of us and Misha Rudnev proved that if $E \subset \mathbb{F}_q^d$, $d \geq 2$, with $|E| > 2q^{(d+1)/2}$, then $\Delta(E) = \mathbb{F}_q$. Hart, Rudnev and two of us [Hart et al. 2011] showed that, in a sense, this result is best possible when d is odd. More precisely, for any $c \in (0, 1)$ and any q sufficiently large with respect to c, they construct subsets $E \subset \mathbb{F}_q^d$ with $|E| > \frac{c}{2}q^{(d+1)/2}$ but $|\Delta(E)| < cq$. This construction does not appear to generalize to the even-dimensional case. In [Chapman et al. 2012], Chapman, Erdoğan, Hart and two of us proved that if q is prime, $q \equiv 3 \pmod 4$ and if $E \subset \mathbb{F}_q^2$ with $|E| \geq Cq^{4/3}$ for a sufficiently large constant $C > 0$, then

$$|\Delta(E)| > \frac{q}{2}.$$

This result was extended to two-dimensional vector spaces over arbitrary finite fields in [Bennett et al. 2017]. In even dimensions $d \geq 2$, it is reasonable to conjecture that if $|E| \geq Cq^{d/2}$ with a sufficiently large C, then $|\Delta(E)| > \frac{1}{2}q$, but this conjecture currently remains open. The exponent $\frac{d}{2}$ cannot be

MSC2010: 11T24, 52C17.
Keywords: quotient set, distance set, finite field.

improved. To see this, let $q = p^2$, p prime, and let $E = \mathbb{F}_p^d \subset \mathbb{F}_q^d$. Then $|E| = q^{d/2}$, yet $\Delta(E) = \mathbb{F}_p$. When q is a prime and $d \geq 4$, the sharpness of $\frac{d}{2}$ can be demonstrated using Lagrangian subspaces [Hart et al. 2011]. In two dimensions, the sharpness of $\frac{d}{2} = 1$ is easily demonstrated by taking a suitable subset of a straight line.

The purpose of this paper is to show that under the assumption $|E| \geq Cq^{d/2}$, taking the quotient of the elements of $\Delta(E)$ recovers all of $\mathbb{F}_q$ for d even, and at least all the square elements of $\mathbb{F}_q$ when d is odd. More precisely, for $E \subset \mathbb{F}_q^d$ we define

$$\frac{\Delta(E)}{\Delta(E)} := \left\{ \frac{a}{b} : a \in \Delta(E),\ b \in \Delta(E)\backslash\{0\} \right\}.$$

Our main results are the following.

Theorem 1.1. *Let $E \subset \mathbb{F}_q^d$, d even. Then if $|E| \geq 9q^{d/2}$, we have*

$$\mathbb{F}_q = \frac{\Delta(E)}{\Delta(E)}.$$

Theorem 1.2. *Let $d \geq 3$ be an odd integer and $E \subset \mathbb{F}_q^d$. Then if $|E| \geq 6q^{d/2}$, we have*

$$\{0\} \cup \mathbb{F}_q^+ \subset \frac{\Delta(E)}{\Delta(E)}.$$

Sharpness of results. The results are in general sharp up to constants. To see this, we once again take $q = p^2$ and $E = \mathbb{F}_p^2$. Then $|E| = q^{d/2}$; yet

$$\left\{ \frac{a}{b} : a \in \Delta(E),\ b \in \Delta(E)\backslash\{0\} \right\} = \mathbb{F}_p.$$

The statement about the squares in Theorem 1.2 is also sharp. The example in [Hart et al. 2011, page 15] that illustrates the sharpness of the exponent $(d+1)/2$ yields a set of size $cq^{(d+1)/2}$, with c sufficiently small, such that $\Delta(E) \subset \{(a-a')^2 : a, a' \in A\}$, where A is a suitable arithmetic progression in $\mathbb{F}_q$. In particular, $\Delta(E)$ is a subset of the squares, so the ratios of the elements of $\Delta(E)$ are also squares.

2. Proof of Theorem 1.1

For $t \in \mathbb{F}_q$, let

$$\nu(t) = \sum_{x,y \in \mathbb{F}_q^d} E(x)E(y)S_t(x-y),$$

where

$$S_t = \{x \in \mathbb{F}_q^d : \|x\| = t\}.$$

It is clear that $0 \in \Delta(E)/\Delta(E)$ unless $\Delta(E) = \{0\}$. Thus it suffices to prove that for each $r \neq 0$ there exists $t \in \Delta(E) \setminus \{0\}$ such that $tr \in \Delta(E)$. Since $t \in \Delta(E)$ if and only if $\nu(t) > 0$, we must show that for any $r \in \mathbb{F}_q^*$,

$$\nu^2(0) < \sum_{t \in \mathbb{F}_q} \nu(t)\nu(rt). \tag{2-1}$$

We shall need the following standard Fourier-analytic preliminaries. Given $f : \mathbb{F}_q^d \to \mathbb{C}$, define the Fourier transform $\hat{f}$ by the formula

$$\hat{f}(m) = q^{-d} \sum_{x \in \mathbb{F}_q^d} \chi(-x \cdot m) f(x),$$

where χ is a nontrivial principal character on $\mathbb{F}_q$. We shall use the following calculation repeatedly.

Lemma 2.1. *With the notation above,*

$$f(x) = \sum_{m \in \mathbb{F}_q^d} \chi(x \cdot m) \hat{f}(m) \qquad \text{(Fourier inversion)}$$

and

$$\sum_{m \in \mathbb{F}_q^d} |\hat{f}(m)|^2 = q^{-d} \sum_{x \in \mathbb{F}_q^d} |f(x)|^2. \qquad \text{(Plancherel)}$$

By Fourier inversion,

$$\nu(t) = q^{2d} \sum_{m \in \mathbb{F}_q^d} \widehat{S}_t(m) |\widehat{E}(m)|^2 = q^{-d} |E|^2 |S_t| + q^{2d} \sum_{m \neq \vec{0}} \widehat{S}_t(m) |\widehat{E}(m)|^2.$$

It follows that for $r \in \mathbb{F}_q^*$,

$$\begin{aligned}
\sum_{t \in \mathbb{F}_q} \nu(t)\nu(rt) &= \sum_{t \in \mathbb{F}_q} \left(q^{-d} |E|^2 |S_t| + q^{2d} \sum_{m \neq \vec{0}} \widehat{S}_t(m) |\widehat{E}(m)|^2 \right) \left(q^{-d} |E|^2 |S_{rt}| + q^{2d} \sum_{m' \neq \vec{0}} \widehat{S}_{rt}(m') |\widehat{E}(m')|^2 \right) \\
&= q^{-2d} |E|^4 \sum_{t \in \mathbb{F}_q} |S_t| |S_{rt}| + q^d |E|^2 \sum_{m' \neq \vec{0}} |\widehat{E}(m')|^2 \sum_{t \in \mathbb{F}_q} |S_t| \widehat{S}_{rt}(m') \\
&\quad + q^d |E|^2 \sum_{m \neq \vec{0}} |\widehat{E}(m)|^2 \sum_{t \in \mathbb{F}_q} |S_{rt}| \widehat{S}_t(m) + q^{4d} \sum_{m, m' \neq \vec{0}} |\widehat{E}(m)|^2 |\widehat{E}(m')|^2 \sum_{t \in \mathbb{F}_q} \widehat{S}_t(m) \widehat{S}_{rt}(m') \\
&= I + II + III + IV. \qquad (2\text{-}2)
\end{aligned}$$

We shall invoke the explicit value of $|S_t|$, which can be deduced by Theorem 6.26 in [Lidl and Niederreiter 1997].

Lemma 2.2. *Let $S_t \subset \mathbb{F}_q^d$ denote the sphere with radius $t \in \mathbb{F}_q$. Then if $d \geq 2$ is even,*

$$|S_t| = q^{d-1} + \lambda(t) q^{(d-2)/2} \eta((-1)^{d/2}),$$

where η is the quadratic character of $\mathbb{F}_q^*$, $\lambda(t) = -1$ for $t \in \mathbb{F}_q^*$, and $\lambda(0) = q - 1$.

We also use the following result, which was given as Lemma 4 in [Iosevich and Koh 2010].

Lemma 2.3. *Let S_j be a sphere in $\mathbb{F}_q^d$, $d \geq 2$. Then for any $m \in \mathbb{F}_q^d$, we have*

$$\widehat{S}_j(m) = q^{-1} \delta_0(m) + q^{-d-1} \eta^d(-1) G^d \sum_{s \in \mathbb{F}_q^*} \eta^d(s) \chi\left(js + \frac{\|m\|}{4s} \right),$$

where G denotes the Gauss sum, η is the quadratic character of $\mathbb{F}_q^*$, and $\delta_0(m) = 1$ if $m = (0, \ldots, 0)$ and $\delta_0(m) = 0$ otherwise.

A lower bound of $\sum_{t\in\mathbb{F}_q} v(t)v(rt)$ for even dimensions $d \geq 2$. Since $\sum_{t\in\mathbb{F}_q}\lambda(rt) = 0$ for $r \neq 0$, it follows from Lemma 2.2 that

$$\begin{aligned} I &:= q^{-2d}|E|^4\sum_{t\in\mathbb{F}_q}|S_t||S_{rt}| = q^{-2d}|E|^4\Big(q^{2d-1}+q^{d-2}\sum_{t\in\mathbb{F}_q}\lambda(t)\lambda(rt)\Big)\\ &= q^{-2d}|E|^4\Big(q^{2d-1}+q^{d-2}\lambda^2(0)+q^{d-2}\sum_{t\neq 0}\lambda(t)\lambda(rt)\Big)\\ &= q^{-2d}|E|^4\big(q^{2d-1}+q^{d-2}(q-1)^2+q^{d-2}(q-1)\big). \end{aligned}$$

Hence, we obtain

$$I = q^{-1}|E|^4 + q^{-d}|E|^4 - q^{-d-1}|E|^4. \tag{2-3}$$

In order to estimate the remaining terms, we need the following calculations.

Lemma 2.4. *Suppose that $m \neq \vec{0}$ in $\mathbb{F}_q^d$, $d \geq 2$. Then for any $r \neq 0$, we have*

$$\sum_{t\in\mathbb{F}_q}\widehat{S}_{rt}(m) = 0, \tag{2-4}$$

$$\sum_{t\in\mathbb{F}_q}\lambda(t)\widehat{S}_{rt}(m) = q\widehat{S}_0(m), \tag{2-5}$$

where $\lambda(t)$ is defined as in Lemma 2.2.

To see this, observe that the left-hand side of (2-4) equals

$$q^{-d}\sum_{t\in\mathbb{F}_q}\sum_{x\in\mathbb{F}_q^d}\chi(-x\cdot m)S_{rt}(x) = q^{-d}\sum_{x\in\mathbb{F}_q^d}\chi(-x\cdot m)\sum_{t\in\mathbb{F}_q}S_{rt}(x) = q^{-d}\sum_{x\in\mathbb{F}_q^d}\chi(-x\cdot m) = 0$$

since $m \neq (0,\dots,0)$. Hence (2-4) follows. By the definition of $\lambda(t)$,

$$\sum_{t\in\mathbb{F}_q}\lambda(t)\widehat{S}_{rt}(m) = (q-1)\widehat{S}_0(m) - \sum_{t\neq 0}\widehat{S}_{rt}(m) = (q-1)\widehat{S}_0(m) - \sum_{t\in\mathbb{F}_q}\widehat{S}_{rt}(m) + \widehat{S}_0(m).$$

Then (2-5) follows by (2-4). This completes the proof of Lemma 2.4.

We shall also need the following orthogonality lemma.

Lemma 2.5. *Suppose that $r \in \mathbb{F}_q^*$ and $m, m' \in \mathbb{F}_q^d$. If $d \geq 2$ is even, then we have*

$$\sum_{t\in\mathbb{F}_q}\widehat{S}_t(m)\widehat{S}_{rt}(m') = \begin{cases} q^{-1}\delta_0(m)\delta_0(m') + q^{-d} - q^{-d-1} & \text{if } \|m\| = r\|m'\|,\\ -q^{-d-1} & \text{if } \|m\| \neq r\|m'\|. \end{cases}$$

The proof shall be given at the end of the paper (see Lemma 4.2). With the lemmas in tow, we are ready to handle terms *II*, *III* and *IV*. In view of Lemmas 2.2 and 2.4, if $m' \neq \vec{0}$, then

$$\sum_{t\in\mathbb{F}_q}|S_t|\widehat{S}_{rt}(m') = q^{d-1}\sum_{t\in\mathbb{F}_q}\widehat{S}_{rt}(m') + q^{(d-2)/2}\eta((-1)^{d/2})\sum_{t\in\mathbb{F}_q}\lambda(t)\widehat{S}_{rt}(m') = q^{d/2}\eta((-1)^{d/2})\widehat{S}_0(m').$$

Using this equation, it follows that

$$II := q^d|E|^2 \sum_{m'\neq\vec{0}} |\widehat{E}(m')|^2 \sum_{t\in\mathbb{F}_q} |S_t|\widehat{S}_{rt}(m') = q^{3d/2}\eta((-1)^{d/2})|E|^2 \sum_{m'\neq\vec{0}} |\widehat{E}(m')|^2\widehat{S}_0(m').$$

By the same argument, it is not difficult to see that $II = III$. Namely, we have

$$II + III = 2q^{3d/2}\eta((-1)^{d/2})|E|^2 \sum_{m\neq\vec{0}} |\widehat{E}(m)|^2\widehat{S}_0(m). \tag{2-6}$$

We now move on to the term

$$IV := q^{4d} \sum_{m,m'\neq\vec{0}} |\widehat{E}(m)|^2|\widehat{E}(m')|^2 \sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m').$$

Using Lemma 2.5, we can write $IV = A + B$, where

$$A = -q^{3d-1} \sum_{\substack{\|m\|\neq r\|m'\| \\ m,m'\neq\vec{0}}} |\widehat{E}(m)|^2|\widehat{E}(m')|^2,$$

$$B = (q^{3d} - q^{3d-1}) \sum_{\substack{\|m\|=r\|m'\| \\ m,m'\neq\vec{0}}} |\widehat{E}(m)|^2|\widehat{E}(m')|^2.$$

It follows that

$$IV = A + B = q^{3d} \sum_{\substack{\|m\|=r\|m'\| \\ m,m'\neq\vec{0}}} |\widehat{E}(m)|^2|\widehat{E}(m')|^2 - q^{3d-1} \sum_{m,m'\neq\vec{0}} |\widehat{E}(m)|^2|\widehat{E}(m')|^2 = A' - B'.$$

Combining this with (2-3), (2-6), we obtain that if $d \geq 2$ is even and $r \neq 0$, then

$$\begin{aligned}\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) &= I + II + III + IV \\ &= (q^{-1}|E|^4 + q^{-d}|E|^4 - q^{-d-1}|E|^4) \\ &\quad + 2q^{3d/2}\eta((-1)^{d/2})|E|^2\Big(\sum_{m\neq\vec{0}} |\widehat{E}(m)|^2\widehat{S}_0(m)\Big) + (A' - B').\end{aligned}$$

Notice that each term above is a real number. It follows that

$$\begin{aligned}\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) &\geq q^{-1}|E|^4 - 2q^{3d/2}|E|^2\big(\max_{m\neq\vec{0}} |\widehat{S}_0(m)|\big)\Big(\sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2\Big) + (A' - B') \\ &= q^{-1}|E|^4 - 2q^{d/2}|E|^3\big(\max_{m\neq\vec{0}} |\widehat{S}_0(m)|\big) + (A' - B'),\end{aligned}$$

where we used the Plancherel theorem, which states

$$\sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2 = q^{-d}|E|.$$

By the definitions of A' and B', we see that

$$A' - B' \geq q^{3d}\bigg(\sum_{\substack{\|m\|=0\\ m\neq\vec{0}}} |\widehat{E}(m)|^2\bigg)^2 - q^{3d-1}\bigg(\sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2\bigg)^2 = q^{3d}\bigg(\sum_{\substack{\|m\|=0\\ m\neq\vec{0}}} |\widehat{E}(m)|^2\bigg)^2 - q^{d-1}|E|^2.$$

We also see from Lemma 2.3 that if $d \geq 2$ is even, then

$$\max_{m\neq\vec{0}} |\widehat{S}_0(m)| \leq q^{-d/2}.$$

Thus we conclude that if $d \geq 2$ is even and $r \neq 0$, then

$$\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) \geq q^{-1}|E|^4 - 2|E|^3 + q^{3d}\bigg(\sum_{\substack{\|m\|=0\\ m\neq\vec{0}}} |\widehat{E}(m)|^2\bigg)^2 - q^{d-1}|E|^2. \tag{2-7}$$

An upper bound of $\nu^2(0)$ for even dimensions $d \geq 2$. It follows that

$$\nu(0) = q^{2d}\sum_{m\in\mathbb{F}_q^d} \widehat{S}_0(m)|\widehat{E}(m)|^2.$$

By Lemma 2.3, notice that if $d \geq 2$ is even, then

$$\widehat{S}_0(m) = q^{-1}\delta_0(m) + q^{-d-1}G^d\sum_{s\in\mathbb{F}_q^*} \chi(s\|m\|).$$

Then we see that

$$\nu(0) = q^{-1}|E|^2 + q^{d-1}G^d\sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2\bigg(-1 + \sum_{s\in\mathbb{F}_q} \chi(s\|m\|)\bigg).$$

By the Plancherel theorem and the orthogonality of χ,

$$\nu(0) = q^{-1}|E|^2 - q^{-1}G^d|E| + q^dG^d\sum_{\|m\|=0} |\widehat{E}(m)|^2.$$

Since $\widehat{E}(\vec{0}) = q^{-d}|E|$, we can write

$$\nu(0) = q^{-1}|E|^2 - q^{-1}G^d|E| + q^{-d}G^d|E|^2 + q^dG^d\sum_{\substack{\|m\|=0\\ m\neq\vec{0}}} |\widehat{E}(m)|^2. \tag{2-8}$$

We shall use the following explicit form of the Gauss sum G.

Lemma 2.6 [Lidl and Niederreiter 1997, Theorem 5.15]. *Let $\mathbb{F}_q$ be a finite field with $q = p^\ell$ for an odd prime p and $\ell \in \mathbb{N}$. Then the Gauss sum G satisfies*

$$G = \begin{cases} (-1)^{\ell-1}q^{1/2} & \text{if } p \equiv 1 \pmod 4,\\ (-1)^{\ell-1}i^\ell q^{1/2} & \text{if } p \equiv 3 \pmod 4.\end{cases}$$

Observe from Lemma 2.6 that if the dimension d is even, then $G^d = \pm q^{d/2}$, where the sign depends on d and q. Combining this observation with (2-8), and considering the sign of each term, we see that if d is even, then

$$\nu(0) \leq \begin{cases} q^{-1}|E|^2 + q^{(d-2)/2}|E| & \text{if } G^d = -q^{d/2}, \\ q^{-1}|E|^2 + q^{-d/2}|E|^2 + q^{3d/2}\sum_{\|m\|=0, m\neq\vec{0}} |\widehat{E}(m)|^2 & \text{if } G^d = q^{d/2}. \end{cases}$$

Assuming that $|E| \geq q^{d/2}$, we see that

$$\nu(0) \leq \begin{cases} 2q^{-1}|E|^2 & \text{if } G^d = -q^{d/2}, \\ 2q^{-1}|E|^2 + q^{3d/2}\sum_{\|m\|=0, m\neq\vec{0}} |\widehat{E}(m)|^2 & \text{if } G^d = q^{d/2}. \end{cases}$$

Since $\nu(0)$ is a nonnegative real number, it follows that if $|E| \geq q^{d/2}$, then

$$\nu^2(0) \leq 4q^{-2}|E|^4 + 4q^{(3d-2)/2}|E|^2 \sum_{\substack{\|m\|=0 \\ m\neq\vec{0}}} |\widehat{E}(m)|^2 + q^{3d}\left(\sum_{\substack{\|m\|=0 \\ m\neq\vec{0}}} |\widehat{E}(m)|^2\right)^2.$$

Since

$$\sum_{\substack{\|m\|=0 \\ m\neq\vec{0}}} |\widehat{E}(m)|^2 \leq \sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2 = q^{-d}|E|,$$

we conclude that if $|E| \geq q^{d/2}$, then

$$\nu^2(0) \leq 4q^{-2}|E|^4 + 4q^{(d-2)/2}|E|^3 + q^{3d}\left(\sum_{\substack{\|m\|=0 \\ m\neq\vec{0}}} |\widehat{E}(m)|^2\right)^2. \tag{2-9}$$

Now we are ready to complete the proof of Theorem 1.1.

Complete proof of Theorem 1.1. We must show that (2-1) holds. By (2-7) and (2-9), it is enough to show that if $|E| \geq 9q^{d/2}$, then

$$q^{-1}|E|^4 - 2|E|^3 - q^{d-1}|E|^2 > 4q^{-2}|E|^4 + 4q^{(d-2)/2}|E|^3.$$

It suffices to show that

$$q^{-1}|E|^4 - 6q^{(d-2)/2}|E|^3 - q^{d-1}|E|^2 > 4q^{-2}|E|^4.$$

If $|E| \geq 9q^{d/2}$, then we see that

$$q^{-1}|E|^4 - 6q^{(d-2)/2}|E|^3 - q^{d-1}|E|^2 \geq \tfrac{1}{3}q^{-1}|E|^4 - q^{d-1}|E|^2,$$

so it is sufficient to show that

$$\tfrac{1}{3}q^{-1}|E|^4 - q^{d-1}|E|^2 > 4q^{-2}|E|^4.$$

Observe that if $|E| \geq 9q^{d/2} (\geq \sqrt{12}q^{d/2})$, then

$$\tfrac{1}{3}q^{-1}|E|^4 - q^{d-1}|E|^2 \geq \tfrac{1}{4}q^{-1}|E|^4.$$

Consequently, it suffices to show that

$$\tfrac{1}{4}q^{-1}|E|^4 > 4q^{-2}|E|^4,$$

which holds if $q > 16$. Therefore, when $q \le 16$, it suffices to prove the statement of Theorem 1.1. More precisely, it remains to show that if $|E| \ge 9q^{d/2}$ and $q \le 16$, then $\mathbb{F}_q = \Delta(E)/\Delta(E)$. Since $9q^{d/2} > 2q^{(d+1)/2}$ for $q \le 16$, it will be enough to prove that if $|E| > 2q^{(d+1)/2}$, then $\Delta(E) = \mathbb{F}_q$. This was proved in [Iosevich and Rudnev 2007]. Thus the proof of Theorem 1.1 is complete.

3. Proof of Theorem 1.2

We proceed as in the proof of Theorem 1.1. As seen in (2-2), for $r \in \mathbb{F}_q^+$, we can write

$$\begin{aligned}\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) &= \sum_{t\in\mathbb{F}_q}\left(q^{-d}|E|^2|S_t| + q^{2d}\sum_{m\neq\vec{0}}\widehat{S}_t(m)|\widehat{E}(m)|^2\right)\left(q^{-d}|E|^2|S_{rt}| + q^{2d}\sum_{m'\neq\vec{0}}\widehat{S}_{rt}(m')|\widehat{E}(m')|^2\right)\\ &= q^{-2d}|E|^4\sum_{t\in\mathbb{F}_q}|S_t||S_{rt}| + q^d|E|^2\sum_{m'\neq\vec{0}}|\widehat{E}(m')|^2\sum_{t\in\mathbb{F}_q}|S_t|\widehat{S}_{rt}(m')\\ &\quad + q^d|E|^2\sum_{m\neq\vec{0}}|\widehat{E}(m)|^2\sum_{t\in\mathbb{F}_q}|S_{rt}|\widehat{S}_t(m) + q^{4d}\sum_{m,m'\neq\vec{0}}|\widehat{E}(m)|^2|\widehat{E}(m')|^2\sum_{t\in\mathbb{F}_q}\widehat{S}_t(m)\widehat{S}_{rt}(m')\\ &= \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}.\end{aligned}$$

The following explicit value of $|S_t|$ is given as Theorem 6.27 in [Lidl and Niederreiter 1997].

Lemma 3.1. *Let $S_t \subset \mathbb{F}_q^d$ denote the sphere with radius $t \in \mathbb{F}_q$. If $d \ge 3$ is odd, then*

$$|S_t| = q^{d-1} + q^{(d-1)/2}\eta((-1)^{(d-1)/2}t),$$

where η denotes the quadratic character of $\mathbb{F}_q^*$ and $\eta(0) = 0$.

We recall from Lemma 2.3 that if $d \ge 3$ is odd, then for any $m \in \mathbb{F}_q^d$,

$$\widehat{S}_j(m) = q^{-1}\delta_0(m) + q^{-d-1}\eta(-1)G^d\sum_{s\in\mathbb{F}_q^*}\eta(s)\chi\left(js + \frac{\|m\|}{4s}\right). \tag{3-1}$$

Estimate of $\sum_{t\in\mathbb{F}_q}\nu(t)\nu(rt)$ for odd dimensions $d \ge 3$. Since $\sum_{t\in\mathbb{F}_q^*}\eta(t) = 0$ (by the orthogonality of η) and $\eta(0) = 0$, it follows from Lemma 3.1 that

$$\begin{aligned}\mathrm{I} &:= q^{-2d}|E|^4\sum_{t\in\mathbb{F}_q}|S_t||S_{rt}|\\ &= q^{-2d}|E|^4\sum_{t\in\mathbb{F}_q}\left(q^{d-1} + q^{(d-1)/2}\eta((-1)^{(d-1)/2}t)\right)\left(q^{d-1} + q^{(d-1)/2}\eta((-1)^{(d-1)/2}rt)\right)\\ &= q^{-2d}|E|^4\left(\sum_{t\in\mathbb{F}_q}q^{2d-2} + \sum_{t\in\mathbb{F}_q}q^{d-1}\eta(r)\eta^2(t)\right) = q^{-2d}|E|^4\left(q^{2d-1} + q^{d-1}\eta(r)(q-1)\right).\end{aligned}$$

Since $\eta(r) = 1$ (by the assumption that $r \in \mathbb{F}_q^+$), we have

$$\mathrm{I} \ge q^{-1}|E|^4.$$

In order to estimate the second term II, we begin by proving the following result.

Lemma 3.2. *Let S_j be the sphere in $\mathbb{F}_q^d$ for odd $d \geq 3$. Then for $r \neq 0$ and $m \neq \vec{0}$, we have*

$$\Omega := \sum_{t\in\mathbb{F}_q} |S_t|\widehat{S}_{rt}(m) = q^{(-d-3)/2}G^{d+1}\eta(r\,(-1)^{(d+1)/2})\Big(-1+\sum_{s\in\mathbb{F}_q}\chi(s\|m\|)\Big).$$

To prove this lemma, recall from (2-4) of Lemma 2.4 that $\sum_{t\in\mathbb{F}_q}\widehat{S}_{rt}(m) = 0$ for $r \neq 0$ and $m \neq \vec{0}$. By Lemma 3.1,

$$\Omega = \sum_{t\in\mathbb{F}_q}\big(q^{d-1}+q^{(d-1)/2}\eta((-1)^{(d-1)/2}t)\big)\widehat{S}_{rt}(m) = q^{(d-1)/2}\eta((-1)^{(d-1)/2})\sum_{t\in\mathbb{F}_q}\eta(t)\widehat{S}_{rt}(m).$$

By using the value of $\widehat{S}_{rt}(m)$ in (3-1), we can write

$$\Omega = q^{(-d-3)/2}\eta((-1)^{(d+1)/2})G^d\sum_{s\neq 0}\eta(s)\chi\left(\frac{\|m\|}{4s}\right)\Big(\sum_{t\in\mathbb{F}_q}\eta(t)\chi(rst)\Big).$$

Since $\eta(0) = 0$ and $\eta(a) = \eta(a^{-1})$ for $a \neq 0$, a simple change of variables yields

$$\sum_{t\in\mathbb{F}_q}\eta(t)\chi(rst) = \eta(rs)G$$

and thus we have

$$\Omega = q^{(-d-3)/2}\eta((-1)^{(d+1)/2})G^{d+1}\eta(r)\sum_{s\neq 0}\chi(s\|m\|),$$

which completes the proof of Lemma 3.2.

By Lemma 3.2 and the orthogonality of χ, we see that

$$\begin{aligned}\mathrm{II} &:= q^d|E|^2\sum_{m'\neq\vec{0}}|\widehat{E}(m')|^2\sum_{t\in\mathbb{F}_q}|S_t|\widehat{S}_{rt}(m')\\ &= q^{(d-1)/2}|E|^2G^{d+1}\eta(r\,(-1)^{(d+1)/2})\sum_{\substack{m'\neq\vec{0}\\ \|m'\|=0}}|\widehat{E}(m')|^2 - q^{(d-3)/2}|E|^2G^{d+1}\eta(r\,(-1)^{(d+1)/2})\sum_{m'\neq\vec{0}}|\widehat{E}(m')|^2\\ &= \mathrm{II}_1-\mathrm{II}_2.\end{aligned}$$

Now observe from Lemma 2.6 that $G^{d+1} \in \mathbb{R}$ for odd d and so both II_1 and II_2 are real numbers. Furthermore, both values are real numbers with the same sign. Hence, $\mathrm{II} = \mathrm{II}_1 - \mathrm{II}_2 \geq \min\{-|\mathrm{II}_1|, -|\mathrm{II}_2|\}$. Since

$$\min\{-|\mathrm{II}_1|, -|\mathrm{II}_2|\} \geq -\big|q^{(d-1)/2}|E|^2G^{d+1}\eta(r\,(-1)^{(d+1)/2})\big|\sum_{m'\in\mathbb{F}_q^d}|\widehat{E}(m')|^2,$$

which is same as $-|E|^3$, we obtain that

$$\mathrm{II} \geq -|E|^3.$$

By the same argument, it is not hard to see that $\mathrm{II} = \mathrm{III}$ and we also have

$$\mathrm{III} \geq -|E|^3.$$

In order to estimate the fourth term IV, we shall need the following orthogonality lemma.

Lemma 3.3. *Suppose that $r \in \mathbb{F}_q^*$ and $m, m' \in \mathbb{F}_q^d$. If $d \geq 3$ is odd, then we have*

$$\sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') = \begin{cases} q^{-1}\delta_0(m)\delta_0(m') + (q^{-d} - q^{-d-1})\eta(r) & \text{if } \|m\| = r\|m'\|, \\ -q^{-d-1}\eta(r) & \text{if } \|m\| \neq r\|m'\|, \end{cases}$$

where η denotes the quadratic character of $\mathbb{F}_q^*$.

The proof shall be given at the end of the paper (see Lemma 4.2). By the definition of the term IV and Lemma 3.3, it follows that

$$\begin{aligned} \mathrm{IV} &:= q^{4d} \sum_{m,m'\neq\vec{0}} |\widehat{E}(m)|^2 |\widehat{E}(m')|^2 \sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') \\ &= -q^{3d-1}\eta(r) \sum_{\substack{m,m'\neq\vec{0} \\ \|m\|\neq r\|m'\|}} |\widehat{E}(m)|^2 |\widehat{E}(m')|^2 + (q^{3d} - q^{3d-1})\eta(r) \sum_{\substack{m,m'\neq\vec{0} \\ \|m\|= r\|m'\|}} |\widehat{E}(m)|^2 |\widehat{E}(m')|^2. \end{aligned}$$

Since $\eta(r) = 1$ (by our assumption that r is a square number in $\mathbb{F}_q^*$), the second term above is positive. Thus we have

$$\mathrm{IV} \geq -q^{3d-1} \sum_{m,m'\in\mathbb{F}_q^d} |\widehat{E}(m)|^2 |\widehat{E}(m')|^2.$$

By the Plancherel theorem,

$$\mathrm{IV} \geq -q^{d-1}|E|^2.$$

Putting this together with all other estimates, we obtain that if $d \geq 3$ is odd and r is a square number, then

$$\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) := \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} \geq q^{-1}|E|^4 - 2|E|^3 - q^{d-1}|E|^2. \tag{3-2}$$

Estimate of $\nu^2(\mathbf{0})$ for odd dimensions $d \geq 3$. Recall that we can write

$$\nu(0) = q^{2d} \sum_{m\in\mathbb{F}_q^d} \widehat{S}_0(m)|\widehat{E}(m)|^2 = q^{2d}\widehat{S}_0(\vec{0})|\widehat{E}(\vec{0})|^2 + q^{2d} \sum_{m\neq\vec{0}} \widehat{S}_0(m)|\widehat{E}(m)|^2 := M + R.$$

Since $|S_0| = q^{d-1}$ for odd $d \geq 3$ (see Lemma 3.1),

$$M = q^{-d}|S_0||E|^2 = q^{-1}|E|^2.$$

To estimate R, observe that

$$R \leq q^{2d}\big(\max_{m\neq\vec{0}} |\widehat{S}_0(m)|\big)\Big(\sum_{m\in\mathbb{F}_q^d} |\widehat{E}(m)|^2\Big) = \big(\max_{m\neq\vec{0}} |\widehat{S}_0(m)|\big)q^d|E|.$$

By (3-1), we see that if $d \geq 3$ is odd and $m \neq \vec{0}$, then

$$\widehat{S}_0(m) = q^{-d-1}\eta(-1)G^d \sum_{s\neq 0} \eta(s)\chi\left(\frac{\|m\|}{4s}\right).$$

Since

$$\left|\sum_{s\neq 0}\eta(s)\chi\left(\frac{\|m\|}{4s}\right)\right| = \sqrt{q}$$

for $\|m\| \neq 0$ and 0 otherwise, we see

$$\max_{m\neq\vec{0}} |\widehat{S}_0(m)| \leq q^{(-d-1)/2}.$$

Hence we obtain

$$R \leq q^{(d-1)/2}|E|.$$

We have seen that $\nu(0) := M + R \leq q^{-1}|E|^2 + q^{(d-1)/2}|E|$, which in turn implies

$$\nu^2(0) \leq q^{-2}|E|^4 + 2q^{(d-3)/2}|E|^3 + q^{d-1}|E|^2, \tag{3-3}$$

since $\nu(0)$ is a nonnegative integer.

Complete proof of Theorem 1.2. Let $d \geq 3$ be odd. Suppose that r is a square number in $\mathbb{F}_q^*$. We must show that if $E \subset \mathbb{F}_q^d$ with $|E| \geq 6q^{d/2}$, then

$$\sum_{t\in\mathbb{F}_q} \nu(t)\nu(rt) > \nu^2(0).$$

By (3-2) and (3-3), it will be enough to show that if $|E| \geq 6q^{d/2}$, then

$$q^{-1}|E|^4 - 2|E|^3 - q^{d-1}|E|^2 > q^{-2}|E|^4 + 2q^{(d-3)/2}|E|^3 + q^{d-1}|E|^2.$$

Note that to prove this it suffices to show that

$$q^{-1}|E|^4 - 4q^{(d-3)/2}|E|^3 - 2q^{d-1}|E|^2 > q^{-2}|E|^4.$$

If $|E| \geq 6q^{d/2} (\geq 6q^{(d-1)/2})$, then we see that

$$q^{-1}|E|^4 - 4q^{(d-3)/2}|E|^3 - 2q^{d-1}|E|^2 \geq \tfrac{1}{3}q^{-1}|E|^4 - 2q^{d-1}|E|^2.$$

Hence it is sufficient to show that if $|E| \geq 6q^{d/2}$, then

$$\tfrac{1}{3}q^{-1}|E|^4 - 2q^{d-1}|E|^2 > q^{-2}|E|^4.$$

Observe that if $|E| \geq 6q^{d/2} (\geq \sqrt{24}q^{d/2})$, then

$$\tfrac{1}{3}q^{-1}|E|^4 - 2q^{d-1}|E|^2 \geq \tfrac{1}{4}q^{-1}|E|^4.$$

In conclusion, it is enough to prove that if $|E| \geq 6q^{d/2}$, then

$$\tfrac{1}{4}q^{-1}|E|^4 > q^{-2}|E|^4,$$

which is clearly true provided that $q > 4$. For this reason, it suffices to prove the statement of Theorem 1.2 in the case when $q \leq 4$ and $|E| \geq 6q^{d/2}$. In other words, our task is to prove that if $|E| \geq 6q^{d/2}$ for $q \leq 4$, then $\mathbb{F}_q = \Delta(E)/\Delta(E)$. Since $6q^{d/2} > 2q^{(d+1)/2}$ for $q \leq 4$, it will be enough to show that if $|E| > 2q^{(d+1)/2}$, then $\Delta(E) = \mathbb{F}_q$. This is a well-known result on the Erdős–Falconer distance problem shown in [Iosevich and Rudnev 2007]. Thus we finish the proof of Theorem 1.2.

4. Proofs of Lemmas 2.5 and 3.3

We begin by proving the following lemma.

Lemma 4.1. *Let $r \in \mathbb{F}_q^*$ and $m, m' \in \mathbb{F}_q^d$. Then we have*

$$\sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') = \begin{cases} q^{-1}\delta_0(m)\delta_0(m') + q^{-2d}G^{2d}\eta^d(-r)(1-q^{-1}) & \text{if } \|m\| = r\|m'\|, \\ -q^{-2d-1}G^{2d}\eta^d(-r) & \text{if } \|m\| \neq r\|m'\|. \end{cases}$$

Proof. By Lemma 2.3, we have

$$\widehat{S}_t(m) = q^{-1}\delta_0(m) + q^{-d-1}\eta^d(-1)G^d \sum_{s\in\mathbb{F}_q^*} \eta^d(s)\chi\left(ts + \frac{\|m\|}{4s}\right) := A(t) + B(t),$$

$$\widehat{S}_{rt}(m') = q^{-1}\delta_0(m') + q^{-d-1}\eta^d(-1)G^d \sum_{s'\in\mathbb{F}_q^*} \eta^d(s')\chi\left(rts' + \frac{\|m'\|}{4s'}\right) := C(t) + D(t).$$

Since $\sum_{t\in\mathbb{F}_q} A(t)D(t) = 0 = \sum_{t\in\mathbb{F}_q} B(t)C(t)$ by the orthogonality of χ, we have

$$\begin{aligned}
\sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') &= \sum_{t\in\mathbb{F}_q} A(t)C(t) + \sum_{t\in\mathbb{F}_q} B(t)D(t) \\
&= q^{-1}\delta_0(m)\delta_0(m') + q^{-2d-2}G^{2d} \sum_{s,s'\in\mathbb{F}_q^*} \eta^d(s)\eta^d(s')\chi\left(\frac{\|m\|}{4s} + \frac{\|m'\|}{4s'}\right) \sum_{t\in\mathbb{F}_q} \chi(t(s+rs')) \\
&= q^{-1}\delta_0(m)\delta_0(m') + q^{-2d-1}G^{2d} \sum_{s\in\mathbb{F}_q^*} \eta^d(-s^2/r)\chi\left(\frac{\|m\|}{4s} - \frac{r\|m'\|}{4s}\right) \\
&= q^{-1}\delta_0(m)\delta_0(m') + q^{-2d-1}G^{2d}\eta^d(-r) \sum_{s\in\mathbb{F}_q^*} \chi(s(\|m\| - r\|m'\|)) \\
&= q^{-1}\delta_0(m)\delta_0(m') + \left[q^{-2d-1}G^{2d}\eta^d(-r) \sum_{s\in\mathbb{F}_q} \chi(s(\|m\| - r\|m'\|))\right] - q^{-2d-1}G^{2d}\eta^d(-r).
\end{aligned}$$

Thus the statement follows by the orthogonality of χ. □

As a corollary of Lemma 4.1, one can deduce Lemmas 2.5 and 3.3 which can be restated as follows.

Lemma 4.2. *Suppose that $r \in \mathbb{F}_q^*$ and $m, m' \in \mathbb{F}_q^d$. If $d \geq 2$ is even, then we have*

$$\sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') = \begin{cases} q^{-1}\delta_0(m)\delta_0(m') + q^{-d} - q^{-d-1} & \text{if } \|m\| = r\|m'\|, \\ -q^{-d-1} & \text{if } \|m\| \neq r\|m'\|. \end{cases}$$

On the other hand, if $d \geq 3$ is odd, then we have

$$\sum_{t\in\mathbb{F}_q} \widehat{S}_t(m)\widehat{S}_{rt}(m') = \begin{cases} q^{-1}\delta_0(m)\delta_0(m') + (q^{-d} - q^{-d-1})\eta(r) & \text{if } \|m\| = r\|m'\|, \\ -q^{-d-1}\eta(r) & \text{if } \|m\| \neq r\|m'\|. \end{cases}$$

Proof. Suppose that $d \geq 2$ is even. Then $\eta^d = 1$. By Lemma 2.6, we see that $G^{2d} = q^d$ for even $d \geq 2$. Thus the statement follows by Lemma 4.1.

Next, assume that $d \geq 3$ is odd. Then $\eta^d = \eta$. Hence, by Lemma 4.1 it suffices to show that $G^{2d}\eta(-1) = q^d$ for odd $d \geq 3$. This equality follows by combining Lemma 2.6 with the facts that $\eta(-1) = 1$ for $q \equiv 1 \pmod 4$, and $\eta(-1) = -1$ for $q \equiv 3 \pmod 4$. □

References

[Bennett et al. 2017] M. Bennett, D. Hart, A. Iosevich, J. Pakianathan, and M. Rudnev, "Group actions and geometric combinatorics in $\mathbb{F}_q^d$", *Forum Math.* **29**:1 (2017), 91–110. MR Zbl

[Bourgain et al. 2004] J. Bourgain, N. Katz, and T. Tao, "A sum-product estimate in finite fields, and applications", *Geom. Funct. Anal.* **14**:1 (2004), 27–57. MR Zbl

[Chapman et al. 2012] J. Chapman, M. B. Erdoğan, D. Hart, A. Iosevich, and D. Koh, "Pinned distance sets, k-simplices, Wolff's exponent in finite fields and sum-product estimates", *Math. Z.* **271**:1-2 (2012), 63–93. MR Zbl

[Hart et al. 2011] D. Hart, A. Iosevich, D. Koh, and M. Rudnev, "Averages over hyperplanes, sum-product theory in vector spaces over finite fields and the Erdős–Falconer distance conjecture", *Trans. Amer. Math. Soc.* **363**:6 (2011), 3255–3275. MR Zbl

[Iosevich and Koh 2010] A. Iosevich and D. Koh, "Extension theorems for spheres in the finite field setting", *Forum Math.* **22**:3 (2010), 457–483. MR Zbl

[Iosevich and Rudnev 2007] A. Iosevich and M. Rudnev, "Erdős distance problem in vector spaces over finite fields", *Trans. Amer. Math. Soc.* **359**:12 (2007), 6127–6142. MR Zbl

[Lidl and Niederreiter 1997] R. Lidl and H. Niederreiter, *Finite fields*, 2nd ed., Encyclopedia of Mathematics and its Applications **20**, Cambridge University Press, 1997. MR Zbl

Received 5 Mar 2018. Revised 24 Nov 2018.

ALEX IOSEVICH:

iosevich@math.rochester.edu

Department of Mathematics, University of Rochester, Rochester, NY, United States

DOOWON KOH:

koh131@chungbuk.ac.kr

Department of Mathematics, Chungbuk National University, Cheongju, South Korea

HANS PARSHALL:

parshall.6@osu.edu

Department of Mathematics, The Ohio State University, Columbus, OH, United States

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.117

Embeddings of weighted graphs in Erdős-type settings

David M. Soukup

Many recent results in combinatorics concern the relationship between the size of a set and the number of distances determined by pairs of points in the set. One extension of this question considers configurations within the set with a specified pattern of distances. In this paper, we use graph-theoretic methods to prove that a sufficiently large set E must contain at least $C_G|E|$ distinct copies of any given weighted tree G, where C_G is a constant depending only on the graph G.

1. Introduction

Many questions in combinatorics involve the behavior of the distance set $\Delta(E)$ of a set E, defined as $\Delta(E) = \{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x}, \boldsymbol{y} \in E\}$ for some distance function d. For instance, Erdős' celebrated distinct distances problem conjectured that for finite sets $E \subset \mathbb{R}^2$, $|\Delta(E)| \gtrsim |E|^{1-\epsilon}$ for any positive ϵ. This conjecture was proven in this form by Guth and Katz [2015]. Distance problems where the ambient space is a finite vector space have also been a subject of much research [Bourgain et al. 2004; Iosevich and Rudnev 2007; Koh and Shen 2012; Vu 2008].

A natural question follows: under what conditions on E can we find not just pairs of points a specified distance apart, but groups of points with some specified pattern of distances? In other words, given some weighted graph G where the edge weights correspond to distances between points, when can we find a "copy" of G inside E? Bennett, Chapman, Covert, Hart, Iosevich, and Pakianathan proved a result in this direction for the Euclidean distance in $\mathbb{Z}_p^d$ and path graphs in [Bennett et al. 2016], and McDonald [2016] gave a similar result for dot products in $\mathbb{Z}_p^d$. In this paper, we give an answer to this question for wide classes of graphs, distances, and ambient sets. Moreover, we show that these questions, and other similar ones, are part of a much more general framework.

These previous results depended on ad hoc methods which made use of Fourier analytic techniques which do not generalize easily. Here, we show that elementary combinatorial arguments can be used to expand results which are only about distances into results which describe larger patterns.

Definition 1.1. Given an ambient set X and some set D of possible distances, a *symmetric distance function* is a function $d : X \times X \to D$ such that $d(x_1, x_2) = d(x_2, x_1)$ for all $x_1, x_2 \in X$. Such a function is *K-surjective* if, for every $Y \subset X$ with $|Y| \geq K$, the restriction of d to $Y \times Y$ is surjective. In other words, every set of size at least K determines every distance.

In other words, a K-surjective distance function has its distances well-mixed enough that one cannot construct large sets missing a particular distance. This means that we cannot avoid patterns of distances simply by constructing a set which does not exhibit some of the specified distances.

MSC2010: 52C10.
Keywords: finite point configurations, distance sets, graphs.

Some motivation for this definition is provided by the following theorem, which will allow us to apply the results of this paper to a common setting for distance problems. Note that what we refer to as a "distance function" does not need to encode distances in a natural sense; we do not require the function to obey a triangle inequality, nor do we require our distances to be real numbers. We call it a distance function in order to make it explicit how our result corresponds to known results. Throughout, we will use $\mathbb{F}_q$ to refer to the unique finite field with q elements for some prime power q, and we will let $\mathbb{F}_q^d$ be a d-dimensional vector space over this field.

Theorem 1.2 (A. Iosevich and M. Rudnev [2007]). *Let $X = \mathbb{F}_q^d$, $D = \mathbb{F}_q$, and define $d(\{x_i\}, \{y_i\}) = \sum(x_i - y_i)^2$. Then d is K-surjective with $K = Cq^{(d+1)/2}$ for some constant C independent of q.*

Now we just need to define a graph embedding, which is done in the natural way:

Definition 1.3. Suppose we have a space X, a set of distances D, and a symmetric distance function d. Then for a weighted graph G with edge weights in D, an *embedding* of G into X is an injective function $f : V(G) \to X$ such that for every edge $(v_1, v_2) \in E(G)$ with weight t,

$$d(f(v_1), f(v_2)) = t.$$

We will typically identify such an embedding with its image in X. A collection of such embeddings $\{f_i\}_{i\in I}$ is *disjoint* if all its images are disjoint subsets of X.

Results. Now we are ready to state our main theorem:

Theorem 1.4 (main theorem). *Let X be a set with a symmetric distance function d to a set of distances D, let d be K-surjective, and let $E \subseteq X$ with $|E| = rK$ for some positive real number r. Then for any weighted tree G with edge weights in D, there exists a disjoint collection A_G of embeddings of G into E with*

$$|A_G| \geq \left(\frac{r}{\sigma(G)} - 1\right)K,$$

where $\sigma(G)$ is a constant depending only on the graph G.

Corollary 1.5. *If $|E| \geq \sigma(G)K$, then there is at least one embedding of G into E.*

We will exhibit an explicit constant $\sigma(G)$, which we define as follows:

Definition 1.6. Let G be a finite nonempty connected graph; let the degrees of the vertices of G be $d_1, d_2, \ldots, d_n$, ordered so that $d_1 \geq d_2 \geq \cdots \geq d_n$. Then the *stringiness* of G, denoted by $\sigma(G)$, is defined to be

$$\sigma(G) = (d_1 + 1)\prod_{i=2}^{n} d_i.$$

For example, the stringiness of the Petersen graph (or any other 10-vertex 3-regular graph) is $4 \cdot 3^9 = 78732$. The following estimate gives bounds on the stringiness of a tree:

Theorem 1.7. *Let G be a nonempty tree with n edges. Then*

$$n + 1 \leq \sigma(G) \leq 2^n.$$

The lower bound is sharp (the star graph $K_{1,n}$ attains it) while the upper bound is sharp up to a constant (the path graph P_{n+1} has stringiness $3 \cdot 2^{n-2}$).

Combining our main result with Theorem 1.2 gives the following result in $\mathbb{F}_q^d$ with respect to a specific distance function:

Corollary 1.8. *Let G be a weighted tree with edge weights in $\mathbb{F}_q$. Then there exists a constant C independent of q and G such that every subset E of $\mathbb{F}_q^d$ with $|E| \geq C\sigma(G)q^{(d+1)/2}$ contains an embedding of G with respect to the distance function $d(\{x_i\}, \{y_i\}) = \sum(x_i - y_i)^2$.*

The motivation for Corollary 1.8 comes from comparing this result to the following result in the literature:

Theorem 1.9 (Bennett, Chapman, Covert, Hart, Iosevich, Pakianathan [Bennett et al. 2016]). *Let G be a weighted path or star graph with edge weights in $\mathbb{F}_q$, and suppose G has k edges. Then there exists a constant C independent of q and G such that every subset E of $\mathbb{F}_q^d$ with $|E| \geq Ckq^{(d+1)/2}$ contains an embedding of G.*

This result is very similar to Corollary 1.8 in the star graph case, but the constant is much stronger than that of Corollary 1.8 in the path graph case. This shows the difference between the Fourier-analytic techniques used in the proof of Theorem 1.9 and the elementary combinatorial arguments used here.

The main gain of Corollary 1.8, however, is the fact that it applies to all trees and not just to the two specific types in Theorem 1.9. The methods used in Theorem 1.9 do not generalize easily to more complex structures. Moreover, Corollary 1.8 also gives similar results for other distances.

The idea of this paper is to show that results of the type of Theorem 1.9 are instances of a more general phenomenon. Corollary 1.8 is an example of this phenomenon applied to the standard distance on $\mathbb{F}_q^d$; it enables these results to be extended to general trees for which no results existed. We also note that proof of Theorem 1.4 is very modular, so it is possible to (for instance) use Theorem 1.9 to get slightly better bounds for embeddings of more complicated trees in $\mathbb{Z}_q^d$. This is because we build embeddings of larger graphs inductively from embeddings of smaller graphs, so if we are able to show better bounds for smaller graphs, this will automatically give better bounds for larger graphs.

We proceed as follows: for illustrative purposes, we first state and prove Theorem 2.2, a weaker version of Theorem 1.4, using Lemma 2.1. We then give the very similar proof of Theorem 1.4, which relies on the analogous Lemma 2.5. Finally we prove Theorem 1.7.

2. Graph embeddings

An easier case of Theorem 1.4. The proof of Theorem 2.2 is by induction; it is convenient to state the base case as a separate lemma.

Lemma 2.1. *Let X be a set with a symmetric distance function d to a set of distances D, let d be K-surjective, and let $E \subseteq X$ with $|E| = rK$. Then for any fixed $t \in D$, there are at least $\frac{1}{2}(r-1)K$ disjoint pairs $\{e_i, f_i\}_{i \in I}$ in E such that $d(e_i, f_i) = t$ for all $i \in I$.*

Proof. Since E is finite, let I be an index set of maximal size. Let F be the union of all the pairs, that is,

$$F = \bigcup_{i \in I} \{e_i, f_i\}.$$

Then by maximality of I, $E\setminus F$ cannot contain any pair of points with distance t. By K-surjectivity, this means $|E\setminus F| < K$; so $|F| \geq (r-1)K$, giving $|I| = \frac{1}{2}|F| \geq \frac{1}{2}(r-1)K$ as required. □

Theorem 2.2. *Let X be a set with a symmetric distance function d to a set of distances D, let d be K-surjective, and let $E \subseteq X$ with $|E| = rK$. Then for any weighted nonempty tree G with edge weights in D, suppose G has n edges. Then there exists a disjoint collection A_G of embeddings of G into E with*

$$|A_G| \geq \left(\frac{r+1}{2^n} - 1\right)K.$$

Proof. We proceed by induction on n. The case $n = 0$ is tautological, and the case $n = 1$ is equivalent to Lemma 2.1.

So assume $n \geq 2$, which means G must have a leaf. Fix any leaf; call it v, its associated edge e, the unique vertex adjacent to it w, and the edge weight t. Then $G - v$ is a tree with strictly fewer edges; so by the inductive hypothesis we have a disjoint collection A_{G-v} of embeddings of $G - v$ into E with

$$|A_{G-v}| \geq \left(\frac{r+1}{2^{n-1}} - 1\right)K.$$

Consider the set $W = \{f(w) \mid f \in A_{G-v}\}$. By Lemma 2.1, there exist at least $\frac{1}{2}(|W|/K - 1)K$ disjoint pairs of points $\{f(w), f'(w)\}$ in W with $d(f(w), f'(w)) = t$. But for each of these pairs we can consider the function $g : V(G) \to E$ given by

$$g(x) = \begin{cases} f'(w), & x = v, \\ f(x), & x \neq v. \end{cases}$$

By construction, these are disjoint embeddings of G into E, and there are at least $\frac{1}{2}(|W|/K - 1)K$ of them; but by disjointness of A_{G-v} we have $|W| = |A_{G-v}|$ so there are at least

$$\frac{|A_{G-v}|/K - 1}{2}K \geq \frac{((r+1)/2^{n-1} - 1) - 1}{2}K = \left(\frac{r+1}{2^n} - 1\right)K$$

disjoint embeddings of G as required. □

Corollary 2.3. *If $|E| \geq 2^{n+1}K$, there is at least one embedding of G into E.*

Note that this proof would have worked equally well even if d were not symmetric, which will not carry over to Theorem 1.4.

Proof of Theorem 1.4. Analogously to the proof of Theorem 2.2, we will state a governing lemma (Lemma 2.5); the structure will be identical except that we are building our graph G out of stars instead of working purely with edges. For technical reasons, the application of the K-surjectivity assumption is more difficult in this case, so we will first state an auxiliary lemma, Lemma 2.4.

Lemma 2.4. *Let X be a set with a symmetric distance function d to a set of distances D, let d be K-surjective, and let $E \subseteq X$ with $|E| = rK$. Then for any fixed $t \in D$, $s \in \mathbb{N}$, there are at most sK points $e \in E$ for which there are fewer than s other distinct points $e_1, e_2, \ldots, e_s \in E$ such that $d(e, e_i) = t$ for all i.*

Proof. Create a graph H whose vertices are the points of E and where two vertices are connected by an edge if and only if the corresponding points of E are distance t. Then consider the subgraph H^* of H generated by only those vertices of degree $< s$. By construction, the maximum degree of vertices in H is less than s, which means H^* can be s-colored by the standard greedy algorithm, that is, partitioned into s independent sets. Since the K-surjectivity condition guarantees that an independent set in H (and thus in H^*) has size $< K$, it follows that $|H^*| < sK$. This means that at most sK of the vertices of H have degree $< s$, proving the lemma. □

Lemma 2.5. *Let X be a set with a symmetric distance function d to a set of distances D, let d be K-surjective, and let $E \subseteq X$ with $|E| = rK$. Then for any weighted nonempty star graph $G \cong K_{1,n}$ with edge weights in D, the set E contains at least $K(r-n)/(n+1)$ disjoint embeddings of G.*

Proof. As in the proof of Lemma 2.1, let

$$I = \{\{g_{1,0}, g_{1,1}, \ldots, g_{1,n}\}, \ldots, \{g_{m,0}, g_{m,1}, \ldots, g_{m,n}\}\}$$

be a maximal (with respect to m) set of embeddings of G into E, and define F to be the union of all the embeddings contained in I, that is,

$$F = \bigcup_{i=1}^{m} \{g_{i,0}, g_{i,1}, \ldots, g_{i,n}\}.$$

Then $E \backslash F$ must have no embeddings of G by maximality of I.

Suppose the set of edge weights of G is $\{t_1, t_2, \ldots, t_a\}$ appearing with multiplicities $\{m_1, m_2, \ldots m_a\}$ respectively. Then by Lemma 2.4, for each fixed i there are at most $m_i K$ points of $E \backslash F$ which are not distance t_i from at least m_i other points of $E \backslash F$. Summing over i we get that there are at most

$$\sum_{i=1}^{a} m_i K = nK$$

points of $E \backslash F$ which are not distance t_i from at least m_i other points of $E \backslash F$ for every i. But if $x \in E \backslash F$ is in fact distance t_i from at least m_i other points of $E \backslash F$ for every i, then there exists an embedding of G into $E \backslash F$ where x corresponds to the nonleaf vertex of G. Thus $|E \backslash F| \leq nK$; so

$$|I| = \frac{|F|}{n+1} \geq \frac{rK - nK}{n+1} = \frac{r-n}{n+1} K$$

as required. □

Note that when $n = 1$, this is exactly the statement of Lemma 2.1, but when $n \geq 2$ we may have to deal with repeated edge weights, which adds the extra complexity.

Now we are ready to prove the main theorem, Theorem 1.4:

Proof. The proof proceeds by strong induction on the number of edges in G. If G contains no edges, $\sigma(G) = 1$ and the theorem is clearly true; if G is a star graph $K_{1,n}$, then $\sigma(G) = n + 1$ and the theorem is just Lemma 2.5.

So assume G is not a star graph. Then we can always find a vertex $w \in G$ such that:

(1) w is not a leaf of G (equivalently, $\deg_G w \geq 2$).

(2) All but one of the vertices connected to w are leaves (call these leaves $v_1, v_2, \dots, v_y$, and the associated distances $t_1, t_2, \dots, t_y$, where $y = \deg_G w - 1$.)

(3) There exists a vertex $v \neq w$ in G such that $\deg_G v \geq \deg_G w$.

To see this, let L be the set of leaves of G. Then since G is a tree which is not a star graph, $G - L$ is a tree which contains at least one edge; any leaf of $G - L$ satisfies conditions (1) and (2), and since $G - L$ has at least two leaves, at least one of these must satisfy condition (3).

Define the graph H to be $G - \{v_1, v_2, \dots, v_y\}$. By conditions (1) and (2), H is a tree with fewer edges than G; by condition (3), $\sigma(H) = \sigma(G)/(y+1)$ (since we can reorder the product in Definition 1.6 to make v correspond to d_1, and all we lose by deleting these leaves is a factor of $\deg_G w$). By the inductive hypothesis we have a disjoint collection A_H of embeddings of H into E with

$$|A_H| \geq \left(\frac{r}{\sigma(H)} - 1\right)K.$$

Consider the set $W = \{f(w) \mid f \in A_H\}$. By Lemma 2.5, there exist at least $(|W|/K - y)/(y+1)$ disjoint sets of points $\{f(w), f_1(w), \dots, f_y(w)\}$ contained in W with $d(f(w), f_i(w)) = t_i$ for every i (i.e., embedddings of a particular star graph). But for each of these sets we can consider the function $g : V(G) \to E$ given by

$$g(x) = \begin{cases} f_i(w), & x = v_i, \\ f(x), & x \notin \{v_1, \dots, v_y\}. \end{cases}$$

By construction, these are disjoint embeddings of G into E, and there are at least $K(|W|/K - y)/(y+1)$ of them; but by disjointness of A_H, we have $|W| = |A_H|$ so there are at least

$$\frac{|A_H|/K - y}{y+1}K \geq \frac{(r/\sigma(H) - 1) - y}{y+1}K = \left(\frac{r}{(y+1)\cdot\sigma(H)} - 1\right)K = \left(\frac{r}{\sigma(G)} - 1\right)K$$

disjoint embeddings of G as required. □

Note that in view of Theorem 1.7, this is stronger than Theorem 2.2.

Proof of Theorem 1.7.

Proof. Let G have n edges.

For the lower bound on $\sigma(G)$, a simple inductive argument suffices. If $n = 1$ then $\sigma(G) = 2$ since there is only one possible graph with one edge. If $n > 1$ then deleting a leaf must decrease the stringiness since the degree of the other vertex decreases, so for a graph with n edges, $\sigma(G) > \sigma(G - v) \geq n$. Thus $\sigma(G) \geq n + 1$ (since $\sigma(G) \in \mathbb{Z}$).

For the upper bound, write the degrees of the vertices of G as $d_1, d_2, \dots, d_{n+1}$; without loss of generality say $d_{n+1} = 1$. Then by the arithmetic-geometric mean inequality

$$\begin{aligned} ((d_1 + 1)\cdot d_2 \cdot d_3 \cdot \dots \cdot d_n)^{1/n} &\leq \frac{d_1 + 1 + d_2 + d_3 + \dots + d_n}{n} \\ \sigma(G)^{1/n} &\leq \frac{\sum d_i}{n} \\ \sigma(G)^{1/n} &\leq \frac{2n}{n} \\ \sigma(G) &\leq 2^n. \end{aligned}$$

□

Acknowledgements

I would like to thank Alex Iosevich and Steven Senger for their help, support and feedback.

References

[Bennett et al. 2016] M. Bennett, J. Chapman, D. Covert, D. Hart, A. Iosevich, and J. Pakianathan, "Long paths in the distance graph over large subsets of vector spaces over finite fields", *J. Korean Math. Soc.* **53**:1 (2016), 115–126. MR Zbl

[Bourgain et al. 2004] J. Bourgain, N. Katz, and T. Tao, "A sum-product estimate in finite fields, and applications", *Geom. Funct. Anal.* **14**:1 (2004), 27–57. MR Zbl

[Guth and Katz 2015] L. Guth and N. H. Katz, "On the Erdős distinct distances problem in the plane", *Ann. of Math.* (2) **181**:1 (2015), 155–190. MR Zbl

[Iosevich and Rudnev 2007] A. Iosevich and M. Rudnev, "Erdős distance problem in vector spaces over finite fields", *Trans. Amer. Math. Soc.* **359**:12 (2007), 6127–6142. MR Zbl

[Koh and Shen 2012] D. Koh and C.-Y. Shen, "The generalized Erdős–Falconer distance problems in vector spaces over finite fields", *J. Number Theory* **132**:11 (2012), 2455–2473. MR Zbl

[McDonald 2016] B. McDonald, *Hinges in $\mathbb{Z}_p^d$ and applications to pinned distance sets*, undergraduate thesis, University of Rochester, 2016, available at http://web.math.rochester.edu/people/faculty/iosevich/mcdonaldhonorsthesis16.pdf.

[Vu 2008] V. H. Vu, "Sum-product estimates via directed expanders", *Math. Res. Lett.* **15**:2 (2008), 375–388. MR Zbl

Received 5 Mar 2018. Revised 10 Aug 2018.

DAVID M. SOUKUP:

soukup@math.ucla.edu
Department of Mathematics, UCLA, Los Angeles, CA, United States

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.125

Identity involving symmetric sums of regularized multiple zeta-star values

Tomoya Machide

An identity involving symmetric sums of regularized multiple zeta-star values of harmonic type was proved by Hoffman. In this paper, we prove an identity of shuffle type. We use Bell polynomials appearing in the study of set partitions to prove the identity.

1. Introduction and statement of results

The multiple zeta value (MZV) and multiple zeta-star value (MZSV, or sometimes referred to as the nonstrict MZV) are real numbers defined by the nested series

$$\zeta(k_1,k_2,\dots,k_r)=\sum_{0<m_1<m_2<\cdots<m_r}\frac{1}{m_1^{k_1}m_2^{k_2}\cdots m_r^{k_r}}, \tag{1-1}$$

$$\zeta^\star(k_1,k_2,\dots,k_r)=\sum_{0<m_1\le m_2\le\cdots\le m_r}\frac{1}{m_1^{k_1}m_2^{k_2}\cdots m_r^{k_r}}, \tag{1-2}$$

respectively, where k_i $(1\le i\le r)$ are arbitrary positive integers with $k_r>1$. MZVs and MZSVs can also be given by integrals. These values have been actively studied for more than two decades, but Euler [1776] already mentioned them in a special case, $r=2$. In this paper, we give an identity involving symmetric sums for a class of regularizations of (1-2).

The two expressions of series and integrals yield two different products $*$ and ш, called *harmonic* (or *stuffle*) and *shuffle*, respectively, for any real value in factored form written in terms of either MZVs or MZSVs. For example, the result of $*$ on MZSV for the value $\zeta^\star(2)\times\zeta^\star(2)$ is

$$\zeta^\star(2)\zeta^\star(2)=\zeta^\star((2)*(2))=\zeta^\star((2,2)+(2,2)-(4))=2\zeta^\star(2,2)-\zeta^\star(4), \tag{1-3}$$

where, for notational simplicity, we think of the product $*$ as taking place among *indices*. (An index means a finite sequence $\boldsymbol{k}=(k_1,\dots,k_r)$ of positive integers.) The result (1-3) follows from series expressions in (1-2) with

$$\Bigg(\sum_{0<m_1}\frac{1}{m_1^2}\Bigg)\Bigg(\sum_{0<m_2}\frac{1}{m_2^2}\Bigg)=\sum_{0<m_1\le m_2}\frac{1}{m_1^2m_2^2}+\sum_{0<m_2\le m_1}\frac{1}{m_2^2m_1^2}-\sum_{0<m}\frac{1}{m^4},$$

or with the division of the summation

$$\sum_{0<m_1,\,0<m_2}=\sum_{0<m_1\le m_2}+\sum_{0<m_2\le m_1}-\sum_{0<m_1=m_2}. \tag{1-4}$$

MSC2010: primary 11M32; secondary 11B73.
Keywords: multiple zeta value, multiple zeta-star value, symmetric sum, Bell polynomial.

The other results for $\zeta(2)\zeta(2)=\zeta^\star(2)\zeta^\star(2)$ are similarly obtained such that $\zeta((2)*(2))=2\zeta(2,2)+\zeta(4)$, $\zeta((2)\text{ш}(2))=2\zeta(2,2)+4\zeta(1,3)$, and $\zeta^\star((2)\text{ш}(2))=2\zeta^\star(1,3)$. The case of $*$ on MZV follows from series expressions in (1-1) with division of the summation as in (1-4),

$$\sum_{0<m_1,\,0<m_2} = \sum_{0<m_1<m_2} + \sum_{0<m_2<m_1} + \sum_{0<m_1=m_2} .$$

The cases of ш on MZV and MZSV follow from integral expressions as

$$\zeta(2)=\int_{0<s<t<1}\frac{ds}{1-s}\frac{dt}{t},$$

with division of domains of integration

$$\int_{\substack{0<s_1<1\\0<s_2<1}} = \int_{0<s_1<s_2<1} + \int_{0<s_2<s_1<1},$$

where we require an extra technique [Kaneko and Yamamoto 2018] of the integral associated to 2-labeled posets for integral expressions of MZSVs. We omit details of $*$ and ш (for which see [Hoffman 1997; Ihara et al. 2006; Kaneko 2018; Kaneko and Yamamoto 2018; Reutenauer 1993; Zudilin 2003]),[1] since many notations are necessary for rigorous statements, though product rules are simply induced from dividing the summation and domain.

MZVs and MZSVs are divergent if $k_r=1$, but recently, the theory of regularization has been established. (For details, see [Ihara et al. 2006] and [Kaneko and Yamamoto 2018] for MZV and MZSV, respectively.) Four polynomials whose coefficients are $\mathbb{Q}$-linear combinations of MZVs and MZSVs, which we denote by

$$\zeta_*(\boldsymbol{k};T),\quad \zeta_{\text{ш}}(\boldsymbol{k};T),\quad \zeta_*^\star(\boldsymbol{k};T),\quad \text{and}\quad \zeta_{\text{ш}}^\star(\boldsymbol{k};T), \tag{1-5}$$

are defined for any index $\boldsymbol{k}$ in the theory: $\zeta_*(\boldsymbol{k};T)$ and $\zeta_{\text{ш}}(\boldsymbol{k};T)$ are generalizations of $\zeta(\boldsymbol{k})$ involving products $*$ and ш, respectively; $\zeta_*^\star(\boldsymbol{k};T)$ and $\zeta_{\text{ш}}^\star(\boldsymbol{k};T)$ are those of $\zeta^\star(\boldsymbol{k})$. (Note that the polynomials in (1-5) are constant and equal to $\zeta(\boldsymbol{k})$ when $k_r>1$.) A key idea of the generalizations is roughly to regard the divergent value $\zeta(1)=\zeta^\star(1)=\frac{1}{1}+\frac{1}{2}+\cdots$ as the variable T when using product rule. For example, using the rule of $*$ on MZSV for $\zeta^\star(2)\zeta^\star(1)$ yields

$$\zeta^\star(2)\zeta^\star(1)=\zeta_*^\star((2)*(1))=\zeta_*^\star((2,1)+(1,2)-(3))=\zeta_*^\star(2,1;T)+\zeta^\star(1,2)-\zeta^\star(3),$$

which, together with $\zeta^\star(2)\zeta^\star(1)=\zeta^\star(2)T$, proves

$$\zeta_*^\star(2,1;T)=\zeta^\star(2)T-\zeta^\star(1,2)+\zeta^\star(3). \tag{1-6}$$

We can obtain

$$\begin{aligned}\zeta_*(2,1;T)&=\zeta(2)T-\zeta(1,2)-\zeta(3),\\ \zeta_{\text{ш}}(2,1;T)&=\zeta(2)T-2\zeta(1,2),\\ \zeta_{\text{ш}}^\star(2,1;T)&=\zeta^\star(2)T-\tfrac{1}{2}\zeta^\star(1,2)\end{aligned}$$

[1] We recommend [Kaneko 2018; Zudilin 2003] for nonspecialists.

in similar ways, where evaluating $\zeta^\star_{\text{Ш}}(2,1;T)$ requires extra computations because $\zeta^\star_{\text{Ш}}(2,1;T)$ is defined by means of the integral associated to 2-labeled posets. The regularized values $\zeta_*(\boldsymbol{k})$, $\zeta_{\text{Ш}}(\boldsymbol{k})$, $\zeta^\star_*(\boldsymbol{k})$, and $\zeta^\star_{\text{Ш}}(\boldsymbol{k})$ are defined by their constant terms; e.g., $\zeta_*(\boldsymbol{k}) = \zeta_*(\boldsymbol{k};0)$.

Fundamental theorems of regularization for MZVs and MZSVs were proved in [Ihara et al. 2006] and [Kaneko and Yamamoto 2018], respectively, which are stated as follows. For any index $\boldsymbol{k}$,

$$\rho(\zeta_*(\boldsymbol{k};T)) = \zeta_{\text{Ш}}(\boldsymbol{k};T) \quad \text{and} \quad \rho^\star(\zeta^\star_*(\boldsymbol{k};T)) = \zeta^\star_{\text{Ш}}(\boldsymbol{k};T), \tag{1-7}$$

where ρ and $\rho^\star$ are $\mathbb{R}$-linear endomorphisms on $\mathbb{R}[T]$ related to the gamma function $\Gamma(u)$. The detailed definition of $\rho^\star$ will be introduced in Section 2, and is necessary to prove our result, Theorem 1.1.

In order to state Theorem 1.1, we will recall Hoffman's identities involving symmetric sums of the polynomials $\zeta_*(\boldsymbol{k};T)$ and $\zeta^\star_*(\boldsymbol{k};T)$, which are shown in [Hoffman 1992; 2015]. Let $\mathcal{P}_r$ be the set of partitions of the set $\{1,\dots,r\}$. For any $\Pi = \{P_1,\dots,P_g\} \in \mathcal{P}_r$, we define integers $c(\Pi) = c_r(\Pi)$ and $c^\star(\Pi) = c^\star_r(\Pi)$ by

$$c_r(\Pi) = (-1)^{r-g}\prod_{i=1}^{g}(|P_i|-1)! \quad \text{and} \quad c^\star_r(\Pi) = \prod_{i=1}^{g}(|P_i|-1)!\,, \tag{1-8}$$

respectively, where $|P|$ is the number of the elements of a set P. We also define

$$H_*(\boldsymbol{k},\Pi;T) := \prod_{i=1}^{g}\eta\Big(\sum_{p\in P_i}k_p;T\Big), \tag{1-9}$$

where[2]

$$\eta(k;T) = \begin{cases}\zeta(k), & k>1,\\ T, & k=1.\end{cases}$$

Let S_r denote the symmetric group of degree r. Hoffman's identities are then

$$\sum_{\sigma\in S_r}\zeta_*(k_{\sigma(1)},\dots,k_{\sigma(r)};T) = \sum_{\Pi\in\mathcal{P}_r}c(\Pi)H_*(\boldsymbol{k},\Pi;T), \tag{1-10}$$

$$\sum_{\sigma\in S_r}\zeta^\star_*(k_{\sigma(1)},\dots,k_{\sigma(r)};T) = \sum_{\Pi\in\mathcal{P}_r}c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T). \tag{1-11}$$

Recently, a shuffle version of (1-10) was proved in [Machide 2017], which is obtained by replacing ζ_* and H_* with $\zeta_{\text{Ш}}$ and $H_{\text{Ш}}$, respectively:

$$\sum_{\sigma\in S_r}\zeta_{\text{Ш}}(k_{\sigma(1)},\dots,k_{\sigma(r)};T) = \sum_{\Pi\in\mathcal{P}_r}c(\Pi)H_{\text{Ш}}(\boldsymbol{k},\Pi;T), \tag{1-12}$$

where $H_{\text{Ш}}$ will be defined in (1-15).

The main result of this paper is the shuffle version of (1-11).

Theorem 1.1. *For any index $\boldsymbol{k}$, we have*

$$\sum_{\sigma\in S_r}\zeta^\star_{\text{Ш}}(k_{\sigma(1)},\dots,k_{\sigma(r)};T) = \sum_{\Pi\in\mathcal{P}_r}c^\star(\Pi)H_{\text{Ш}}(\boldsymbol{k},\Pi;T), \tag{1-13}$$

[2]We note that $\eta(k;T) = \zeta_*(k;T) = \zeta_{\text{Ш}}(k;T) = \zeta^\star_*(k;T) = \zeta^\star_{\text{Ш}}(k;T)$.

where $H_{\text{Ш}}(\boldsymbol{k}, \Pi; T)$ *is similar to* $H_*(\boldsymbol{k}, \Pi; T)$, *but the characteristic function*

$$\chi_{\text{Ш}}(\boldsymbol{k}, P_i) := \begin{cases} 0 & \text{if } |P_i| > 1, \text{ and } k_p = 1 \text{ for all } p \in P_i, \\ 1 & \text{otherwise} \end{cases} \tag{1-14}$$

is added in each multiplicand; *that is*,

$$H_{\text{Ш}}(\boldsymbol{k}, \Pi; T) := \prod_{i=1}^{g} \chi_{\text{Ш}}(\boldsymbol{k}, P_i)\eta\bigg(\sum_{p\in P_i} k_p; T\bigg). \tag{1-15}$$

We give some examples of (1-13). The number of the terms on its right-hand side decreases as the number of k_i equal to 1 increases, because of (1-14).

Example 1.2. Let k and l be integers at least 2. Then

$$\zeta^\star(1,k) + \zeta^\star_{\text{Ш}}(k,1;T) = \zeta(k)T + \zeta(k+1),$$

$$\begin{aligned}\zeta^\star(1,k,l) + \zeta^\star(1,l,k) + \zeta^\star(k,1,l) + \zeta^\star(l,1,k) + \zeta^\star_{\text{Ш}}(k,l,1;T) + \zeta^\star_{\text{Ш}}(l,k,1;T)& \\ = (\zeta(k)\zeta(l) + \zeta(k+l))T + \zeta(k)\zeta(l+1) + \zeta(l)\zeta(k+1) + 2\zeta(k+l+1),&\end{aligned}$$

$$2(\zeta^\star(1,1,k) + \zeta^\star_{\text{Ш}}(1,k,1;T) + \zeta^\star_{\text{Ш}}(k,1,1;T)) = \zeta(k)T^2 + 2\zeta(k+1)T + 2\zeta(k+2).$$

In particular, we have a simple equation (1-16) when the number of k_i equal to 1 is $r-1$ (or equivalently, there is just one k_j that is greater than 1): the right-hand side is written in terms of only single zeta values.

Corollary 1.3. *For integers* $k \geq 2$ *and* $r \geq 1$, *we have*

$$\sum_{i=0}^{r-1} \zeta^\star_{\text{Ш}}(\{1\}^i, k, \{1\}^{r-1-i}; T) = \sum_{j=0}^{r-1} \zeta(k+r-1-j)\frac{T^j}{j!}, \tag{1-16}$$

where $\{1\}^i$ *means* i *repetitions of* 1.

The method of the proof of Theorem 1.1 is an improvement to that used in [Machide 2017]. We will use complete exponential Bell polynomials to show Proposition 2.3, which are defined by

$$B_r(x_1, \dots, x_r) := r! \sum_{\substack{i_1,i_2,\dots,i_r\geq 0 \\ 1\cdot i_1+2\cdot i_2+\cdots+r\cdot i_r=r}} \frac{1}{i_1!\,i_2!\cdots i_r!}\prod_{a=1}^{r}\bigg(\frac{x_a}{a!}\bigg)^{i_a}. \tag{1-17}$$

Bell polynomials [1927/28] first appear in the study of set partitions. Currently it is known that they have many relations to combinatorial numbers and applications to other areas; see, e.g., [Comtet 1974; Roman 1980]. We will mention an identity involving $\zeta^\star_*(1,1,\dots,1;T)$ in Remark 2.5, which is a variation of the identity $r! = B_r(0!, 1!, \dots, (r-1)!)$.

This paper is organized as follows. We prepare some propositions in Section 2, and prove Theorem 1.1 and Corollary 1.3 in Section 3.

2. Propositions

In this section, we introduce Propositions 2.1, 2.2, and 2.3, which will be used to prove Theorem 1.1. We will omit the proofs of Propositions 2.1 and 2.2 because these are almost the same as Lemmas 4.7 and 4.8 in [Machide 2017], respectively, where some notation and terminology are modified.

Let $[r]$ denote the set $\{1, \dots, r\}$, and let A and B be its subsets. We denote by $\mathcal{P}(A)$ the set of partitions of A (i.e., $\mathcal{P}([r]) = \mathcal{P}_r$), and we define a subset $\mathcal{P}_B(A)$ in $\mathcal{P}(A)$ by

$$\mathcal{P}_B(A) := \{\Pi = \{P_1, \dots, P_m\} \in \mathcal{P}(A) : P_i \not\subset B \text{ for all } i\}.$$

For example, if $(r, A, B) = (4, \{1, 2, 3\}, \{3, 4\})$, then

$$\mathcal{P}(A) = \{1|2|3, 12|3, 13|2, 23|1, 123\} \quad \text{and} \quad \mathcal{P}_B(A) = \{13|2, 23|1, 123\},$$

where $a_1 \cdots a_p | b_1 \cdots b_q | \cdots$ means a partition such as $12|3 = \{\{1, 2\}, \{3\}\}$.

Let $\Xi = \{P_1, \dots, P_g\} \in \mathcal{P}(A)$, and let $s = |A|$. We will define a partition $\sigma_A(\Xi)$ in $\mathcal{P}_s$ as follows. Let $a_1 < \cdots < a_s$ be the increasing sequence of integers such that

$$A = \{a_1, \dots, a_s\}$$

and let σ_A be the permutation of S_r that is uniquely determined by

$$\sigma_A^{-1}(i) = a_i \quad (i = 1, \dots, s) \quad \text{and} \quad \sigma_A^{-1}(s+1) < \cdots < \sigma_A^{-1}(r);$$

by the definition,

$$\sigma_A(A) = \{\sigma_A(a_1), \dots, \sigma_A(a_s)\} = [s].$$

We then define

$$\sigma_A(\Xi) := \{\sigma_A(P_1), \dots, \sigma_A(P_g)\} \in \mathcal{P}_s.$$

For convenience, $\sigma_A(\Xi) = \phi$ if $A = \Xi = \phi$.

The propositions are as follows.

Proposition 2.1 [Machide 2017, Lemma 4.7]. *For any subset $B \subsetneqq [r]$, we have*

$$\bigsqcup_{A \subset B} \{\Xi \sqcup \Delta : (\Xi, \Delta) \in \mathcal{P}(A) \times \mathcal{P}_B([r] \setminus A)\} = \mathcal{P}_r, \tag{2-1}$$

where $\sqcup$ denotes the disjoint union, and $\bigsqcup_{A \subset B}$ ranges over all subsets in B which include ϕ.

Proposition 2.2 [Machide 2017, Lemma 4.8]. *Let A and B be subsets such that $A \subset B \subsetneqq [r]$, and let (Ξ, Δ) be in $\mathcal{P}(A) \times \mathcal{P}_B([r] \setminus A)$. Let the symbol $\bullet$ mean either $*$ or ш:*

(i) We define $c^\star(\phi) = 1$. We have

$$c^\star(\Xi \cup \Delta) = c^\star(\Xi) c^\star(\Delta). \tag{2-2}$$

(ii) We define $H_\bullet(\phi, \phi; T) = 1$ and

$$\mathbf{1}_s = (\underbrace{1, \dots, 1}_{s}).$$

Suppose that $\boldsymbol{k} = (k_1, \dots, k_r)$ is an index satisfying $B = \{a \in [r] : k_a = 1\}$ (and $\boldsymbol{k} \neq \mathbf{1}_r$). Then we have

$$H_\bullet(\boldsymbol{k}, \Xi \cup \Delta; T) = \left(\prod_{i=1}^{h} \zeta(k_{Q_i}) \right) H_\bullet(\mathbf{1}_{|A|}, \sigma_A(\Xi); T), \tag{2-3}$$

where $Q_1, \dots, Q_h$ are the blocks of Δ (i.e., $\Delta = \{Q_1, \dots, Q_h\}$), and

$$k_{Q_i} = \sum_{q \in Q_i} k_q \quad (i = 1, \dots, h).$$

Note that $k_{Q_i} > 1$ and $\zeta(k_{Q_i})$ is not infinity for any i, because $\Delta \in \mathcal{P}_B([r] \setminus A)$ and $Q_i \not\subset B$.

Proposition 2.3 [Machide 2017, Lemma 4.9]. *For any positive integer r, we have*

$$\sum_{\Pi \in \mathcal{P}_r} c^\star(\Pi) H_*(\mathbf{1}_r, \Pi; T) = \rho^{\star -1}(T^r), \tag{2-4}$$

$$\sum_{\Pi \in \mathcal{P}_r} c^\star(\Pi) H_{\text{Ш}}(\mathbf{1}_r, \Pi; T) = T^r. \tag{2-5}$$

The condition $B \neq [r]$ in the first two propositions is necessary for taking an element in $\mathcal{P}_B([r] \setminus A)$; see [Machide 2017, Remark 4.6] for details.

To prove Proposition 2.3, we need Lemma 2.4, which is the *star*-version of [Machide 2017, Lemma 4.10] in terms of Bell polynomials.

Lemma 2.4. *For any positive integer r, we have*

$$\sum_{\Pi \in \mathcal{P}_r} c^\star(\Pi) H_*(\mathbf{1}_r, \Pi; T) = B_r\big(0!\,\eta(1;T), 1!\,\eta(2;T), \dots, (r-1)!\,\eta(r;T)\big). \tag{2-6}$$

We will now prove Proposition 2.3, and then prove Lemma 2.4.

Proof of Proposition 2.3. We first recall the definition of $\rho^\star$, which is an $\mathbb{R}$-linear endomorphism on $\mathbb{R}[T]$ determined by the equality

$$\rho^\star(e^{Tt}) = A(-t)^{-1} e^{Tt} \tag{2-7}$$

in the formal power series algebra $\mathbb{R}[T][\![t]\!]$ on which $\rho^\star$ acts coefficientwise, see [Kaneko and Yamamoto 2018, Section 4], where

$$A(t) = \exp\left(\sum_{m=2}^{\infty} \frac{(-1)^m \zeta(m)}{m} t^m\right).$$

Note that $A(t) = e^{\gamma t}\Gamma(1+t)$, where γ is Euler's constant. We can see from (2-7) that the inverse endomorphism $\rho^{\star -1}$ exists and it satisfies

$$\rho^{\star -1}(e^{Tt}) = A(-t)e^{Tt} = \exp\left(Tt + \sum_{m=2}^{\infty} \frac{\zeta(m)}{m} t^m\right) = \exp\left(\sum_{m=1}^{\infty} \frac{\eta(m;T)}{m} t^m\right). \tag{2-8}$$

The exponential partial Bell polynomials can be defined by use of the generating function (see [Comtet 1974, Chapter 3]):

$$\exp\left(\sum_{m=1}^{\infty} x_m \frac{t^m}{m!}\right) = \sum_{r=0}^{\infty} B_r(x_1, \dots, x_r) \frac{t^r}{r!}. \tag{2-9}$$

Combining (2-8) and (2-9) with $x_m = (m-1)!\,\eta(m;T)$, we obtain

$$\rho^{\star -1}(e^{Tt}) = \sum_{r=0}^{\infty} B_r\big(0!\,\eta(1;T), 1!\,\eta(2;T), \dots, (r-1)!\,\eta(r;T)\big) \frac{t^r}{r!},$$

which, together with (2-6), gives

$$\rho^{\star -1}(e^{Tt}) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \sum_{\Pi \in \mathcal{P}_r} c^\star(\Pi) H_*(\mathbf{1}_r, \Pi; T).$$

Identity (2-4) follows from comparing the coefficients of t^r on both sides of this equation. We will give a proof of (2-5), which is a modification to that of (4-58) in [Machide 2017, Lemma 4.9].[3] Let $\Lambda = \Lambda_r$ be the partition in $\mathcal{P}_r$ defined by

$$\Lambda_r := 1|2|\cdots|r = \{\{1\}, \{2\}, \ldots, \{r\}\}.$$

We see from (1-14) and (1-15) that $H_{ш}(\mathbf{1}_r, \Pi; T) = 0$ for any $\Pi \in \mathcal{P}_r$ with $\Pi \neq \Lambda$, and so

$$\sum_{\Pi \in \mathcal{P}_r} c^\star(\Pi) H_{ш}(\mathbf{1}_r, \Pi; T) = c^\star(\Lambda) H_{ш}(\mathbf{1}_r, \Lambda; T).$$

Since

$$c^\star(\Lambda) = \prod_{i=1}^{r} 0! = 1 \quad \text{and} \quad H_{ш}(\mathbf{1}_r, \Lambda; T) = \prod_{i=1}^{r} \eta(1; T) = T^r,$$

we obtain (2-5). □

We will need the partial exponential Bell polynomials to prove Lemma 2.4, which we denote by $B_{r,k}(x_1, \ldots, x_{r-k+1})$ for integers r and k with $1 \le k \le r$. Complete and partial Bell polynomials have the relations

$$B_r(x_1, \ldots, x_r) = \sum_{k=1}^{r} B_{r,k}(x_1, \ldots, x_{r-k+1}). \tag{2-10}$$

Let $b_{r,k}(i_1, \ldots, i_{r-k+1})$ be the coefficients of $B_{r,k}(x_1, \ldots, x_{r-k+1})$ such that

$$B_{r,k}(x_1, \ldots, x_{r-k+1}) = \sum_{\substack{i_1, \ldots, i_{r-k+1} \ge 0 \\ i_1 + i_2 + \cdots + i_{r-k+1} = k \\ 1 \cdot i_1 + 2 \cdot i_2 + \cdots + (r-k+1) \cdot i_{r-k+1} = r}} b_{r,k}(i_1, \ldots, i_{r-k+1}) \prod_{a=1}^{r-k+1} x_a^{i_a}.$$

From combinatorial considerations, see, e.g., [Comtet 1974, Chapter 3], we know that $b_{r,k}(i_1, \ldots, i_{r-k+1})$ is the number of partitions with total k blocks in $\mathcal{P}_r$ which consist of i_a blocks of size a for $a \in [r-k+1]$. For instance,

$$b_{4,2}(1, 0, 1) = 4, \quad b_{4,2}(0, 2, 0) = 3, \quad \text{and} \quad B_{4,2}(x_1, x_2, x_3) = 4x_1x_3 + 3x_2^2,$$

since we have four partitions with blocks of size 1 and 3 and three partitions with 2 blocks of size 2 when a set with four elements is divided into two blocks. We note that $B_{r,k} = B_{r,k}(1, \ldots, 1)$ are Stirling numbers of the second kind; that is, they count the number of ways to partition a set of r elements into k nonempty subsets.

[3] There is a misprint in the proof of [Machide 2017, (4-58)]: $\underbrace{\{\{1\}, \ldots, \{1\}\}}_{n}$ should be $\{\{1\}, \ldots, \{n\}\}$.

Proof of Lemma 2.4. For any partition $\Pi = \{P_1, \dots, P_g\}$ in $\mathcal{P}_r$ and integer a in $[r]$, let $N_a(\Pi)$ be the number of the blocks whose cardinalities equal a; i.e.,

$$N_a(\Pi) := |\{j \in [g] : |P_j| = a\}|.$$

We see from the definition that

$$g = N_1(\Pi) + \cdots + N_r(\Pi) \quad \text{and} \quad r = 1 \cdot N_1(\Pi) + \cdots + r \cdot N_r(\Pi),$$

so that

$$c^\star(\Pi) H_*(\mathbf{1}_r, \Pi; T) = \prod_{i=1}^{g} (|P_i| - 1)!\, \eta(|P_i|; T) = \prod_{a=1}^{r} ((a-1)!\, \eta(a; T))^{N_a(\Pi)}.$$

It follows from the combinatorial meaning of $b_{r,k}(i_1, \dots, i_{r-k+1})$ that

$$\sum_{\substack{\Pi \in \mathcal{P}_r \\ N_a(\Pi) = i_a (\forall a)}} 1 = b_{r,k}(i_1, \dots, i_{r-k+1})$$

for nonnegative integers $i_1, \dots, i_{r-k+1}$ with $\sum_{a=1}^{r-k+1} i_a = k$ and $\sum_{a=1}^{r-k+1} a \cdot i_a = r$, and so

$$\begin{aligned}
&\sum_{\substack{i_1, i_2, \dots, i_r \geq 0 \\ 1 \cdot i_1 + 2 \cdot i_2 + \cdots + r \cdot i_r = r}} \sum_{\substack{\Pi \in \mathcal{P}_r \\ N_a(\Pi) = i_a (\forall a)}} c^\star(\Pi) H_*(\mathbf{1}_r, \Pi; T) \\
&\quad = \sum_{\substack{i_1, i_2, \dots, i_r \geq 0 \\ 1 \cdot i_1 + 2 \cdot i_2 + \cdots + r \cdot i_r = r}} \left(\prod_{a=1}^{r} ((a-1)! \eta(a; T))^{i_a} \right) \sum_{\substack{\Pi \in \mathcal{P}_r \\ N_a(\Pi) = i_a (\forall a)}} 1 \\
&\quad = \sum_{k=1}^{r} \sum_{\substack{i_1, i_2, \dots, i_{r-k+1} \geq 0 \\ i_1 + i_2 + \cdots + i_{r-k+1} = k \\ 1 \cdot i_1 + 2 \cdot i_2 + \cdots + (r-k+1) \cdot i_{r-k+1} = r}} \left(\prod_{a=1}^{r-k+1} ((a-1)! \eta(a; T))^{i_a} \right) b_{r,k}(i_1, \dots, i_{r-k+1}) \\
&\quad = \sum_{k=1}^{r} B_{r,k}\big(0!\, \eta(1; T), 1!\, \eta(2; T), \dots, (r-k)!\, \eta(r-k+1; T)\big). \qquad (2\text{-}11)
\end{aligned}$$

We thus obtain (2-6), since the first line of (2-11) is equal to the left-hand side of (2-6) because of [Machide 2017, (4-74)], or

$$\mathcal{P}_r = \bigsqcup_{\substack{i_1, i_2, \dots, i_r \geq 0 \\ 1 \cdot i_1 + 2 \cdot i_2 + \cdots + r \cdot i_r = r}} \{\Pi \in \mathcal{P}_r : N_a(\Pi) = i_a, \text{where } a \in [r]\},$$

and since the last line of (2-11) is equal to the right-hand side of (2-6) because of (2-10). □

Remark 2.5. Bell polynomials are related to many combinatorial numbers. It may be worth noting the relation to the unsigned Stirling numbers of the first kind, which can be expressed as

$$c(r, k) = B_{r,k}(0!, \dots, (r-k)!).$$

The unsigned Stirling numbers are defined as coefficients of the rising factorial; that is,

$$x(x+1)\cdots(x+r-1)=\sum_{k=1}^{r} B_{r,k}(0!,\dots,(r-k)!)x^k. \tag{2-12}$$

Substituting $x=1$ in this equation and combining it with (2-10), we thus obtain

$$r!=B_r(0!,1!,\dots,(r-1)!). \tag{2-13}$$

We also have

$$r!\zeta^\star_*(\mathbf{1}_r;T)=B_r\big(0!\,\eta(1;T),1!\,\eta(2;T),\dots,(r-1)!\,\eta(r;T)\big), \tag{2-14}$$

which follows from Hoffman's identity (1-11) with $\boldsymbol{k}=\mathbf{1}_r$ and (2-6). Equation (2-14) is a variation of (2-13) in the sense that we can obtain (2-14) from (2-13) by replacing $r!$ in the left-hand side with $r!\,\zeta^\star_{\text{Ш}}(\mathbf{1}_r;T)$ and $j!$ in the right-hand side with $j!\,\eta(j+1;T)$. We note that (2-14) corresponds to an identity in terms of symmetric functions (see [Hoffman 1997, Theorem 5.1] and [Kaneko and Yamamoto 2018, Lemma 5.1]):

$$r!h_r=B_r(0!\,p_1,1!\,p_2,\dots,(r-1)!\,p_r), \tag{2-15}$$

where h_i is the complete symmetric function of degree i and p_i is the i-th power sum symmetric function.

3. Proof

We will need (3-1) to prove (1-13), which is the star version of [Machide 2017, (4-51)].

Proposition 3.1. *For any index* $\boldsymbol{k}$,

$$\rho^\star\Big(\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T)\Big)=\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_{\text{Ш}}(\boldsymbol{k},\Pi;T). \tag{3-1}$$

We can prove (3-1) in a quite similar way, as we see below.

Proof of Proposition 3.1. Let $B=\{j\in[r]:k_j=1\}\subset[r]$. We suppose that $B=[r]$. Then, $\boldsymbol{k}=\mathbf{1}_r$, and so we can see from Proposition 2.3 that

$$\rho^\star\Big(\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T)\Big)\overset{(2\text{-}4)}{=}\rho^\star(\rho^{\star-1}(T^r))=T^r\overset{(2\text{-}5)}{=}\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_{\text{Ш}}(\boldsymbol{k},\Pi;T), \tag{3-2}$$

which proves (3-1) for this case.

We suppose that $B\neq[r]$. Let A be a subset in B. Then we have

$$\{\sigma_A(\Xi):\Xi\in\mathcal{P}(A)\}=\{\Xi':\Xi'\in\mathcal{P}_{|A|}\}, \tag{3-3}$$

because the restriction of σ_A to A is a bijection from A to $[|A|]$. From (1-8) we easily see that $c^\star(\Xi)=c^\star(\sigma_A(\Xi))$. Thus,

$$\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T)\overset{\text{Prop. 2.1}}{=}\sum_{A\subset B}\sum_{\substack{\Xi\in\mathcal{P}(A)\\ \Delta\in\mathcal{P}_B([r]\setminus A)}} c^\star(\Xi\cup\Delta)H_*(\boldsymbol{k},\Xi\cup\Delta;T)$$

$$\overset{\text{Prop. 2.2}}{=}\sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)\sum_{\Xi\in\mathcal{P}(A)} c^\star(\Xi)H_*(\mathbf{1}_{|A|},\sigma_A(\Xi);T)$$

$$\overset{(3\text{-}3)}{=} \sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)\sum_{\Xi'\in\mathcal{P}_{|A|}} c^\star(\Xi')H_*(\mathbf{1}_{|A|},\Xi';T)$$

$$\overset{(2\text{-}4)}{=} \sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)\rho^{\star -1}(T^{|A|}),$$

where $Q_1,\dots,Q_h$ mean the blocks of Δ. Therefore,

$$\rho^\star\Big(\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T)\Big) = \sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)\rho^\star(\rho^{\star -1}(T^{|A|}))$$

$$= \sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)T^{|A|}. \tag{3-4}$$

By using Propositions 2.1 and 2.2, and (3-3), and by using (2-5) instead of (2-4), we can similarly prove

$$\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_{\text{ш}}(\boldsymbol{k},\Pi;T) = \sum_{A\subset B}\sum_{\Delta\in\mathcal{P}_B([r]\setminus A)} c^\star(\Delta)\Big(\prod_{i=1}^{h}\zeta(k_{Q_i})\Big)T^{|A|}. \tag{3-5}$$

Equating (3-4) and (3-5), we obtain (3-1) for $B\neq[r]$, and complete the proof. □

We are now in a position to prove Theorem 1.1.

Proof of Theorem 1.1. Recall the *star-regularization theorem*, the second identity of (1-7), which we tag as (s-reg) here. We obtain

$$\sum_{\sigma\in S_r}\zeta^\star_{\text{ш}}(k_{\sigma(1)},\dots,k_{\sigma(r)};T) \overset{\text{(s-reg)}}{=} \rho^\star\Big(\sum_{\sigma\in S_r}\zeta^\star_*(k_{\sigma(1)},\dots,k_{\sigma(r)};T)\Big)$$

$$\overset{(1\text{-}11)}{=} \rho^\star\Big(\sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_*(\boldsymbol{k},\Pi;T)\Big)$$

$$\overset{(3\text{-}1)}{=} \sum_{\Pi\in\mathcal{P}_r} c^\star(\Pi)H_{\text{ш}}(\boldsymbol{k},\Pi;T),$$

which proves (1-13). □

Finally, we deduce Corollary 1.3 from Theorem 1.1.

Proof of Corollary 1.3. Let $\boldsymbol{k}=(k_1,\dots,k_r)$ be the index $(k,1,\dots,1)$, and let $\Pi=\{P_1,\dots,P_m\}$ denote a partition in $\mathcal{P}_r$. Assume $1\in P_1$ through this proof, which does not lose the generality. We see from (1-14) that

$$H_{\text{ш}}(\boldsymbol{k},\Pi;T)=\begin{cases}0 & \text{if } |P_i|>1 \text{ for some } i\geq 2,\\ \zeta(k+|P_1|-1)T^{m-1} & \text{otherwise.}\end{cases}$$

Since

$$m-1=\sum_{i=2}^{m}|P_i|=r-|P_1|$$

if $|P_i| = 1$ for all $i \geq 2$, it follows from (1-13) that

$$(r-1)! \sum_{i=0}^{r-1} \zeta^{\star}_{\text{Ш}}(\{1\}^i, k, \{1\}^{r-1-i}; T) = \sum_{j=1}^{r} \sum_{\Pi \in \mathcal{X}_j} (j-1)!\, \zeta(k+j-1) T^{r-j}$$
$$= \sum_{j=1}^{r} (j-1)!\, \zeta(k+j-1) T^{r-j} \sum_{\Pi \in \mathcal{X}_j} 1, \tag{3-6}$$

where

$$\mathcal{X}_j = \{\{P_1, P_2, \ldots, P_{r+1-j}\} \in \mathcal{P}_r : |P_1| = j, |P_2| = \cdots = |P_{r+1-j}| = 1\}.$$

Noting the assumption $1 \in P_1$, we have

$$\mathcal{X}_j = \big\{\{[r] \setminus \{a_2, \ldots, a_{r+1-j}\}, \{a_2\}, \ldots, \{a_{r+1-j}\}\} : 2 \leq a_2 < \cdots < a_{r+1-j} \leq r\big\},$$

where $[r] \setminus \{a_2, \ldots, a_{r+1-j}\}$ corresponds to P_1 and $\{a_i\}$ $(2 \leq i \leq r)$ correspond to P_i. Thus $|\mathcal{X}_j|$ is the number of $(r-j)$-combinations of $\{2, \ldots, r\}$, or

$$\sum_{\Pi \in \mathcal{X}_j} 1 = \binom{r-j}{r-1}. \tag{3-7}$$

Combining (3-6) and (3-7), we obtain

$$\sum_{i=0}^{r-1} \zeta^{\star}_{\text{Ш}}(\{1\}^i, k, \{1\}^{r-1-i}; T) = \sum_{j=1}^{r} \frac{1}{(r-j)!} \zeta(k+j-1) T^{r-j}.$$

Replacing j with $r-j$ in the right-hand side of this equation gives (1-16). □

Acknowledgements

The author would like to thank referees for useful comments, which improved Section 1 and Remark 2.5. This work was supported by JST ERATO, grant number JPMJER1201, Japan.

References

[Bell 1927/28] E. T. Bell, "Partition polynomials", *Ann. of Math.* (2) **29**:1-4 (1927/28), 38–46. MR Zbl

[Comtet 1974] L. Comtet, *Advanced combinatorics: the art of finite and infinite expansions*, D. Reidel, Dordrecht, 1974. MR Zbl

[Euler 1776] L. Euler, "Meditationes circa singulare serierum genus", *Novi Comm. Acad. Sci. Petropol.* **20** (1776), 140–186. Reprinted in *Opera Omnia Ser. I* **15** (1911), 217–267. JFM

[Hoffman 1992] M. E. Hoffman, "Multiple harmonic series", *Pacific J. Math.* **152**:2 (1992), 275–290. MR Zbl

[Hoffman 1997] M. E. Hoffman, "The algebra of multiple harmonic series", *J. Algebra* **194**:2 (1997), 477–495. MR Zbl

[Hoffman 2015] M. E. Hoffman, "Quasi-symmetric functions and mod p multiple harmonic sums", *Kyushu J. Math.* **69**:2 (2015), 345–366. MR Zbl

[Ihara et al. 2006] K. Ihara, M. Kaneko, and D. Zagier, "Derivation and double shuffle relations for multiple zeta values", *Compos. Math.* **142**:2 (2006), 307–338. MR Zbl

[Kaneko 2018] M. Kaneko, "An introduction to classical and finite multiple zeta values", preprint, 2018, available at http://www2.math.kyushu-u.ac.jp/~mkaneko/papers/Lyon_Proc_rev.pdf. To appear in *Publications Mathématiques de Besançon*.

[Kaneko and Yamamoto 2018] M. Kaneko and S. Yamamoto, "A new integral-series identity of multiple zeta values and regularizations", *Selecta Math.* (*N.S.*) **24**:3 (2018), 2499–2521. MR Zbl

[Machide 2017] T. Machide, "Identities involving cyclic and symmetric sums of regularized multiple zeta values", *Pacific J. Math.* **286**:2 (2017), 307–359. MR Zbl

[Reutenauer 1993] C. Reutenauer, *Free Lie algebras*, London Mathematical Society Monographs (N.S.) **7**, Oxford University Press, New York, 1993. MR Zbl

[Roman 1980] S. Roman, "The formula of Faà di Bruno", *Amer. Math. Monthly* **87**:10 (1980), 805–809. MR Zbl

[Zudilin 2003] V. V. Zudilin, "Algebraic relations for multiple zeta values", *Uspekhi Mat. Nauk* **58**:1 (2003), 3–32. In Russian; translated in *Russian Math. Surveys* **58**:1 (2003), 1–29. MR Zbl

Received 10 Apr 2018.

TOMOYA MACHIDE:

machide@nii.ac.jp
National Institute of Informatics, Tokyo, Japan

and

JST, ERATO, Kawarabayashi Large Graph Project, Global Research Center for Big Data Mathematics, Tokyo, Japan

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.137

Matiyasevich-type identities for hypergeometric Bernoulli polynomials and poly-Bernoulli polynomials

Ken Kamano

We give a Matiyasevich-type identity for hypergeometric Bernoulli polynomials and their generalizations. By using this identity, we also give an identity for poly-Bernoulli polynomials.

1. Introduction and main theorem

The Bernoulli polynomials $B_n(x)$ are defined by the generating function

$$\frac{te^{xt}}{e^t-1}=\sum_{n=0}^{\infty}\frac{B_n(x)}{n!}t^n. \tag{1}$$

When $x=0$, the numbers $B_n(0)=B_n$ are called Bernoulli numbers.

A well-known convolution identity for Bernoulli numbers is the following Euler's formula:

$$\sum_{i=0}^{n}\binom{n}{i}B_iB_{n-i}=-nB_{n-1}-(n-1)B_n \quad (n\geq 1).$$

There are many generalizations of this identity. For example, Dilcher [1996] gave an identity for sums of m products of Bernoulli polynomials ($m=2,3,\ldots$).

On the other hand, by a p-adic method, Miki [1978] proved the following interesting identity which relates two types of convolutions of Bernoulli numbers:

$$\sum_{i=2}^{n-2}\beta_i\beta_{n-i}-\sum_{i=2}^{n-2}\binom{n}{i}\beta_i\beta_{n-i}=2H_n\beta_n \quad (n\geq 4), \tag{2}$$

where $\beta_m:=B_m/m$ and $H_m:=\sum_{i=1}^{m}1/i$. Many alternative proofs and generalizations of this identity have been discovered by several authors; see, e.g., [Crabb 2005; Dilcher and Vignat 2016; Gessel 2005]. Matiyasevich [1997, Identity #0202] discovered the following identity, which also relates two types of convolutions of Bernoulli numbers:

$$(n+2)\sum_{i=2}^{n-2}B_iB_{n-i}-2\sum_{i=2}^{n-2}\binom{n+2}{i}B_iB_{n-i}=n(n+1)B_n \tag{3}$$

for any $n\geq 4$. We note that the identity (3) becomes trivial for odd $n>4$. It is known that Miki's and Matiyasevich's identities can be proved by using a difference operator [Pan and Sun 2006; Artamkin 2007]; see also [Sun and Pan 2006].

MSC2010: 11B68.
Keywords: Matiyasevich's identity, poly-Bernoulli numbers.

Let N be a positive integer and $Q(t) \in t^N\mathbb{R}[\![t]\!]$. We introduce polynomials $f_{N,n}(x; Q) \in \mathbb{R}[x]$ $(n = 0, 1, 2, \ldots)$ by the generating function

$$\frac{Q(t)}{e^t - \sum_{i=0}^{N-1} t^i/i!} e^{xt} = \sum_{n=0}^{\infty} \frac{f_{N,n}(x; Q)}{n!} t^n.$$

When $Q(t) = t^N/N!$, the polynomials $f_{N,n}(x; Q)$ are nothing but the hypergeometric Bernoulli polynomials $B_{N,n}(x)$, which were first introduced by Howard [1967a; 1967b]. We note that $B_{1,n}(x)$ is the ordinary n-th Bernoulli polynomial $B_n(x)$ defined by (1). We denote $f_{N,n}(x; Q)$ by $f_n(x; Q)$ if there is no fear of confusion.

By definition, we have

$$f_n(x + y; Q) = \sum_{i=0}^{n} \binom{n}{i} f_i(y; Q) x^{n-i} \quad (n \geq 0), \tag{4}$$

$$\frac{d}{dx} f_n(x; Q) = n f_{n-1}(x; Q) \quad (n \geq 1). \tag{5}$$

The purpose of this paper is to give a Matiyasevich-type identity for $f_{N,n}$ by using the difference operator. The following is the main theorem of this paper.

Theorem 1.1. *Let N, m and n be integers with $N, m \geq 1$ and $n \geq 0$. For $Q_u(t) \in t^N\mathbb{R}[\![t]\!]$ $(1 \leq u \leq m)$, we have*

$$\begin{aligned}
&\binom{n+N+m-1}{N} \sum_{\substack{i_1,\ldots,i_m \geq 0 \\ i_1+\cdots+i_m=n}} \prod_{u=1}^{m} f_{N,i_u}(x + y_u; Q_u) \\
&\quad = \sum_{p_1,\ldots,p_m \geq 0} \binom{n+N+m-1}{P_m+m-1} B_{N,n-P_m+N}(x) \\
&\qquad \times \left(\left(\prod_{u=1}^{m} f_{N,p_u}(y_u + 1; Q_u) \right) - \sum_{\substack{j_1,\ldots,j_m \geq 0 \\ 0 \leq j_1+\cdots+j_m \leq N-1}} \prod_{l=1}^{m} \binom{p_l}{j_l} f_{N,p_l-j_l}(y_l, Q_l) \right),
\end{aligned} \tag{6}$$

where P_m means $p_1 + \cdots + p_m$.

In Section 2, we give a proof of Theorem 1.1. In Section 3, we see that the identity (6) is a generalization of Matiyasevich's identity (3). Moreover, as an example of identity (6), we give an identity for poly-Bernoulli polynomials.

2. Proof of Theorem 1.1

For an integer $N \geq 1$, let us define a kind of difference operator Δ_N as

$$\Delta_N f(x) := f(x + 1) - \sum_{i=0}^{N-1} \frac{f^{(i)}(x)}{i!} \quad (f(x) \in \mathbb{R}[\![x]\!]), \tag{7}$$

where $f^{(i)}$ is the i-th derivative of f. It is clear that Δ_1 is the ordinary difference operator.

Since

$$\Delta_N\left(\frac{e^{xt}}{e^t-\sum_{i=0}^{N-1}t^i/i!}\right)=e^{xt},$$

we have

$$\Delta_N B_{N,n+N}(x)=\binom{n+N}{N}x^n \quad (n\ge 0). \tag{8}$$

By definition, we have

$$\Delta_N x^m=\begin{cases}\sum_{i=N}^{m}\binom{m}{i}x^{m-i} & \text{for } m\ge N,\\ 0 & \text{for } 0\le m\le N-1.\end{cases}$$

Hence $\{\Delta_N x^N, \Delta_N x^{N+1},\ldots\}$ provides a basis of the vector space $\mathbb{R}[x]$ over $\mathbb{R}$. Therefore $\Delta_N f(x)=0$ implies that $f(x)$ is a polynomial of degree $N-1$ and we obtain the following lemma.

Lemma 2.1. *Let $f(x), g(x)\in\mathbb{R}[x]$. If $\Delta_N f(x)=\Delta_N g(x)$, then $f(x)$ and $g(x)$ agree in their coefficients of x^j for $j\ge N$.*

By the identity

$$\sum_{i=0}^{\infty}\binom{i}{p}x^i=\frac{x^p}{(1-x)^{p+1}} \quad (p\ge 0),$$

we have

$$\sum_{i_1=0}^{\infty}\binom{i_1}{p_1}x^{i_1}\cdots\sum_{i_m=0}^{\infty}\binom{i_m}{p_m}x^{i_m}=\frac{x^{p_1+\cdots+p_m}}{(1-x)^{p_1+\cdots+p_m+m}}$$

for $m\ge 1$. By comparing the coefficients of both sides, we obtain the following lemma.

Lemma 2.2. *For integers $m\ge 1$, $n\ge 0$ and $p_1,\ldots,p_m\ge 0$, we have*

$$\sum_{\substack{i_1,\ldots,i_m\ge 0\\ i_1+\cdots+i_m=n}}\binom{i_1}{p_1}\cdots\binom{i_m}{p_m}=\binom{n+m-1}{p_1+\cdots+p_m+m-1}.$$

Now we prove our main theorem.

Proof of Theorem 1.1. For integers $i_1,\ldots,i_m\ge 0$, we have by (7)

$$\begin{aligned}&\Delta_N\left(\prod_{u=1}^{m}f_{i_u}(x+y_u;Q_u)\right)\\ &=\left(\prod_{u=1}^{m}f_{i_u}(x+1+y_u;Q_u)\right)-\sum_{j=0}^{N-1}\frac{1}{j!}\frac{d^j}{dx^j}\prod_{u=1}^{m}f_{i_u}(x+y_u;Q_u)\\ &=\prod_{u=1}^{m}\left(\sum_{p_u=0}^{i_u}\binom{i_u}{p_u}f_{p_u}(y_u+1;Q_u)x^{i_u-p_u}\right)-\sum_{\substack{j_1,\ldots,j_m\ge 0\\ 0\le j_1+\cdots+j_m\le N-1}}\frac{f_{i_1}^{(j_1)}(x+y_1;Q_1)\cdots f_{i_m}^{(j_m)}(x+y_m;Q_m)}{j_1!\cdots j_m!},\end{aligned} \tag{9}$$

where we have used the general Leibniz rule. For any $i,j\ge 0$ we have, by (4) and (5),

$$\frac{f_i^{(j)}(x+y;Q)}{j!}=\binom{i}{j}f_{i-j}(x+y;Q)=\binom{i}{j}\sum_{p=0}^{i-j}\binom{i-j}{p}f_p(y;Q)x^{i-j-p}=\sum_{p=j}^{i}\binom{i}{p}\binom{p}{j}f_{p-j}(y;Q)x^{i-p},$$

where an empty sum is taken to be zero. Hence

$$\sum_{\substack{j_1,\dots,j_m\ge 0\\ 0\le j_1+\cdots+j_m\le N-1}} \frac{f_{i_1}^{(j_1)}(x+y_1;\,Q_1)\cdots f_{i_m}^{(j_m)}(x+y_m;\,Q_m)}{j_1!\cdots j_m!} = \sum_{\substack{j_1,\dots,j_m\ge 0\\ 0\le j_1+\cdots+j_m\le N-1}} \prod_{u=1}^{m}\Bigg(\sum_{p_u=j_u}^{i_u}\binom{i_u}{p_u}\binom{p_u}{j_u} f_{p_u-j_u}(y_u;\,Q_u)x^{i_u-p_u}\Bigg).$$

Therefore, by Lemma 2.2, we have

$$\begin{aligned}
&\sum_{\substack{i_1,\dots,i_m\ge 0\\ i_1+\cdots+i_m=n}} \Delta_N\Bigg(\prod_{u=1}^{m} f_{i_u}(x+y_u;\,Q_u)\Bigg)\\
&\quad= x^n \sum_{p_1,\dots,p_m\ge 0}\binom{n+m-1}{P_m+m-1}\Bigg(\prod_{u=1}^{m} f_{p_u}(y_u+1;\,Q_u)x^{-p_u}\Bigg)\\
&\qquad\qquad - x^n \sum_{0\le j_1+\cdots+j_m\le N-1}\ \sum_{p_1,\dots,p_m\ge 0}\binom{n+m-1}{P_m+m-1}\prod_{u=1}^{m}\binom{p_u}{j_u} f_{p_u-j_u}(y_u;\,Q_u)x^{-p_u}\\
&\quad= \sum_{p_1,\dots,p_m\ge 0} x^{n-P_m}\binom{n+m-1}{P_m+m-1}\Bigg(\prod_{u=1}^{m} f_{p_u}(y_u+1;\,Q_u) - \sum_{0\le j_1+\cdots+j_m\le N-1}\ \prod_{u=1}^{m}\binom{p_u}{j_u} f_{p_u-j_u}(y_u;\,Q_u)\Bigg).
\end{aligned}$$

By the relation

$$x^{n-P_m} = \frac{\Delta_N B_{N,n-P_m+N}(x)}{\binom{n-P_m+N}{N}},$$

which comes from (8), we have, for $n \ge 0$,

$$\begin{aligned}
&\Delta_N \sum_{\substack{i_1,\dots,i_m\ge 0\\ i_1+\cdots+i_m=n}} \prod_{u=1}^{m} f_{i_u}(x+y_u;\,Q_u)\\
&\quad= \Delta_N \sum_{p_1,\dots,p_m\ge 0} \frac{1}{\binom{n-P_m+N}{N}} B_{N,n-P_m+N}(x)\binom{n+m-1}{P_m+m-1}\\
&\qquad\qquad \times\Bigg(\Bigg(\prod_{u=1}^{m} f_{p_u}(y_u+1;\,Q_u)\Bigg) - \sum_{0\le j_1+\cdots+j_m\le N-1}\ \prod_{u=1}^{m}\binom{p_u}{j_u} f_{p_u-j_u}(y_u;\,Q_u)\Bigg).
\end{aligned}$$

Applying Lemma 2.1 to this last identity, with

$$\frac{1}{\binom{n-P_m+N}{N}}\binom{n+m-1}{P_m+m-1} = \frac{\binom{n+N+m-1}{P_m+m-1}}{\binom{n+N+m-1}{N}},$$

we see that (6) holds up to a polynomial in x of degree $N-1$. Finally, for any $n \ge 0$, by replacing n by $n+N$ in (6) and differentiating with respect to x both sides N times, we obtain (6) for n. □

3. Identities for poly-Bernoulli polynomials

In this section, we give some identities derived from Theorem 1.1. Firstly, we give identities for the ordinary Bernoulli polynomials.

Corollary 3.1. *The following identities hold*:

$$(n+2)\sum_{i_1+i_2=n} B_{i_1}(x)B_{i_2}(x) = \binom{n+2}{3}B_{n-1}(x)+2\sum_{p\geq 0}\binom{n+2}{p+2}B_pB_{n-p}(x) \qquad (n\geq 1), \quad (10)$$

$$(n+2)\sum_{i_1+i_2=n} B_{i_1}(y_1)B_{i_2}(y_2) = \sum_{p_1,p_2\geq 0}\binom{n+2}{p_1+p_2+1}B_{n-p_1-p_2+1} \times\big(B_{p_1}(y_1+1)B_{p_2}(y_2+1)-B_{p_1}(y_1)B_{p_2}(y_2)\big) \quad (n\geq 0). \quad (11)$$

Proof. We apply $N=1$, $m=2$, $Q_1(t)=Q_2(t)=t$ and $y_1=y_2=0$ in Theorem 1.1. Since $f_{1,n}(x;t)=B_n(x)$, we have

$$(n+2)\sum_{i_1+i_2=n} B_{i_1}(x)B_{i_2}(x) = \sum_{p_1,p_2\geq 0}\binom{n+2}{p_1+p_2+1}B_{n-p_1-p_2+1}(x)(B_{p_1}(1)B_{p_2}(1)-B_{p_1}B_{p_2}). \quad (12)$$

It is well known that $B_p(1)=B_p+\delta_{1p}$ $(p\geq 0)$, where δ_{ij} is Kronecker's delta function. Therefore the right-hand side of (12) equals

$$\binom{n+2}{3}B_{n-1}(x)+2\sum_{p\geq 0}\binom{n+2}{p+2}B_{n-p}(x)B_p,$$

and this proves (10). Equation (11) can be also proved by applying $x=0$ in Theorem 1.1. □

Remark 3.2. (i) Matiyasevich's identity (3) can be obtained by setting $x=0$ in (10).

(ii) Agoh and Dilcher [2014, Theorem 1] gave an identity which includes (10). Pan and Sun [2006, Theorem 2.1] gave an identity for $\sum B_{i_1}(y_1)B_{i_2}(y_2)$ with $y_1\neq y_2$, but our identity (11) is different from theirs.

For any integer k, poly-Bernoulli polynomials $C_n^{(k)}(x)$ are defined by the generating function

$$\frac{\mathrm{Li}_k(1-e^{-t})}{e^t-1}e^{xt}=\sum_{n=0}^{\infty}\frac{C_n^{(k)}(x)}{n!}t^n;$$

see, e.g., [Imatomi 2014, Chapter 6]. Here $\mathrm{Li}_k(z)$ is the k-th polylogarithm defined by $\mathrm{Li}_k(z)=\sum_{n=1}^{\infty}z^n/n^k$. The numbers $C_n^{(k)}(1)$ and $C_n^{(k)}(0)$ are poly-Bernoulli numbers $B_n^{(k)}$ and $C_n^{(k)}$ introduced by Kaneko [1997] and Arakawa and Kaneko [1999], respectively. When $k=1$, it can be checked that $C_n^{(1)}(x)=B_n(x)$ where $B_n(x)$ are the ordinary Bernoulli polynomials defined by (1). When $N=1$ and $Q(t)=\mathrm{Li}_k(1-e^{-t})$, we have $f_n(x;Q)=C_n^{(k)}(x)$. Hence the following corollary is obtained from Theorem 1.1.

Corollary 3.3. *For integers k_1, k_2 and n with $n\geq 0$, we have*

$$(n+2)\sum_{i_1+i_2=n} C_{i_1}^{(k_1)}(x)C_{i_2}^{(k_2)}(x) = \sum_{p_1,p_2\geq 0}\binom{n+2}{p_1+p_2+1}B_{n-p_1-p_2+1}(x)(B_{p_1}^{(k_1)}B_{p_2}^{(k_2)}-C_{p_1}^{(k_1)}C_{p_2}^{(k_2)}).$$

It is known that $B_n^{(k)} = C_n^{(k)} + C_{n-1}^{(k-1)}$ for $n \geq 0$. Here, when $n = 0$, we set $C_{-1}^{(k-1)} = 0$ for any k. Hence the identity above can be rewritten in the form using only $C_n^{(k)}$:

Corollary 3.4. *For integers k_1, k_2 and n with $n \geq 0$, we have*

$$(n+2) \sum_{i_1+i_2=n} C_{i_1}^{(k_1)}(x) C_{i_2}^{(k_2)}(x) = \sum_{p_1,p_2\geq 0} \binom{n+2}{p_1+p_2+1} B_{n-p_1-p_2+1}(x)(C_{p_1}^{(k_1)} C_{p_2-1}^{(k_2-1)} + C_{p_2}^{(k_2)} C_{p_1-1}^{(k_1-1)} + C_{p_1-1}^{(k_1-1)} C_{p_2-1}^{(k_2-1)}).$$

Acknowledgement

This work was supported by Grant-in-Aid for Young Scientists (B) from JSPS KAKENHI (16K17583).

References

[Agoh and Dilcher 2014] T. Agoh and K. Dilcher, "Higher-order convolutions for Bernoulli and Euler polynomials", *J. Math. Anal. Appl.* **419**:2 (2014), 1235–1247. MR Zbl

[Arakawa and Kaneko 1999] T. Arakawa and M. Kaneko, "On poly-Bernoulli numbers", *Comment. Math. Univ. St. Paul.* **48**:2 (1999), 159–167. MR Zbl

[Artamkin 2007] I. V. Artamkin, "An elementary proof of the Miki–Zagier–Gessel identity", *Uspekhi Mat. Nauk* **62**:6 (2007), 165–166. In Russian; translated in *Russian Math. Surveys* **62**:6 (2007), 1194–1196. MR Zbl

[Crabb 2005] M. C. Crabb, "The Miki–Gessel Bernoulli number identity", *Glasg. Math. J.* **47**:2 (2005), 327–328. MR Zbl

[Dilcher 1996] K. Dilcher, "Sums of products of Bernoulli numbers", *J. Number Theory* **60**:1 (1996), 23–41. MR Zbl

[Dilcher and Vignat 2016] K. Dilcher and C. Vignat, "General convolution identities for Bernoulli and Euler polynomials", *J. Math. Anal. Appl.* **435**:2 (2016), 1478–1498. MR Zbl

[Gessel 2005] I. M. Gessel, "On Miki's identity for Bernoulli numbers", *J. Number Theory* **110**:1 (2005), 75–82. MR Zbl

[Howard 1967a] F. T. Howard, "A sequence of numbers related to the exponential function", *Duke Math. J.* **34** (1967), 599–615. MR Zbl

[Howard 1967b] F. T. Howard, "Some sequences of rational numbers related to the exponential function", *Duke Math. J.* **34** (1967), 701–716. MR Zbl

[Imatomi 2014] K. Imatomi, *Multiple zeta values and multi-poly-Bernoulli numbers*, Ph.D. thesis, Kyushu University, 2014, available at https://catalog.lib.kyushu-u.ac.jp/opac_download_md/1441041/math162.pdf.

[Kaneko 1997] M. Kaneko, "Poly-Bernoulli numbers", *J. Théor. Nombres Bordeaux* **9**:1 (1997), 221–228. MR Zbl

[Matiyasevich 1997] Y. Matiyasevich, "Identities with Bernoulli numbers", web page, 1997, available at http://logic.pdmi.ras.ru/~yumat/Journal/Bernoulli/bernulli.htm.

[Miki 1978] H. Miki, "A relation between Bernoulli numbers", *J. Number Theory* **10**:3 (1978), 297–302. MR Zbl

[Pan and Sun 2006] H. Pan and Z.-W. Sun, "New identities involving Bernoulli and Euler polynomials", *J. Combin. Theory Ser. A* **113**:1 (2006), 156–175. MR Zbl

[Sun and Pan 2006] Z.-W. Sun and H. Pan, "Identities concerning Bernoulli and Euler polynomials", *Acta Arith.* **125**:1 (2006), 21–39. MR Zbl

Received 18 Apr 2018. Revised 19 Aug 2018.

KEN KAMANO:

ken.kamano@oit.ac.jp
Department of Robotics, Osaka Institute of Technology, Osaka, Japan

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.143

A family of four-variable expanders with quadratic growth

Mehdi Makhul

We prove that if $g(x, y)$ is a polynomial of degree d that is not a polynomial of only y, then for any finite set $A \subset \mathbb{R}$

$$|X| \gg_d |A|^2, \quad \text{where } X := \left\{ \frac{g(a_1, b_1) - g(a_2, b_2)}{b_2 - b_1} : a_1, a_2, b_1, b_2 \in A \right\}.$$

We will see this bound is also tight for some polynomial $g(x, y)$.

1. Introduction

Throughout this paper, when we write $X \gg Y$, this means that $X \geq cY$ for some absolute constant $c > 0$.

The sum set of a subset $A \subset \mathbb{R}$ is defined as $A + A := \{a + b : a, b \in A\}$. The product set is defined in a similar way, $AA := \{ab : a, b \in A\}$.

The Erdős–Szemerédi conjecture [1983] states that, for all $\epsilon > 0$ and for any finite set $A \subset \mathbb{N}$,

$$\max\{|A + A|, |AA|\} \geq c(\epsilon)|A|^{2-\epsilon}.$$

It is natural to extend this conjecture to other settings (such as $\mathbb{R}$), and also to change the polynomials $F(x, y) = x + y$ and $F(x, y) = xy$ defining the sum and product sets to other polynomials or rational functions. In recent years much research has been done in this direction.

For many such functions, the images of sets are known to always grow. For example, the authors of [Murphy et al. 2015] have studied several multivariable polynomials, including the function

$$G(x_1, x_2, x_3, x_4, x_5) = x_1(x_2 + x_3 + x_4 + x_5).$$

More precisely they showed that, for any finite set $A \subset \mathbb{R}$,

$$|A(A + A + A + A)| \gg \frac{|A|^2}{\log |A|},$$

where $A(A + A + A + A) := \{x_1(x_2 + x_3 + x_4 + x_5) : x_i \in A\}$.

In [Murphy et al. 2017], the authors studied a more complicated function, namely

$$H(x_1, x_2, x_3, x_4, x_5) = (x_1 + x_2 + x_3 + x_4)^2 + \log x_5.$$

Makhul was supported by the Austrian Science Fund (FWF): W1214-N15, Project DK9.
MSC2010: primary 11B30; secondary 11B75.
Keywords: Bisector, expander functions.

They showed that, for any finite $A \subset \mathbb{R}$,

$$|\{(a_1 + a_2 + a_3 + a_4)^2 + \log a_5 : a_i \in A\}| \gg \frac{|A|^2}{\log |A|}.$$

In the same circle of ideas, Balog and Roche-Newton [2015] investigated the rational function

$$F(x_1, x_2, x_3, x_4) = \frac{x_1 + x_2}{x_3 + x_4},$$

showing that for any finite set $A \subset \mathbb{R}$, we have

$$|F(A, A, A, A)| \geq 2|A|^2 - 1.$$

Our result is a generalization of the method of [Murphy et al. 2015, Corollary 3.1], where they used the Szemerédi–Trotter theorem to prove that for any finite set $A \subset \mathbb{R}$

$$\left|\frac{A - A}{A - A}\right| \gg |A|^2.$$

A stronger version of this result, with a multiplicative constant 1, follows from an earlier geometric result of Ungar [1982].

In this article we consider a certain class of rational functions of four variables. Suppose that $g(x, y)$ is a polynomial of two variables of degree d. Let

$$F(x_1, x_2, y_1, y_2) = \frac{g(x_1, y_1) - g(x_2, y_2)}{y_2 - y_1}$$

be a four-variable rational function in terms of x_1, x_2, y_1, y_2. The main theorem of this paper is the following result concerning the growth of F.

Theorem 1.1. *Suppose that $g(x, y)$ is a polynomial of degree d, that it is not a polynomial of only y, and that $A \subset \mathbb{R}$ is a finite set. Then*

$$|X| \gg_d |A|^2, \quad \text{where } X := \left\{\frac{g(a_1, b_1) - g(a_2, b_2)}{b_2 - b_1} : a_1, a_2, b_1, b_2 \in A\right\}.$$

Notice that the following example shows that the condition that $g(x, y)$ cannot be a polynomial of only y is necessary.

Example 1.2. Suppose that $g(x, y) = y^2$ and $A = \{1, 2, \ldots, n\}$. Then

$$X = \left\{\frac{b_1^2 - b_2^2}{b_2 - b_1} : b_1, b_2 \in A\right\}$$

equals $-\{b_2 + b_1 : b_i \in A\}$ and has cardinality $O(n)$.

On the other hand, it is known that for some polynomials g, the result of Theorem 1.1 is tight. For example, if we define $g(x_1, y_1) = x_1$ then Theorem 1.1 recovers the result of [Murphy et al. 2015; Ungar

1982]. This is known to be tight, since for the set $A = \{1, \dots, N\}$,

$$\left|\frac{A-A}{A-A}\right| = O(N^2).$$

However, we are not aware of any other polynomials g for which the bound in Theorem 1.1 is tight, and whether or not the bound can be improved for some particular g is an interesting question.

Our main result has some similarities with a result of Raz, Sharir and Solymosi [Raz et al. 2015] concerning the growth of two-variable polynomials. Their result states that, if F is a two-variable polynomial with bounded degree, then for any $A, B \subset \mathbb{R}$ with $|A| = |B| = n$,

$$|F(A, B)| \gg_d n^{4/3},$$

provided that F satisfies a nondegeneracy condition. This condition states that F cannot be of one of the following forms:

(1) $F(u, v) = f(g(u) + h(v))$.

(2) $F(u, v) = f(g(u) \cdot h(v))$.

This result gave an improvement upon an earlier result of Elekes and Ronyai [2000].

The Szemerédi–Trotter theorem. The essential ingredient used to prove our result is a corollary of the Szemerédi–Trotter theorem [1983], which gives a bound for the number of lines in the plane containing at least a fixed number of points k from a given finite set, that is, the number of *k-rich* lines.

Theorem 1.3. *Let P be a finite set of points and let L be a finite set of lines. Then the number of incidences $I(P, L) := \{(p, \ell) \in P \times L : p \in \ell\}$ has the upper bound*

$$I(P, L) \ll |P|^{2/3}|L|^{2/3} + |P| + |L|.$$

More precisely,

$$I(P, L) \le 4|P|^{2/3}|L|^{2/3} + 4|P| + |L|.$$

If each line in L appears at most d times for some constant d, then a generalization of the Szemerédi–Trotter theorem states that

$$I(P, L) \le 4d|P|^{2/3}|L|^{2/3} + 4d|P| + d|L|.$$

The main idea of the following corollary is known in literature; we present here a slightly improved version which we could not find in the literature in the form we need.

Corollary 1.4. *Let $k, n \ge 2$ be natural numbers and fix $d \in \mathbb{N}$ such that $4d + 1 \le k \le d\sqrt{n}$. Let $\mathcal{L}$ be a set of n lines in the plane, and let $t_{\ge k}$ denote the number of points in the plane contained in at least k lines of $\mathcal{L}$, where each line appears with multiplicity at most d. Then*

$$t_{\ge k} = O_d\left(\frac{n^2}{k^3}\right).$$

Notice that if $\mathcal{L}$ is a set of n lines in the plane such that each line appears at most d times for some constant d, then for computing t_k, k should be greater than or equal to $4d + 1$. To see this, suppose that

P_k is the set of k-rich points. Then we have $k|P_k| \le 4d|P_k|^{2/3}|L|^{2/3} + 4d|P_k| + d|L|$. This implies

$$(k - 4d)|P_k| \le 4d|P_k|^{2/3}|L|^{2/3} + d|L|.$$

Hence we may assume $k \ge 4d + 1$, otherwise the inequality gives nothing.

Proof of Corollary 1.4. Let P_k be the set of k-rich points. Since each line appears at most d times we have

$$\frac{k|P_k|}{d} \ll |P_k|^{2/3}|L|^{2/3} + |L|,$$

so $k^3|P_k| \ll d^3|L|^2$ or otherwise $|P_k| \ll d|L|/k$. Plugging these bounds back into the Szemerédi–Trotter theorem gives

$$I(P_k, L) \ll |L|^{2/3}\left(\frac{d^3|L|^2}{k^3}\right)^{2/3} + |L|^{2/3}\left(\frac{d|L|}{k}\right)^{2/3} + |L| + \frac{d^3|L|^2}{k^3} + \frac{d|L|}{k}.$$

Since $k > 4d$ we can ignore last two summands and we obtain

$$I(P_k, L) \ll \frac{d^2|L|^2}{k^2} + \left(\frac{d}{k}\right)^{2/3}|L|^{4/3} + |L|.$$

Note that we have

$$\frac{d^2|L|^2}{k^2} \ge \left(\frac{d}{k}\right)^{2/3}|L|^{4/3}$$

if $k \le d\sqrt{n}$. □

2. Main results

Suppose that $A, B \subset \mathbb{R}$ are finite, and $g(x_1, y_1)$ is a polynomial of degree d. We associate an element of $A \times B$ with a line via

$$A \times B \ni (a, b) \longleftrightarrow l_{a,b} : y = bx - g(a, b).$$

Consider $\mathcal{L} = \{\ell_{a,b} : a, b \in A \times B\}$ as a multiset. Then $\mathcal{L}$ is a set of $|A||B|$ lines, such that each line appears at most d times. We also define the quantity

$$n(x, y) = |\{(a, b) \in A \times B : (x, y) \in l_{a,b}\}|,$$

which is interpreted geometrically as the number of lines of $\mathcal{L}$ that pass through (x, y).

Lemma 2.1. *Suppose that $d \in \mathbb{N}$ is fixed. Suppose that $A, B, X \subset \mathbb{R}$ are finite and satisfy*

$$|X| \le \frac{|A||B|}{4d^2},$$

with $0 \notin X$. Then

$$\sum_{x \in X}\sum_{y} n^2(x, y) \ll |A|^{3/2}|B|^{3/2}|X|^{1/2}. \tag{1}$$

Proof. The set of t-rich points is given by

$$R_t := \{(x, y) \in \mathbb{R}^2 : n(x, y) \ge t\}.$$

We first show that

$$|R_t| \ll \frac{|A|^2|B|^2}{t^3}.$$

We begin by bounding $n(x, y)$ for a given point (x, y). For fixed $b_0 \in B$ we obtain a line with slope b_0 passing through (x, y) and a one-variable polynomial equation $g(a, b_0)$. Since each line is determined uniquely, by its slope and one point on it (for fixed b_0 and (x, y) the equation $g(a, b_0) = 0$ has at most d distinct solutions), we have

$$n(x, y) \le d|B|.$$

With a similar argument for fixed $a \in A$ we obtain a univariate polynomial equation. Since each line is determined uniquely by its y-intercept and one point on it we have

$$n(x, y) \le d|A|.$$

These together imply

$$n(x, y) \le d(\min\{|A|, |B|\}) \le (d|A|d|B|)^{1/2} = d|\mathcal{L}|^{1/2}.$$

This implies there are no points incident to more than $d\sqrt{|\mathcal{L}|}$ lines in $\mathcal{L}$, and by applying Corollary 1.4 we get

$$|R_t| \ll \frac{|\mathcal{L}|^2}{t^3} \le \frac{|A|^2|B|^2}{t^3}.$$

Let $\Delta > 2d$ be an integer to be specified later. We have

$$\sum_{x\in X}\sum_{y} n^2(x, y) \le \sum_{x\in X}\sum_{n(x,y)\le\Delta} n^2(x, y) + \sum_{\substack{(x,y)\\ n(x,y)>\Delta}} n^2(x, y). \tag{2}$$

The first term is bounded by $\Delta|A||B||X|$; in fact

$$\sum_{x\in X}\sum_{n(x,y)\le\Delta} n^2(x, y) \le \Delta\sum_{x\in X}\sum_{y} n(x, y) = \Delta|A||B|\sum_{x\in X} 1 = \Delta|A||B||X|. \tag{3}$$

For the second term we have

$$\begin{aligned}\sum_{\substack{(x,y)\\ n(x,y)>\Delta}} n^2(x, y) &= \sum_{j\ge1}\sum_{2^{j-1}\Delta\le n(x,y)\le 2^j\Delta} n^2(x, y)\\ &\ll \sum_{j\ge1}\frac{|A|^2|B|^2}{(2^j\Delta)^3}\cdot(2^j\Delta)^2 = \frac{|A|^2|B|^2}{\Delta}\sum_{j\ge1}\frac{1}{2^j} = \frac{|A|^2|B|^2}{\Delta}.\end{aligned} \tag{4}$$

For an optimal choice, set

$$\Delta = \left\lceil \frac{(|A||B|)^{1/2}}{|X|^{1/2}} \right\rceil > 2d.$$

Combining the bounds from (2) and (3) and (4), it follows that

$$\sum_{x}\sum_{y} n^2(x, y) \ll |A|^{3/2}|B|^{3/2}|X|^{1/2}. \qquad \square$$

Proof of Theorem 1.1. Consider

$$\begin{aligned}|Y| &= \left|\left\{(x, a_1, a_2, b_1, b_2) : x = \frac{g(a_1, b_1) - g(a_2, b_2)}{b_1 - b_2}, a_i, b_i \in A\right\}\right| \\ &= \left|\left\{(x, a_1, a_2, b_1, b_2) : b_1 x - g(a_1, b_1) = b_2 x - g(a_2, b_2)\right\}\right| \\ &= \sum_{x \in X} \sum_{y} n^2(x, y) \ll |A|^3 |X|^{1/2}.\end{aligned}$$

On the other hand, $|Y| \geq |A|^4$. Thus we obtain

$$|A|^4 \ll |A|^3 |X|^{1/2}, \quad \text{and hence} \quad |X| \gg |A|^2. \qquad \square$$

Notice that the proof of Theorem 1.1 fails when $g(x, y)$ is a polynomial of only y. In fact if $g(x, y) = h(y)$ for some polynomial h, then $\mathcal{L} = \{l_{a,b} : a, b \in A \times B\}$ is a set of $|A||B|$ lines such that each line appears at least $|A|$ times (and at most $d|A|$ times). On the other hand the generalization of the Szemerédi–Trotter theorem and its corollary hold when each line appears at most d times, where d is independent of $|P|$ and $|L|$ in Theorem 1.3.

Corollary 2.2. *Suppose that $P = A \times A$ is a set of $|A|^2$ points. Let l be the y-axis. Suppose that $B(P)$ is the set of all bisectors determined by P. Then $|B \cap l| \gg |A|^2$.*

Proof. By a simple calculation we can see that the equation of the bisector determined by two points (x_1, y_1) and (x_2, y_2) in the (s, t)-plane is

$$s = \frac{2(x_1 - x_2)t + (x_2^2 - x_1^2) + (y_2^2 - y_1^2)}{2(y_2 - y_1)}.$$

Inserting $t = 0$, the hitting point has coordinate

$$\left(0, \frac{(x_2^2 - x_1^2) + (y_2^2 - y_1^2)}{2(y_2 - y_1)}\right).$$

Setting $g(x, y) = -\frac{1}{2}(x^2 + y^2)$, we obtain the result by Theorem 1.1. $\square$

As we mentioned, this bound is tight for some polynomials, for instance $g(x, y) = x$. However, we expect that if $F(x_1, x_2, y_1, y_2)$ is a generic rational function satisfying the condition of Theorem 1.1 we have $|X| \gg |A|^3$.

Acknowledgements

I would like to thank Oliver Roche-Newton for bringing this problem to my attention and for several helpful conversations.

References

[Balog and Roche-Newton 2015] A. Balog and O. Roche-Newton, "New sum-product estimates for real and complex numbers", *Discrete Comput. Geom.* **53**:4 (2015), 825–846. MR Zbl

[Elekes and Rónyai 2000] G. Elekes and L. Rónyai, "A combinatorial problem on polynomials and rational functions", *J. Combin. Theory Ser. A* **89**:1 (2000), 1–20. MR Zbl

[Erdős and Szemerédi 1983] P. Erdős and E. Szemerédi, "On sums and products of integers", pp. 213–218 in *Studies in pure mathematics*, edited by P. Erdős, Birkhäuser, Basel, 1983. MR Zbl

[Murphy et al. 2015] B. Murphy, O. Roche-Newton, and I. Shkredov, "Variations on the sum-product problem", *SIAM J. Discrete Math.* **29**:1 (2015), 514–540. MR Zbl

[Murphy et al. 2017] B. Murphy, O. Roche-Newton, and I. D. Shkredov, "Variations on the sum-product problem, II", *SIAM J. Discrete Math.* **31**:3 (2017), 1878–1894. MR Zbl

[Raz et al. 2015] O. E. Raz, M. Sharir, and J. Solymosi, "On triple intersections of three families of unit circles", *Discrete Comput. Geom.* **54**:4 (2015), 930–953. MR Zbl

[Szemerédi and Trotter 1983] E. Szemerédi and W. T. Trotter, Jr., "Extremal problems in discrete geometry", *Combinatorica* **3**:3-4 (1983), 381–392. MR Zbl

[Ungar 1982] P. Ungar, "$2N$ noncollinear points determine at least $2N$ directions", *J. Combin. Theory Ser. A* **33**:3 (1982), 343–347. MR Zbl

Received 11 May 2018. Revised 17 Jul 2018.

MEHDI MAKHUL:

mmakhul@risc.jku.at

Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz and Research Institute for Symbolic Computation (RISC), Johannes Kepler University, Linz, Austria

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.151

The Lind–Lehmer Constant for $\mathbb{Z}_2^r \times \mathbb{Z}_4^s$

Michael J. Mossinghoff, Vincent Pigno and Christopher Pinner

For a finite abelian group the Lind–Lehmer constant is the minimum positive logarithmic Lind–Mahler measure for that group. Finding this is equivalent to determining the minimal nontrivial group determinant when the matrix entries are integers.

For a group of the form $G = \mathbb{Z}_2^r \times \mathbb{Z}_4^s$ with $|G| \geq 4$ we show that this minimum is always $|G| - 1$, a case of sharpness in the trivial bound. For $G = \mathbb{Z}_2 \times \mathbb{Z}_{2^n}$ with $n \geq 3$ the minimum is 9, and for $G = \mathbb{Z}_3 \times \mathbb{Z}_{3^n}$ the minimum is 8. Previously the minimum was only known for 2- and 3-groups of the form $G = \mathbb{Z}_p^k$ or $\mathbb{Z}_{p^k}$. We also show that a congruence satisfied by the group determinant when $G = \mathbb{Z}_p^r$ generalizes to other abelian p-groups.

1. Introduction

Recall that for a polynomial $F(x_1, \ldots, x_k)$ in $\mathbb{Z}[x_1, \ldots, x_k]$, one defines the traditional logarithmic Mahler measure by

$$m(F) = \int_0^1 \cdots \int_0^1 \log |F(e^{2\pi i x_1}, \ldots, e^{2\pi i x_k})| \, dx_1 \cdots dx_k.$$

Lind [2005] viewed $[0, 1]^k$ as the group $(\mathbb{R}/\mathbb{Z})^k$ and generalized the Mahler measure to arbitrary compact abelian groups. In particular, for the finite abelian group

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k} \tag{1}$$

and $F \in \mathbb{Z}[x_1, \ldots, x_k]$, we define the *logarithmic Lind–Mahler measure* by

$$m_G(F) = \frac{1}{|G|} \sum_{x_1=1}^{n_1} \cdots \sum_{x_k=1}^{n_k} \log |F(e^{2\pi i x_1/n_1}, \ldots, e^{2\pi i x_k/n_k})|.$$

Writing

$$w_n := e^{2\pi i/n},$$

we plainly have

$$m_G(F) = \frac{1}{|G|} \log |M_G(F)|,$$

This work was supported in part by a grant from the Simons Foundation (#426694 to Mossinghoff).
MSC2010: primary 11R06; secondary 11B83, 11C08, 11G50, 11T22, 43A40.
Keywords: Lind–Lehmer constant, Mahler measure, group determinant.

where

$$M_G(F) := \prod_{j_1=1}^{n_1} \cdots \prod_{j_k=1}^{n_k} F(w_{n_1}^{j_1}, \ldots, w_{n_k}^{j_k}) \in \mathbb{Z}.$$

The close connection of the Lind–Mahler measure to the group determinant was explored by Vipismakul [2013]. Recall that for a finite group $G = \{g_1, \ldots, g_N\}$ one assigns a variable x_g for each g in G and defines the group determinant, $\mathscr{D}_G(x_{g_1}, \ldots, x_{g_N})$, to be the determinant of the $N \times N$ matrix whose ij-th entry is $x_{g_i g_j^{-1}}$, a homogeneous polynomial of degree N in the x_g. From Dedekind's factorization of the group determinant of an abelian group in terms of the group characters [Dedekind 1968, pp. 420–421] (see also [Lang 1978, p. 89], or the historical survey [Conrad 1998]), it is readily seen that for a group of the form (1) we have

$$\mathscr{D}_G(a_{g_1}, \ldots, a_{g_N}) = M_G(F), \quad F(x_1, \ldots, x_k) := \sum_{g=(m_1,\ldots,m_k)\in G} a_g\, x_1^{m_1} \cdots x_k^{m_k}. \tag{2}$$

Analogous to the classical Lehmer problem [1933], we can ask for the minimal positive $m_G(F)$, and to this end we define the *Lind–Lehmer constant* for G by

$$\lambda(G) := \min\{|M_G(F)| > 1 : F \in \mathbb{Z}[x_1, \ldots, x_k]\}.$$

We use $|M_G(F)|$ rather than $m_G(F)$ or $|M_G(F)|^{1/|G|}$ so that we are dealing with integers; of course the minimal positive logarithmic measure will be $(1/|G|)\log\lambda(G)$. As Lind observed, for $|G| \geq 3$ we always have the trivial bound

$$\lambda(G) \leq |G| - 1, \tag{3}$$

achieved, for example, by

$$F(x_1, \ldots, x_k) = -1 + \prod_{i=1}^{k} \left(\frac{{x_i}^{n_i} - 1}{x_i - 1} \right).$$

Lind also showed that for prime powers p^α with $\alpha \geq 1$ we have

$$\lambda(\mathbb{Z}_{p^\alpha}) = \begin{cases} 3 & \text{if } p = 2, \\ 2 & \text{if } p \geq 3, \end{cases} \tag{4}$$

achieved with $x^2 + x + 1$ if $p = 2$ and $x + 1$ if $p \geq 3$. Lind's results for cyclic groups were extended by Kaiblinger [2010] and Pigno and Pinner [2014] so that $\lambda(\mathbb{Z}_m)$ is now known if $892\,371\,480 \nmid m$. The value for the p-group $\mathbb{Z}_p^k$ was recently established by De Silva and Pinner [2014], but little is known for direct products involving at least one term $\mathbb{Z}_{p^\alpha}$ with $\alpha \geq 2$.

Here we are principally interested in the case of 2-groups

$$G = \mathbb{Z}_{2^{\alpha_1}} \times \cdots \times \mathbb{Z}_{2^{\alpha_k}}. \tag{5}$$

It was shown in [DeSilva and Pinner 2014] that for all $k \geq 2$

$$\lambda(\mathbb{Z}_2^k) = 2^k - 1, \tag{6}$$

a case of equality in (3). We establish two main results regarding the Lind–Lehmer constant for groups of the form (5). First, we prove that equality occurs in (3) whenever G is a 2-group whose factors are all $\mathbb{Z}_2$ or $\mathbb{Z}_4$.

Theorem 1.1. *If $G = \mathbb{Z}_2^r$ or $\mathbb{Z}_4^s$ or $\mathbb{Z}_2^r \times \mathbb{Z}_4^s$, then*

$$\lambda(G) = \max\{3, |G| - 1\}.$$

Second, we show that this is not true for all 2-groups: if we allow $\alpha_i \geq 3$ in (5) then (3) need not be sharp.

Theorem 1.2. *For $n \geq 3$*

$$\lambda(\mathbb{Z}_2 \times \mathbb{Z}_{2^n}) = 9,$$

achieved with $F(x, y) = y^2 + y + 1$.

Crucial to our proofs of these statements will be a congruence satisfied by $M_G(F)$ when G is an abelian p-group. This generalizes a result [DeSilva and Pinner 2014, Lemma 2.1] for groups of the form $G = \mathbb{Z}_p^k$; see also [Vipismakul 2013, Theorem 2.1.2].

Lemma 1.3. *If p is a prime and*

$$G = \mathbb{Z}_{p^{\alpha_1}} \times \cdots \times \mathbb{Z}_{p^{\alpha_k}}, \tag{7}$$

then

$$M_G(F) \equiv F(1, \ldots, 1)^{|G|} \bmod p^k.$$

By the correspondence (2) this gives us a congruence satisfied by the group determinant, when the variables are integers and G is of the form (7),

$$\mathscr{D}_G(a_{g_1}, \ldots, a_{g_N}) \equiv \left(\sum_{g \in G} a_g\right)^N \bmod p^k.$$

Notice that for the p-group (7) we have

$$M_G(F) = \prod_{t_1=0}^{\alpha_1} \cdots \prod_{t_k=0}^{\alpha_k} N_{t_1,\ldots,t_k}(F),$$

where

$$N_{t_1,\ldots,t_k}(F) = \prod_{\substack{j_1=1 \\ (j_1, p^{\alpha_1}) = p^{t_1}}}^{p^{\alpha_1}} \cdots \prod_{\substack{j_k=1 \\ (j_k, p^{\alpha_k}) = p^{t_k}}}^{p^{\alpha_k}} F(w_{p^{\alpha_1}}^{j_1}, \ldots, w_{p^{\alpha_k}}^{j_k}) \in \mathbb{Z}.$$

Since $|1 - w_{p^\alpha}^j|_p < 1$ and the $N_{t_1,\ldots,t_k}(F)$ are integers, we have

$$N_{t_1,\ldots,t_k}(F) \equiv F(1, \ldots, 1)^{\varphi(p^{\alpha_1 - t_1}) \cdots \varphi(p^{\alpha_k - t_k})} \bmod p. \tag{8}$$

In particular if $p \mid F(1, \ldots, 1)$ we have $p \mid N_{t_1,\ldots,t_k}(F)$ for all t_i and $|G| p^k \mid M_G(F)$. So, in view of (3), we can assume for the p-group (7) that $p \nmid F(1, \ldots, 1)$ for any F achieving $\lambda(G)$.

Thus, in the case of 2-groups we can assume an F with minimal measure has $F(1,\dots,1)$ odd, and by Lemma 1.3 we see that

$$M_G(F) \equiv 1 \bmod 2^k. \tag{9}$$

Note this immediately produces (6).

Similarly for 3-groups we can assume that an F with minimal measure has $3 \nmid F(1,\dots,1)$ and $M_G(F) \equiv \pm 1 \bmod 3^k$. This produces another case of equality in (3):

$$\lambda(\mathbb{Z}_3^k) = 3^k - 1,$$

as observed in [DeSilva and Pinner 2014]. For $G = \mathbb{Z}_3 \times \mathbb{Z}_{3^n}$, we have $M_G(F) \equiv \pm 1 \bmod 9$ and so we immediately obtain the minimal measure for an additional family of 3-groups.

Theorem 1.4. *For $n \ge 1$*

$$\lambda(\mathbb{Z}_3 \times \mathbb{Z}_{3^n}) = 8,$$

achieved with $F(x, y) = y + 1$.

Section 2 of this article is devoted to the proof of Lemma 1.3, Section 3 establishes Theorem 1.1, and Section 4 proves Theorem 1.2.

2. Proof of Lemma 1.3

We proceed by induction on $\alpha_1 + \cdots + \alpha_k$. If $G = \mathbb{Z}_p$ then, as in (8), we can just use that $|w_p - 1|_p = p^{-1/(p-1)} < 1$; since $M_G(F) \in \mathbb{Z}$ and $M_G(F) \equiv F(1)^p \bmod (1 - w_p)$ we see that $M_G(F) \equiv F(1)^p \bmod p$.

Set

$$g(x_1,\dots,x_k) = \prod_{l_1=1}^{p^{\alpha_1}} \cdots \prod_{l_k=1}^{p^{\alpha_k}} F(x_1^{l_1},\dots,x_k^{l_k})$$

and let I be the ideal in $\mathbb{Z}[x_1,\dots,x_n]$ generated by $x_1^{p^{\alpha_1}} - 1,\dots,x_k^{p^{\alpha_k}} - 1$. Expanding, we have

$$g(x_1,\dots,x_k) = \sum_{0 \le \ell_1 < p^{\alpha_1}} \cdots \sum_{0 \le \ell_k < p^{\alpha_k}} a(\ell_1,\dots,\ell_k)\, x_1^{\ell_1} \cdots x_k^{\ell_k} \bmod I.$$

We set

$$\begin{aligned} S &:= \sum_{j_1=1}^{p^{\alpha_1}} \cdots \sum_{j_k=1}^{p^{\alpha_k}} g(w_{p^{\alpha_1}}^{j_1},\dots,w_{p^{\alpha_k}}^{j_k}) \\ &= \sum_{0 \le \ell_1 < p^{\alpha_1}} \cdots \sum_{0 \le \ell_k < p^{\alpha_k}} a(\ell_1,\dots,\ell_k) \sum_{j_1=1}^{p^{\alpha_1}} \cdots \sum_{j_k=1}^{p^{\alpha_k}} w_{p^{\alpha_1}}^{j_1\ell_1} \cdots w_{p^{\alpha_k}}^{j_k\ell_k} \\ &= a(0,\dots,0)\, p^{\alpha_1+\cdots+\alpha_k}. \end{aligned}$$

If $(j_1, p^{\alpha_1}) = \cdots = (j_k, p^{\alpha_k}) = 1$, then for these $\varphi(p^{\alpha_1}) \cdots \varphi(p^{\alpha_k})$ values we have

$$g(w_{p^{\alpha_1}}^{j_1},\dots,w_{p^{\alpha_k}}^{j_k}) = M_G(F).$$

Suppose that $(j_1, p^{\alpha_1}) = p^{t_1}, \ldots, (j_k, p^{\alpha_k}) = p^{t_k}$ with at least one $t_j \neq 0$, and with $t_i = \alpha_i$ for exactly $L \geq 0$ of the t_i. Suppose without loss of generality that $t_i = \alpha_i$ for any $i = 1, \ldots, L$ and $t_i < \alpha_i$ for any $i = L+1, \ldots, k$. For these $\varphi(p^{\alpha_{L+1}-t_{L+1}}) \cdots \varphi(p^{\alpha_k - t_k})$ values, applying the induction hypothesis to $G' = \mathbb{Z}_{p^{\alpha_{L+1}-t_{L+1}}} \times \cdots \times \mathbb{Z}_{p^{\alpha_k - t_k}}$, we have

$$\begin{aligned} g(w_{p^{\alpha_1}}^{j_1}, \ldots, w_{p^{\alpha_k}}^{j_k}) &= M_{G'}(F(1, \ldots, 1, x_{L+1}, \ldots, x_k))^{p^{t_1+\cdots+t_k}} \\ &= (F(1, \ldots, 1)^{p^{(\alpha_{L+1}-t_{L+1})+\cdots+(\alpha_k-t_k)}} + hp^{k-L})^{p^{t_1+\cdots+t_k}} \\ &\equiv F(1, \ldots, 1)^{|G|} \bmod p^{k-L+\alpha_1+\cdots+\alpha_L+t_{L+1}+\cdots+t_k}. \end{aligned}$$

Hence these $(p-1)^{k-L} p^{(\alpha_{L+1}-t_{L+1}-1)+\cdots+(\alpha_k-t_k-1)}$ terms contribute

$$\varphi(p^{\alpha_{L+1}-t_{L+1}}) \cdots \varphi(p^{\alpha_k-t_k}) F(1, \ldots, 1)^{|G|} \bmod p^{\alpha_1+\cdots+\alpha_k}$$

to S. Thus

$$\begin{aligned} 0 &\equiv \varphi(p^{\alpha_1}) \cdots \varphi(p^{\alpha_k}) M_G(F) + (p^{\alpha_1+\cdots+\alpha_k} - \varphi(p^{\alpha_1}) \cdots \varphi(p^{\alpha_k})) F(1, \ldots, 1)^{|G|} \\ &\equiv (p-1)^k p^{\alpha_1+\cdots+\alpha_k-k} (M_G(F) - F(1, \ldots, 1)^{|G|}) \bmod p^{\alpha_1+\cdots+\alpha_k} \end{aligned}$$

and the statement follows. □

3. Proof of Theorem 1.1

To prove Theorem 1.1, we require the following lemma.

Lemma 3.1. *Suppose that $F \in \mathbb{Z}[x_1, \ldots, x_n]$, and let I denote the ideal of $\mathbb{Z}[x_1, \ldots, x_n]$ generated by $x_1^{n_1} - 1, \ldots, x_k^{n_k} - 1$. Then $F(w_{n_1}^{j_1}, \ldots, w_{n_k}^{j_k}) = 0$ for all $1 \leq j_i \leq n_i$ if and only if $F \in I$.*

Proof. Plainly any F in I will have $F(w_{n_1}^{j_1}, \ldots, w_{n_k}^{j_k}) = 0$ for all $0 \leq j_i < n_i$. Conversely, suppose that $F(w_{n_1}^{j_1}, \ldots, w_{n_k}^{j_k}) = 0$ for all $0 \leq j_i < n_i$. Clearly any F can be reduced mod I to a polynomial of degree less than n_i in each x_i:

$$F(x_1, \ldots, x_k) = \sum_{t_1=0}^{n_1-1} \cdots \sum_{t_k=0}^{n_k-1} a(t_1, \ldots, t_k)\, x_1^{t_1} \cdots x_k^{t_k} \bmod I.$$

Since $\sum_{j_i=0}^{n_i-1} w_{n_i}^{(t_i-T_i)j_i} = 0$ if $t_i \not\equiv T_i \bmod n_i$ (and n_i otherwise) we have

$$a(T_1, \ldots, T_k) = \frac{1}{n_1 \cdots n_k} \sum_{j_1=0}^{n_1-1} \cdots \sum_{j_k=0}^{n_k-1} F(w_{n_1}^{j_1}, \ldots, w_{n_k}^{j_k})\, w_{n_1}^{-T_1 j_1} \cdots w_{n_k}^{-T_k j_k}.$$

So $a(T_1, \ldots, T_k) = 0$ for all $0 \leq T_i < n_i$ and $F = 0 \bmod I$. □

We now proceed to the proof of our first principal result.

Proof of Theorem 1.1. Suppose that $G = \mathbb{Z}_{2^{\alpha_1}} \times \cdots \times \mathbb{Z}_{2^{\alpha_k}}$ with $2^{\alpha_i} = 4$ for $1 \leq i \leq s$ and $2^{\alpha_i} = 2$ for $s+1 \leq i \leq k$. We write $r = k - s$. In view of (4) and (6) we may assume that $k \geq 2$ and $s \geq 1$. Suppose that $F(x_1, \ldots, x_k)$ has

$$1 < |M_G(F)| < |G| - 1 = 2^{k+s} - 1,$$

where

$$M_G(F) = \prod_{\substack{u_1,\dots,u_s=\pm 1,\pm i\\ u_{s+1},\dots,u_k=\pm 1}} F(u_1,\dots,u_k). \tag{10}$$

Suppose that one of the nonunits $F(u_1,\dots,u_k)$ in the product (10) has at least one of its u_j complex, say $u_1 = \pm i$, and set $G' = \mathbb{Z}_{2^{\alpha_2}} \times \cdots \times \mathbb{Z}_{2^{\alpha_k}}$. Plainly we may write

$$M_G(F) = AB,$$

with

$$A := M_{\mathbb{Z}_2 \times G'}(F), \quad B := M_{G'}(F(i,x_2,\dots,x_k)F(-i,x_2,\dots,x_k)).$$

From (9) we know that $M_G(F)$ and A, and hence B, are all congruent to 1 mod 2^k. Also B will be of the form $|a+ib|^2$ and hence cannot be negative. Since it contains a nonunit we have $B > 1$; hence $B \ge 2^k + 1$. If $A \ne 1$ then $|A| \ge 2^k - 1$ and $|M_G(F)| \ge (2^k-1)(2^k+1) = 4^k - 1 \ge |G| - 1$, so we must have $A = 1$. Thus if $F(u_1,u_2,\dots,u_k)$ is a nonunit with $u_j = \pm i$, we may assume that the $F(u_1,\dots,u_k)$ with $u_j = \pm 1$ are units. We have two possibilities for the $F(u_1,\dots,u_k)$ in the product (10):

(a) There is at least one nonunit $F(u_1,\dots,u_k)$ with some $u_j = \pm i$.

(b) $F(u_1,\dots,u_k)$ is a unit whenever any of the $u_j = \pm i$.

With I denoting the ideal generated by the $x_j^{2^{\alpha_j}} - 1$, and splitting the x_1-dependence into even and odd exponents $p(x_1) = \alpha(x_1^2) + x_1\beta(x_1^2)$, we can write

$$F(x_1,\dots,x_k) = \sum_{\substack{0\le \varepsilon_2,\dots,\varepsilon_s \le 3,\\ 0\le \varepsilon_1,\varepsilon_{s+1},\dots,\varepsilon_k \le 1}} a(\varepsilon_1,\varepsilon_2,\dots,\varepsilon_k)(x_1^2)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2}\cdots x_k^{\varepsilon_k} \bmod I.$$

Since $F(1,\dots,1) = \sum a(\varepsilon_1,\varepsilon_2,\dots,\varepsilon_k)(1)$ is odd, we know that at least one of the $a(\varepsilon_1,\varepsilon_2,\dots,\varepsilon_k)(1)$ is odd. Replacing F by $x_1^{\delta_1}\cdots x_n^{\delta_n} F$ with $0 \le \delta_1, \delta_{s+1},\dots,\delta_k \le 1$ and $0 \le \delta_2,\dots,\delta_s \le 3$, and reducing mod I, to reshuffle the $a(\varepsilon_1,\dots,\varepsilon_k)(x_1^2)$, and replacing F by $-F$ as necessary, we can assume that

$$F(1,\dots,1) \equiv 1 \bmod 4, \quad a(0,\dots,0)(1) \text{ is odd.} \tag{11}$$

Case (a): Suppose we have nonunits in the product (10) with complex u_j. Reordering and taking $x_j \mapsto \pm x_1 x_j$ for $2 \le j \le s$ and $x_j \mapsto \pm x_j$ for $s < j \le k$ as necessary, we assume that the first of these is $\gamma_1 = F(i,1,\dots,1)$. If (after the transformations) there are other nonunits with complex entries in positions other than the first, by reordering and substituting x_j with $x_j x_2$ as necessary for $j \ge 3$, we may assume that $\gamma_2 = F(\pm i, i, \pm 1,\dots,\pm 1)$. If there are still nonunits with $u_j = \pm i$, $j \ge 3$, then, after reordering and substitutions, we have a nonunit $\gamma_3 = F(\pm i, \pm i, i, \pm 1,\dots,\pm 1)$. We repeat this, h times say, until there are no new nonunits with a complex u_j, $j > h$. That is, for some $1 \le h \le s$, we have h nonunits $\gamma_j = F(a_{j1},\dots,a_{jk})$ with $a_{jj} = i$, $a_{j\ell} = \pm i$ for $1 \le \ell < j$ and $a_{j\ell} = \pm 1$ for $h < \ell \le k$, and $F(u_1,\dots,u_k)$ is a unit whenever $u_\ell = \pm i$ with $h < \ell \le s$ if $h < s$. Adjusting as above we can assume that (11) holds.

Since the $F(\pm 1, u_2, \dots, u_k)$ are all units, with $F(1, \dots, 1) = 1$, and

$$a(0, \dots, 0)(1) = \frac{2}{|G|} \sum_{\substack{u_2, \dots, u_s = \pm i, \pm 1 \\ u_1, u_{s+1}, \dots, u_k = \pm 1}} F(u_1, \dots, u_k)$$

is odd, plainly the $F(\pm 1, u_2, \dots, u_k)$ must all be 1. Applying Lemma 3.1, we may therefore assume that

$$F(x_1, \dots, x_k) = 1 + (x_1^2 - 1) \sum_{\substack{0 \le \varepsilon_2, \dots, \varepsilon_s \le 3, \\ 0 \le \varepsilon_1, \varepsilon_{s+1}, \dots, \varepsilon_k \le 1}} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

Notice that the $F(\pm i, u_2, \dots, u_k) \in \mathbb{Z}[i]$ will all have odd real part and even imaginary part. Moreover, writing $\pi = (1 - i)$, where $\pi^2 \mid 2$, we have $u_j \equiv 1 \bmod \pi$ for any $u_j = \pm 1$ or $\pm i$, and the $F(\pm i, u_2, \dots, u_k)$ must all be congruent mod π^3 in $\mathbb{Z}[i]$. Since $|\pi|_2 = 2^{-1/2}$, plainly two units $\pm 1, \pm i$ in $\mathbb{Z}[i]$ cannot be congruent mod π^3 unless they are equal. If $h \ge 2$ then we know that the $F(\pm i, \pm 1, u_3, \dots, u_k)$ will all be units and so must be all 1 or all -1. Replacing F by $x_1^2 F$ we can assume that they are all 1. Applying Lemma 3.1 we get

$$F(x_1, \dots, x_k) = 1 + (x_1^2 - 1)(x_2^2 - 1) \sum_{\substack{0 \le \varepsilon_3, \dots, \varepsilon_s \le 3, \\ 0 \le \varepsilon_1, \varepsilon_2, \varepsilon_{s+1}, \dots, \varepsilon_k \le 1}} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

Likewise, if $h \ge 3$ we have that $F(\pm i, \pm i, \pm 1, u_4, \dots, u_k)$ are all units and congruent to 1 mod 4, so these must all equal 1. Applying the lemma and repeating up to $F(\pm i, \dots, \pm i, \pm 1, u_{h+1}, \dots, u_k)$, we deduce that

$$F(x_1, \dots, x_k) = 1 + \prod_{j=1}^{h} (x_j^2 - 1) \sum_{\substack{0 \le \varepsilon_{h+1}, \dots, \varepsilon_s \le 3, \\ 0 \le \varepsilon_1, \dots, \varepsilon_h, \varepsilon_{s+1}, \dots, \varepsilon_k \le 1}} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

If $s > h$, we further have that the $F(\pm i, \dots, \pm i, u_{h+2}, \dots, u_k)$ are all units. If $h \ge 2$ they will all be congruent to 1 mod 4 and so must all equal 1. If $h = 1$ then they are all 1 or all -1 and, by replacing F by $x_1^2 F$ if necessary, we may assume they are all 1. Writing

$$F(x_1, \dots, x_k) = 1 + \prod_{j=1}^{h} (x_j^2 - 1)\big(f(x_{h+2}, \dots, x_k) + x_{h+1} g(x_{h+2}, \dots, x_k) \bmod (x_{h+1}^2 + 1)\big),$$

separating into real and imaginary parts and applying Lemma 3.1 to f and g, we get that $f, g = 0 \bmod I$. Repeating for each variable, we find that

$$F(x_1, \dots, x_k) = 1 + \prod_{j=1}^{h} (x_j^2 - 1) \prod_{j=h+1}^{s} (x_j^2 + 1) \sum_{0 \le \varepsilon_1, \dots, \varepsilon_k \le 1} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

Suppose that there are $t \ge 1$ conjugate pairs of nonunits $F(a_{j1}, \dots, a_{jk}) = \gamma_j$. Then plainly

$$\gamma_j = a_j + i b_j, \qquad a_j \equiv 1 \bmod 2^s, \quad b_j \equiv 0 \bmod 2^s. \tag{12}$$

Trivially we have $|\gamma_j|^2 \geq 5$, and if $t \geq r+s$ then

$$|M_G(F)| \geq 5^t \geq 5^r \cdot 5^s > 2^r \cdot 4^s - 1,$$

so we can assume that

$$t \leq r+s-1. \tag{13}$$

If $t \leq r$ then, using $x_i \mapsto -x_i$ as necessary for γ_1, and for the subsequent γ_j reordering and using the transformation $x_\ell \mapsto x_\ell x_j$ if $u_j = -1$ to remove any $u_\ell = -1$ with $\ell > j$, we can assume that the r-tuples $(u_{s+1}, \ldots, u_k)$ achieving the γ_j take the form

$$(1,\ldots,1), \quad (\pm 1,1,\ldots,1), \quad (\pm 1,\pm 1,1,\ldots,1), \quad \ldots, \quad (\overbrace{\pm 1,\ldots,\pm 1}^{t-1},1,\ldots,1)$$

(here we are focusing on the u_j with $j > s$, which recall are taking the values ± 1). In particular, $F(u_1,\ldots,u_k)$ will be a unit if $u_j = -1$ for any $s+t \leq j \leq k$. (If $s \geq 2$, the units will all be 1; if $s = 1$ we may need to take $x_1^2 F$ to make the value when $u_{s+t} = -1$ and hence the rest equal 1.) Successively applying the lemma again, we find

$$F(x_1,\ldots,x_k) = 1 + \prod_{j=1}^{h}(x_j^2-1) \prod_{j=h+1}^{s}(x_j^2+1) \prod_{j=s+t}^{k}(x_j+1)R$$

with

$$R = \sum_{0 \leq \varepsilon_1,\ldots,\varepsilon_{s+t-1} \leq 1} a(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_{s+t-1})\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_{s+t-1}^{\varepsilon_{s+t-1}}.$$

Hence we obtain

$$\gamma_j = a_j + i b_j, \qquad a_j \equiv 1 \bmod 2^{s+r+1-t}, \quad b_j \equiv 0 \bmod 2^{s+r+1-t}.$$

From (13) and (12) this is plainly also valid if $t > r$. Thus, we have

$$|M_G(F)| = |\gamma_1| \cdots |\gamma_t| \geq (2^{r+s+1-t}-1)^{2t} > 2^{2t(r+s+1/2-t)} \geq 2^{2(r+s-1/2)} \geq 2^{r+2s}$$

for $r \geq 1$. If $r = 0$ and $t \geq 2$ then we have $s \geq 2$, and from (12) we obtain

$$|M_G(F)| \geq (2^s-1)^{2t} > 2^{2t(s-1/2)} \geq 2^{4s-2} > 4^s.$$

Finally if $t = 1$ and $r = 0$ then, since $F(i,1,\ldots 1)$ and its conjugate are the only nonunits, we know that $F(\pm i,-1,u_3,\ldots,u_k)$ are all units and so equal 1. Hence we can add an extra factor (x_2+1) to get

$$|M_G(F)| \geq (2^{s+1}-1)^2 > 2^{2s}.$$

Case (b): Since $a(0,\ldots,0)(1)$ is odd, we know that $a(0,\ldots,0)(-1)$ is odd. Since the $F(\pm i,u_2,\ldots,u_k)$ are all units and

$$a(0,\ldots,0)(-1) = \frac{1}{|G|/2} \sum_{\substack{u_1=\pm i \\ u_2,\ldots,u_s=\pm i,\pm 1 \\ u_{s+1},\ldots,u_k=\pm 1}} F(u_1,\ldots,u_k)$$

is odd, plainly the $F(\pm i, u_2, \dots, u_k)$ must all be 1 or all be -1. Replacing F by $x_1^2 F$ we assume $F(\pm i, u_2, \dots, u_k) = 1$. Applying Lemma 3.1 to the real and imaginary parts we can assume that

$$F(x_1, \dots, x_k) = 1 + (x_1^2 + 1) \sum_{\substack{0 \le \varepsilon_2, \dots, \varepsilon_s \le 3, \\ 0 \le \varepsilon_1, \varepsilon_{s+1}, \dots, \varepsilon_k \le 1}} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

Notice that all the $F(\pm 1, u_2, \dots, u_k)$ satisfy $F(\pm 1, u_2, \dots, u_k) \equiv F(1, \dots, 1) \equiv 1 \bmod \pi^3$. Hence if $s > 1$, the units $F(\pm 1, \pm i, u_3, \dots, u_k)$ are all 1. Applying the lemma and repeating we obtain

$$F(x_1, \dots, x_k) = 1 + \prod_{j=1}^{s} (x_j^2 + 1) \sum_{0 \le \varepsilon_1, \dots, \varepsilon_k \le 1} a(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} x_2^{\varepsilon_2} \cdots x_k^{\varepsilon_k}.$$

Hence we have

$$M_G(F) = M_{\mathbb{Z}_2^k}(f),$$

where

$$f(x_1, \dots, x_k) = 1 + 2^s \sum_{0 \le \varepsilon_1, \dots, \varepsilon_k \le 1} A(\varepsilon_1, \dots, \varepsilon_k)\, x_1^{\varepsilon_1} \cdots x_k^{\varepsilon_k}.$$

Suppose that there are t elements $f(\pm 1, \dots, \pm 1)$ that are not ± 1. If $t \ge k + s - 1$ then plainly $|M_G(F)| \ge 3^t \ge 3^{k+s-1} > 2^{k+s}$ since $k + s \ge 3$, so we assume that $t \le k + s - 2$. Sending $x_j \mapsto -x_j$ we assume that one of them is $f(1, \dots, 1) = \gamma_1$. If $t > 1$ then, reordering and mapping x_ℓ to $x_\ell x_j$ if we have $\ell > j$ with $u_\ell = u_j = -1$, we can assume that the remaining values are $\gamma_2 = f(-1, 1, \dots, 1)$, $\gamma_3 = f(a_{31}, a_{32}, 1, \dots, 1)$, $\dots$, $\gamma_t = f(a_{t1}, \dots a_{t(t-1)}, 1, \dots, 1)$. If $t \le k$ then we will have $f(u_1, \dots, u_k) = 1$ whenever $u_j = -1$ for some $t \le j \le k$, and applying the lemma we find

$$f(x_1, \dots, x_k) = 1 + 2^s \prod_{j=t}^{k} (x_j + 1) \sum_{0 \le \varepsilon_1, \dots, \varepsilon_{t-1} \le 1} A(\varepsilon_1, \dots, \varepsilon_{t-1})\, x_1^{\varepsilon_1} \cdots x_{t-1}^{\varepsilon_{t-1}}.$$

Thus

$$\gamma_j \equiv 1 \bmod 2^{s+k-t+1}$$

(with this trivially holding if $k \le t - 1$), and

$$|M_G(F)| \ge (2^{s+k+1-t} - 1)^t.$$

For $t = 1$ this gives

$$|M_G(F)| \ge 2^{s+k} - 1 = |G| - 1,$$

and for $t \ge 2$

$$|M_G(F)| \ge 2^{t(s+k+1/2-t)} \ge 2^{2s+2k-3} \ge 2^{s+k}. \qquad \square$$

4. Proof of Theorem 1.2

Using $\Phi_j(x)$ to denote the j-th cyclotomic polynomial and recalling, see [Apostol 1970; Lehmer 1930], that for $j > k$ the resultant satisfies $|\mathrm{Res}(\Phi_j, \Phi_k)| = q^{\varphi(k)}$ if $j = kq^\alpha$ for some prime q and 1 otherwise,

we see that

$$M_{\mathbb{Z}_2\times\mathbb{Z}_{2^n}}(1+y+y^2) = M_{\mathbb{Z}_{2^n}}(\Phi_3(y))^2 = \left(\prod_{j=0}^{n}|\mathrm{Res}(\Phi_3,\Phi_{2^j})|\right)^2 = 9.$$

Let $G = \mathbb{Z}_2 \times \mathbb{Z}_{2^n}$. Reducing mod x^2-1, we can write our $F(x,y)$ in $\mathbb{Z}[x,y]$ in the form

$$F(x,y) = A_0(y^2) + xA_1(y^2) + yA_2(y^2) + xyA_3(y^2).$$

Plainly,

$$M_G(F(x,y)) = M_{\mathbb{Z}_{2^n}}(F(1,y))M_{\mathbb{Z}_{2^n}}(F(-1,y)),$$

where each of these measures is a product of $n+1$ integers,

$$M_{\mathbb{Z}_{2^n}}(f(y)) = \prod_{j=0}^{n} N_j(f), \quad N_j(f) := \mathrm{Res}(f,\Phi_{2^j}),$$

that is,

$$N_0(f) = f(1), \quad N_1(f) = f(-1), \quad N_2(f) = f(i)f(-i) = |f(i)|^2,$$

and, writing $w_j := e^{2\pi i/2^j}$, for any $j = 3,\dots,n$, we have

$$N_j(f) = \prod_{\substack{k=1\\ k \text{ odd}}}^{2^j} f(w_j^k) = \prod_{\substack{k=1\\ k \text{ odd}}}^{2^{j-1}} f(w_j^k)f(-w_j^k) = |R_j(f)|^2,$$

where

$$R_j(f) := \prod_{k\equiv 1 \bmod 4}^{2^{j-1}} f(w_j^k)f(-w_j^k) \in \mathbb{Z}[i], \quad 3 \le j \le n.$$

Note $N_j(f)$ and $R_j(f)$ represent the norms of $f(w_j^k)$ from $\mathbb{Q}(w_j)$ to $\mathbb{Q}$ and $\mathbb{Q}(i)$ respectively, and since they are algebraic integers they will be in $\mathbb{Z}$ and $\mathbb{Z}[i]$, respectively.

Since $|1-w_j|_2 = 2^{-1/\varphi(2^j)}$, for all $j = 3,\dots,n$ we have $N_j(F(\pm 1,y)) \equiv F(1,1)^{2^{j-1}} \bmod 2$, and if $M_G(F) < 2^{2n+2}$ we can assume $F(1,1)$ and all the $N_j(F(\pm 1,y))$ are odd. Note that for all the $j \ge 2$ we have

$$N_j(F(\pm 1,y)) = |a+ib|^2 = a^2+b^2 \equiv 1 \bmod 4.$$

If $|M_G(F)| < 9$ then $|M_{\mathbb{Z}_{2^n}}(F(1,y))|$ or $|M_{\mathbb{Z}_{2^n}}(F(-1,y))|$ must be 1. Replacing x with $-x$ as necessary we assume that

$$1 < |M_{\mathbb{Z}_{2^n}}(F(1,y))| < 9, \quad |M_{\mathbb{Z}_{2^n}}(F(-1,y))| = 1.$$

Since

$$F(1,1) = A_0(1) + A_1(1) + A_2(1) + A_3(1)$$

is odd, we can assume that at least one of the $A_i(1)$ is odd. Replacing F by xF or yF or xyF and reducing by x^2-1 as necessary, we may assume that $A_0(1)$ is odd. Replacing y by $-y$ and F by $-F$ as necessary, we may further assume that $|F(1,1)| \ge |F(1,-1)|$ and $F(1,1) > 0$.

Since

$$F(1,-1) = A_0(1) + A_1(1) - A_2(1) - A_3(1),$$
$$F(-1,1) = A_0(1) - A_1(1) + A_2(1) - A_3(1),$$
$$F(-1,-1) = A_0(1) - A_1(1) - A_2(1) + A_3(1),$$

we have

$$A_0(1) = \tfrac{1}{4}\big(F(1,1) + F(1,-1) + F(-1,1) + F(-1,-1)\big),$$
$$A_1(1) = \tfrac{1}{4}\big(F(1,1) + F(1,-1) - F(-1,1) - F(-1,-1)\big),$$
$$A_2(1) = \tfrac{1}{4}\big(F(1,1) - F(1,-1) + F(-1,1) - F(-1,-1)\big),$$
$$A_3(1) = \tfrac{1}{4}\big(F(1,1) - F(1,-1) - F(-1,1) + F(-1,-1)\big).$$

Observe that

$$F(1,w_j^k)F(1,-w_j^k) = \big(A_0(w_j^{2k}) + A_1(w_j^{2k})\big)^2 - w_j^{2k}\big(A_2(w_j^{2k}) + A_3(w_j^{2k})\big)^2$$

and

$$F(-1,w_j^k)F(-1,-w_j^k) = \big(A_0(w_j^{2k}) - A_1(w_j^{2k})\big)^2 - w_j^{2k}\big(A_2(w_j^{2k}) - A_3(w_j^{2k})\big)^2$$

differ by

$$4\big(A_0(w_j^{2k})A_1(w_j^{2k}) - w_j^{2k}A_2(w_j^{2k})A_3(w_j^{2k})\big) \in 4\mathbb{Z}[w_{j-1}].$$

Hence $R_j(F(1,y))$ and $R_j(F(-1,y))$ differ by an element of $4\mathbb{Z}[w_{j-1}]$ and, since both are in $\mathbb{Z}[i]$, we conclude that

$$R_j(F(1,y)) - R_j(F(-1,y)) \in 4\mathbb{Z}[i].$$

Since $N_j(F(-1,y)) = 1$, we have $R_j(F(-1,y)) = \pm 1$ or $\pm i$, and either $R_j(F(1,y)) = R_j(F(-1,y))$ and $N_j(F(1,y)) = 1$, or $N_j(F(1,y)) \geq (4-1)^2 = 9$.

Thus if $|M_G(F)| < 9$ then we must have $N_j(F(1,y)) = N_j(F(-1,y)) = 1$ for $j = 3,\dots,n$ and $M_G(F) = M_{\mathbb{Z}_2\times\mathbb{Z}_4}(F)$. By Theorem 1.1 and Lemma 1.3, we have $|M_{\mathbb{Z}_2\times\mathbb{Z}_4}(F)| \geq 7$ and $M_{\mathbb{Z}_2\times\mathbb{Z}_4}(F) \equiv 1 \bmod 4$, and so

$$M_G(F) = M_{\mathbb{Z}_2\times\mathbb{Z}_4}(F) = -7.$$

Since $N_j(f) \equiv 1 \bmod 4$ for $j \geq 2$ we must have $|F(1,1)F(1,-1)| = 7$ and $N_2(F(1,y)) = 1$ and

$$F(1,1) = 7, \quad F(1,-1), \quad F(-1,\pm 1) = \pm 1, \quad F(\pm 1,\pm i) = \pm 1 \text{ or } \pm i,$$

with $R_j(F(1,y)) = R_j(F(-1,y)) = \pm 1$ or $\pm i$ for $j = 3,\dots,n$.

We have

$$A_0(1) = \tfrac{1}{4}\big(F(1,1) + F(1,-1) + F(-1,1) + F(-1,-1)\big) = \tfrac{1}{4}(7 \pm 1 \pm 1 \pm 1)$$

and, since $A_0(1)$ is odd, we must have $F(1,-1) = F(-1,\pm 1) = -1$ and $A_0(1) = 1$ and $A_1(1) = A_2(1) = A_3(1) = 2$. Hence

$$F(x,y) = 1 + 2x + 2y + 2xy + (y^2-1)\big(B_0(y^2) + xB_1(y^2) + yB_2(y^2) + xyB_3(y^2)\big).$$

Thus

$$F(1,i) = 3 + 4i - 2\big(B_0(-1) + B_1(-1) + iB_2(-1) + iB_3(-1)\big),$$
$$F(-1,i) = -1 - 2\big(B_0(-1) - B_1(-1) + iB_2(-1) - iB_3(-1)\big),$$

and since $F(\pm 1, i)$ are units with odd real part and difference in $4\mathbb{Z}[i]$ they must both be 1 or -1. By replacing F by y^2F as necessary, we may assume $F(\pm 1, i) = -1$. Solving, we obtain $B_0(-1) = B_1(-1) = B_2(-1) = B_3(-1) = 1$ and

$$F(x, y) = -1 + (1+x)(1+y)(1+y^2) + (y^4-1)\big(C_0(y^2) + xC_1(y^2) + yC_2(y^2) + xyC_3(y^2)\big).$$

Therefore

$$F(1, w_3)F(1, -w_3) = (1 + 2i - 2C_0(i) - 2C_1(i))^2 - 4i(1 + i - C_2(i) - C_3(i))^2$$

and

$$F(-1, w_3)F(-1, -w_3) = (-1 - 2C_0(i) + 2C_1(i))^2 - 4i(C_2(i) - C_3(i))^2.$$

Since both are units and are members of $1 + 4\mathbb{Z}[i]$, these must both equal 1. However, their difference

$$4\big((i - 2C_0(i))(1 + i - 2C_1(i)) - i(1 + i - 2C_3(i))(1 + i - 2C_2(i))\big) \in 4(1 + i + 2\mathbb{Z}[i])$$

is not zero. □

References

[Apostol 1970] T. M. Apostol, "Resultants of cyclotomic polynomials", *Proc. Amer. Math. Soc.* **24** (1970), 457–462. MR Zbl

[Conrad 1998] K. Conrad, "The origin of representation theory", *Enseign. Math.* (2) **44**:3-4 (1998), 361–392. MR Zbl

[Dedekind 1968] R. Dedekind, *Gesammelte mathematische Werke, II*, Chelsea, New York, 1968. MR

[DeSilva and Pinner 2014] D. DeSilva and C. Pinner, "The Lind Lehmer constant for $\mathbb{Z}_p^n$", *Proc. Amer. Math. Soc.* **142**:6 (2014), 1935–1941. MR Zbl

[Kaiblinger 2010] N. Kaiblinger, "On the Lehmer constant of finite cyclic groups", *Acta Arith.* **142**:1 (2010), 79–84. MR Zbl

[Lang 1978] S. Lang, *Cyclotomic fields*, Graduate Texts in Mathematics **59**, Springer, 1978. MR Zbl

[Lehmer 1930] E. T. Lehmer, "A numerical function applied to cyclotomy", *Bull. Amer. Math. Soc.* **36**:4 (1930), 291–298. MR Zbl

[Lehmer 1933] D. H. Lehmer, "Factorization of certain cyclotomic functions", *Ann. of Math.* (2) **34**:3 (1933), 461–479. MR Zbl

[Lind 2005] D. Lind, "Lehmer's problem for compact abelian groups", *Proc. Amer. Math. Soc.* **133**:5 (2005), 1411–1416. MR Zbl

[Pigno and Pinner 2014] V. Pigno and C. Pinner, "The Lind–Lehmer constant for cyclic groups of order less than 892,371,480", *Ramanujan J.* **33**:2 (2014), 295–300. MR Zbl

[Vipismakul 2013] W. Vipismakul, *The stabilizer of the group determinant and bounds for Lehmer's conjecture on finite abelian groups*, Ph.D. thesis, University of Texas at Austin, 2013, available at http://hdl.handle.net/2152/21685.

Received 21 Jun 2018.

MICHAEL J. MOSSINGHOFF:
mimossinghoff@davidson.edu
Department of Mathematics & Computer Science, Davidson College, Davidson, NC, United States

VINCENT PIGNO:
vincent.pigno@csus.edu
Department of Mathematics & Statistics, California State University, Sacramento, CA, United States

CHRISTOPHER PINNER:
pinner@math.ksu.edu
Department of Mathematics, Kansas State University, Manhattan, KS, United States

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.163

Lattices with exponentially large kissing numbers

Serge Vlăduț

We construct a sequence of lattices $\{L_{n_i} \subset \mathbb{R}^{n_i}\}$ for $n_i \to \infty$ with exponentially large kissing numbers, namely, $\log_2 \tau(L_{n_i}) > 0.0338 \cdot n_i - o(n_i)$. We also show that the maximum lattice kissing number τ_n^l in n dimensions satisfies $\log_2 \tau_n^l > 0.0219 \cdot n - o(n)$ for any n.

1. Introduction

In this paper we consider lattice packings of spheres in real n-dimensional space $\mathbb{R}^n$ and their kissing numbers. Recall that the maximum kissing number is known only in a handful of dimensions, the largest being $n = 24$ for which the Leech lattice Λ_{24} gives the optimal kissing number $\tau(\Lambda_{24}) = 196560$. Recall also that the random choice procedure guarantees, see [Chabauty 1953; Shannon 1959; Wyner 1965], the existence of nonlattice packings P_n with

$$\frac{\log_2 \tau(P_n)}{n} \geq \log_2 \frac{2}{\sqrt{3}} \simeq 0.2075\ldots.$$

More precisely, it gives the existence of local arrangements of spheres touching one sphere which can be included then into a nonlattice packing. Note also that the upper bound of Kabatiansky and Levenstein [1978] is

$$\frac{\log_2 \tau(P_n)}{n} \leq 0.4041\ldots.$$

However, for lattice packings this procedure does not work, and as far as we know, no reasonable lower bound for the maximum lattice kissing number τ_n^l is known for $n \to \infty$. For instance, the Barnes–Wall lattices BW_n with $n = 2^m$ give the quasipolynomial bound $\tau_n^l \geq n^{c \log n}$, i.e., $\log \tau_n^l \geq c \log^2 n$, which can hardly be characterized as "reasonable". The main purpose of the present paper is to give an exponential lower bound for τ_n^l (however, these lattices are worse than nonlattice packing guaranteed by random choice). This is achieved by applying Constructions D and E from [Barnes and Sloane 1983] and [Bos et al. 1982], respectively, to codes from [Ashikhmin et al. 2001] having exponentially many light vectors. In order to apply Constructions D and E we need specific good curves (the curves in the Garcia–Stichtenoth towers [1995; 1996] do not perfectly match our construction) and some Drinfeld modular curves [Gekeler 2001; Elkies 2001] perfectly suit our purposes.

Our main result is:

MSC2010: 11H31, 11H71, 14G15, 52C17.
Keywords: lattices, algebraic geometry codes, kissing numbers, Drinfeld modular curves.

Theorem 1.1. *We have*

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{20}\big(1-\tfrac{2}{31}\log 33\big) - \frac{2+2\log N}{N} \tag{1-1}$$

for $N = 5\cdot 2^{10n+2}$ *and any* $n \geq 2$,

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{24}\big(1-\tfrac{2}{63}\log 65\big) - \frac{2+2\log N}{N} \tag{1-2}$$

for $N = 3\cdot 2^{12n+3}$ *and any* $n \geq 2$,

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{28}\big(1-\tfrac{2}{127}\log 129\big) - \frac{2+2\log N}{N} \tag{1-3}$$

for $N = 7\cdot 2^{14n+2}$ *and any* $n \geq 2$, *where*

$$\tfrac{1}{20}\big(1-\tfrac{2}{31}\log 33\big)\simeq 0.033727\ldots, \quad \tfrac{1}{24}\big(1-\tfrac{2}{63}\log 65\big)\simeq 0.033700\ldots, \quad \tfrac{1}{28}\big(1-\tfrac{2}{127}\log 129\big)\simeq 0.0317709\ldots.$$

All our logarithms are binary.

Corollary 1.2. *We have*

$$\frac{\log(\tau_n^l)}{n} \geq c_0 \tag{1-4}$$

for some $c_0 > 0$ *and any* $n \geq 1$.

The exact value of c_0 is not clear, but $c_0 = 0.02$ is probably sufficient.

It is possible to ameliorate the constants slightly, if we do not insist on the effectiveness of results:

Theorem 1.3. *We have*

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{20}\big(\tfrac{21}{31} - \log\tfrac{1024}{1023}\big) - o(1) \simeq 0.033800\ldots - o(1) \tag{1-5}$$

for $N = 5\cdot 2^{10n+2}$,

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{24}\big(\tfrac{17}{21} - \log\tfrac{4096}{4095}\big) - o(1) \simeq 0.033715\ldots - o(1) \tag{1-6}$$

for $N = 3\cdot 2^{12n+3}$,

$$\frac{\log(\tau_N^l)}{N} \geq \tfrac{1}{28}\big(\tfrac{113}{127} - \log\tfrac{16384}{16383}\big) - o(1) \simeq 0.031774\ldots - o(1) \tag{1-7}$$

for $N = 7\cdot 2^{14n+2}$.

In fact, the implied functions in $o(1)$ terms can be made explicit, but they decrease slowly and their precise calculation is not justified.

Note also that using other finite fields $\mathbb{F}_q$ with a square q one can obtain infinitely many series of similar lattices in the corresponding dimensions, but for all of them the ratio $\log(\tau_N^l)/N$ is less than 0.03.

Corollary 1.4. *We have*

$$\limsup_{n\to\infty} \frac{\log(\tau_n^l)}{n} \geq \tfrac{1}{20}\big(\tfrac{21}{31} - \log\tfrac{1024}{1023}\big).$$

For the lower limit we can prove:

Theorem 1.5. *Let* $A = \log \frac{4096}{4095}$. *We have then*

$$\lim \inf_{n\to\infty} \frac{\log(\tau_n^l)}{n} \geq \tfrac{1}{504}(17 - 21A)\delta_0 \simeq 0.021937\ldots, \tag{1-8}$$

where $\delta_0 \simeq 0.6506627\ldots$ *is the unique root of the equation*

$$21H(\delta) = 2\delta(4 + 21A + (17 - 21A)\delta)$$

in the interval (0.5, 1).

One can think that c_0 in (1-4) can be chosen rather close to that value.

The rest of the paper is organized as follows: in Section 2 we recall some basic definitions and results on lattices and error-correcting codes. Section 3 is devoted to Constructions D and E from [Barnes and Sloane 1983] and [Bos et al. 1982], respectively, while Section 4 recalls and slightly modifies the constructions from [Ashikhmin et al. 2001]. We describe some known good curve families in Section 5 and prove our results in Section 6.

2. Preliminaries

In this section we recall some basic definitions and results on lattices and linear error-correcting codes.

2A. ***Lattice packings.*** A sphere packing is a configuration of nonintersecting equal open spheres in $\mathbb{R}^N$. Let d be the diameter of the spheres; then the distance between any two sphere centers is at least d. Thus a packing is a set of points P in $\mathbb{R}^N$ such that the minimum distance between any two of them is at least d. If P is an additive subgroup of $\mathbb{R}^N$, it is called a lattice or a lattice packing; below we are concerned mainly with such packings. For any packing P its density $\Delta(P)$ is defined as the fraction of space covered by spheres (which can be defined as the upper limit of this fraction inside a large cube whose size tends to infinity).

If L is a lattice then a choice of basis gives an embedding $e_L : \mathbb{Z}^n \to \mathbb{R}^n$; its matrix is called a generating matrix of the lattice. For the diameter of spheres one can take $d(L) = \min\{|v| : v \in L, v \neq 0\}$. For any packing $P \subset \mathbb{R}^n$ the ratio $\nu(P) = \Delta(P)/V_n$ is called its center density, where

$$V_n = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$$

is the volume of the unit sphere.

The ratio $\lambda(P) = \log \Delta(P)/n$ is called the density exponent of P; thus, $\Delta(P) = 2^{-\lambda(P)n}$. The Minkowski bound, which is a corollary of the Minkowski–Hlawka theorem, says that some lattice families $\{L_n \subset \mathbb{R}^n\}$ satisfy $\lambda(L_n) \leq 1$; however, no construction is known for such families. On the other hand, the Kabatiansky–Levenstein bound says that $\lambda(P_n) \geq 0.599\ldots - o(1)$ for any family of packings $\{P_n \subset \mathbb{R}^n\}$. Families of packings with $\liminf_{n\to\infty} \lambda(P_n) < \infty$ are called *asymptotically good.* It is not easy to construct such families, especially for lattice packings. The best known results in that direction use algebraic geometry codes and similar constructions; see [Litsyn and Tsfasman 1987; Rosenbloom and Tsfasman 1990].

Another important parameter of a packing $P \subset \mathbb{R}^n$ is its kissing number

$$\tau(P) = \max_{x \in P} |\{y \in P : |x - y| = d\}|.$$

A random choice argument gives, see [Chabauty 1953; Shannon 1959], the existence of (nonlattice) packings $P_n \subset \mathbb{R}^n$ with

$$\liminf_{n\to\infty} \frac{\log \tau(P_n)}{n} \geq \log \frac{2}{\sqrt{3}} \simeq 0.2075\ldots,$$

whereas the Kabatiansky–Levenstein bound [1978] for τ says that

$$\limsup_{n\to\infty} \frac{\log \tau(P_n)}{n} \leq 0.4041\ldots.$$

We will say that a family of packings $P_n \subset \mathbb{R}^n$ is *τ-asymptotically good* whenever

$$\limsup_{n\to\infty} \frac{\log \tau(P_n)}{n} > 0.$$

Since the random choice argument does not work for lattices, it is not clear whether τ-asymptotically good lattice families exist, and our main purpose is to prove their existence.

2B. ***Error-correcting codes.*** Let us recall several facts about (linear error-correcting) codes; for additional information we refer to [MacWilliams and Sloane 1977a; 1977b]; see also [Tsfasman et al. 2007, Chapter 1]. We fix a finite field $\mathbb{F}_q$.

A q-ary linear code is simply a subspace $C \subseteq \mathbb{F}_q^n$, where n is called the length of C, and the ratio $R = k/n$ for $k = \dim C$ is called the rate of C. The minimum distance $d = d(C)$ is the minimum Hamming weight $\mathrm{wt}(c)$, i.e., the number of nonzero coordinates, of $c \in C \setminus \{0\}$; the ratio $\delta = d/n$ is called the relative minimum distance. We say in this case that C is an $[n, k, d]_q$-code. A choice of basis in C defines a linear map $\varphi_C : \mathbb{F}_q^k \to \mathbb{F}_q^n$ and its matrix is called a generating matrix of C. A set of codes $C_1 \subset \cdots \subset C_m \subseteq \mathbb{F}_q^n$ is called a nested family. For $C \subseteq \mathbb{F}_q^n$ its dual code $C^\perp$ is the orthogonal complement of C:

$$C^\perp = \{v \in \mathbb{F}_q^n : v \cdot c = 0 \text{ for all } c \in C\},$$

where $v \cdot c = v_1c_1 + \cdots + v_nc_n$; $C^\perp$ is an $[n, n-k, d^\perp]_q$-code for some $d^\perp$.

A random choice argument shows that asymptotically for $n \to \infty$ and fixed δ the rate R of the best linear codes satisfies the Gilbert–Varshamov bound

$$R = R_q(\delta) \geq 1 - H_q(\delta) = 1 - \frac{\delta \log(q-1) + H(\delta)}{\log q},$$

where $H(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta)$ is the binary entropy function.

2C. ***Algebraic geometry codes.*** All our curves here and below are smooth projective absolutely irreducible over a finite field $\mathbb{F}_q$; let X be such a curve of genus g, let D be an $\mathbb{F}_q$-rational divisor of degree $a \geq g - 1$, and let, see, e.g., [Tsfasman et al. 2007, Section 2.2],

$$L(D) = \{f \in \mathbb{F}_q(X) : (f) + D \geq 0\}$$

be the associated function space. For a set $\mathcal{P}=\{P_1,\dots,P_n\}$ of $\mathbb{F}_q$-rational points on X with $\mathcal{P}\cap\operatorname{Supp} D=\varnothing$ the evaluation map

$$\mathrm{ev}_{\mathcal{P}} : L(D)\to\mathbb{F}_q^n,\quad \mathrm{ev}_{\mathcal{P}}(f)=(f(P_1),\dots,f(P_n)),$$

is well-defined. Whenever $a<n$, this map is injective and its image is a linear q-ary code $C(X,D,\mathcal{P})$ of length n, dimension $k\ge a-g+1$ (by the Riemann–Roch theorem), and distance $d>n-a$ (since the number of zeros of a function cannot exceed the number of poles). If $D=aP_0$ for an $\mathbb{F}_q$-rational point $P_0\ne P_i,\ i=1,\dots,n$, we get a nested family of codes C_a for $a=n-1,n-2,\dots,g-1$. In the particular case $g=0,\ a\ge 0,\ P_0=\infty$ (i.e., X is the projective line), we get nested Reed–Solomon codes with parameters $n=q,\ k=a+1,\ d=q-a$.

Algebraic geometry codes (AG-codes below) have good parameters when the ratio of the number of $\mathbb{F}_q$-rational points on the curve to its genus is high enough. The Drinfeld–Vlăduţ bound says that asymptotically this ratio cannot exceed $\sqrt{q}-1$. For $q=p^{2h}$ there exist many families of curves over $\mathbb{F}_q$ attaining this bound (see, e.g., Section 5 below), which implies the lower bound

$$R_q(\delta)\ge 1-\frac{1}{\sqrt{q}-1}$$

for the best asymptotical rate of $\mathbb{F}_q$-linear codes; see, e.g., [Tsfasman et al. 2007, Section 4.5]. If $q\ge 49$, it improves (on some interval) the Gilbert–Varshamov bound.

One can dispense with the above condition $\mathcal{P}\cap\operatorname{Supp} D=\varnothing$ without spoiling the parameters of the codes $C(X,D,\mathcal{P})$; for instance, if $P_i\in\operatorname{Supp} D$ we can replace the term $f(P_i)$ in $\mathrm{ev}_{\mathcal{P}}$ by $f_i(P_i)$ with $f_i=t_i^s f$, where t_i is some fixed local parameter at P_i and s is a suitable integer (see [Tsfasman et al. 2007, Section 4.1, pp. 194–197], where the H- and P-constructions are discussed).

3. Constructions D and E

We recall now two constructions from [Barnes and Sloane 1983] and [Bos et al. 1982] (see also Chapter 8 in [Conway and Sloane 1988]), which permit us to construct good lattices from good codes.

3A. ***Construction D.*** Let $C_0=\mathbb{F}_2^n\supset C_1\supset\cdots\supset C_a$, $a\ge 1$ be a finite decreasing family of linear binary codes with parameters $[n,k_i,d_i]$ for $C_i,\ i=0,\dots,a$, where $d_i=4^i$ (we will need only the case $n=2^{2a+1}$ and thus $\delta_a=d_a/n=\frac{1}{2}$). We can and will consider C_0 as a subset of $\mathbb{R}^n$. We choose a basis $c_1,\dots,c_n$ for $\mathbb{F}_2^n$ such that $c_1,\dots,c_{k_i}$ span C_i for $i=0,\dots,a$ and define L as the lattice in $\mathbb{R}^n$ generated by $(2\mathbb{Z})^n$ and the vectors $\{c_j\cdot 2^{1-i}\}$ for $i=1,\dots,a,\ k_{i+1}+1\le j\le k_i$. Then we have [Barnes and Sloane 1983, Theorem 1]:

Proposition 3.1. *The lattice L has minimum distance $d_L=2$ and its center density satisfies*

$$\delta\ge 2^{K-n}$$

for $K=\sum_{i=1}^a k_i$.

Note that we will need only the statement $d_L=2$, which is easy in view of the minimum distances d_i of C_i for $i=0,\dots,a$.

3B. ***Construction E.*** Here we need more elaborate techniques.

First we define T-lattices as follows [Barnes and Sloane 1983; Bos et al. 1982]; see also [Litsyn and Tsfasman 1987]. A lattice $\Lambda \subset \mathbb{R}^m$ is a *T-lattice* if it satisfies the following four conditions:

(i) The minimal vectors of Λ span Λ.

(ii) There is a linear map T from $\mathbb{R}^m$ to $\mathbb{R}^m$ that sends all the minimal vectors of Λ into elements of Λ which have norm R^2 and are at a distance R from Λ for some $R > 0$.

(iii) There is a positive integer ν dividing m and an element $A \in \mathrm{Aut}(\Lambda)$ such that

(iii)$_1$ $T^\nu = \frac{1}{2}A$ and

(iii)$_2$ $\frac{1}{2}(A^2 - A) = \sum_{i=0}^{\nu-1} a_i T^i$, $a_i \in \mathbb{Z}$.

We set $b = m/\nu$ and $q = 2^b$.

(iv) $\Lambda \subseteq T\Lambda$ and

(iv)$_1$ $[T\Lambda : \Lambda] = q$.

It follows from (iii)$_1$ that $T = tP$, where $t = 2^{1/\nu}$ and P is an orthogonal transformation satisfying $P^\nu = A$. If M is the minimal square norm of Λ, we have $t = R/\sqrt{M}$, and from (iv)$_1$ we get

(v) $t^m = |\det T| = 2^{-b} = q^{-1}$.

Note that the square lattice $\mathbb{Z}^2$ is a T-lattice with $T = (1/\sqrt{2})R_{\pi/4}$ for the rotation $R_{\pi/4}$ through the angle $\pi/4 = 45°$.

Construction E produces from a T-lattice, together with a nested family of linear codes $C_0 = \mathbb{F}_{2^b}^n \supset C_1 \supset \cdots \supset C_a$ over $\mathbb{F}_{2^b}$, another T-lattice $L \subset \mathbb{R}^{mn}$ in the following way.

We suppose that the parameters of the code C_i, $0 \le i \le a$ are $[n, k_i, d_i]$ and we choose a basis $c_1, \ldots, c_n$ for $\mathbb{F}_{2^b}^n$ such that $c_1, \ldots, c_{k_i}$ span C_i for $i = 0, \ldots, a$. Define then the lattices Λ_i as follows. Let $v_i, \ldots, v_m$ be minimal vectors of Λ that span Λ. Then $Tv_i, \ldots, Tv_m$ span $T\Lambda$ and $T\Lambda/\Lambda$ is an elementary abelian group of order q, so that there are b vectors $u_i^{(1)} = Tv_{r_1}, \ldots, u_b^{(1)} = Tv_{r_b}$, for appropriate $r_1, \ldots, r_b$, such that $T\Lambda/\Lambda$ is isomorphic to the $\mathbb{F}_2$-span of $u_i^{(1)}, \ldots, u_b^{(1)}$. Let

$$\Lambda_i = T^i\Lambda, \qquad u_j^{(i)} = T^i v_{r_j}, \quad j = 1, \ldots, b, \qquad \text{for all } i \in \mathbb{Z}.$$

The lattice Λ_i has minimal square norm $t^{2i}M$, and $\mathrm{dist}(u_i^{(1)}, \Lambda_i) \ge t^{i-1}R$.

Define now the maps $\sigma_i : \mathbb{F}_q \to \Lambda_i$ by

$$\sigma_i\left(\sum_{j=1}^{b} \alpha_j \omega_j\right) = \sum_{j=1}^{b} \alpha_j u_j^{(i)}$$

for some generators $\omega_1, \ldots, \omega_b$ for $\mathbb{F}_q$ over $\mathbb{F}_2$ and any $\alpha_j \in \mathbb{F}_2$, $j = 1, \ldots, b$; those maps define the maps $\sigma_i : \mathbb{F}_q^n \to \mathbb{R}^{mn}$.

The construction. The lattice $L \subset \mathbb{R}^{mn}$ consists of all vectors of the form

$$x = l + \sum_{i=1}^{a} \sum_{j=1}^{bk_i} \alpha_j^{(i)} \sigma_i(c_j)$$

for $l \in \Lambda^n$, $\alpha_j^{(i)} \in \mathbb{F}_2$. Note that L is a T-lattice, since it inherits T from Λ; the parameter t remains the same, while b becomes nb; see also Proposition 3.2 below. The main property of this Construction E, which coincides with Construction D for $\Lambda = 2\mathbb{Z}$, is [Barnes and Sloane 1983, Theorem 3]:

Proposition 3.2. *The lattice L is fixed under the transformation $\hat{A}$, which applies A simultaneously to each component, and its minimum distance equals*

$$\sqrt{\overline{M}} \quad \text{for } \overline{M} = \min_{i=1,\dots,a} \{M, d_i R^{2i} M^{1-i}\}.$$

Theorem 3 of [Barnes and Sloane 1983] gives also the density of L, but we do not need it.

Applying Construction E to $\mathbb{Z}^2$ with $a = 1$, $M = 4$, $R = \sqrt{2}$ and the single parity check $[2, 1, 2]_q$ code C_1, we get successfully the T-lattices D_4, E_8, Λ_{16}, $\bar{\Lambda}_{32}$ in the corresponding dimensions; one can take this description as a definition for those lattices. Moreover, applying Construction E to D_4 and the single parity check $[m, m-1, 2]_4$ code for any $m \geq 2$ we get a T-lattice $\tilde{\Lambda}_{4m}$ in $4m$ dimensions. The Leech lattice Λ_{24} is also a T-lattice [Bos et al. 1982, p. 177]; note, however that $\Lambda_{24} \neq \tilde{\Lambda}_{24}$.

4. Codes with many light vectors

Recall the following principal result of [Ashikhmin et al. 2001].

Denote by A_d is the number of minimum weight vectors in an $[n, k, d]_q$-code C_n, and let E_s for $s \in \mathbb{N}$, $s \geq 3$ be the function

$$E_s(\delta) = H(\delta) - \frac{2s}{2^s - 1} - \log \frac{2^{2s}}{2^{2s} - 1}, \tag{4-1}$$

which has two zeros $0 < \delta_1 < \delta_2 < 1 - 2^{-2s}$ and is positive for $\delta_1 < \delta < \delta_2$. In particular, for $s = 3$, $q = 64$, $\delta = \frac{1}{2}$ we have

$$E_3(0.5) = \tfrac{1}{7} - \log \tfrac{64}{63} \simeq 0.1201\ldots, \qquad \tfrac{1}{64} E_3(0.5) \simeq 0.001877\ldots.$$

Theorem 4.1. *Let $q = 2^{2s}$, $s = 3, 4, \ldots$ be fixed. Then for any $\delta_1 < \delta < \delta_2$ there exists a sequence of binary linear codes $\{C_n\}$ of length $n = qN$, $N \to \infty$ and distance $d_n = n\delta/2$ such that*

$$\frac{\log A_{d_n}}{n} \geq \frac{E_s(\delta)}{2^s} - o(1). \tag{4-2}$$

Theorem 4.1 is a simple consequence of the following result concerning AG codes. Consider a curve X of genus g over $\mathbb{F}_q$, where $q = 2^{2s}$, $s \geq 3$. Suppose that $N \geq (2^s - 1)g$, where $N = |X(\mathbb{F}_q)|$ is the number of $\mathbb{F}_q$-rational points of X (e.g., X is a curve from Subsections 5A, 5B below). Let D be an $\mathbb{F}_q$-rational positive divisor of degree $a > 0$, and let $C = C(X, D, X(\mathbb{F}_q))$ be the corresponding AG code of length N, dimension $k(C) \geq a - g + 1$, and distance $d(C) \geq N - a$.

Proposition 4.2. *Let $\delta = (N - a)/N$ satisfy the inequality $\delta_1 < \delta < \delta_2$. Then there exists an $\mathbb{F}_q$-rational positive divisor with $\deg(D) = a$ such that the corresponding AG code C has the minimum distance $d = N - a = \delta N$ and for the number A_d of vectors of weight d we have*

$$\log A_d \geq N E_s(d) - o(N).$$

Recall that this is proved using an averaging procedure applied to the set of linearly equivalent classes of $\mathbb{F}_q$-rational positive divisors D with $\deg(D) = a$ which form the set $J_X(\mathbb{F}_q)$ of $\mathbb{F}_q$-rational points on the Jacobian J_X of X. This result is based on the estimate

$$\frac{\log |J_X(\mathbb{F}_q)|}{g} = q + (\sqrt{q} - 1) \log \frac{q}{q-1} + o(1). \tag{4-3}$$

In order to deduce Theorem 4.1 from Proposition 4.2 we take the binary simplex code, that is, the linear code dual to the $[n = q - 1, n - 2s, 3]$ Hamming code and lengthen each vector of this simplex code by a zero coordinate. This gives a binary linear $[q, 2s, q/2]$-code C_0 in which every nonzero vector has Hamming weight $q/2$. Using then a linear bijection $\varphi : \mathbb{F}_q \to C_0$ and replacing every coordinate by its image, we obtain from $C(D)$ a linear binary code C_n in Theorem 4.1.

Remark. Proposition 4.2 is valid for any even prime power $q \geq 49$, but we do not use this below. Note also that its proof guarantees in general only the existence of *one* divisor class D satisfying the conclusion (and not of exponentially many such divisor classes); however, when the bound is strictly bigger than $k(C)$, we get exponentially many such divisor classes in $J_X(\mathbb{F}_q)$.

Effective version. Note that at the expense of a small decline in parameters the above estimate can be made completely explicit, namely, we have:

Theorem 4.3. *Let $q = p^h$ be a prime power, let X be a curve of genus g over $\mathbb{F}_q$, let $S \subseteq X(\mathbb{F}_q)$, $|S| = N$, and let $a \in \mathbb{N}$ with $1 \leq a \leq N - 1$. Then there exists an $\mathbb{F}_q$-rational positive divisor $D \geq 0$, $\deg(D) = a$, such that the corresponding AG code $C = C(X, D, S)$ has the minimum distance $d = N - a = \delta N$ and we have*

$$A_d \geq \frac{\binom{N}{a}}{(\sqrt{q} + 1)^{2g}}.$$

The proof simply replaces the asymptotic inequality (4-3) by a simpler effective inequality

$$|J_X(\mathbb{F}_q)| \leq (\sqrt{q} + 1)^{2g}.$$

Applying Stirling's formula, we get:

Corollary 4.4. *We have*

$$\frac{\log A_d}{N} \geq H(\delta) - \frac{2g}{N} \log(\sqrt{q} + 1) - \frac{\log(2\pi ad)}{2N} - \frac{1}{12ad}.$$

In particular, if $N = 2a = 2d \geq (\sqrt{q} - 1)g$, then

$$\frac{\log A_d}{N} > 1 - \frac{2 \log(\sqrt{q} + 1)}{\sqrt{q} - 1} - \frac{2 + 2 \log N}{N}.$$

Note, that Theorem 4.3 and Corollary 4.4 are applicable, e.g., for $g = 0$, where we get an estimate for the Reed–Solomon codes.

5. Some good families of curves

We recall now some constructions of curves over $\mathbb{F}_q$ with many rational points. Let q be a prime power (we will be interested only by the case $q = p^{2h}$), and let

$$N_q(g) := \max\{|C(\mathbb{F}_q)| : C \text{ is a curve of genus } g \text{ over } \mathbb{F}_q\}.$$

Define then

$$A(q) := \lim_{g\to\infty} \sup \frac{N_q(g)}{g} \le \sqrt{q} - 1, \quad A^-(q) := \lim_{g\to\infty} \inf \frac{N_q(g)}{g}$$

as the corresponding upper and lower asymptotic quantities. We begin with some families attaining the bound for $A(q)$ (the Drinfeld–Vlăduț bound).

5A. ***Garcia–Stichtenoth tower.*** The tower X_n, $n = 1, 2, \ldots$, from [Garcia and Stichtenoth 1996] is defined recursively by the equations

$$x_{i+1}^q + x_{i+1} = \frac{x_i^q}{x_i^{q-1} + 1} \quad \text{for } i = 1, \ldots, n-1. \tag{5-1}$$

Therefore, the function field $T_n := \mathbb{F}_{q^2}(X_n)$ of the curve X_n is given by $T_n = \mathbb{F}_{q^2}(x_1, \ldots, x_n)$, where x_i, $i = 1, \ldots, n$, are related by (5-1). The main result of [Garcia and Stichtenoth 1996] gives the parameters of that tower.

Theorem 5.1. *We have for the genus* $g_n = g(X_n)$

$$g_n = (q^m - 1)^2 \qquad \text{for } n = 2m,$$
$$g_n = (q^m - 1)(q^{m-1} - 1) \quad \text{for } k = 2m - 1,$$

and the number $N(n) = |X_n(\mathbb{F}_{q^2})|$ *of* $\mathbb{F}_{q^2}$*-rational points of* X_n *satisfies*

$$N(n) \ge (q-1)q^n.$$

Let us then describe an optimal tower of Drinfeld curves closely related to the tower X_n.

5B. ***Drinfeld modular curves.*** The general reference for Drinfeld modular curves is [Gekeler 1986], but we use a particular case from [Elkies 2001]; see also [Gekeler 2001].

A tower of Drinfeld curves. For any field $L \supseteq \mathbb{F}_q$, we denote by $L\{\tau\}$ the noncommutative L-algebra generated by τ and satisfying the relation $\tau a = a^q \tau$ for all $a \in L$. Let $A = \mathbb{F}_q[T]$; then a rank-2 Drinfeld module φ over A is an $\mathbb{F}_q$-algebra homomorphism from A to $L\{\tau\}$ such that

$$\varphi(T) = l_0 + l_1\tau + l_2\tau^2 = l_0 + g\tau + \Delta\tau^2 \in L\{\tau\}, \tag{5-2}$$

with nonzero *discriminant* $\Delta = \Delta(\varphi)$. The map $\gamma : A \to L$ taking any $a \in A$ to the constant term of a is a ring homomorphism; thus, $\gamma(T) = l_0$ in (5-2).

If φ, ψ are two Drinfeld modules, an isogeny from φ to ψ is an element $u \in \bar{L}\{\tau\}$ such that

$$u \circ \varphi_a = \psi_a \circ u$$

for all $a \in A$, and its kernel is the A-submodule of $\bar{L}$ given by

$$\ker(u) := \{x \in \bar{L} : u(x) = 0\},$$

which is of finite dimension over $\mathbb{F}_q$ unless $u = 0$. In particular, if $u = \varphi_a$ then u is an isogeny from φ to itself, called multiplication by a, and its kernel is isomorphic with $(A/aA)^2$ as an A-module for $\gamma(a) \neq 0$; elements of $\ker(a)$ are called a-torsion points of φ. If γ is not injective then $\ker \gamma = Ab$ for

some irreducible $b \in A$; φ is then said to be supersingular if $\ker(b) = \{0\}$, and for $\deg(b) = 1$ we have $\varphi_b = g\tau + \Delta\tau^2$ and φ_b is supersingular if and only if $g = 0$. An isomorphism between Drinfeld modules is simply an element $u \in \bar{L}^*$, and it multiplies each coefficient l_i in (5-2) by u^{1-q^i}. Let

$$J(\varphi) = \frac{g^{q+1}}{\Delta}.$$

Then φ and ψ with the same γ are isomorphic over $\bar{L}$ if and only if $J(\varphi) = J(\psi)$. Thus, we can refer to the J-line as the Drinfeld modular curve $X(1)$ for a given γ. Moreover, for $N \in A$ with $\gamma(N) \neq 0$, we have Drinfeld modular curves $X_0(N)$ parametrizing Drinfeld modules with a choice of torsion subgroup $G \simeq A/NA$ (and fixed γ). If $\gamma(T) \in \mathbb{F}_q$, we may regard the curves $X(1)$ and $X_0(N)$ as the "reduction mod $(T - \gamma(T))$" of the corresponding modular curves for $\gamma(T) = T$. Below we suppose that $\gamma(T) = 1$ and we say that a point on $X_0(N)$ is supersingular if the corresponding Drinfeld module is supersingular; such points are $\mathbb{F}_{q^2}$-rational.

Let us consider the case $N = T^{k+1}$; for the curve $\widetilde{X}_k := X_0(T^{k+1})$ of genus $\tilde{g}_k = g(\widetilde{X}_k)$ we have [Gekeler 2001, Example 10.2]

$$\begin{aligned}
\tilde{g}_k &= \frac{(q^m-1)^2}{q-1} && \text{for } k = 2m,\\
\tilde{g}_k &= \frac{(q^{m+1}-1)(q^m-1)}{q-1} && \text{for } k = 2m+1,\\
\widetilde{N}(k) &= \left|\widetilde{X}_k(\mathbb{F}_{q^2})\right| \geq q^k + 4 && \text{for } k \geq 2;
\end{aligned}$$

thus,

$$\widetilde{N}(k) \geq (q-1)\tilde{g}_k \quad \text{for } k \geq 2$$

and the number of supersingular points on $\widetilde{X}_k$ equals q^k.

Elkies [2001] proved that the function field $\widetilde{K}_k = \mathbb{F}_q(\widetilde{X}_k)$, $k \geq 2$, is given by

$$\widetilde{K}_k = \mathbb{F}_q(x_1, \ldots, x_k) \quad \text{with } x_{j+1}(x_{j+1}+1)^{q-1}(x_j+1)^{q-1} = x_j^q,\ j = 1, \ldots, k-1,$$

and the set of q^k supersingular points of $\widetilde{X}_k(\mathbb{F}_{q^2})$ is determined by the conditions $\Phi_{q+1}(x_j) = 0$ for $j = 1, \ldots, k$, where $\Phi_{q+1}(t) = (t^{q+1}-1)/(t-1)$.

Note also that the Garcia–Stichtenoth curve X_n is a cyclic covering of $\widetilde{X}_n$ of degree $q+1$, but we do not need this fact.

More general Drinfeld curves. We will need also more general Drinfeld modular curves which do not form a tower and as yet have no explicit equations. However, the family of those curves is optimal and their genera are explicitly known [Gekeler 2001]. Let M be a monic element of A with $M(1) \neq 0$, $\deg M \geq 3$, and let $M = \prod_{i=1}^s P_i^{r_i}$ be its prime factorization; thus each $P_i \in A$ is a monic irreducible polynomial of degree l_i and $r_i \geq 1$ for $1 \leq i \leq s$. We put $q_i := q^{l_i}$ and define the arithmetic functions

$$\varepsilon = \varepsilon(M) = \prod_{i=1}^s q_i^{r_i-1}(q_i+1), \quad \kappa = \kappa(M) = \prod_{i=1}^s (q_i^{[r_i/2]} + q_i^{[(r_i-1)/2]}).$$

Consider the curve $\widetilde{X}_0(M)$ over $\mathbb{F}_q$ which is the Drinfeld modular curve $X_0(M)$ with $\gamma(T)=1$. We have then [Gekeler 1986, Sections 8–10]:

Proposition 5.2. *Suppose that at least one degree l_i is odd. Then*:

(i) *The curve $\widetilde{X}_0(M)$ is smooth of genus $g_0(M)$ given by*

$$g_0(M) = 1 + \frac{\varepsilon - (q+1)\kappa - 2^{s-1}(q+1)(q-2)}{q^2-1} \le \frac{\varepsilon}{q^2-1}.$$

(ii) $$|\widetilde{X}_0(M)(\mathbb{F}_{q^2})| \ge \frac{\varepsilon}{q+1} \ge (q-1)g_0(M).$$

Therefore, for any sequence M_i with $\deg(M_i)\to\infty$ the family $\widetilde{X}_0(M_i)$ is asymptotically optimal over $\mathbb{F}_{q^2}$.

5C. ***Curves of every genus with many points.*** Note the genera of curves in Subsections 5A–5B are of a special form and thus they give no estimate for the quantity $A^-(q)$ measuring the maximal number of points on curves of every genus. However, in [Elkies et al. 2004] it was shown that $A^-(q)\ge c\log q$ for any prime power q and a positive constant c. Moreover, for an even square q the result gets much better:

Theorem 5.3. *For $q=2^{2h}$ we have*

$$A^-(q) \ge \frac{\sqrt{q}-1}{2+1/\log q} = \frac{2^h-1}{2+1/(2h)}.$$

Thus $A^-(q)$ is, roughly speaking, only half as small as $A(q)$; a similar result holds also for the odd squares.

6. Proofs

We begin with an easy construction which gives a small positive constant lower bound for the ratio $\log(\tau_n^l)/n$, ensuring thus the existence of τ-asymptotically good lattice families. Indeed, let us take $N=2^{K+1}$, $d=a=N/2=2^K$ for some $K\ge 2$, and let us apply Theorem 4.1 with $s=3$, $q=64$ and the Drinfeld curves $\widetilde{X}_k$ over $\mathbb{F}_8$ having at least $8^k=2^{K+1}$, $K=3k-1$, points rational over the field $\mathbb{F}_{64}$. We get then a binary $[N,k,d]$-code C_K with

$$\log A_d \ge \tfrac{1}{64}E_3(0.5)N - o(N) = \tfrac{1}{64}\big(\tfrac{1}{7}-\log\tfrac{64}{63}\big)N - o(N).$$

We can construct then a decreasing family $C_0=\mathbb{F}_2^N\supset C_1\supset\cdots\supset C_K$ defining inductively C_{K-i} for $i=1,\dots,K-1$ as generated by C_{K-i+1} and c_i for some binary vector $c_i\in\mathbb{F}_2^N$ with $\mathrm{wt}(c_i)=2^{K-i}$. Applying then Construction D we get a lattice $L_N\subset\mathbb{R}^N$ with $d_L=2$, and each minimum weight vector of C_K produces a minimum norm vector in L. Therefore we have

$$\frac{\log\tau(L_N)}{N} \ge \frac{\log A_d}{N} \ge \tfrac{1}{64}\big(\tfrac{1}{7}-\log\tfrac{64}{63}\big) - o(1) > 0.00187 - o(1).$$

This formula implies Corollary 1.2, albeit with a very small c_0.

Remark. We do not care here about the density of L, but the constructed family is still asymptotically good, albeit very poor for its density; however, it is easy to modify the construction to get a better (yet rather poor) family while conserving the ratio $\log\tau(L_N)/N$.

Remark. If we replace in the above construction the Drinfeld curve $\widetilde{X}_k$ by the Garcia–Stichtenoth curve X_k over $\mathbb{F}_{64}$ which has $63\cdot 64^k + O(1)$ points rational over $\mathbb{F}_{64}$, we can use $\delta = \frac{32}{63}$, since the minimum distance should be a power of 2. This leads to the bound $\frac{1}{64}\big(H\big(\frac{32}{63}\big) - \frac{6}{7} - \log\frac{64}{63}\big) \simeq 0.001874\ldots$ instead of $\frac{1}{64}\big(\frac{1}{7} - \log\frac{64}{63}\big) \simeq 0.001877\ldots$, and in that sense the Garcia–Stichtenoth tower is not optimal for our construction. The same remark applies to the constructions below, but the deterioration of the parameters is always very small.

It is then clear how to proceed: we can replace Construction D by Construction E applied to suitable T-lattices and codes from Theorem 4.3, which we complete in an appropriate manner. The best results are obtained using the T-lattices $\tilde{\Lambda}_{20}$, Λ_{24} (or $\tilde{\Lambda}_{24}$), and $\tilde{\Lambda}_{28}$, which give the lattice families in Theorem 4.1.

More precisely, in the case of Λ_{24} we take $q = 2^{12} = 4096$, the curve $\widetilde{X}_k$ over $\mathbb{F}_{64}$ having $N = 2^{12k} = 4^{6k}$ points rational over $\mathbb{F}_{2^{12}}$, put $d = a = N/2$ and apply Construction E to Λ_{24} and the family $C_0 = \mathbb{F}_2^N \supset C_1 \supset \cdots \supset C_{6k}$ of $[N, k_i, 4^i]$-codes over $\mathbb{F}_{2^{12}}$ for $i = 0, \ldots, 6k$, where $d_i = 4^i$, $d_{6k} = d = N/2$ and C_{6k-i} is defined inductively for $i = 1, \ldots, 6k-1$ as generated by C_{6k-i+1} and c_i for some vector $c_i \in \mathbb{F}_{4096}^N$ with $\mathrm{wt}(c_i) = 4^{6k-i}$. Exactly as above, each minimum-weight vector of C_{6k} gives rise to a minimum-norm vector of the resulting lattice L_{24N} and applying Corollary 4.4 we get (1-2). If we apply the same construction to $\tilde{\Lambda}_{4m}$, $q = 2^{2m}$ and the curve $\widetilde{X}_k$ over $\mathbb{F}_q$ having $N = 2^{2mk} = 4^{mk}$ points rational over $\mathbb{F}_q$, we get a lattice with

$$\frac{\log(\tau_N^l)}{N} \geq \frac{1}{4m}\left(1 - \frac{2\log(2^m+1)}{2^m-1}\right) - \frac{2+2\log N}{N}, \tag{6-1}$$

which gives (1-1)–(1-3) for $m = 5$, 6 and 7, respectively (the result is < 0.03 for any other value of m).

Applying in the same way Proposition 4.2 instead of Corollary 4.4 we get the lattices with

$$\frac{\log(\tau_N^l)}{N} \geq \frac{1}{4m}\left(1 - \frac{2m}{2^m-1} - \log\frac{2^{2m}}{2^{2m}-1}\right) - o(1) \tag{6-2}$$

and thus Theorem 1.3 for $m = 5$, 6 and 7.

We begin the proof of Theorem 1.5 with the following:

Proposition 6.1. *For any $q = p^h$ there exist monic polynomials $M_i \in \mathbb{F}_q[T]$ for $i = 1, 2, \ldots,$ with* $\deg M_{i+1} \geq \deg M_i$, *satisfying*

$$\lim_{i\to\infty} \frac{\tilde{g}_{i+1}}{\tilde{g}_i} = 1, \quad \tilde{g}_i < \tilde{g}_{i+1},$$

for $\tilde{g}_i := g(\widetilde{X}_0(M_i)) > 0$.

To prove this we "densify" the tower $\{\widetilde{X}_k\}$, inserting between its consecutive levels some curves from the family $\{\widetilde{X}_0(M)\}$. Indeed, let us consider two consecutive curves $\widetilde{X}_{2m}$ of genus $\tilde{g}_{2m} = (q^m-1)^2/(q-1)$ and $\widetilde{X}_{2m+1}$ of genus

$$\tilde{g}_{2m+1} = \frac{(q^{m+1}-1)(q^m-1)}{q-1} = q\,\tilde{g}_{2m} + O(\sqrt{\tilde{g}_{2m}}),$$

say, for $k = 2m \geq 100$. Set $s = s(k)$ for a suitable nondecreasing unbounded function $s : \mathbb{N} \to \mathbb{N}$ (to be chosen afterwards); then the number $P(s)$ of monic irreducible polynomials in A of degree s satisfies

$$\frac{q^s - q^{s/2}}{s} \leq P(s) \leq \frac{q^s}{s}.$$

We consider then the curves $\widetilde{X}_{k,j},\ j=1,\dots,l_k$ for $l_k=\min\{P(s),\lfloor k/s\rfloor\}$, defined by

$$\widetilde{X}_{k,j}=\overline{X}_0(T^{k+1-js}M_{s,j})\quad\text{for } M_{s,j}=\prod_{i=1}^{j}M_i^{(s)},$$

where $\{M_1^{(s)},\dots,M_{P(s)}^{(s)}\}$ is the list of all monic degree-s irreducible polynomials in A. The genus of $\widetilde{X}_{k,j}$ equals

$$\tilde{g}_{k,j}=\frac{q^{2m-sj}(q^s+1)^j}{q-1}+O(\sqrt{\tilde{g}_{2m}}),$$

which is increasing with j and $\tilde{g}_{k,j+1}/\tilde{g}_{k,j}$ tends to 1 for growing k. If $\tilde{g}_{k,l_k}$ is still less than $q^{2m+1}/(q-1)$, we can increase further the genus, taking $s+1$ instead of s and continuing to replace the factors T^{s+1} consecutively by irreducible polynomials of degree $s+1$, until we run out of such polynomials. If $k-sP(s)-(s+1)P(s+1)>0$ we can continue with the polynomials of degree $s+2$ and so on. The procedure stops when either we reach the genus $\tilde{g}_{2m+1}$ and we have densified our level, or there are no factors T^l to replace by the next polynomial of degree, say, $s+h,\ h\ge 1$. We want to show that choosing $s(k)$ appropriately, we can always reach $\tilde{g}_{2m+1}$ and thus densify our initial tower, which will end the proof. Indeed, for a given s, using all $P(s)$ degree-s irreducible polynomials, we multiply the genus by the factor $(1+q^{-s})^{P(s)}\simeq\exp(1/s)$. Therefore, using all irreducible polynomials of degrees from s to, say $s+t$, we can multiply the genus by

$$\exp\left(\frac{1}{s}+\cdots+\frac{1}{s+t}\right)\simeq 1+\frac{t}{s},$$

where this is possible whenever $sP(s)+\cdots+(s+t)P(s+t)\simeq q^s+\cdots+q^{s+t}\le k$. It is then sufficient to take $t/s>q,\ (s+t)q^{s+t}\le k$; for example, we can choose $t=(q+1)s,\ s=\log k/(2q\log q)$ to guarantee those inequalities for sufficiently large k, and the proof is finished (the case of an odd k is similar).

Remark. This proof can replace the sketchy proof of Claim (3.2)–(3.3) in [Shparlinski et al. 1992], equivalent to Proposition 6.1.

Let us deduce Theorem 1.5 from Proposition 6.1. Let $q=2^{12}=4096$, and let $k\in\mathbb{N}$ satisfy $\tilde{g}_k<n/24\le\tilde{g}_{k+1}$ for a given large dimension n; moreover, let $2^a\tilde{g}_k<n/24\le 2^{a+1}\tilde{g}_k$ for some $0\le a\le 11$ (recall that $\tilde{g}_{k+1}/\tilde{g}_k\simeq q$). Let us take the curve $X_0(M_i)$ from Proposition 6.1 of genus closest to $2^a\tilde{g}_k$ and the curve $X_0(M_j)$ of genus closest to $2^{a+1}\tilde{g}_k$. Then we construct, by Proposition 4.2, an $[N_i,k_i,2^{a+12k}=d_i]$-code C_i on $X_0(M_i)$ with exponentially many light vectors and the same with an $[N_j,k_j,2^{a+1+12k}=d_j]$-code C_j on $X_0(M_j)$; note that relative distances of both codes are asymptotic to $\frac{1}{2}$ and the ratio N_j/N_i is asymptotic to 2. We can then construct the lattices L_{24N_i} and L_{24N_j} in dimensions $24N_i$ and $24N_j$ using Construction E for the Leech lattice Λ_{24} (or $\tilde{\Lambda}_{24}$) and nested families of codes beginning, respectively, by C_i and C_j. The lattices L_{24N_i} and L_{24N_j} have then kissing numbers satisfying (1-6). Since $24N_i\le n\le 24N_j\simeq 48N_i$, the kissing number of the lattice L_{24N_i} gives the estimate

$$\frac{\log(\tau_n^l)}{n}\ge\tfrac{1}{24}\left(\tfrac{17}{21}-\log\tfrac{4096}{4095}\right)\delta \tag{6-3}$$

for $\delta=24N_i/n\in[0.5,1]$, and thus we can shorten the code C_j by deleting some $\mathbb{F}_q$-rational points from the corresponding curve to get a code of length $n/24$ and then apply Construction E with Λ_{24}. This gives

the estimate

$$\frac{\log(\tau_n^l)}{n} \geq \tfrac{1}{24}\Big(\lambda H\Big(\frac{1}{2\lambda}\Big) - \tfrac{4}{21} - \log\tfrac{4096}{4095}\Big), \tag{6-4}$$

with $\lambda \simeq 1/(2\delta) = n/(24N_j) \in [0.5, 1]$, and taking the minimax we get (1-8).

Remark. Using the lattices $\tilde{\Lambda}_{4m}$ together with the codes over $\mathbb{F}_{2^{2m}}$ with similar properties constructed on the curves from Theorem 5.3, instead of the above "densified" curves, we get the lattices with somewhat worse parameters, which are optimal for $m = 7$ and give the estimate

$$\liminf_{n\to\infty} \frac{\log(\tau_N^l)}{N} \geq 0.020715\ldots.$$

Acknowledgement

I thank G. Kabatiansky for drawing my attention to the problem of asymptotics for lattice kissing numbers.

References

[Ashikhmin et al. 2001] A. Ashikhmin, A. Barg, and S. Vlăduț, "Linear codes with exponentially many light vectors", *J. Combin. Theory Ser. A* **96**:2 (2001), 396–399. MR Zbl

[Barnes and Sloane 1983] E. S. Barnes and N. J. A. Sloane, "New lattice packings of spheres", *Canad. J. Math.* **35**:1 (1983), 117–130. MR Zbl

[Bos et al. 1982] A. Bos, J. H. Conway, and N. J. A. Sloane, "Further lattice packings in high dimensions", *Mathematika* **29**:2 (1982), 171–180. MR Zbl

[Chabauty 1953] C. Chabauty, "Résultats sur l'empilement de calottes égales sur une périsphère de R^n et correction à un travail antérieur", *C. R. Acad. Sci. Paris* **236** (1953), 1462–1464. MR Zbl

[Conway and Sloane 1988] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*, Grundlehren der Mathematischen Wissenschaften **290**, Springer, 1988. MR Zbl

[Elkies 2001] N. D. Elkies, "Explicit towers of Drinfeld modular curves", pp. 189–198 in *European Congress of Mathematics, II* (Barcelona, 2000), edited by C. Casacuberta et al., Progr. Math. **202**, Birkhäuser, Basel, 2001. MR Zbl

[Elkies et al. 2004] N. D. Elkies, E. W. Howe, A. Kresch, B. Poonen, J. L. Wetherell, and M. E. Zieve, "Curves of every genus with many points, II: Asymptotically good families", *Duke Math. J.* **122**:2 (2004), 399–422. MR Zbl

[Garcia and Stichtenoth 1995] A. Garcia and H. Stichtenoth, "A tower of Artin–Schreier extensions of function fields attaining the Drinfel'd–Vlăduț bound", *Invent. Math.* **121**:1 (1995), 211–222. MR Zbl

[Garcia and Stichtenoth 1996] A. Garcia and H. Stichtenoth, "On the asymptotic behaviour of some towers of function fields over finite fields", *J. Number Theory* **61**:2 (1996), 248–273. MR Zbl

[Gekeler 1986] E.-U. Gekeler, *Drinfeld modular curves*, Lecture Notes in Mathematics **1231**, Springer, 1986. MR Zbl

[Gekeler 2001] E.-U. Gekeler, "Invariants of some algebraic curves related to Drinfeld modular curves", *J. Number Theory* **90**:1 (2001), 166–183. MR Zbl

[Kabatiansky and Levenstein 1978] G. A. Kabatiansky and V. I. Levenstein, "Bounds for packings on the sphere and in space", *Problemy Peredači Informacii* **14**:1 (1978), 3–25. In Russian; translated in Probl. Inf. Transm. 14 (1978) 1–17. MR

[Litsyn and Tsfasman 1987] S. N. Litsyn and M. A. Tsfasman, "Constructive high-dimensional sphere packings", *Duke Math. J.* **54**:1 (1987), 147–161. MR Zbl

[MacWilliams and Sloane 1977a] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes, I*, North-Holland Mathematical Library **16**, North-Holland, Amsterdam, 1977. MR Zbl

[MacWilliams and Sloane 1977b] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes, II*, North-Holland Mathematical Library **16**, North-Holland, Amsterdam, 1977. MR Zbl

[Rosenbloom and Tsfasman 1990] M. Y. Rosenbloom and M. A. Tsfasman, "Multiplicative lattices in global fields", *Invent. Math.* **101**:3 (1990), 687–696. MR Zbl

[Shannon 1959] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel", *Bell System Tech. J.* **38** (1959), 611–656. MR

[Shparlinski et al. 1992] I. E. Shparlinski, M. A. Tsfasman, and S. G. Vladut, "Curves with many points and multiplication in finite fields", pp. 145–169 in *Coding theory and algebraic geometry* (Luminy, 1991), edited by H. Stichtenoth and M. A. Tsfasman, Lecture Notes in Math. **1518**, Springer, 1992. MR Zbl

[Tsfasman et al. 2007] M. Tsfasman, S. Vlăduţ, and D. Nogin, *Algebraic geometric codes: basic notions*, Mathematical Surveys and Monographs **139**, American Mathematical Society, Providence, RI, 2007. MR Zbl

[Wyner 1965] A. D. Wyner, "Capabilities of bounded discrepancy decoding", *Bell Systems Tech. J.* **44** (1965), 1061–1122. MR

Received 22 Aug 2018. Revised 3 Oct 2018.

SERGE VLĂDUŢ:

serge.vladuts@univ-amu.fr
Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Marseille, France

and

IITP RAS, Moscow, Russia

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.179

A note on the set $A(A+A)$

Pierre-Yves Bienvenu, François Hennecart and Ilya Shkredov

Let p be a large enough prime number. When A is a subset of $\mathbb{F}_p \setminus \{0\}$ of cardinality $|A| > (p+1)/3$, then an application of the Cauchy–Davenport theorem gives $\mathbb{F}_p \setminus \{0\} \subset A(A+A)$. In this note, we improve on this and we show that $|A| \geq 0.3051p$ implies $A(A+A) \supseteq \mathbb{F}_p \setminus \{0\}$. In the opposite direction we show that there exists a set A such that $|A| > \left(\frac{1}{8} + o(1)\right)p$ and $\mathbb{F}_p \setminus \{0\} \not\subseteq A(A+A)$.

1. Introduction

The aim of this note is to study the size of the set $A(A+A) = \{a(b+c) : a, b, c \in A\}$ for a subset $A \subseteq \mathbb{F}_p \setminus \{0\}$. This sort of problem belongs to the realm of expanding polynomials and sum-product problems. In the literature, they are usually discussed in the sparse set regime; for instance, Roche-Newton et al. [2016] and Aksoy Yazici et al. [2017] proved that in the regime where $|A| \ll p^{2/3}$, one has $\min(|A+AA|, |A(A+A)|) \gg |A|^{3/2}$ (see also [Stevens and de Zeeuw 2017]). This implies in particular that as soon as $|A| \gg p^{2/3}$, both sets $A(A+A)$ and $A+AA$ occupy a positive proportion of $\mathbb{F}_p$.

Now we focus on the case where $A \subseteq \mathbb{F}_p$ occupies already a positive proportion of $\mathbb{F}_p$. Let $\alpha = |A|/p$, so we suppose that $\alpha > 0$ is bounded below by a positive constant, while p tends to infinity. We will see that in this case the set $A(A+A)$ contains all but a finite number of elements. Additionally, we prove that this finite number of elements may be strictly larger than 1, unless α is large enough.

Here are our main results.

Theorem 1.1. *Let $A \subseteq \mathbb{F}_p$ so that $|A| = \alpha p$ with $\alpha \geq 0.3051$. Then for any large enough prime p, we have $A(A+A) \supseteq \mathbb{F}_p \setminus \{0\}$.*

For smaller densities, we have the following result.

Theorem 1.2. *Let $A \subseteq \mathbb{F}_p \setminus \{0\}$ and $0 < \alpha < 1$ satisfy $|A| \geq \alpha p$. Then one has*

$$|A(A+A)| > p - 1 - \alpha^{-3}(1-\alpha)^2 + o(1).$$

We note that similar results were obtained [Hegyvári and Hennecart 2018] for the set $AA+A$. However, the constant 0.3051 is replaced by the larger $\frac{1}{3}$ in Theorem 1.1, and the term $\alpha^{-3}(1-\alpha)^2$ is replaced by the larger α^{-3}. Further, the slightly weaker bound $|A(A+A)| \geq p - \alpha^{-3}$ may be extracted from [Sárközy 2005].

In the opposite direction, we have the following result.

This work was performed within the framework of the Labex MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).
MSC2010: 11B75.
Keywords: sum-product estimates, arithmetic combinatorics, finite fields.

Theorem 1.3. *There exists $A \subseteq \mathbb{F}_p \setminus \{0\}$ such that $|A| > \left(\frac{1}{8} + o(1)\right)p$ and $A(A+A) \subsetneq \mathbb{F}_p \setminus \{0\}$ for any large prime p. Additionally, for any $\epsilon > 0$ there exists a set of size $O(p^{3/4+\epsilon})$ such that $A(A+A)$ misses $\Omega(p^{1/4-\epsilon})$ elements.*

2. Proof of Theorem 1.1

In this section, we shall need the Cauchy–Davenport theorem, which we now state. See for instance [Nathanson 1996, Theorem 2.2] for a proof.

Lemma 2.1. *Let A and B be subsets of $\mathbb{F}_p$. Then $|A+B| \geq \min(|A|+|B|-1, p)$.*

In particular, if $|A|+|B| > p$, then $A+B = \mathbb{F}_p$, which is also obvious because A and $x-B$ cannot be disjoint for any x.

First, we note that if $\alpha > \frac{1}{2}$, then $|A+A| \geq |A| > p/2$ so that $A(A+A) = \mathbb{F}_p$. But as soon $\alpha < \frac{1}{2}$, we can easily have $A(A+A) \subsetneq \mathbb{F}_p^*$, for instance by taking $A = \{1, \dots, \lfloor (p-1)/2 \rfloor\}$.

Here is another almost equally immediate corollary.

Corollary 2.2. *Let $A \subseteq \mathbb{F}_p \setminus \{0\}$ satisfy $|A| > (p+1)/3$. Then either $A(A+A) = \mathbb{F}_p$ or $\mathbb{F}_p \setminus \{0\}$.*

Proof. Let $B = (A+A) \setminus \{0\}$. Using Lemma 2.1, we have $|A+A| > (2p-1)/3$ so $|B| > (2p-4)/3$, whence $|A|+|B| > p-1$. We infer that for any $x \in \mathbb{F}_p \setminus \{0\}$ we have

$$xB^{-1} \cap A \neq \varnothing,$$

which yields $AB = \mathbb{F}_p \setminus \{0\}$. □

We now prove Theorem 1.1, which reveals that we can lower the density requirement from $\frac{1}{3}$ to 0.3051 while maintaining $A(A+A) \supset \mathbb{F}_p \setminus \{0\}$.

To start with, we recall the famous Freiman's $3k-4$ theorem for the integers, which gives precise structural information on a set which has quite small, but not necessarily minimal, doubling [Nathanson 1996, Theorem 1.16].

Proposition 2.3. *If $A \subset \mathbb{Z}$ satisfies $|A+A| \leq 3|A|-4$ then A is contained in an arithmetic progression of length at most $|A+A|-|A|+1$.*

An analogue of this proposition has been developed in $\mathbb{F}_p$, and it is known as the *Freiman 2.4-theorem*. A useful lemma in [Freiman 1962] (see also [Nathanson 1996, Theorem 2.9]) was derived in the proof thereof, and we will need it here. We also include an improvement due to Lev.

We first define the Fourier transform of a function $f : \mathbb{F}_p \to \mathbb{C}$ by

$$\hat{f}(t) = \sum_{x \in \mathbb{F}_p} f(x) e_p(tx)$$

for any $t \in \mathbb{F}_p$, where $e_p(x) = \exp(2i\pi x/p)$. The Parseval identity is

$$\sum_{x \in \mathbb{F}_p} f(x)\overline{g(x)} = \frac{1}{p} \sum_{h \in \mathbb{F}_p} \hat{f}(h)\overline{\hat{g}(h)}. \tag{1}$$

The characteristic function of a subset A of $\mathbb{F}_p$ is denoted by 1_A and for $r \in \mathbb{F}_p$ we let $rA = \{ra : a \in A\}$.

Lemma 2.4. *Let $A \subseteq \mathbb{F}_p$ with $|A| = \alpha p$ and $0 < \gamma < 1$ satisfy $|\hat{1}_A(r)| \geq \gamma |A|$ for some $r \in \mathbb{F}_p \setminus \{0\}$. Then there exists an interval modulo p of length at most $p/2$ that contains at least $\alpha_1 p$ elements of rA where α_1 can be freely chosen as*

(i) $\alpha_1 = (1+\gamma)\alpha/2$ (see [Freiman 1962]), *or*

(ii) $\alpha_1 = \alpha/2 + 1/(2\pi)\arcsin(\pi\gamma\alpha)$ (see [Lev 2005]).

There a few other basic results about Fourier transforms that we will need in the sequel.

Lemma 2.5. *Let P be an arithmetic progression in $\mathbb{F}_p$. Then*

$$\sum_{r\in\mathbb{F}_p} |\hat{1}_P(r)| \ll p \log p.$$

We now recall Weil's bound [1948] for Kloosterman sums.

Lemma 2.6. *For any $(a,b) \neq (0,0)$, we have*

$$\left| \sum_{k\in\mathbb{F}_p\setminus\{0\}} e_p(ak + bk^{-1}) \right| \leq 2\sqrt{p}.$$

We will also need a bound for so-called incomplete Kloosterman sums, whose proof follows easily from the last two lemmas.

Lemma 2.7. *Let $P \subseteq \mathbb{F}_p \setminus \{0\}$ be an arithmetic progression. Then for any $r \neq 0$ we have*

$$|\hat{1}_{P^{-1}}(r)| \ll \sqrt{p} \log p.$$

Now we start the proof of Theorem 1.1 itself. Let $\alpha \geq 0.3051$, let $A \subseteq \mathbb{F}_p \setminus \{0\}$ of size $|A| = \alpha p$ and set $B = (A+A) \setminus \{0\}$. We assume that there exists $x \in \mathbb{F}_p \setminus \{0\}$ such that $x \notin A(A+A)$. Then

$$xB^{-1} \cap A = \varnothing, \quad (xA^{-1} - A) \cap A = \varnothing. \tag{2}$$

It follows that $|A| + |B| \leq p - 1$, since otherwise $AB = \mathbb{F}_p \setminus \{0\}$. Hence $|A+A| \leq |B| + 1 \leq p - |A|$.

We define

$$r_1(y) = |\{(a,b) \in A \times A : y = xa^{-1} - b\}|,$$
$$r_2(y) = |\{(c,d) \in A \times A : c + d \neq 0 \text{ and } y = x(c+d)^{-1}\}|,$$

and $E_i = \sum_{y\in\mathbb{F}_p} r_i(y)^2$, $i = 1, 2$, the corresponding energies. Observe from (2) that

$$\sum_{\substack{y\in\mathbb{F}_p \\ r_1(y)+r_2(y)>0}} 1 \leq p - |A|.$$

By Cauchy–Schwarz we get

$$4|A|^4 = \left(\sum_{y\in\mathbb{F}_p} (r_1(y) + r_2(y)) \right)^2 \leq (p - |A|) \times \sum_{y\in\mathbb{F}_p} (r_1(y) + r_2(y))^2. \tag{3}$$

Expanding the later inner sum gives

$$\sum_{y\in\mathbb{F}_p} (r_1(y) + r_2(y))^2 = E_1 + E_2 + 2\sum_{y\in\mathbb{F}_p} r_1(y) r_2(y).$$

Let

$$\gamma = \max_{h\neq 0} \frac{|\hat{1}_A(h)|}{|A|}.$$

We have by Parseval

$$pE_2 = \sum_h |\hat{1}_A(h)|^4 = |A|^4 + \sum_{h\neq 0} |\hat{1}_A(h)|^4 \leq |A|^4 + \gamma^2 |A|^2 (p|A| - |A|^2)$$

and

$$\begin{aligned} pE_1 &= \sum_h |\hat{1}_{xA^{-1}}(h)|^2 |\hat{1}_A(h)|^2 = |A|^4 + \sum_{h\neq 0} |\hat{1}_{xA^{-1}}(h)|^2 |\hat{1}_A(h)|^2 \\ &\leq |A|^4 + \gamma^2 |A|^2 (p|A| - |A|^2). \end{aligned}$$

Moreover

$$\begin{aligned} p \sum_{y\in\mathbb{F}_p} r_1(y) r_2(y) &= \sum_h \hat{1}_{xA^{-1}}(h) \hat{1}_A(-h) \hat{r}_2(h) \\ &\leq |A|^4 + \max_{h\neq 0} |\hat{r}_2(h)| \sum_{h\neq 0} |\hat{1}_{xA^{-1}}(h)| |\hat{1}_A(h)| \\ &\leq |A|^4 + \max_{h\neq 0} |\hat{r}_2(h)| (p|A| - |A|^2), \end{aligned}$$

by Parseval and Cauchy–Schwarz. For $h \neq 0$,

$$\hat{r}_2(h) = \sum_{\substack{c,d\in A \\ c+d\neq 0}} e_p(hx(c+d)^{-1}) = \frac{1}{p} \sum_r \sum_{z\neq 0} \sum_{c,d\in A} e_p(r(c+d-z)) e_p(hxz^{-1});$$

hence by the Parseval identity (1) and Lemma 2.6

$$|\hat{r}_2(h)| \leq \frac{1}{p} \sum_r |\hat{1}_A(r)|^2 \left| \sum_{z\neq 0} e_p(hxz^{-1}) \right| \ll \sqrt{p}|A|;$$

similar arguments were used in [Moshchevitin 2007, Theorem 4]. We thus obtain from (3) and the above bounds

$$2\alpha \leq (1-\alpha)(2\alpha + \gamma^2(1-\alpha) + o(1)).$$

This finally gives the lower bound

$$\gamma \geq \frac{\sqrt{2\alpha}}{1-\alpha} + o(1).$$

We are in position to apply Lemma 2.4(i). Let $A_1 \subset A$ be such that $|A_1| \geq (1+\gamma)|A|/2$ and rA_1 is included in an interval of length $p/2$ for some $r \neq 0$. This shows that A_1 is 2-Freiman isomorphic[1] to a subset A_1' of $\mathbb{Z}$. So we seek to apply Proposition 2.3 to A_1'. We get

$$\alpha_1 = \frac{|A_1|}{p} \geq f(\alpha) + o(1) := \frac{(1 + (\sqrt{2}-1)\alpha)\alpha}{2(1-\alpha)} + o(1), \tag{4}$$

$$c_1 = \frac{|A_1 + A_1|}{|A_1|} \leq \frac{|A+A|}{|A_1|} \leq \frac{(1-\alpha)p}{\alpha_1 p} \leq \frac{1-\alpha}{f(\alpha)} + o(1). \tag{5}$$

[1]That is, there exists a bijection $f : A_1 \to A_1'$ such that $a+b=c+d \iff f(a)+f(b)=f(c)+f(d)$ for all $a, b, c, d \in A_1$.

In order to have $c_1 < 3$, it is sufficient to have

$$\alpha > \frac{7-\sqrt{9+24\sqrt{2}}}{10-6\sqrt{2}} = 0.29513\ldots,$$

which is satisfied since we have assumed $\alpha \geq 0.3051$. We thus obtain that A_1 (resp. A_1+A_1) is contained inside an arithmetic progression P_1 (resp. $Q_1 = P_1+P_1$) of length $|P_1| = |A_1+A_1| - |A_1| + 1$ (resp. $2|P_1|-1$).

We define $B_1 = (A_1+A_1)\setminus\{0\}$ and $Q_1^* = Q_1 \setminus \{0\}$. We need to estimate

$$T = \frac{1}{p}\sum_{r \bmod p}\sum_{\substack{a\in P_1\\ b\in Q_1^*}} e_p(r(a-b^{-1}x)) \geq \frac{|P_1||Q_1^*|}{p} - \frac{1}{p}\sum_{0<|r|<p/2} |\hat{1}_{P_1}(r)||\hat{1}_{Q_1^{*-1}}(rx)|,$$

which counts the solutions $(a,b)\in P_1\times Q_1^*$ to the equation $a = b^{-1}x$.

Now $|\hat{1}_{P_1}(r)| \ll p/|r|$ by Lemma 2.5 and $|\hat{1}_{Q_1^{*-1}}(rx_0)| \ll \sqrt{p}\log p$ by Lemma 2.7 because Q_1^* is the union of at most two arithmetic progressions.

As a result, we have

$$T \geq \frac{|P_1||Q_1^*|}{p} + O(\sqrt{p}(\log p)^2).$$

The number of solutions to $a = b^{-1}x$ with $a\in P_1\setminus A_1$ or $b\in Q_1^*\setminus B_1$ is at most $|P_1|-|A_1|+|Q_1^*|-|B_1|$. Since by assumption there is no solution to $a = b^{-1}x$ with $(a,b)\in A_1\times B_1$ we get

$$T \leq |P_1| - |A_1| + |Q_1^*| - |B_1|$$

yielding

$$\frac{|P_1||Q_1^*|}{p} \leq |P_1| - |A_1| + |Q_1^*| - |B_1| + O(\sqrt{p}(\log p)^2).$$

This implies

$$\frac{(|B_1|-|A_1|)^2}{p} \leq |B_1| - 2|A_1| + O(\sqrt{p}(\log p)^2),$$

whence

$$\alpha_1(c_1-1)^2 \leq c_1 - 2 + o(1).$$

Because of (4), this gives

$$f(\alpha)\times(c_1-1)^2 - c_1 + 2 \leq o(1). \tag{6}$$

The left-hand side of this inequality defines a function of c_1 which is decreasing in the range $2\leq c_1 \leq 1+1/(2f(\alpha))$, a contradiction. We check easily that $\alpha + f(\alpha) \geq \frac{1}{2}$ whenever $\alpha \geq 0.3$. Hence for such α

$$\frac{1-\alpha}{f(\alpha)} \leq 1 + \frac{1}{2f(\alpha)}.$$

We thus obtain from (5) and (6)

$$f(\alpha)\left(\frac{1-\alpha}{f(\alpha)} - 1\right)^2 - \frac{1-\alpha}{f(\alpha)} + 2 \leq o(1),$$

which reduces to

$$(1-\alpha-f(\alpha))^2-(1-\alpha-2f(\alpha))\le o(1).$$

In view of the definition of $f(\alpha)$ in (4), we get by expanding the above formula

$$(11-6\sqrt{2})\alpha^3-(22-6\sqrt{2})\alpha^2+17\alpha-4\le o(1),$$

giving $\alpha < 0.305091 + o(1)$, a contradiction for all p large enough. This concludes the proof of Theorem 1.1. □

Remark 2.8. Using instead the sharpest result (ii) of Lemma 2.4 leads to a slight improvement: if $|A|\ge 0.30065p$ then $\mathbb{F}_p\setminus\{0\}\subseteq A(A+A)$ for any large p. The improvement is very small and uses nonalgebraic expressions, which is why we decided not to exploit it.

3. Proof of Theorem 1.2

We will now use multiplicative characters of $\mathbb{F}_p$. We denote by $\mathfrak{X}$ the set of all multiplicative characters modulo p and by χ_0 the trivial character. In this context Parseval's identity is the statement that

$$\frac{1}{p-1}\sum_{\chi\in\mathfrak{X}}\left|\sum_{x\in\mathbb{F}_p\setminus\{0\}}f(x)\chi(x)\right|^2=\sum_{x\in\mathbb{F}_p\setminus\{0\}}|f(x)|^2. \tag{7}$$

We state and prove a lemma which is a multiplicative analogue of a lemma of Vinogradov [1955], see also [Sárközy 2005, Lemma 7], according to which

$$\left|\sum_{(x,y)\in A\times B}e_p(xy)\right|\le\sqrt{p|A||B|}. \tag{8}$$

Lemma 3.1. *For any subsets A, B of $\mathbb{F}_p\setminus\{0\}$ and any nontrivial character $\chi\in\mathfrak{X}$, we have*

$$\left|\sum_{(y,z)\in A\times B}\chi(y+z)\right|\le(|A||B|p)^{1/2}\left(1-\frac{|B|}{p}\right)^{1/2}.$$

We now prove Theorem 1.2. Let A be a subset of $\mathbb{F}_p\setminus\{0\}$ and $\alpha=|A|/p$. We estimate the number of nonzero elements in $A(A+A)$ by estimating the number N of solutions to

$$x(y+z)=x'(y'+z')\ne 0,\quad x,y,z,x',y',z'\in A,$$

which we can rewrite as $x'x^{-1}(y+z)^{-1}(y'+z')=1$. This number is

$$\begin{aligned}N&=\frac{1}{p-1}\sum_{\chi\in\mathfrak{X}}\left|\sum_{y,z\in A}\chi(z+y)\sum_{x\in A}\chi(x)\right|^2\\&\le\frac{|A|^6}{p-1}+\max_{\chi\ne\chi_0}\left|\sum_{y,z\in A}\chi(y+z)\right|^2\times\frac{1}{p-1}\sum_{\chi\ne\chi_0}\left|\sum_{x\in A}\chi(x)\right|^2;\end{aligned}$$

hence by Lemma 3.1 and Parseval's identity (7)

$$\begin{aligned} N &\le \frac{|A|^6}{p-1} + p|A|^2(1-\alpha)\left(|A| - \frac{|A|^2}{p-1}\right) \\ &\le \frac{|A|^6}{p-1} + p|A|^3(1-\alpha)^2 \\ &\le \frac{|A|^6}{p-1}(1+p^2|A|^{-3}(1-\alpha)^2) \\ &\le \frac{|A|^6}{p-1}(1+p^{-1}\alpha^{-3}(1-\alpha)^2). \end{aligned}$$

We let $\rho(w) = |\{(x, y, z) \in A \times A \times A : w = x(y+z)\}|$ for $w \in \mathbb{F}_p$. Then

$$N = \sum_{w \in A(A+A)\setminus\{0\}} \rho(w)^2 \quad \text{and} \quad \sum_{w \in A(A+A)\setminus\{0\}} \rho(w) \ge |A|^6 - |A|^4.$$

Finally N is related to $|A(A+A)|$ by the Cauchy–Schwarz inequality as follows:

$$\begin{aligned} |A(A+A)| \ge |A(A+A)\setminus\{0\}| &\ge (|A|^6 - |A|^4)N^{-1} \\ &\ge (p-1)(1-\alpha^{-2}p^{-2})(1+p^{-1}\alpha^{-3}(1-\alpha)^2)^{-1} \\ &> p - 1 - \alpha^{-3}(1-\alpha)^2 + o(1). \end{aligned}$$

This concludes the proof of Theorem 1.2. □

4. Proof of Theorem 1.3

First we need a lemma.

Lemma 4.1. *Let $c < \frac{1}{2}$ and p be large enough. Let $P = \{1, \dots, \lceil cp \rceil\}$. Then the set $(P+P)^{-1}$ of the inverses (modulo p) of nonzero elements of $P+P$ has at most $2c^2p + O(\sqrt{p}(\log p)^2)$ common elements with P; that is,*

$$|(P+P)^{-1} \cap P| \le 2c^2p + O(\sqrt{p}(\log p)^2).$$

Proof. We note that $P + P = \{2, \dots, 2\lceil cp \rceil\} \subset \mathbb{F}_p \setminus \{0\}$.

Now we observe that

$$|P \cap (P+P)^{-1}| = \sum_{\substack{x \in P \\ y \in P+P \\ x = y^{-1}}} 1 = \frac{1}{p} \sum_{t \in \mathbb{F}_p} \sum_{\substack{x \in P \\ y \in P+P}} e_p(t(x - y^{-1})) = \frac{1}{p} \sum_{t \in \mathbb{F}_p} \sum_{x \in P} e_p(tx) \sum_{y \in P+P} e_p(-ty^{-1}).$$

Using Lemmas 2.5 and 2.7, we find that

$$\begin{aligned} |P \cap (P+P)^{-1}| &= \frac{|P||P+P|}{p} + \frac{1}{p} \sum_{t \in \mathbb{F}_p \setminus \{0\}} \hat{1}_P(t) \hat{1}_{(P+P)^{-1}}(-t) \\ &= 2c^2p + O(\sqrt{p}(\log p)^2). \end{aligned}$$

□

Now we prove Theorem 1.3.

Let $c < \frac{1}{2}$ (to be determined later) and p be large enough. Let $P = \{1, \dots, \lceil cp \rceil\}$. Let $A = P \setminus (P+P)^{-1}$. It satisfies $A \cap (A+A)^{-1} = \varnothing$, i.e., $1 \notin A(A+A)$, and has cardinality at least $cp - 2c^2 p - O(\sqrt{p}(\log p)^2)$. To optimise, we take $c = \frac{1}{4}$, in which case $|A| \geq p/8 - O(\sqrt{p}(\log p)^2)$. For any $\epsilon > 0$, for p large enough, this is at least $\left(\frac{1}{8} - \epsilon\right)p$, whence the first part of the theorem.

For the second part, we note that Lemma 4.1 provides a bound for the cardinality $|P \cap x(P+P)^{-1}|$ for any x, so for any $k \leq p-1$ we can get a set a of size $cp - 2kc^2 p - O(k\sqrt{p}(\log p)^2)$ so that $A(A+A)$ misses 0 and k nonzero elements. The main term is optimised for $c = 1/(4k)$, where it is worth $p/(8k)$. Taking k of size $p^{1/4}(\log p)^{-3/2}$, the error term is significantly smaller than the main term (for large p), so we obtain a set A of size $\Omega(p^{3/4}(\log p)^{3/2})$ for which $A(A+A)$ misses at least $p^{1/4}(\log p)^{-3/2}$ elements. This is even a slightly stronger statement than claimed. □

5. Final remarks

5A. Let p be an odd prime, $a, b \in \mathbb{F}_p \setminus \{0\}$ and assume that $ba^{-1} = c^2$ is a square. Let $A \subset \mathbb{F}_p \setminus \{0\}$. Then $a \notin A(A+A)$ if and only if $b \notin cA(cA+cA) = c^2 A(A+A)$. Moreover $|cA| = |A|$.

We define

$$m_p = \max\{|A| : A \subseteq \mathbb{F}_p \setminus \{0\} \text{ and } A(A+A) \not\supseteq \mathbb{F}_p \setminus \{0\}\}.$$

From the above remark we have

$$m_p = \max\{|A| : A \subseteq \mathbb{F}_p \setminus \{0\} \text{ and } 1 \notin A(A+A) \text{ or } r \notin A(A+A)\},$$

where r is any fixed nonsquare residue modulo p. By Theorems 1.1 and 1.3 we have

$$3.277\ldots \leq \liminf_{p\to\infty} \frac{p}{m_p} \leq \limsup_{p\to\infty} \frac{p}{m_p} \leq 8.$$

5B. Let $p > 3$ be a prime number. The set I of residues modulo p in the interval $\{r \in \mathbb{F}_p : p/3 < r < 2p/3\}$ is sum-free (i.e., $a + b \neq c$ for any $a, b, c \in I$) and achieves the largest cardinality for those sets, namely $|I| = \lfloor (p+1)/3 \rfloor$, as it can be deduced from the Cauchy–Davenport theorem combined with the fact that $|I \cap (I+I)| = 0$.

Let

$$A = \{x \in I : x^{-1} \in I\}.$$

Then $A = A^{-1}$ and A is sum-free. It readily follows that $1 \notin A(A+A)$. Moreover, since I is an arithmetic progression, the events $x \in I$ and $x^{-1} \in I$ are independent, so we may observe that A has cardinality $\sim p/9$ as p tends to infinity (it can be formally proved using Fourier analysis). This raises the next question:

What is the largest size of a sum-free set $A \subset \mathbb{F}_p \setminus \{0\}$ such that $A = A^{-1}$?

From Theorem 1.1, we deduce the following statement.

Corollary 5.1. *Let $A \subset \mathbb{F}_p \setminus \{0\}$ be a sum-free set such that $A = A^{-1}$. Then $|A| < 0.3051p$ for any sufficiently large prime number p.*

This is related to the question of how large a sum-free multiplicative subgroup of $\mathbb{F}_p^*$ can be. Alon and Bourgain [2014] showed that it can be at least $\Omega(p^{1/3})$.

5C. Let $A \subset \mathbb{F}_p \setminus \{0\}$ with $\alpha = |A|/p \gg 1$, and let us set $A_s = A \cap (A+s)$. Let $0 < \epsilon < 1$ be defined by

$$E^+(A) = \sum_{s \in A-A} |A_s|^2 = (1-\epsilon)|A|^3,$$

and S be the subset of $A - A$ given by

$$S = \{s \in A - A : |A_s| > (1 - \epsilon - p^{-1/3})|A|\}.$$

Then

$$E^+(A) \le (1-\epsilon-p^{-1/3})|A| \sum_{s \notin S} |A_s| + |A|^2|S| = (1-\epsilon-p^{-1/3})|A|^3 + |A|^2|S|,$$

from which we deduce

$$|S| \ge |A| p^{-1/3}. \tag{9}$$

Assume that $A = A^{-1}$ and let N be the number of solutions to the equation

$$(a-s)(b-t) = 1, \quad (s,a,t,b) \in S \times A_s \times S \times A_t.$$

For fixed $s, t \in S$, we have

$$\begin{aligned} |(A-s) \cap (A_t - t)^{-1}| &= |A_s| + |A_t| - |(A-s) \cap (A_t - t)^{-1}| \\ &\ge 2(1-\epsilon-o(1))|A| - |A| = (1-2\epsilon-o(1))|A| \end{aligned}$$

since $A_s - s \subset A$ and $(A_t - t)^{-1} \subset A^{-1} = A$. This yields

$$N \ge (1-2\epsilon-o(1))|A||S|^2. \tag{10}$$

On the other hand, defining $r(x) = |\{(a,s) \in A \times S : x(a-s) = 1\}|$, we have

$$N \le \frac{1}{p} \sum_h \hat{1}_A(h) \hat{1}_S(-h) \hat{r}(-h) \le \frac{|A|^2|S|^2}{p} + \max_{h \ne 0} |\hat{r}(h)| \times \frac{1}{p} \sum_h |\hat{1}_A(h) \hat{1}_S(h)|.$$

By adapting (8) we get $\max_{h \ne 0} |\hat{r}(h)| \le \sqrt{p|A||S|}$ and by Cauchy–Schwarz and Parseval we derive $N \le |A|^2|S|^2/p + O(\sqrt{p}|A||S|)$. Combined with (10), this gives

$$\alpha + O(\sqrt{p}|S|^{-1}) \ge 1 - 2\epsilon - o(1),$$

yielding by (9) that $\epsilon \ge (1-\alpha)/2 + o(1)$. Hence when $A = A^{-1}$,

$$E^+(A) \le \frac{1+\alpha+o(1)}{2}|A|^3.$$

Together with Theorem 1.1, this implies the following result.

Proposition 5.2. *Let $A \subset \mathbb{F}_p^*$ be as in Corollary 5.1. Then for large enough p the additive energy satisfies*

$$E^+(A) \le 0.6526|A|^3.$$

By considering similarly the multiplicative energy of A, it is possible to get the following sum-product upper bound for an arbitrary $A \subset \mathbb{F}_p$:

$$2E^+(A) + E^\times(A) \le (2+\alpha+o(1))|A|^3.$$

References

[Aksoy Yazici et al. 2017] E. Aksoy Yazici, B. Murphy, M. Rudnev, and I. Shkredov, "Growth estimates in positive characteristic via collisions", *Int. Math. Res. Not.* **2017**:23 (2017), 7148–7189. MR Zbl

[Alon and Bourgain 2014] N. Alon and J. Bourgain, "Additive patterns in multiplicative subgroups", *Geom. Funct. Anal.* **24**:3 (2014), 721–739. MR Zbl

[Freiman 1962] G. A. Freiman, "Inverse problems of additive number theory, VII: The addition of finite sets, IV: The method of trigonometric sums", *Izv. Vyssš. Učebn. Zaved. Matematika* **1962**:6 (1962), 131–144. MR

[Hegyvári and Hennecart 2018] N. Hegyvári and F. Hennecart, "A note on the size of the set $A^2 + A$", *Ramanujan J.* **46**:2 (2018), 357–372. MR Zbl

[Lev 2005] V. F. Lev, "Distribution of points on arcs", *Integers* **5**:2 (2005), art. id. A11. MR Zbl

[Moshchevitin 2007] N. G. Moshchevitin, "Sets of the form $\mathcal{A} + \mathcal{B}$ and finite continued fractions", *Mat. Sb.* **198**:4 (2007), 95–116. In Russian; translated in *Sb. Math* **198**:4 (2007), 537–557. MR Zbl

[Nathanson 1996] M. B. Nathanson, *Additive number theory*, Graduate Texts in Mathematics **165**, Springer, 1996. MR Zbl

[Roche-Newton et al. 2016] O. Roche-Newton, M. Rudnev, and I. D. Shkredov, "New sum-product type estimates over finite fields", *Adv. Math.* **293** (2016), 589–605. MR Zbl

[Sárközy 2005] A. Sárközy, "On sums and products of residues modulo p", *Acta Arith.* **118**:4 (2005), 403–409. MR Zbl

[Stevens and de Zeeuw 2017] S. Stevens and F. de Zeeuw, "An improved point-line incidence bound over arbitrary fields", *Bull. Lond. Math. Soc.* **49**:5 (2017), 842–858. MR Zbl

[Vinogradov 1955] I. M. Vinogradov, *An introduction to the theory of numbers*, Pergamon Press, London & New York, 1955. MR

[Weil 1948] A. Weil, "On some exponential sums", *Proc. Nat. Acad. Sci. U. S. A.* **34** (1948), 204–207. MR Zbl

Received 21 Nov 2018. Revised 14 Dec 2018.

PIERRE-YVES BIENVENU:

pbienvenu@math.univ-lyon1.fr
Université Lyon 1, CNRS, ICJ UMR 5208, Villeurbanne, France

FRANÇOIS HENNECART:

francois.hennecart@univ-st-etienne.fr
Université Jean-Monnet, CNRS, ICJ UMR 5208, Saint-Étienne, France

ILYA SHKREDOV:

ilya.shkredov@gmail.com
Steklov Mathematical Institute, Divison of Algebra and Number Theory, Moscow, Russia

and

IITP RAS, Moscow, Russia

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 2, 2019
dx.doi.org/10.2140/moscow.2019.8.189

On a theorem of Hildebrand

Carsten Dietzel

We give a short proof that for each multiplicative subgroup H of finite index in $\mathbb{Q}^+$, the set of integers a with $a, a+1 \in H$ is an IP-set. This generalizes a theorem of Hildebrand concerning completely multiplicative functions taking values in the k-th roots of unity.

A theorem of Hildebrand [1991, Theorem 2], which was essential in answering a question of Lehmer, Lehmer and Mills [Lehmer et al. 1963] on consecutive power residues can be formulated as follows:

Theorem 1 (Hildebrand). *Fix some $k \in \mathbb{Z}^+$. If $f : \mathbb{Z}^+ \to \mathbb{C}$ is a completely multiplicative function (i.e., $f(mn) = f(m)f(n)$ for all $m, n \in \mathbb{Z}^+$) taking its values in the k-th roots of unity then the set of $a \in \mathbb{Z}^+$ fulfilling $f(a) = f(a+1) = 1$ is nonempty.*

Remark 2. Hildebrand actually proved more; i.e., there is a constant $c(k)$, independent of the specific multiplicative function f, and an $a \in \mathbb{Z}^+$ such that $a \le c(k)$ and $f(a) = f(a+1) = 1$. By a standard compactness argument, these versions can be seen to be equivalent. It should, however, be noted that from Hildebrand's proof one can get an effective value for $c(k)$ (as was pointed out by the anonymous referee).

It makes sense to restate Hildebrand's result as follows:

Theorem 3 (Hildebrand). *Let $H \le \mathbb{Q}^+$ be a (multiplicative) subgroup such that $\mathbb{Q}^+/H$ is cyclic of finite order. Let $H^* := H \cap \mathbb{Z}^+$. Then $H^* \cap (H^* - 1)$ is nonempty.*

The original proof made use of analytic methods and was rather long. We will give a short elementary proof of a more general theorem.

However, before we can state (and prove) our generalization we need some notation and the set-theoretical version of Hindman's theorem:

We denote by $\mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+)$ the set of finite, nonempty subsets of $\mathbb{Z}^+$.

For $A, B \in \mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+)$ write $A \prec B$ if $\max A < \min B$.

Furthermore, for a sequence $A_1 \prec A_2 \prec \cdots$ in $\mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+)$, we define

$$\mathrm{FU}((A_i)_{i\in\mathbb{Z}^+}) = \left\{ \bigcup_{i\in I} A_i : I \subseteq \mathbb{Z}^+,\ 0 < |I| < \infty \right\}.$$

Similarly, for a sequence $a_1, a_2, \ldots$ in $\mathbb{Z}^+$, we define

$$\mathrm{FS}((a_i)_{i\in\mathbb{Z}^+}) = \left\{ \sum_{i\in I} a_i : I \subseteq \mathbb{Z}^+,\ 0 < |I| < \infty \right\}.$$

MSC2010: 11B75.
Keywords: IP-set, multiplicative subgroup.

We call a set $M \subseteq \mathbb{Z}^+$ an *IP-set* [Hindman and Strauss 2012, Definition 16.3] if there is a sequence $a_1, a_2, \ldots$ in $\mathbb{Z}^+$ such that $\mathrm{FS}((a_i)_{i\in\mathbb{Z}^+}) \subseteq M$.

If a set A is the disjoint union of subsets $B_1, \ldots, B_n \subseteq A$, that is, $B_1 \cup \cdots \cup B_n = A$ and $B_i \cap B_j = \varnothing$ for $1 \le i < j \le n$, we denote this relation by $A = B_1 \sqcup \cdots \sqcup B_n$.

Now Hindman's theorem on partitions of $\mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+)$ [Hindman and Strauss 2012, Corollary 5.17] can be stated as follows:

Theorem 4 (Hindman). *For any finite partition $\mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+) = M_1 \sqcup M_2 \sqcup \cdots \sqcup M_n$ there are sets $A_1 \prec A_2 \prec \cdots$ and $1 \le j \le k$ such that*

$$\mathrm{FU}((A_i)_{i\in\mathbb{Z}^+}) \subseteq M_j.$$

We can now state our generalization of Hildebrand's theorem:

Theorem 5. *Let $H \le \mathbb{Q}^+$ be a (multiplicative) subgroup of finite index.*[1] *Let $H^* := H \cap \mathbb{Z}^+$. Then $H^* \cap (H^* - 1)$ is an IP-set.*

Hildebrand's proof of Theorem 3 is an application of Ramsey's theorem on *special* sets, i.e., finite sets $\{n_1 < n_2 < \cdots < n_r\}$ such that $n_j - n_i = \gcd(n_i, n_j)$ holds for $1 \le i < j \le r$.

We will use a similar concept:

Definition 6. For a sequence s_n and a finite subset $A \subset \mathbb{Z}^+$, set

$$s_A := \sum_{n\in A} s_n.$$

A *block-divisible sequence* is a strictly decreasing sequence s_n in $\mathbb{Z}^+$ such that for $A, B \in \mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+)$, s_A divides s_B whenever $A \prec B$.

For our proof, *any* block-divisible sequence will work. Thus, we only need to confirm the existence of block-divisible sequences:

Lemma 7. *There is a block-divisible sequence in $\mathbb{Z}^+$.*

Proof. We construct a sequence as follows:

$$s_0 := 1, \quad s_{n+1} := \prod_{\substack{A\subseteq\{0,\ldots,n\}\\ A\neq\varnothing}} s_A.$$

Ignoring the s_0 at the beginning, we end up with a strictly increasing sequence fulfilling the desired divisibility condition. □

Now we can show our main result:

Proof of Theorem 5. Let N_i' $(1 \le i \le k)$ be the (multiplicative) cosets of H in $\mathbb{Q}^+$.

These give a finite partition $\mathbb{Z}^+ = N_1 \sqcup N_2 \sqcup \cdots \sqcup N_k$, where $N_i = N_i' \cap \mathbb{Z}^+$.

We now fix a block-divisible sequence s_n (whose existence is guaranteed by Lemma 7) and define a partition $\mathcal{P}^{\mathrm{fin}}(\mathbb{Z}^+) = M_1 \sqcup M_2 \sqcup \cdots \sqcup M_k$ by declaring $A \in M_i$ if and only if $s_A \in N_i$.

By Theorem 4 there is a sequence $A_1 \prec A_2 \prec \cdots$ such that $\mathrm{FU}(A_1, A_2, \ldots)$ is contained in one M_i for some $1 \le i \le k$.

[1] Note that we do not require $\mathbb{Q}^+/H$ to be cyclic.

By the definition of block-divisibility, s_{A_1} divides s_A for all $A \in \mathrm{FU}(A_2, A_3, \ldots)$ and, consequently, for all $A \in \mathrm{FU}(A_1, A_2, \ldots)$, too.

Thus, defining $b_i := s_{A_i}$, the members of $\mathrm{FS}(b_1, b_2, \ldots)$ all lie in the same coset of H and are divisible by b_1. Therefore, setting $a_i := b_i/b_1$, one has

$$\mathrm{FS}(a_1, a_2, \ldots) = \mathrm{FS}(1, a_2, a_3, \ldots) \subseteq H^*.$$

Furthermore, $\mathrm{FS}(1, a_2, a_3, \ldots) = \mathrm{FS}(a_2, a_3, \ldots) \cup (\mathrm{FS}(a_2, a_3, \ldots) + 1) \subseteq H^*$.

We conclude that $\mathrm{FS}(a_2, a_3, \ldots) \subseteq H^* \cap (H^* - 1)$. □

Remark 8. We use the terminology of Theorem 5 to summarize the state of possible generalizations:

There are (multiplicative) subgroups H of arbitrary even index in $\mathbb{Q}^+$ such that $H^* \cap (H^* - 1) \cap (H^* - 2)$ is empty, as has been shown by Lehmer and Lehmer [1962, p. 103].

Graham [1964] proved that there are subgroups of arbitrary (finite) index in $\mathbb{Q}^+$ such that $H^* \cap \cdots \cap (H^* - 3)$ is empty.

However, if $\mathbb{Q}^+/H$ is of odd order k, it is still an open question if $H^* \cap (H^* - 1) \cap (H^* - 2)$ is necessarily nonempty. Only in the case $k = 3$ is this set known to be always nonempty, as has been shown computationally by Lehmer, Lehmer, Mills and Selfridge [Lehmer et al. 1962]. Maybe the combinatorial methods presented in this article may help in resolving this problem!

Remark 9. Some ideas shown in this article are based on notes of the author, [Dietzel 2013], which have not been submitted to any journal.

References

[Dietzel 2013] C. Dietzel, "A generalization of Schur's theorem and its application to consecutive power residues", preprint, 2013. arXiv

[Graham 1964] R. L. Graham, "On quadruples of consecutive kth power residues", *Proc. Amer. Math. Soc.* **15** (1964), 196–197. MR Zbl

[Hildebrand 1991] A. Hildebrand, "On consecutive kth power residues, II", *Michigan Math. J.* **38**:2 (1991), 241–253. MR Zbl

[Hindman and Strauss 2012] N. Hindman and D. Strauss, *Algebra in the Stone–Čech compactification: theory and applications*, 2nd ed., Walter de Gruyter & Co., Berlin, 2012. MR Zbl

[Lehmer and Lehmer 1962] D. H. Lehmer and E. Lehmer, "On runs of residues", *Proc. Amer. Math. Soc.* **13** (1962), 102–106. MR Zbl

[Lehmer et al. 1962] D. H. Lehmer, E. Lehmer, W. H. Mills, and J. L. Selfridge, "Machine proof of a theorem on cubic residues", *Math. Comp.* **16** (1962), 407–415. MR Zbl

[Lehmer et al. 1963] D. H. Lehmer, E. Lehmer, and W. H. Mills, "Pairs of consecutive power residues", *Canad. J. Math.* **15** (1963), 172–177. MR Zbl

Received 29 Jan 2019. Revised 7 Feb 2019.

CARSTEN DIETZEL:

carstendietzel@gmx.de

Institute of algebra and number theory, University of Stuttgart, Stuttgart, Germany

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles are usually in English or French, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not refer to bibliography keys. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables.

White Space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Moscow Journal of Combinatorics and Number Theory
vol. 8 no. 2 2019