ON THE HEREDITARY DISCREPANCY OF HOMOGENEOUS ARITHMETIC PROGRESSIONS

ALEKSANDAR NIKOLOV AND KUNAL TALWAR

Abstract. We show that the hereditary discrepancy of homogeneous arithmetic progressions is lower bounded by $n^{1/O(\log\log n)}$. This bound is tight up to the constant in the exponent. Our lower bound goes via proving an exponential lower bound on the discrepancy of set systems of subcubes of the boolean cube $\{0,1\}^d$.

1. Introduction

Circa 1932 Paul Erdős made the following conjecture:

Conjecture 1.1 ([2])**.** For any function $f : \mathbb{N} \to \{-1, +1\}$ and for any constant C, there exist positive integers n and a such that

$$|\sum_{i=1}^{\lfloor n/a \rfloor} f(ia)| > C.$$

In modern terms, the maximum value of $\sum_{i=1}^{k} f(ia)$ over all a and all $k \le n/a$ is the *discrepancy* of f over arithmetic progressions of the form $(ia)_{i=1}^{\lfloor n/a \rfloor}$. As is customary, we shall call such arithmetic progressions *homogeneous*. The minimum discrepancy over all functions f that take values in $\{-1, 1\}$ is the discrepancy of homogeneous arithmetic progressions up to n. In this language, Conjecture 1.1 states that the discrepancy of homogeneous arithmetic progressions is unbounded.

This problem is now known as the Erdős discrepancy problem, and stands as a major open problem in discrepancy theory and combinatorial number theory. It is known that the discrepancy of homogeneous arithmetic progressions is at least 2. On the other hand, the function f which takes value $f(i) = -1$ if and only if the last nonzero digit of i in ternary representation is 2 has discrepancy $O(\log n)$. For references and other partial results related to the Erdős discrepancy problem, see [3, 1].

The Erdős discrepancy problem recently received attention as the subject of the fifth polymath project [1]. Our note is motivated by recent results of Alon and Kalai, announced and sketched in the weblog post [4]. Using the Beck-Fiala theorem, they showed that even for homogeneous arithmetic progressions restricted to an arbitrary subset of the integers up to n, the discrepancy is no more than $n^{1/\Omega(\log\log n)}$. Also, for infinitely many n, they constructed a set of integers W_n all bounded by n, so that there is a set of homogeneous arithmetic progressions which, when restricted to W_n, form a known high discrepancy set system (the Hadamard set system). This construction showed that the minimum discrepancy for homogeneous arithmetic progressions restricted to W_n is at least $\Omega(\sqrt{\log n}/\sqrt{\log\log n})$. Since their discrepancy upper bound only uses a bound on the number of distinct

arithmetic progressions an integer can belong to, it was reasonable to guess that the lower bound is closer to the truth.

In this note we show that in fact it is the upper bound of Alon and Kalai which is tight up to the constant in the exponent. Namely, for infinitely many n, we construct a set of integers W_n all bounded by n, so that the discrepancy of homogeneous arithmetic progressions restricted to W_n is at least $n^{1/O(\log\log n)}$. Our construction of the sets W_n is inspired by the construction of Alon and Kalai. Instead of the Hadamard set system, we embed a set system of subcubes of the boolean cube inside the set of homogeneous arithmetic progressions. We prove a lower bound on the set system of boolean subcubes using the determinant lower bound on hereditary discrepancy due to Lovász, Spencer, and Vesztergombi [6]. Such systems of boolean subcubes were previously considered in computer science in the context of private data analysis [5, 8].

Our construction produces sets W_n of square free integers with a large number of prime divisors, suggesting that such integers are a chief obstacle in achieving bounded discrepancy for homogeneous arithmetic progressions.

2. Preliminaries

For a positive integer n, let $[n]$ be the set $\{1,\ldots,n\}$. Given a set S, let $\binom{S}{k}$ be the set of cardinality k subsets of S. The expression $\langle\cdot,\cdot\rangle_2$ denotes the inner product over the field $\mathbb{F}_2^d$, and $|\cdot|$ — the hamming weight of a 0-1 vector.

A *set system* is defined as a pair $(\mathcal{S},U)$, where $\mathcal{S}=\{S_1,\ldots,S_m\}$ and $\forall j\in[m]: S_j\subseteq U$. The *restriction* $(\mathcal{S}|_W,W)$ of a set system $(\mathcal{S},U)$ to some $W\subseteq U$ is defined by $\mathcal{S}|_W=\{S_1\cap W,\ldots,S_m\cap W\}$.

The discrepancy and hereditary discrepancy of a set system $\mathcal{S}$ are defined as

$$\operatorname{disc}(\mathcal{S}) = \min_{f:U\to\{-1,+1\}} \max_{j\in[m]} \left|\sum_{i\in S_j} f(i)\right|$$
$$\operatorname{herdisc}(\mathcal{S}) = \max_{W\subseteq U} \operatorname{disc}(\mathcal{S}|_W)$$

The definitions of discrepancy and hereditary discrepancy can be extended to matrices $A\in\mathbb{R}^{m\times n}$ in a natural way. Analogously to the definition of a restriction of a set system, we define a restriction $A|_W$ of $A\in\mathbb{R}^{m\times n}$ for $W\subseteq[n]$ as the submatrix of columns of A indexed by elements of S. Then discrepancy and hereditary discrepancy for matrices are defined as

$$\operatorname{disc}(A) = \min_{x\in\{-1,+1\}^n} \|Ax\|_\infty$$
$$\operatorname{herdisc}(A) = \max_{W\subseteq[n]} \operatorname{disc}(A|_W)$$

We shall need the determinant lower bound on hereditary discrepancy, due to Lovász, Spencer, and Vesztergombi.

Theorem 2.1 ([6])**.** *For any real $m\times n$ matrix A,*

$$\operatorname{herdisc}(A) \geq \frac{1}{2}\max_k \max_B |\det(B)|^{1/k}$$

where, for any k, B ranges over all $k\times k$ submatrices of A.

3. Main Theorem

Theorem 3.1. *For infinitely positive integers n, there exists a set $W \subseteq \{1, \ldots, n\}$ of square free integers such that the following holds. For any $f : W \to \{-1, +1\}$ there exists a positive integer a so that*

$$|\sum_{\substack{b \in W \\ a|b}} f(b)| = n^{1/O(\log\log n)}.$$

Theorem 3.1 is a consequence of a lower bound on the hereditary discrepancy of the set system of subcubes of the boolean cube. Next we define this set system formally. For a positive integer d, we define the set system $(\mathcal{S}^d, \{0,1\}^d)$, where $\mathcal{S}^d = \{S_v\}_{v \in \{0,1,*\}^d}$ is defined by

$$S_v = \{u \in \{0,1\}^d : v_i \neq * \Rightarrow u_i = v_i\}.$$

Similar set systems were studied in computer science in relation to computing conjunction queries on a binary database under the constraint of differential privacy [5, 8]. The least singular value lower bounds in [5, 8] imply discrepancy lower bounds for subsets of $\mathcal{S}^d$ corresponding to the subcubes with constant co-dimension. In this note we need a quantitatively stronger lower bound on the discrepancy of the entire set system $\mathcal{S}^d$.

Lemma 3.2. *For all positive integers d,* $\mathrm{herdisc}(\mathcal{S}^d) = 2^{\Omega(d)}$.

In the remainder of this section we prove that Lemma 3.2 implies Theorem 3.1. We prove Lemma 3.2 in the subsequent section.

Proof of Theorem 3.1. For each positive integer d, we will construct a set of integers B such that the hereditary discrepancy of homogeneous arithmetic progressions restricted to B is lower bounded by the hereditary discrepancy of $\mathcal{S}^d$. Then Theorem 3.1 will follow from Lemma 3.2.

Let $p_{1,0} < p_{1,1} < \ldots < p_{d,0} < p_{d,1}$ be the first $2d$ primes. We define B to be the following set of square free integers

$$B = \{\prod_{i=1}^{d} p_{i,u_i} : u \in \{0,1\}^d\}.$$

In other words, B is the set of all integers that are divisible by exactly one prime $p_{i,b}$ from each pair $(p_{i,0}, p_{i,1})$ and no other primes. By the prime number theorem, $p_{d,1} = \Theta(d \log d)$. Let $n = n(d)$ be the largest integer in B. The crude bound $n(d) = 2^{O(d \log d)}$ will suffice for our purposes. Notice that $d = \Omega(\log n / \log\log n)$.

There is a natural one to one correspondence between the set B and the set $\{0,1\}^d$: to each $u \in \{0,1\}^d$ we associate the integer $b_u = \prod_{i=1}^{d} p_{i,u_i}$. By this correspondence, we can think of any assignment $f : \{0,1\}^d \to \{-1,+1\}$ as an assignment $f : B \to \{-1,+1\}$. We also claim that each set in the set system $\mathcal{S}^d$ corresponds to a homogeneous arithmetic progression restricted to B. With any $S_v \in \mathcal{S}^d$ (where $v \in \{0,1,*\}^d$) associate the integer $a_v = \prod_{i:v_i \neq *} p_{i,v_i}$. Observe that for any $b_u \in B$, a_v divides b_u if and only if $u \in S_v$. We have the following implication for any assignment f, any $U \subseteq \{0,1\}^d$, and the corresponding $W = \{b_u : u \in U\}$:

$$\exists S_v : |\sum_{u \in S_v \cap U} f(u)| \geq D \quad \Leftrightarrow \quad \exists a \in \mathbb{N} : |\sum_{\substack{b \in W \\ a|b}} f(b)| \geq D. \tag{3.1}$$

Notice again that we treat f as an assignment both to elements of $\{0,1\}^d$ and to integers in B by the correspondence $u \leftrightarrow b_u$. Lemma 3.2 guarantees the existence of some U such that the left hand side of (3.1) is satisfied with $D = 2^{\Omega(d)} = n^{1/O(\log\log n)}$ for any f. Theorem 3.1 follows from the right hand side of (3.1). □

4. Lower Bounding the Discrepancy of $\mathcal{S}^d$

It is convenient to first prove an easier lower bound on the hereditary discrepancy of low-weight characters of $\mathbb{F}_2^d$. Then we show that an exponential (in d) lower bound on the discrepancy of characters of weight $d/8$ implies an exponential lower bound on $\mathcal{S}^d$. This approach is inspired by the noise lower bounds on differential privacy in [5].

As usual, for $v \in \{0,1\}^d$ we define the character χ_v by

$$\forall u \in \{0,1\}^d : \chi_v(u) = (-1)^{\langle v,u\rangle_2}.$$

We refer to $|v|$ as the *weight* of the character χ_v. The matrix of the Walsh-Hadamard transform is defined as $H_d = (\chi_v)_{v\in\{0,1\}^d}$, where each χ_v is written as a row vector of dimension 2^d. Notice that for any $v \neq w$, $\sum_{u\in\{0,1\}^d} \chi_v(u)\chi_w(u) = 0$, i.e. H_d is an orthogonal matrix; each row of H_d has squared euclidean norm $\sum_{u\in\{0,1\}^d} \chi_v(u)^2 = 2^d$.

We will be interested in a submatrix of H_d. For the remainder of this note we assume that d is divisible by 8; this is purely for notational convenience: our arguments can easily be adapted to the case when d is not divisible by 8. Let $G_d = (\chi_v)_{v:|v|=d/8}$. Notice that $G_d G_d^T = 2^d I_M$ where $M = \binom{d}{d/8}$ and I_M is the M-dimensional identity matrix. Therefore,

$$\det(G_d G_d^T) = (2^d)^{\binom{d}{d/8}} \tag{4.1}$$

Given (4.1) and using the determinant lower bound, we can derive a lower bound on the hereditary discrepancy of G_d.

Lemma 4.1. *For positive integers d,*

$$\operatorname{herdisc}(G_d) \geq \frac{2^{3d/16}}{2e}$$

Proof. Let $N = 2^d$ and let $M = \binom{d}{d/8}$. By (4.1) and the Binet-Cauchy formula for the determinant, we have

$$N^M = \det(G_d G_d^T) = \sum_{W\in\binom{[N]}{M}} \det(G_d|_W)^2$$

By averaging, there exists a set $W \in \binom{[N]}{M}$ so that

$$|\det(G_d|_W)|^{1/M} \geq \sqrt{N}\binom{N}{M}^{-1/2M} \geq \sqrt{\frac{M}{e}} \tag{4.2}$$

For the second inequality above we used the bound $\binom{N}{M} \leq (Ne/M)^M$. Plugging in the lower bound $M = \binom{d}{d/8} \geq 2^{3d/8}$ in (4.2), we have $|\det(G_d|_W)|^{1/M} \geq 2^{3d/16}e^{-1}$. The proof is completed by an application of Theorem 2.1. □

We are now ready to prove Lemma 3.2 by exhibiting a connection between the discrepancy of G_d and the discrepancy of $\mathcal{S}^d$.

Proof of Lemma 3.2. It is enough to prove the following inequality:

$$\operatorname{herdisc}(G_d) \le 2^{d/8} \operatorname{herdisc}(\mathcal{S}^d) \tag{4.3}$$

The key observation is that we can express the character χ_v as a linear combination of the indicator functions of $2^{d/8}$ sets in $\mathcal{S}^d$. Moreover, the coefficients of the linear combination are ± 1. Next we make this observation precise.

Fix arbitrary $U \subseteq \{0,1\}^d$, and $v \in \{0,1\}^d$ such that $|v| = d/8$. Also, for $w \in \{0,1\}^{d/8}$, let its extension $w' \in \{0,1,*\}^d$ be defined by

$$w'_i = \begin{cases} w_i & v_i = 1 \\ * & \text{otherwise} \end{cases}$$

We use the notation $\mathbf{e}$ for the $d/8$-dimensional all-ones vector $(1,\dots,1)$ and $\mathbf{1}_{w'}$ for the indicator function of the set $S_{w'}$. Notice that the value $\chi_v(u) = (-1)^{\langle v,u\rangle_2}$ only depends on those components u_i of u where $v_i \neq 0$. Therefore, we can express $\chi_v(u)$ as the linear combination

$$\forall u \in \{0,1\}^d : \chi_v(u) = \sum_{w \in \{0,1\}^{d/8}} (-1)^{\langle \mathbf{e}, w\rangle_2} \mathbf{1}_{w'}(u). \tag{4.4}$$

Using (4.4) we can write $(G_d|_U)f$ in terms of discrepancy values of sets in $\mathcal{S}^d$:

$$\begin{aligned} \sum_{u\in U} \chi_v(u) f(u) &= \sum_{u\in U} \left(\sum_{w \in \{0,1\}^{d/8}} (-1)^{\langle \mathbf{e}, w\rangle_2} \mathbf{1}_{w'}(u) \right) f(u) \\ &= \sum_{w \in \{0,1\}^{d/8}} (-1)^{\langle \mathbf{e}, w\rangle_2} \left(\sum_{u \in S_{w'} \cap U} f(u) \right). \end{aligned} \tag{4.5}$$

Each of the $2^{d/8}$ terms on the right hand side of (4.5) is bounded by $\operatorname{herdisc}(\mathcal{S}^d)$. Since the choice of U and v was arbitrary, the lemma follows. □

5. Conclusion

We presented a tight (up to the constant in the exponent) lower bound on the hereditary discrepancy of homogeneous arithmetic progressions. Our lower bound instances are given by a set of integers in the interval $[1,n]$ with a large number $\Theta(\log n/\log\log n)$ of distinct prime factors. This suggests that integers with many distinct factors are the main obstacle to achieving bounded discrepancy for homogeneous arithmetic progressions.

Our discrepancy lower bound follows from a lower bound on the discrepancy of a set system of subcubes of the boolean cube. Such set systems have applications in the theory of differential privacy. The proof of Lemma 3.2 together with the connection between discrepancy and differential privacy formalized in [7] can be used to give simpler proofs of the noise lower bounds of the type considered in [5]. It is an interesting question whether discrepancy bounds on set systems of boolean subcubes can find other applications in combinatorics and computer science. We leave open the question of characterizing the exact discrepancy of such set systems.

References

1. Multiple Authors, *Erdős discrepancy problem: Polymath wiki*, http://michaelnielsen.org/polymath1/index.php?title=The_Erd%C5%91s_discrepancy_problem.
2. P. Erdős, *Some unsolved problems.*, The Michigan Mathematical Journal **4** (1957), no. 3, 291–300.
3. Steven Finch, *Two-colorings of positive integers*, http://www.people.fas.harvard.edu/~sfinch/csolve/ec.pdf.
4. Gil Kalai, *Erdős discrepancy problem 22*, http://gowers.wordpress.com/2012/08/22/edp22-first-guest-post-from-gil-kalai 09 2012.
5. S.P. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman, *The price of privately releasing contingency tables and the spectra of random matrices with correlated rows*, Proceedings of the 42nd ACM symposium on Theory of computing, ACM, 2010, pp. 775–784.
6. L. Lovász, J. Spencer, and K. Vesztergombi, *Discrepancy of set-systems and matrices*, European Journal of Combinatorics **7** (1986), no. 2, 151–160.
7. S. Muthukrishnan and Aleksandar Nikolov, *Optimal private halfspace counting via discrepancy*, Proceedings of the 44th symposium on Theory of Computing (New York, NY, USA), STOC '12, ACM, 2012, pp. 1285–1292.
8. M. Rudelson, *Row products of random matrices*, Arxiv preprint arXiv:1102.1947 (2011).