The discrepancy of permutation families[*]

J. H. Spencer[†] A. Srinivasan[‡] P. Tetali[§]

Abstract

In this note, we show that the discrepancy of any family of ℓ permutations of $[n] = \{1, 2, \ldots, n\}$ is $O(\sqrt{\ell} \log n)$, improving on the $O(\ell \log n)$ bound due to Bohus (*Random Structures & Algorithms*, 1:215–220, 1990). In the case where $\ell \geq n$, we show that the discrepancy is $\Theta(\min\{\sqrt{n \log(2\ell/n)}, n\})$.

Key Words and Phrases. Discrepancy, probabilistic method, permutations, geometric discrepancy.

1 Introduction

Discrepancy theory, the study of *uniform distributions and irregularities of distribution*, arises in many branches of mathematics and has a rich combinatorial aspect; see the chapter by Beck & Sós [3]. The discrepancy $\mathrm{disc}(A)$ of an $m \times n$ matrix A is defined to be $\min\{\|A\chi\|_\infty : \chi \in \{-1, 1\}^n\}$. The discrepancy $\mathrm{disc}(H)$ of a set-system H is defined to be the discrepancy of its vertex-edge incidence matrix. For any positive integer $\ell \leq n!$ and any set $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_\ell\}$ of ℓ permutations of $[n] = \{1, 2, \ldots, n\}$, define $P_\sigma(n)$ to be the set-system $\{\{\sigma_k(i), \sigma_k(i+1), \ldots, \sigma_k(j)\} : k \in [\ell],\ 1 \leq i \leq j \leq n\}$, defined on the ground set $[n]$. Define $D_\ell(n) = \max_\sigma \mathrm{disc}(P_\sigma(n))$. It is known that $D_2(n) \leq 2$, and a major open question, due to Beck, is whether $D_3(n) = O(1)$; this is problem 1.9 of [3]. Classical discrepancy results show constructively that $D_\ell(n) = O(\sqrt{n \log(\ell + n)})$ [8]. The best known result prior to our work is that $D_\ell(n) = O(\ell \log n)$, due to Bohus [4]. We improve this to show that

[*]A preliminary version of part of this work appeared in the *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 692–701, 1997.

[†]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012. Email: spencer@cs.nyu.edu; research supported in part by NSF grant DMS 9970822.

[‡]Department of Computer Science and University of Maryland Institute for Advanced Computer Studies, University of Maryland at College Park, College Park, MD 20742. E-mail: srin@cs.umd.edu. Parts of this work were done: (i) while at the School of Computing, National University of Singapore, Singapore 119260, Republic of Singapore, and (ii) while visiting the Department of Computer Science, University of Melbourne, Victoria 3052, Australia, sponsored by a "Travel Grants for Young Asian Scholars" scheme of the University of Melbourne.

[§]School of Mathematics and College of Computing, Georgia Institute of Technology, Atlanta, GA 30332. Email: tetali@math.gatech.edu; research supported in part by NSF grant DMS 0100298.

$D_\ell(n) = O(\sqrt{\ell}\log n)$. We also show that $D_\ell(n) = \Theta(\min\{\sqrt{n\log(2\ell/n)}, n\})$ if $\ell \geq n$; thus, the dependence on ℓ in our "$O(\sqrt{\ell}\log n)$" bound cannot be improved significantly.

2 Discrepancy upper bounds for permutation families

There is an elegant approach to discrepancy via entropy and the pigeonhole principle [1, 7, 5]. Let $A \in \{0,1\}^{m\times n}$, and let cols$(A)$ denote the set of columns of A. For any $i \in [m]$, let nz(i, A) denote the number of nonzero entries in row i of A. A *partial coloring* is a map $\chi : \text{cols}(A) \to \{-1, 0, 1\}$. For any $v \in \text{cols}(A)$, we call v uncolored if $\chi(v) = 0$, and colored otherwise. Let $\exp(x) \doteq 2^x$, and let $\log x$ denote the logarithm of x to the base 2. For a certain absolute constant $c > 0$ and for any $\lambda > 0$, define $G(\lambda)$ to be: $c\cdot\exp(-\lambda^2/9)$, if $\lambda > 10$; c, if $0.1 \leq \lambda \leq 10$, and $c\cdot\log(\lambda^{-1})$, if $\lambda \in (0, 0.1)$.

We get, from Corollary 2.4 of [5]:

Theorem 2.1 ([5]) *Suppose $A \in \{0,1\}^{m\times n}$ has at most $f(s)$ rows with s nonzero entries, for each $s \geq 1$. If $b(s) = \sigma(s)\sqrt{s}$ where $\sigma(s)$ satisfies $\sum_{s\geq 1} f(s)G(\sigma(s)) \leq n/5$, then there is a partial coloring χ of cols(A) with $|(A\chi)_i| \leq b(nz(i, A))$ for each $i \in [m]$, and with at least half of the columns of A colored.*

Theorem 2.2 *Given ℓ permutations on n points, there exists a partial coloring with discrepancy $O(\sqrt{\ell})$, and with at most half the points uncolored.*

Proof. For convenience, we assume here that both ℓ and n are powers of 2. (It is straightforward to remove this assumption at the loss of at most a multiplicative constant in the discrepancy.) Given the ℓ permutations, for each permutation we make canonical sets - partitioning into $n2^{-i}$ sets of size 2^i, for each i. More precisely, for each given permutation σ and for every integer $1 \leq i \leq \log n$, we construct $n2^{-i}$ sets of size 2^i each, as follows: each such set is defined to be the set of points ordered from $(j-1)\cdot 2^i + 1$ to $j\cdot 2^i$ by σ, for some positive integer j. Summing over all the ℓ permutations, we get a total of at most $n2^{-j}$ sets of size $\ell 2^j$ for each integer j such that $-\log\ell \leq j \leq \log(n/\ell)$.

We want a coloring so that for some constant $C > 0$, sets of size $\ell 2^j$ have discrepancy at most $C(j+1)^{-2}\sqrt{\ell}$ for $j \geq 0$, and sets of size $\ell 2^{-k}$ have discrepancy at most $Ck^{-2}\sqrt{\ell}$ for $k \geq 1$. Then any interval is the difference of two intervals beginning at the beginning, each of which is the union of canonical sets of *distinct* sizes of the form $\ell\cdot 2^j$; so, the discrepancy of an interval would be at most $O(\sum_{j\geq 1} j^{-2}\sqrt{\ell}) = O(\sqrt{\ell})$, using the convergence of $\sum_j j^{-2}$. In view of Theorem 2.1, it suffices to show that

$$[\sum_{j\geq 0} n2^{-j}G(\frac{C}{2^{j/2}(j+1)^2})] + \sum_{k\geq 1} n2^k G(\frac{C2^{k/2}}{k^2}) \leq n/5. \quad (1)$$

We now show that this holds if C is large enough.

We will ensure that $C > 100$. Let j_1 be the smallest non-negative integer such that $\frac{C}{2^{j/2}(j+1)^2} < \sqrt{C}$; we have $j_1 \geq c_0 \log C$ for some absolute constant $c_0 > 0$ (which, in particular, is independent of C). Since $\sqrt{C} > 10$, we see that $G(\frac{C}{2^{j/2}(j+1)^2}) \leq \exp(-C/9)$ for $0 \leq j < j_1$. Also, it is easy to verify that for $j \geq j_1$, $G(\frac{C}{2^{j/2}(j+1)^2}) \leq c_1 j$ for some absolute constant c_1. Finally, since $C > 100$, $G(\frac{C2^{k/2}}{k^2}) \leq \exp(-C^2 2^k/(9k^4))$ for all $k \geq 1$. These ideas show that the l.h.s. of (1) is at most

$$[\sum_{0 \leq j < j_1} n2^{-j} \exp(-C/9)] + [\sum_{j \geq c_0 \log C} n2^{-j} \cdot c_1 j] + \sum_{k \geq 1} n2^k \exp(-C^2 2^k/(9k^4)),$$

which is at most $n/5$ if C is large enough. □

So, if we remove the colored points and iterate, we will get a coloring in $O(\log n)$ iterations. Thus we have:

Corollary 2.1 *$D_\ell(n) = O(\sqrt{\ell} \log n)$.*

An application to geometric discrepancy. As usual, define a rectangle in $\Re^k$ to be a cross-product of k real intervals. Given a set S of n points lying in $\Re^k$, let E_S be the set of rectangles that enclose the n points, and let R_S be the set-system $\{(S \cap R) : R \in E_S\}$, defined over the ground-set S. (Note that R_S has at most n^{2k} distinct nonempty elements.) Define $G_k(n) = \sup_S \text{disc}(R_S)$. $G_k(n)$ has been studied for fixed k, for a while; see [3]. It is conjectured in [3] that $G_k(n)$ has order of magnitude $\log^{k-1} n$. It follows from the work of [4] that for any fixed k, $G_k(n) = O(D_\ell(n) \log^{k-1} n)$, where $\ell = (1 + \log n)^{k-1}$. So, we get that $G_k(n) = O(\log^{(3k-1)/2} n)$ for fixed k.

2.1 The case $\ell \geq n$

Theorem 2.3 *If $\ell \geq n$, then $D_\ell(n) = O(\min\{\sqrt{n \log(2\ell/n)}, n\})$.*

Proof. Suppose $\ell \geq n$. Clearly, it suffices to show that $D_\ell(n) = O(\sqrt{n \log(2\ell/n)})$. As in [7] (and as in [5]), we now repeat the argument of Theorem 2.2, taking into account the fact that at each step the number of points remaining to be colored is decreasing by a constant fraction.

For the general step, suppose that we have ℓ permutations on $2^{-s}n$ points, for $s \geq 0$. As before, consider the partitioning of each permutation into canonical sets: we have ℓ sets of size $n2^{-s}$, 2ℓ sets of size $n2^{-s-1}$, and in general, $2^j\ell$ sets of size $n2^{-s-j}$ for $j \geq 0$. We want a coloring so that sets of size $n2^{-s-j}$ have discrepancy at most

$$C\sqrt{(s+j)n2^{-s-j}\log(2\ell/n)};$$

then the final discrepancy of an interval (at the end of the $O(\log n)$ iterations) would be at most

$$O(\sum_{j,s \geq 0} \sqrt{(s+j)n2^{-s-j}\log(2\ell/n)}) = O(\sqrt{n \log(2\ell/n)}).$$

Once again, in view of Theorem 2.1, we get the above discrepancy as long as we have $\sum_{j\geq 0} \ell 2^j G(C\sqrt{(s+j)\log(2\ell/n)}) \leq n2^{-s}/5$. Choosing $C > 10$, it suffices to have

$$\sum_{j\geq 0} \ell 2^j \exp(-C^2(s+j)\log(2\ell/n)/9) \leq n2^{-s}/5,$$

which is true for C large enough. □

We next present a simple probabilistic proof to show that Theorem 2.3 is tight:

Theorem 2.4 *If $\ell \geq n$, then $D_\ell(n) = \Omega(\min\{\sqrt{n\log(2\ell/n)}, n\})$.*

Proof. We will assume throughout that n is large enough, and for notational convenience, will avoid the use of the floor and ceiling signs in several places.

First, for a certain constant $\epsilon \in (0,1)$ that will be appropriately chosen later, we may assume that $\ell \leq 2^{\epsilon n}$, for the following reason. There is some constant $\delta \in (0, 1/2)$ such that for large enough n, $\binom{n}{\delta n} \leq 2^{\epsilon n}$. (Note that for a fixed choice of δ, we may choose ϵ accordingly, e.g., $\epsilon > H(\delta)$ suffices, where $H(.)$ denotes the binary entropy function.) Thus, if $\ell > 2^{\epsilon n}$, we can construct ℓ permutations such that the set of first δn elements of these permutations equals the set of all δn–sized subsets of $[n]$. Now, since $\delta \leq 1/2$, it is easy to see that any two-coloring will lead to at least one of the ℓ permutations having its first δn elements monochromatic. Thus, if $\ell > 2^{\epsilon n}$, then $D_\ell(n) \geq \delta n = \Omega(n)$, and we will be done. Thus we can assume that $\ell \leq 2^{\epsilon n}$, and our approach will be the following simple probabilistic one.

Let $a > 0$ be a constant that will be chosen small enough later, and let $v_1, v_2, \ldots, v_n$ be arbitrary elements of $\{-1, 1\}$. Suppose we can show, for a subset S of $[n]$ chosen at random from the uniform distribution, that

$$\Pr[|\sum_{j\in S} v_{\rho_j}| \geq a\sqrt{n\ln(2\ell/n)}] > (\ln 2)\cdot n/\ell. \tag{2}$$

Next suppose we choose subsets $S_1, S_2, \ldots, S_\ell$ of $[n]$ uniformly at random and independently. Then, (2) shows that

$$\Pr[\bigwedge_{i=1}^{\ell}(|\sum_{j\in S_i} v_{\rho_j}| < a\sqrt{n\ln(2\ell/n)})] < (1-(\ln 2)\cdot n/\ell)^\ell < 2^{-n}.$$

So, since there are only 2^n choices for $(v_1, v_2, \ldots, v_n)$, there exist subsets $S_1^*, S_2^*, \ldots, S_\ell^*$ of $[n]$ such that for each choice of $(v_1, v_2, \ldots, v_n)$, there is some S_i^* such that $|\sum_{j\in S_i^*} v_{\rho_j}| \geq a\sqrt{n\ln(2\ell/n)}$. Now construct permutations $P_1, P_2, \ldots P_\ell$ such that each S_i^* is a prefix (under an arbitrary ordering of the elements of S_i^*) of P_i; this will show that $D_\ell(n) \geq a\sqrt{n\ln(2\ell/n)}$. So, we now proceed to show that (2) holds for a small enough, using appropriate estimates of the binomial coefficients; these calculations are quite routine.

For $n/2 \leq k \leq n$, a standard estimate (see, e.g., equation (4.10) in [6]) shows that

$$\binom{n}{k} \geq \frac{c2^n}{\sqrt{n}}\cdot e^{-2(k-n/2)^2/n - d(k-n/2)^3/n^2},$$

where $c, d > 0$ are absolute constants. Thus, there is a constant $\epsilon > 0$ such that if $\ell \leq 2^{\epsilon n}$, $a \leq 1/2$ and n is large enough, then

$$\forall k \in [n/2 + (a/2) \cdot \sqrt{n \ln(2\ell/n)},\ n/2 + (a/2) \cdot \sqrt{n \ln(2\ell/n)} + \sqrt{n}],\ \binom{n}{k} \geq \frac{c 2^n}{\sqrt{n}} \cdot e^{-4(k-n/2)^2/n}. \tag{3}$$

We will assume that $\ell \leq 2^{\epsilon n}$; as mentioned above, this is without loss of generality.

Suppose some s of the v_i equal 1, and that $n - s$ of the v_i are -1. Now, to show (2), we may assume that $s \geq n/2$, since $|\sum_{j \in S} v_{\rho_j}| = |\sum_{j \in S}(-v_{\rho_j})|$. Now,

$$\Pr[|\sum_{j \in S} v_{\rho_j}| \geq a\sqrt{n \ln(2\ell/n)}] \geq \Pr[\sum_{j \in S} v_{\rho_j} \geq a\sqrt{n \ln(2\ell/n)}];$$

since $s \geq n/2$, the r.h.s. is minimized when $s = n/2$. Thus,

$$\begin{aligned}
\Pr[|\sum_{j \in S} v_{\rho_j}| \geq a\sqrt{n \ln(2\ell/n)}] &\geq 2^{-n} \cdot \sum_{k=n/2+(a/2)\cdot\sqrt{n\ln(2\ell/n)}}^{n} \binom{n}{k} \\
&\geq 2^{-n} \cdot \sum_{k=n/2+(a/2)\cdot\sqrt{n\ln(2\ell/n)}}^{n/2+(a/2)\cdot\sqrt{n\ln(2\ell/n)}+\sqrt{n}} \binom{n}{k} \\
&\geq \frac{c}{\sqrt{n}} \cdot \sum_{i=0}^{\sqrt{n}} e^{-4((a/2)\cdot\sqrt{n\ln(2\ell/n)}+i)^2/n} \quad \text{(by (3))} \\
&\geq \frac{c'}{\sqrt{n}} \cdot e^{-a^2 \ln(2\ell/n)} \cdot \sum_{i=0}^{\sqrt{n}} e^{-4a\sqrt{\ln(2\ell/n)/n}\cdot i} \\
&= \frac{c'}{\sqrt{n}} \cdot (n/(2\ell))^{a^2} \cdot \sum_{i=0}^{\sqrt{n}} e^{-4a\sqrt{\ln(2\ell/n)/n}\cdot i}, \qquad (4)
\end{aligned}$$

where $c' > 0$ is some absolute constant. Now set $a \ll 1$; say, $a \leq 1/20$. Then, there is an absolute constant $c_1 > 0$ such that for all $a \in (0, 1/20]$, all large enough n and all $\ell \geq n^{c_1}$, the inequality $\frac{c'}{\sqrt{n}} \cdot (n/(2\ell))^{a^2} > (\ln 2) \cdot n/\ell$ holds. So, we will be done if $\ell \geq n^{c_1}$; suppose $\ell < n^{c_1}$. Then, a simple lower bound on the geometric series in (4) using the fact that $\ln(2\ell/n) \geq \Omega(1)$ shows the following: there is an absolute constant $c_2 > 0$ such that (4) is at least

$$\frac{c_2}{a \cdot \sqrt{\ln(2\ell/n)}} \cdot (n/(2\ell))^{a^2}.$$

Now, it is easy to verify that if we choose a small enough, then for any $\ell \in [n, n^{c_1}]$, this quantity is greater than $(\ln 2) \cdot n/\ell$. □

3 Conclusion

Our results motivate us in making the following conjecture:

Conjecture. For all positive integers ℓ, $D_\ell(n) = O(\sqrt{\ell})$.

As another interesting open problem we wonder whether an $O(\sqrt{\ell}\log n)$–discrepancy result can be achieved in time polynomial in $n + \ell$.

Acknowledgements. We thank Alan Frieze, Jirka Matoušek and David Shmoys for helpful discussions.

References

[1] J. Beck. Roth's estimate on the discrepancy of integer sequences is nearly sharp. *Combinatorica*, 1:319–325, 1981.

[2] J. Beck. Irregularities of distribution II. In *Proc. London Math. Soc.*, Volume 56, pages 1–50, 1988.

[3] J. Beck and V. T. Sós. Discrepancy theory. In *Handbook of combinatorics, Volume II*, chapter 26, pages 1405–1446. Elsevier Science B.V. and the MIT Press, 1995.

[4] G. Bohus. On the discrepancy of 3 permutations. *Random Structures & Algorithms*, 1:215–220, 1990.

[5] J. Matoušek and J. H. Spencer. Discrepancy in arithmetic progressions. *Journal of the American Mathematical Society*, 9:195–204, 1996.

[6] A. M. Odlyzko. Asymptotic enumeration methods. In *Handbook of combinatorics, Volume II*, chapter 22, pages 1063–1229. Elsevier Science B.V. and the MIT Press, 1995.

[7] J. H. Spencer. Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289:679–706, 1985.

[8] J. H. Spencer. *Ten Lectures on the Probabilistic Method, Second Edition.* SIAM, Philadelphia, 1994.