Low Discrepancy Sets Yield Approximate Min-Wise Independent Permutation Families*

Michael Saks† Aravind Srinivasan‡ Shiyu Zhou§ David Zuckerman¶

Abstract

Motivated by a problem of filtering near-duplicate Web documents, Broder, Charikar, Frieze & Mitzenmacher defined the following notion of *ϵ-approximate min-wise independent permutation families.* A multiset $\mathcal{F}$ of permutations of $\{0, 1, \ldots, n-1\}$ is such a family if for all $K \subseteq \{0, 1, \ldots, n-1\}$ and any $x \in K$, a permutation π chosen uniformly at random from $\mathcal{F}$ satisfies

$$\left| \ \Pr[\min\{\pi(K)\} = \pi(x)] - \frac{1}{|K|} \ \right| \leq \frac{\epsilon}{|K|}.$$

We show connections of such families with *low discrepancy sets for geometric rectangles*, and give explicit constructions of such families $\mathcal{F}$ of size $n^{O(\sqrt{\log n})}$ for $\epsilon = 1/n^{\Theta(1)}$, improving upon the previously best-known bound of Indyk. We also present polynomial-size constructions when the min-wise condition is required only for $|K| \leq 2^{O(\log^{2/3} n)}$, with $\epsilon \geq 2^{-O(\log^{2/3} n)}$.

Keywords: Combinatorial problems; min-wise independent permutations; information retrieval; document filtering; pseudorandom permutations; explicit constructions.

---

*This work appears in the *Third International Workshop on Randomization and Approximation Techniques in Computer Science*, 1999.

†Department of Mathematics, Rutgers University, Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854. Research supported in part by NSF grant CCR-9700239. `saks@math.rutgers.edu`

‡Bell Laboratories, Lucent Technologies, 600–700 Mountain Ave., Murray Hill, NJ 07974-0636. `srin@research.bell-labs.com`. Parts of this work were done: (i) while at the School of Computing, National University of Singapore, Singapore 119260, and (ii) while visiting DIMACS (Center for Discrete Mathematics and Theoretical Computer Science). DIMACS is an NSF Science and Technology Center, funded under contract STC-91-19999; and also receives support from the New Jersey Commission on Science and Technology.

§Department of Computer & Information Science, University of Pennsylvania. `shiyu@cis.upenn.edu`

¶Computer Science Division, University of California, Berkeley, CA 94720. `diz@cs.berkeley.edu`. On leave from the University of Texas at Austin. Supported in part by NSF NYI Grant No. CCR-9457799, a David and Lucile Packard Fellowship for Science and Engineering, and an Alfred P. Sloan Research Fellowship. Part of this research was done while the author attended the DIMACS Workshop on Randomization Methods in Algorithm Design.

1 Introduction

Constructing pseudorandom permutation families is often more difficult than constructing pseudorandom function families. For example, there are polynomial size constructions of k-wise independent function families for constant k [8, 9, 1, 12]. On the other hand, although there are polynomial-size 3-wise independent permutation families (see, e.g. [14]), there are only exponential size constructions known for higher k. In fact, the only subgroups of the symmetric group that are 6-wise independent are the alternating group and the symmetric group itself; for 4-wise and 5-wise independence there are only finitely many besides these (see [4]). There are constructions of almost k-wise independent permutation families with error $\epsilon = O(k^2/n)$ [13], again not as good as is known for function families.

We address a different type of pseudorandom permutation family, called a *min-wise independent permutation family.* Motivated by a problem of filtering near-duplicate Web documents, Broder, Charikar, Frieze & Mitzenmacher [3] defined them as follows:

Definition 1.1 ([3]) *Let* $[n]$ *denote* $\{0, 1, \ldots, n-1\}$*, and* S_n *denote the set of permutations of* $[n]$*. A multiset* $\mathcal{F}$ *contained in* S_n *is called* min-wise independent *if for all* $K \subseteq [n]$ *and any* $x \in K$*, when a permutation* π *is chosen uniformly at random from* $\mathcal{F}$ *we have that* $\Pr[\min\{\pi(K)\} = \pi(x)] = \frac{1}{|K|}$. *($\pi(K)$ denotes the set* $\{\pi(y) : y \in K\}$*.)*

While $\mathcal{F} = S_n$ of course satisfies the above, even indexing from such an $\mathcal{F}$ is difficult, as some applications have n of the order of magnitude of 2^{64} [3]. Furthermore, it is shown in [3] that any min-wise independent family must have exponential size: more precisely, its cardinality is at least $\operatorname{lcm}(1, 2, \ldots, n) \geq e^{n-o(n)}$. (This lower bound of $\operatorname{lcm}(1, 2, \ldots, n)$ is in fact tight [15].) This motivates one to study families that are only *approximately* min-wise independent; moreover, in practice, we may also have an upper bound d on the cardinality of the sets K of Definition 1.1, such that $d \ll n$. Thus, the following notion is also introduced in [3]; we use slightly different terminology here.

Definition 1.2 ([3]) *Suppose a multi-set* $\mathcal{F}$ *is contained in* S_n*; let* π *be as in Definition 1.1.* $\mathcal{F}$ *is called an* (n, d, ϵ)*-mwif (for* d-wise ϵ-approximate min-wise independent family*) if for all* $K \subseteq [n]$ *with* $|K| \leq d$ *and any* $x \in K$*, we have*

$$\left| \ \Pr[\min\{\pi(K)\} = \pi(x)] - \frac{1}{|K|} \ \right| \leq \frac{\epsilon}{|K|}.$$

Using a random construction, Broder *et. al.* showed the *existence* of an (n, d, ϵ)-mwif of cardinality $O(d^2 \log(2n/d)/\epsilon^2)$ [3]. Indyk presented an explicit construction of an (n, n, ϵ)-mwif of cardinality $n^{O(\log(1/\epsilon))}$ in [7]. In this paper, we show a connection between the construction of approximate min-wise independent families and the construction of low discrepancy sets for geometric rectangles, and use this connection to give a new construction of an (n, d, ϵ)-mwif.

To state our main result we first need some definitions. Let m, d and n be integers with $d \leq n$. We denote by $\mathcal{GR}(m, d, n)$ the set of *(geometric) rectangles* $[a_1, b_1) \times [a_2, b_2) \times \cdots [a_n, b_n)$ such that:

- For all i, $a_i, b_i \in \{0, 1, \ldots, m-1\}$ with $a_i \leq b_i$;

- $a_i = 0$ and $b_i = m-1$ simultaneously hold for at least $n-d$ indices i (i.e., the rectangle is "nontrivial" in at most d dimensions).

Given such a rectangle $R \in \mathcal{GR}(m,d,n)$, its *volume* $\text{vol}(R)$ is defined to be $(\prod_{i=1}^{n}(b_i - a_i))/m^n$. A set $D \subseteq [0,m)^n$ is called a *δ-discrepant set* for $\mathcal{GR}(m,d,n)$ if:

$$\forall R \in \mathcal{GR}(m,d,n), \ \Big| \ \frac{|D \cap R|}{|D|} - \text{vol}(R) \ \Big| \leq \delta. \tag{1}$$

For an element $r = (r_1, r_2, \ldots, r_n) \in [0,m)^n$, define $\Gamma(r)$ to be the induced permutation $\pi_r \in S_n$ such that for any $0 \leq i,j \leq n-1$, $\pi_r(i) < \pi_r(j)$ if and only if $r_i < r_j$, or $r_i = r_j$ but $i < j$. For a subset $D \subseteq [0,m)^n$, $\Gamma(D)$ is defined to be the multiset of $\Gamma(r)$ where $r \in D$.

Our main theorem is the following:

Theorem 1.1 *Let m be arbitrary. Suppose $D \subseteq [0,m)^n$ is any δ-discrepant set for $\mathcal{GR}(m,d,n)$. Then for any $\frac{1}{m} \leq \alpha < 1$, $\Gamma(D)$ is an (n,d,ϵ)-mwif, where $\epsilon = (\alpha + \frac{\delta}{\alpha})d^2$.*

Lu [11] gave an explicit construction of δ-discrepant sets for $\mathcal{GR}(m,d,n)$ of cardinality

$$(mn)^{O(1)} \cdot (1/\delta)^{O(\sqrt{\log(\max\{2, d/\log(1/\delta)\})})}.$$

Therefore, setting $m = 2d^2/\epsilon$, $\alpha = 1/m$ and $\delta = 1/m^2$ in the main theorem and invoking Lu's construction, we obtain the following corollary:

Corollary 1.1 *There exists an explicit construction of an (n,d,ϵ)-mwif of cardinality*

$$L = n^{O(1)} \cdot (d/\epsilon)^{O(\sqrt{\log(\max\{2, d/\log(1/\epsilon)\})})}.$$

Note that this size is poly(n) if $d \leq 2^{O(\log^{2/3} n)}$ and $\epsilon \geq 2^{-O(\log^{2/3} n)}$. Also, when $d = n$, our bound is better than that of [7] if $\epsilon \leq 2^{-c_0\sqrt{\log n}}$, where $c_0 > 0$ is a certain absolute constant. We remark that Lu's construction builds on earlier work of [2, 5, 6, 10]. Given $\log L$ random bits to index a random element π of the permutation family guaranteed by Corollary 1.1, and given any $i \in [n]$, we can deterministically construct $\pi(i)$ in time polylogarithmic in L.

2 Proof of Main Theorem

Fix an arbitrary set $K \subseteq [n]$ of any size $k \leq d$, and choose any $x \in K$. We want to show that

$$\Big| \ \Pr[\min\{\pi(K)\} = \pi(x)] - \frac{1}{k} \ \Big| \leq \frac{\epsilon}{k},$$

where π is chosen uniformly at random from $\Gamma(D)$.

Assume without loss of generality that $t = 1/\alpha$ and αm are integers. Given x and K, we will define a sequence of pairwise disjoint rectangles $\{R_i = R_i(K,x) : 1 \leq i \leq t-1\}$ such that the permutations corresponding to points in $R = \cup_i R_i$ all satisfy $\min\{\pi(K)\} = \pi(x)$, and such that $\text{vol}(R)$ is approximately $\frac{1}{k}$. Using the fact that D is a good discrepant set for each R_i we will conclude that $\Gamma(D)$ has the required property.

We define R_i as follows.

$$R_i = \{(r_1, r_2, \ldots, r_n) \quad | \quad (i-1)\alpha m \le r_x < i\alpha m;\ i\alpha m \le r_y < m \text{ for all } y \in (K - \{x\}); \text{ and } 0 \le r_z < m \text{ for } z \notin K\}.$$

The following facts are easily seen:

1. For any $1 \le i < j \le t-1$, $R_i \cap R_j = \phi$.
2. $\mathrm{vol}(R_i) = \alpha(1 - i\alpha)^{k-1}$.
3. For any $\pi \in \Gamma(R_i)$, $\min\{\pi(K)\} = \pi(x)$.

Define $R = \cup_{i=1}^{t-1} R_i$. Using the first two facts, we can lower bound the volume of R as follows:

$$\begin{aligned}
\mathrm{vol}(R) &= \sum_{i=1}^{t-1} \mathrm{vol}(R_i) \\
&= \sum_{i=1}^{t-1} \alpha(1 - i\alpha)^{k-1} \\
&\ge \int_1^t \alpha(1 - \alpha x)^{k-1} dx \\
&= -\frac{1}{k}(1 - \alpha x)^k \mid_1^{1/\alpha} \\
&= (1 - \alpha)^k / k \\
&\ge \frac{1}{k} - \alpha.
\end{aligned}$$

Since D is a δ-discrepant set for $\mathcal{GR}(m, d, n)$, (1) shows that for each $1 \le i \le t-1$,

$$\left| \ \frac{|D \cap R_i|}{|D|} - \mathrm{vol}(R_i) \ \right| \le \delta.$$

Therefore,

$$\begin{aligned}
\frac{|D \cap R|}{|D|} &= \sum_{i=1}^{t-1} \frac{|D \cap R_i|}{|D|} \\
&\ge \sum_{i=1}^{t-1} (\mathrm{vol}(R_i) - \delta) \\
&= \mathrm{vol}(R) - (t-1)\delta \\
&\ge \frac{1}{k} - (\alpha + \frac{\delta}{\alpha}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr[\min\{\pi(K)\} = \pi(x)] &\ge \frac{|D \cap R|}{|D|} \\
&\ge \frac{1}{k} - (\alpha + \frac{\delta}{\alpha}).
\end{aligned}$$

Since this holds for any $x \in K$, an upper bound on this probability can be derived as follows:

$$\begin{aligned}\Pr[\min\{\pi(K)\} = \pi(x)] &\leq 1 - (k-1)(\frac{1}{k} - (\alpha + \frac{\delta}{\alpha})) \\ &\leq \frac{1}{k} + k(\alpha + \frac{\delta}{\alpha}).\end{aligned}$$

Since $k \leq d$, this completes the proof of the theorem.

Acknowledgments. We thank Andrei Broder and Michael Mitzenmacher for helpful discussions. We also thank Leonard Schulman and Monica Vazirani for the references about k-wise independent permutations and for interesting discussions about them. Our thanks also to the RANDOM '99 and journal referees for their helpful comments.

References

[1] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. Journal of Algorithms (7), 1986, pp. 567–583.

[2] R. Armoni, M. Saks, A. Wigderson, and S. Zhou. Discrepancy sets and pseudorandom generators for combinatorial rectangles. In: Proc. IEEE Symposium on Foundations of Computer Science, 1996, pp. 412–421.

[3] A. Z. Broder, M. Charikar, A. Frieze and M. Mitzenmacher. Min-wise independent permutations. In: Proc. ACM Symposium on Theory of Computing, 1998, pp. 327–336.

[4] P. J. Cameron. Finite permutation groups and finite simple groups. Bull. London Math. Soc. (13), 1981, pp. 1–22.

[5] G. Even, O. Goldreich, M. Luby, N. Nisan, and B. Veličković. Approximations of general independent distributions. In: Proc. ACM Symposium on Theory of Computing, 1992, pp. 10–16.

[6] R. Impagliazzo, N. Nisan, and A. Wigderson. Pseudorandomness for network algorithms. In: Proc. ACM Symposium on Theory of Computing, 1994, pp. 356–364.

[7] P. Indyk. A small approximately min-wise independent family of hash functions. In: Proc. ACM-SIAM Symposium on Discrete Algorithms, 1999, pp. 454–456.

[8] A. Joffe. On a set of almost deterministic k-independent random variables. Annals of Probability 2(1), 1974, pp. 161–162.

[9] R. M. Karp and A. Wigderson. A fast parallel algorithm for the maximal independent set problem. Journal of the ACM (32), 1985, pp. 762–773.

[10] N. Linial, M. Luby, M. Saks, and D. Zuckerman. Efficient construction of a small hitting set for combinatorial rectangles in high dimension. Combinatorica (17), 1997, pp. 215–234.

[11] C.-J. Lu. Improved pseudorandom generators for combinatorial rectangles. In: Proc. International Conference on Automata, Languages and Programming, 1998, pp. 223–234.

[12] M. Luby. A simple parallel algorithm for the maximal independent set problem. SIAM J. Comput. 15(4), 1986, pp. 1036–1053.

[13] M. Naor and O. Reingold. On the construction of pseudo-random permutations: Luby-Rackoff revisited. J. of Cryptology (12), 1999, pp. 29–66.

[14] E. G. Rees. Notes on Geometry, Springer Verlag, 1983.

[15] Y. Takei, T. Itoh and T. Shinozaki. An optimal construction of exactly min-wise independent permutations. Technical Report COMP98-62, IEICE, 1998.