On (ε, k)-min-wise independent permutations

Noga Alon[*]
nogaa@post.tau.ac.il

Toshiya Itoh[†]
titoh@dac.gsic.titech.ac.jp

Tatsuya Nagatani[‡]
Nagatani.Tatsuya@aj.MitsubishiElectric.co.jp

Abstract

A family of permutations $\mathcal{F}$ of $[n] = \{1, 2, \ldots, n\}$ is (ε, k)-min-wise independent if for every nonempty subset X of at most k elements of $[n]$, and for any $x \in X$, the probability that in a random element π of $\mathcal{F}$, $\pi(x)$ is the minimum element of $\pi(X)$, deviates from $1/|X|$ by at most $\varepsilon/|X|$. This notion can be defined for the uniform case, when the elements of $\mathcal{F}$ are picked according to a uniform distribution, or for the more general, biased case, in which the elements of $\mathcal{F}$ are chosen according to a given distribution D. It is known that this notion is a useful tool for indexing replicated documents on the web.

We show that even in the more general, biased case, for all admissible k and ε and all large n, the size of $\mathcal{F}$ must satisfy

$$|\mathcal{F}| \geq \Omega\left(\frac{k}{\varepsilon^2 \log(1/\varepsilon)} \log n\right),$$

as well as

$$|\mathcal{F}| \geq \Omega\left(\frac{k^2}{\varepsilon \log(1/\varepsilon)} \log n\right).$$

This improves the best known previous estimates even for the uniform case.

[*] School of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Science, Tel Aviv University, Tel Aviv 69978, Israel. Research supported in part by a USA Israeli BSF grant, by a grant from the Israel Science Foundation, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

[†] Global Scientific Information and Computing Center, Tokyo Institute of Technology, 2–12–1 O-okayama, Meguro-ku, Tokyo 152–8550, Japan. Research supported in part by Japan Society of the Promotion of Science, Grant-in-Aid for Scientific Research on Priority Areas, 16092205, 2006.

[‡] Advanced Technology R&D Center, Mitsubishi Electric Corporation, 8–1–1, Tsukaguchi-Honmachi, Amagasaki, Hyogo 661–8661, Japan.

1 Introduction

1.1 Background

A family $\mathcal{F}$ of permutations of $[n] = \{1, 2, \ldots, n\}$ is an *ε-approximate k-restricted min-wise independent family* (or an *(ε, k)-min-wise independent family*, for short) if for every nonempty subset X of at most k elements of $[n]$, and for any $x \in X$, the probability that in a random element π of $\mathcal{F}$, $\pi(x)$ is the minimum element of $\pi(X)$, deviates from $1/|X|$ by at most $\varepsilon/|X|$. This notion can be defined for the uniform case, when the elements of $\mathcal{F}$ are picked according to a uniform distribution, or for the more general, biased case, in which the elements of $\mathcal{F}$ are chosen according to a given distribution D.

The notion of (ε, k)-min-wise independent families, as well as the related ones of *k-restricted min-wise independent families* (corresponding to the case $\varepsilon = 0$), *min-wise independent families* (corresponding to $\varepsilon = 0$, $k = n$), and *ε-min-wise independent families* (corresponding to $k = n$), were introduced by Broder et al. [5] and further investigated in [6, 7, 8, 9, 11, 12, 13]. These form a basic tool to estimate resemblance between documents [4] and have applications of detecting almost identical documents on the Web [4] and of reducing the amount of randomness used by probabilistic algorithms [10]. Among the several variants of min-wise independence, we focus on the notion of (ε, k)-min-wise independent families defined above, and investigate the minimum possible size of families $\mathcal{F} \subseteq S_n$ of (ε, k)-min-wise independent permutations.

1.2 Known results

For families $\mathcal{F} \subseteq S_n$ of (ε, k)-min-wise independent permutations, Broder et al. [5] proved upper and lower bounds for $|\mathcal{F}|$ as in Table 1.

Table 1: Known results on (ε, k)-min-wise independent permutations

Constructive Upper Bound (uniform)	$2^{4k+o(k)}k^{2\log\log(n/\varepsilon)}$	[5, Theorem 5]
Non-constructive Upper Bound (uniform)	$O\left(\frac{k^2}{\varepsilon^2}\log\left(\frac{n}{k}\right)\right)$	[5, Theorem 4]
Lower Bound (uniform)	$\Omega\left(k^2(1-\sqrt{8\varepsilon})\right)$	[5, Theorem 6]
Lower Bound (biased)	$\Omega\left(\min\left\{k2^{k/2}\log\left(\frac{n}{k}\right), \frac{\log(1/\varepsilon)(\log n - \log\log(1/\varepsilon))}{\varepsilon^{1/3}}\right\}\right)$	[5, Theorem 9]

Note that there is a large gap between the upper and lower bounds. The lower bound for the uniform distribution depends only on k and ε and does not grow with n, which seems unnatural. The lower bound for biased distributions gives $|\mathcal{F}| = \Omega(k2^{k/2}\log(n/k))$ if $k < \frac{2}{3}\log(1/\varepsilon)$ and $|\mathcal{F}| = \Omega(\frac{\log(1/\varepsilon)(\log n - \log\log(1/\varepsilon))}{\varepsilon^{1/3}})$ otherwise. Thus for any fixed $\varepsilon > 0$, if k is large, then the lower bound for biased distributions depends only on n and ε and does not grow with k. Therefore, it seems that there should be a tighter lower bound for $|\mathcal{F}|$ that depends on all variables n, k, and ε.

1.3 The main results

Our main results in the present paper are the following two improved lower bounds for the minimum possible cardinality of (ε, k)-min-wise independent permutation families.

Theorem 1.1 *For any $0 < \varepsilon < \frac{1}{8}$ and $k \geq 3$, and all sufficiently large n, the following holds. Let $\mathcal{F} \subset S_n$ be an (ε, k)-min-wise independent family of permutations of $[n]$, with respect to a distribution D on $\mathcal{F}$. Then*

$$|\mathcal{F}| \geq \Omega\left(\frac{k}{\varepsilon^2 \log(1/\varepsilon)} \log n\right).$$

Theorem 1.2 *For any $0 < \varepsilon < \frac{1}{11}$ and $k \geq 3$, and all sufficiently large n, the following holds. Let $\mathcal{F} \subset S_n$ be an (ε, k)-min-wise independent family of permutations of $[n]$, with respect to a distribution D on $\mathcal{F}$. Then*

$$|\mathcal{F}| \geq \Omega\left(\frac{k^2}{\varepsilon \log(1/\varepsilon)} \log n\right).$$

The proofs combine a linear-algebra approach with a geometric lemma proved in [1] and a few additional combinatorial arguments. Throughout the proofs we omit all floor and ceiling signs whenever these are not crucial, and make no attempt to optimize the absolute constants.

2 Proofs

An m by m real matrix is *diagonally dominant* if the absolute value of each diagonal entry in the matrix exceeds the sum of the absolute values of all other entries in its row. It is easy and well known that every matrix of this type is nonsingular. In particular, if each diagonal entry is at least 1, and the absolute value of every other entry is smaller than $1/m$, then the matrix has full rank. The basic tool applied in our proofs is an extension of this result which shows that even if the assumptions are relaxed and one only assumes that each diagonal entry is at least $1/2$ and the absolute value of each other entry is at most δ there is still a meaningful lower bound for the rank. A statement of this form is proved in [1], where it is shown that it can be interpreted as a geometric result, supplying a lower bound for the minimum possible dimension of an Euclidean space in which one can embed a simplex of m equilateral points with low distortion.

Lemma 2.1 ([1, Theorem 9.3]) *Let $B = (b_{i,j})$ be an m by m real matrix with $b_{i,i} = 1$ for all i and $|b_{i,j}| \leq \delta$ for all $i \neq j$. If the rank of B is r, and $\frac{1}{\sqrt{m}} \leq \delta < 1/2$, then*

$$r \geq \Omega\left(\frac{1}{\delta^2 \log(1/\delta)} \log m\right).$$

Corollary 2.2 *Let B be an m by m real matrix with $b_{i,i} \geq 1/2$ for all i and $|b_{i,j}| \leq \delta$ for all $i \neq j$. If the rank of B is r, and $\frac{1}{2\sqrt{m}} \leq \delta < 1/4$, then*

$$r \geq \Omega\left(\frac{1}{\delta^2 \log(1/\delta)} \log m\right).$$

Proof: Let $C = (c_{i,j})$ be the m by m diagonal matrix defined by $c_{i,i} = 1/b_{i,i}$ for all i. Then every diagonal entry of CB is 1 and every off-diagonal entry is of absolute value at most 2δ. The result thus follows from Lemma 2.1. □

Proof of Theorem 1.1: Let $\mathcal{F}$ be an (ε, k)-min-wise independent family of permutations of $[n]$, with respect to the distribution D, where $\varepsilon > 0$, $k \geq 3$ and n is large. Put $s = k/3$, $L = n/s$ and partition $[n]$ into L pairwise disjoint sets $X_0, X_1, \ldots, X_{L-1}$, each of size s, where $X_0 = \{1, 2, \ldots, s\}$. Put $\mathcal{F} = \{\pi_1, \pi_2, \ldots, \pi_d\}$, $m = L - 1$, and define, for each $h \in [s]$, an m by d matrix $U^{(h)} = (u_{ij}^{(h)})$ as follows:

$$u_{ij}^{(h)} = \begin{cases} \sqrt{\Pr_D(\pi_j)} & \text{if } \min(\pi_j(X_0 \cup X_i)) = \pi_j(h) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Define $V^{(h)} = (v_{ij}^{(h)}) = U^{(h)}(U^{(h)})^T$ and observe that $v_{ii}^{(h)}$ is precisely the probability that h is the minimum element of $X_0 \cup X_i$ (according to the distribution D on $\mathcal{F}$), whereas for $i \neq j$, $v_{ij}^{(h)}$ is the probability that h is the minimum element of $X_0 \cup X_i \cup X_j$ according to the same distribution. By the assumption on $\mathcal{F}$ and D, each $v_{ii}^{(h)}$ deviates from $\frac{1}{2s}$ by at most $\frac{\varepsilon}{2s}$, and each $v_{ij}^{(h)}$ for $i \neq j$ deviates from $\frac{1}{3s}$ by at most $\frac{\varepsilon}{3s}$. In addition, by the definition of the matrices $U^{(h)}$, for any distinct $h, g \in [s]$, $U^{(h)}(U^{(g)})^T = 0$.

Let U be the ms by d matrix defined by $U^T = [(U^{(1)})^T, (U^{(2)})^T, \ldots, (U^{(s)})^T]$. Then $V = UU^T$ is a block-diagonal matrix whose blocks are the matrices $V^{(h)}$, implying that its rank is the sum of ranks of the matrices $V^{(h)}$.

The crucial claim now is that the rank of each matrix $V^{(h)}$ is at least $\Omega(\frac{1}{\varepsilon^2 \log(1/\varepsilon)} \log m)$. Indeed, if we subtract from $V^{(h)}$ the rank-one matrix in which every entry is exactly $\frac{1}{3s}$, and multiply the result by $6s$, then from the assumption that $0 < \varepsilon < \frac{1}{8}$, we get a matrix in which each diagonal entry is at least

$$6s\left(\frac{1-\varepsilon}{2s} - \frac{1}{3s}\right) = 1 - 3\varepsilon > \frac{5}{8} > \frac{1}{2},$$

and each off-diagonal entry is in absolute value at most

$$6s\left(\frac{1+\varepsilon}{3s} - \frac{1}{3s}\right) = 2\varepsilon < \frac{1}{4}.$$

As the above subtraction and multiplication can change the rank by at most 1, the assertion of the claim follows from Corollary 2.2. Combining this with the fact that for all large n ($n > k^2$ will suffice here), $\log m > 0.5 \log n$, and the fact that $|\mathcal{F}| = d \geq \text{rank}(V)$, the assertion of the theorem follows. $\Box$

In order to prove Theorem 1.2 we need an additional simple lemma.

Lemma 2.3 *Let $\varepsilon > 0$ be a real and let $s \geq 1$ and $M \geq 2s^2$ be integers. If $\varepsilon s > 1$, then there is a collection of more than $N = (\frac{M}{2s^2})^{\varepsilon s}$ subsets of cardinality s of $[M] = \{1, 2, \ldots, M\}$, so that the intersection of any two of them has cardinality at most εs.*

Proof : Starting with the family of all s-subsets of $[M]$, pick an arbitrary s-subset, and omit all s-subsets that intersect it by more than εs elements. As long as there is a yet unchosen s-subset that has not been omitted, pick it, and omit all those that intersect it by more than εs elements. Clearly this produces a collection of at least

$$\frac{\binom{M}{s}}{\binom{s}{\varepsilon s}\binom{M-s}{(1-\varepsilon)s} + 1}$$

s-subsets satisfying the desired intersection condition, and a simple calculation shows that for $M \geq 2s^2$ this quantity is at least $(\frac{M}{2s^2})^{\varepsilon s}$ (with room to spare). $\Box$

Proof of Theorem 1.2: Let $\varepsilon, \mathcal{F}$ and D be as in the assumption of the theorem. Put $s = k/3$ and let $X_0 = [s]$ consist of the first s elements of $[n]$. Put $M = n - s$, let $N = (\frac{M}{2s^2})^{\varepsilon s}$, and let $X_1, X_2, \ldots, X_N$ be a collection of subsets of cardinality s of $\{s+1, \ldots, n\}$, so that each pair of them has at most εs common elements. (The existence of such a collection follows from Lemma 2.3.) For every $h \in [s]$ define an N by d matrix $U^{(h)}$ by (1), and an N by N matrix $V^{(h)} = U^{(h)}(U^{(h)})^T$.

By the reasoning described in the proof of Theorem 1.1 every diagonal matrix of $V^{(h)}$ deviates from $\frac{1}{2s}$ by at most $\frac{\varepsilon}{2s}$, whereas for $i \neq j$, $v_{ij}^{(h)}$ deviates from $\frac{1}{|X_0 \cup X_i \cup X_j|}$ by at most $\frac{\varepsilon}{|X_0 \cup X_i \cup X_j|}$. However, since the intersection of X_i and X_j is of size at most εs, it follows that each such off-diagonal entry is at least $\frac{1-\varepsilon}{3s}$ and at most $\frac{1+\varepsilon}{(3-\varepsilon)s}$. We can thus subtract, as before, a constant $\frac{1}{3s}$ from each entry of $V^{(h)}$ and multiply the resulting matrix by $6s$ to get a matrix in which every diagonal entry is at least

$$6s\left(\frac{1-\varepsilon}{2s} - \frac{1}{3s}\right) = 1 - 3\varepsilon > \frac{8}{11} > \frac{1}{2},$$

and every off-diagonal entry is, in absolute value, at most

$$6s\left\{\frac{1+\varepsilon}{(3-\varepsilon)s} - \frac{1}{3s}\right\} = \frac{8\varepsilon}{3-\varepsilon} < \frac{1}{4},$$

using the assumption that $0 < \varepsilon < \frac{1}{11}$. By Corollary 2.2 the rank of such a matrix is at least $\Omega(\frac{1}{\varepsilon^2 \log(1/\varepsilon)} \log N)$, and as for large n,

$$\log N = \varepsilon s \log \frac{n-s}{2s^2} \geq 0.5 \varepsilon s \log n = \Omega(\varepsilon k \log n),$$

the assertion of Theorem 1.2 follows, as in the proof of Theorem 1.1. □

3 Concluding remarks

Although our results provide an asymptotically optimal bound, up to a constant factor, for the minimum possible cardinality of (ε, k)-min-wise independent families of permutations for **fixed** ε and all admissible $n \gg k$, for small ε there is still a gap between the known upper and lower bounds, and it will be interesting to close it.

Lemma 2.1 seems useful in obtaining lower bounds for the smallest possible size of a sample space supporting random variables exhibiting some approximate independence properties. One example is given here. In a similar way one can apply the lemma to show that the minimum possible size of a sample space supporting n k-wise ε-biased binary random variables is $\Omega(\frac{k}{\varepsilon^2 \log(1/\varepsilon)} \log n)$, which is tight, up to the $\log(1/\varepsilon)$-term. The lemma can also be applied to derive a quick, nearly tight upper bound for the maximum possible rate of a binary error correcting code in which every code-word is of Hamming weight that deviates from $n/2$ by at most εn, a nearly tight lower bound for the minimum possible distortion in an embedding of a simplex in a low-dimensional Euclidean space, and a nontrivial upper bound for the maximum possible number of real vectors of length n so that the ℓ_1-distance between every pair is exactly 1. See [1], [3], [2] for some further examples and details.

Acknowledgment: We would like to thank Yoshinori Takei for his valuable comments and for insightful discussions on an earlier version of the paper.

References

[1] N. Alon, Problems and results in extremal combinatorics, I, *Discrete Math.*, 273:31-53, 2003.

[2] N. Alon, Perturbed identity matrices have high rank: proof and applications, to appear.

[3] N. Alon and P. Pudlak, Equilateral sets in l_p^n, Geometric and Functional Analysis 13:467-482, 2003.

[4] A. Broder, On the Resemblance and Containment of Documents, in *Proc. of Compression and Complexity of Sequences*, 21-29, 1998.

[5] A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher, Min-wise independent permutations, *J. Comput. Sys. Sci.*, 60:630-659, 2000.

[6] P. Indyk, A Small Approximately Min-Wise Independent Family of Hash Functions, *J. of Algorithms*, 38:84-90, 2001.

[7] T. Itoh, Y. Takei, and J. Tarui, On Permutations with Limited Independence, in *Proc. of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, 137-146, 2000.

[8] T. Itoh, Y. Takei, and J. Tarui, On the Sample Size k-Restricted Min-Wise Independent Permutations and Other k-Wise Distributions, in *Proc. of the 35th Annual ACM Symposium on Theory of Computing*, 710-719, 2003.

[9] J. Matoušek and M. Stojaković, On Restricted Min-Wise Independence of Permutations, *Random Structures & Algorithms*, 23(4):397-408, 2003.

[10] K. Mulmuley, Randomized Geometric Algorithms and Pseudorandom Generators, *Algorithmica*, 16:450-463, 1996.

[11] S. Norin, A Polynomial Lower Bound for the Size of any k-Min-Wise Independent Set of Permutation, *Zapiski Nauchnyh Seminarov (POMI)*, 277:104-116, 2001 (in Russian). Available at http://www.pdmi.ras.ru/znsl/

[12] M. Saks, A. Srinivasan, S. Zhou, and D. Zuckerman, Low Discrepancy Sets Yield Approximate Min-Wise Independent Permutation Families, *Inform. Process. Lett.*, 73:29-32, 2000.

[13] J. Tarui, T. Itoh, and Y. Takei, A Nearly Linear Size 4-Min-Wise Independent Permutation Family by Finite Geometries, in *Proc. of RANDOM-APPROX'03*, Lecture Notes in Computer Science 2764, Springer, 396-408, 2003.