# Min-Wise independent linear permutations

Tom Bohman* Colin Cooper† Alan Frieze‡

January 13, 2000

## 1 Introduction

Broder, Charikar, Frieze and Mitzenmacher [3] introduced the notion of a set of min-wise independent permutations. We say that $\mathcal{F} \subseteq S_n$ is *min-wise independent* if for any set $X \subseteq [n]$ and any $x \in X$, when $\pi$ is chosen at random in $\mathcal{F}$ we have

$$\mathbb{P}(\min\{\pi(X)\} = \pi(x)) = \frac{1}{|X|}. \tag{1}$$

The research was motivated by the fact that such a family (under some relaxations) is essential to the algorithm used in practice by the AltaVista web index software to detect and filter near-duplicate documents. A set of permutations satisfying (1) needs to be exponentially large [3]. In practice we can allow certain relaxations. First, we can accept small relative errors. We say that $\mathcal{F} \subseteq S_n$ is *approximately min-wise independent with relative error* $\epsilon$ (or just approximately min-wise independent, where the meaning is clear) if for any set $X \subseteq [n]$ and any $x \in X$, when $\pi$ is chosen at random in $\mathcal{F}$ we have

$$\left|\mathbb{P}\big(\min\{\pi(X)\} = \pi(x)\big) - \frac{1}{|X|}\right| \leq \frac{\epsilon}{|X|}. \tag{2}$$

In other words we require that all the elements of any fixed set $X$ have only an almost equal chance to become the minimum element of the image of $X$ under $\pi$.

*Linear permutations* are an important class of permutations. Let $p$ be a (large) prime and let $\mathcal{F}_p = \{\pi_{a,b} :\ 1 \leq a \leq p-1,\ 0 \leq b \leq p-1\}$ where for $x \in [p] = \{0, 1, \ldots, p-1\}$,

$$\pi_{a,b}(x) = ax + b \bmod p,$$

*Department of Mathematical Sciences,Carnegie Mellon University, Pittsburgh PA15213, U.S.A.

†School of Mathematical Sciences, University of North London, London N7 8DB, UK. Research supported by the STORM Research Group

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, U.S.A., Supported in part by NSF grant CCR-9818411

where for integer $n$ we define $n \bmod p$ to be the non-negative remainder on division of $n$ by $p$.

For $X \subseteq [p]$ we let

$$F(X) = \max_{x \in X} \{\mathbb{P}_{a,b}(\min\{\pi(X)\} = \pi(x))\}$$

where $\mathbb{P}_{a,b}$ is over $\pi$ chosen uniformly at random from $\mathcal{F}_p$. The natural questions to discuss are what are the extremal and average values of $F(X)$ as $X$ ranges over $\mathcal{A}_k = \{X \subseteq [p] : |X| = k\}$. The following results were some of those obtained in [3]:

**Theorem 1**
**(a)** *Consider the set* $X_k = \{0, 1, 2 \ldots k-1\}$, *as a subset of* $[p]$. *As* $k, p \to \infty$, *with* $k^2 = o(p)$,

$$\mathbb{P}_{a,b}(\min\{\pi(X_k)\} = \pi(0)) = \frac{3}{\pi^2}\frac{\ln k}{k} + O\left(\frac{k^2}{p} + \frac{1}{k}\right).$$

**(b)** *As* $k, p \to \infty$, *with* $k^4 = o(p)$,

$$\frac{1}{2(k-1)} \leq \mathbb{E}_X[F(X)] \leq \frac{\sqrt{2}+1}{\sqrt{2}k} + O\left(\frac{1}{k^2}\right),$$

*where* $\mathbb{E}_X$ *denotes expectations over* $X$ *chosen uniformly at random from* $\mathcal{A}_k$.

In this paper we improve the second result and prove

**Theorem 2**
*As* $k, p \to \infty$,

$$\mathbb{E}_X[F(X)] = \frac{1}{k} + O\left(\frac{(\log k)^3}{k^{3/2}}\right).$$

Thus for most sets, simply chosen, random linear permutations, will suffice as (near) min-wise independent. Other results on min-wise independence have been obtained by Indyk [6], Broder, Charikar and Mitzenmacher [4] and Broder and Feige [5].

# 2 Proof of Theorem 2

Let $X = \{x_0, x_1, \ldots, x_{k-1}\} \subseteq [p]$. Let $\beta_i = ax_i \bmod p$ for $i = 0, 1, \ldots, k-1$. Let

$$i = i(X, a) = \min\{\beta_0 - \beta_j \bmod p : j = 1, 2, \ldots, k-1\}. \tag{3}$$

Let

$$A_i = A_i(X) = \{a \in [p] : i(X, a) = i\}$$

and note that

$$|A_i| \leq k-1, \qquad i = 1, 2, \dots, p-1.$$

Then

$$\min\{\pi(X)\} = \pi(x_0) \text{ iff } 0 \in \{\beta_0 + b, \beta_0 + b - 1, \dots, \beta_0 + b - i + 1\} \bmod p.$$

Thus if

$$Z = Z(X) = \sum_{i=1}^{p-1} i|A_i|,$$

$$\mathbb{P}_{a,b}(\min\{\pi(X)\} = \pi(x_0)) = \frac{Z}{p(p-1)}. \tag{4}$$

Fix $a \in \{1, 2, \dots, p-1\}$ and $x_0$. Then

$$\mathbb{P}(a \in A_i) = (k-1) \cdot \frac{1}{p-1} \prod_{t=1}^{k-2} \left(1 - \frac{i+t}{p-1-t}\right) \tag{5}$$

We write $Z = Z_0 + Z_1$ where $Z_0 = \sum_{i=1}^{i_0} i|A_i|$ where $i_0 = \frac{4p \log k}{k}$. Now, by symmetry,

$$\mathbb{E}_X(\mathbb{P}_{a,b}(\min\{\pi(X)\} = \pi(x_0)) = \frac{1}{k} \tag{6}$$

and so

$$\mathbb{E}_X(Z) = \frac{p(p-1)}{k}.$$

It follows from (5) that

$$\begin{aligned} \mathbb{E}(Z_1) &\leq (k-1) \sum_{i=i_0+1}^{p-1} i \exp\left\{-\frac{4(k-2)\log k}{k}\right\} \\ &\leq \frac{p^2}{k^3} \end{aligned} \tag{7}$$

for large $k, p$.

We continue by using the Azuma-Hoeffding Martingale tail inequality – see for example [1, 2, 7, 8, 9]. Let $x_0$ be fixed and for a given $X$ let $\hat{X}$ be obtained from $X$ by replacing $x_j$ by randomly chosen $\hat{x}_j$. For $j \geq 1$ let

$$d_j = \max_X\{|\mathbb{E}_{\hat{x}_j}(Z(X) - Z(\hat{X}))|\}.$$

Then for any $t > 0$ we have

$$\mathbb{P}(|Z_0 - \mathbb{E}(Z_0)| \geq t) \leq 2\exp\left\{-\frac{2t^2}{d_1^2 + \cdots + d_{k-1}^2}\right\}. \tag{8}$$

We claim that

$$d_j \leq \sum_{i=1}^{i_0} i + \sum_{i=1}^{i_0} \frac{(k-1)i^2}{p} \tag{9}$$

$$\leq \frac{i_0^2}{2} + \frac{i_0^3 k}{3p} + O(p)$$

$$\leq \frac{30(\log k)^3 p^2}{k^2} \tag{10}$$

**Explanation for (9):** If $a \in A_i(X)$ because $ax_j = ax_0 - i \bmod p$ then changing $x_j$ to $\hat{x}_j$ changes $|A_i|$ by one. This explains the first summation. The second accounts for those $a \in A_i(X)$ for which $ax_0 - a\hat{x}_j \bmod p < i$, changing the minimum in (3). We then use $|A_i| \leq k-1$ and $\mathbb{P}(ax_0 - a\hat{x}_j \bmod p < i) = \frac{i}{p}$.

Using (10) in (8) with $t = \varepsilon\frac{p^2}{k}$ we see that

$$\mathbb{P}\left(|Z_0 - \mathbb{E}(Z_0)| \geq \varepsilon\frac{p^2}{k}\right) \leq \exp\left\{-\frac{\varepsilon^2 k}{450(\log k)^6}\right\}.$$

It now follows from (4), (6), (7) and the above that

$$\mathbb{E}_X[F(X)] = \frac{1}{k} + O\left(\frac{1}{k^2} + \frac{1}{k}\int_{\varepsilon=0}^{\infty} \min\left\{1, k\exp\left\{-\frac{\varepsilon^2 k}{450(\log k)^6}\right\}\right\} d\varepsilon\right)$$

and the result follows. □

# References


[1] N.Alon and J.H.Spencer, *The Probabilistic Method,* Wiley, 1992.

[2] B.Bollobás, *Martingales, isoperimetric inequalities and random graphs*, in Combinatorics, A.Hajnal, L.Lovász and V.T.Sós Ed., Colloq. Math. Sci. Janos Bolyai 52, North Holland 1988.

[3] A.Z.Broder, M.Charikar, A.M.Frieze and M.Mitzenmacher, *Min-Wise Independent permutations*

[4] A.Z.Broder, M.Charikar and M.Mitzenmacher, *A derandomization using min-wise independent permutations*, Proceedings of Second International Workshop RANDOM '98 (M.Luby, J.Rolim, M.Serna Eds.) (1998) 15-24.

[5] A.Z.Broder and U.Feige, *Min-Wise versus Linear Independence*, Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (2000).

[6] P.Indyk, *A small approximately min-wise independent family of hash-functions*, Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (1999) 454-456.

[7] C.J.H.McDiarmid, *On the method of bounded differences*, Surveys in Combinatorics, 1989, Invited papers at the Twelfth British Combinatorial Conference, Edited by J.Siemons, Cambridge University Press, 148-188.

[8] C.J.H.McDiarmid, *Concentration*, Probabilistic methods for algorithmic discrete mathematics, (M.Habib, C.McDiarmid, J.Ramirez-Alfonsin, B.Reed, Eds.), Springer (1998) 195-248.

[9] M.J.Steele, Probability theory and combinatorial optimization, CBMS-NSF Regional Conference Series in Applied Mathematics 69, 1997.