On Restricted Min-Wise Independence of Permutations

JIŘÍ MATOUŠEK*

Department of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University, Malostranské nám. 25
118 00 Praha 1, Czech Republic
and
Institut für Theoretische Informatik
ETH Zentrum, Zürich, Switzerland

MILOŠ STOJAKOVIĆ†

Institut für Theoretische Informatik
ETH Zentrum, Zürich, Switzerland
and
Institute of Mathematics
University of Novi Sad
Trg D. Obradovića 4
21000 Novi Sad, Yugoslavia

Rev. 21/XI/2002 JM

Abstract

A family of permutations $\mathcal{F} \subseteq S_n$ with a probability distribution on it is called *k-restricted min-wise independent* if we have $\Pr[\min \pi(X) = \pi(x)] = \frac{1}{|X|}$ for every subset $X \subseteq [n]$ with $|X| \leq k$, every $x \in X$, and $\pi \in \mathcal{F}$ chosen at random. We present a simple proof of a result of Norin: every such family has size at least $\binom{n-1}{\lfloor \frac{k-1}{2} \rfloor}$. Some features of our method might be of independent interest.

*Supported by Project LN00A056 of the Ministry of Education of the Czech Republic and by ETH Zürich.

†Supported by the joint Berlin/Zurich graduate program Combinatorics, Geometry and Computation, financed by ETH Zurich and the German Science Foundation (DFG).

The best available upper bound for the size of such family is $1 + \sum_{j=2}^{k}(j-1)\binom{n}{j}$. We show that this bound is tight if the goal is to imitate not the uniform distribution on S_n, but a distribution given by assigning suitable priorities to the elements of $[n]$ (the stationary distribution of the *Tsetlin library*, or self-organizing lists). This is analogous to a result of Karloff and Mansour for k-wise independent random variables.

We also investigate the cases where the min-wise independence condition is required only for sets X of size *exactly* k (where we have only an $\Omega(\log\log n + k)$ lower bound), or for sets of size k and $k-1$ (where we already obtain a lower bound of $n-k+2$).

1 Introduction

Let S_n denote the set of all permutations of $[n] = \{1, 2, \ldots, n\}$. A family $\mathcal{F} \subseteq S_n$ of permutations with a probability distribution on it is called *min-wise independent* if, for any set $X \subseteq [n]$, the following condition holds: For every $x \in X$ and for π chosen at random from $\mathcal{F}$, we have

$$\Pr[\min \pi(X) = \pi(x)] = \frac{1}{|X|}.$$

That is, we require that all elements of X have equal chance of becoming the minimum element of the image of X under π. We call this condition the *min-uniform condition for* X.

As for the probability distribution on $\mathcal{F}$, two basic cases can be distinguished: Either the distribution is required to be uniform (then we call $\mathcal{F}$ a *uniform* family), or it can be arbitrary (then we speak of a *biased* family).

Min-wise independent families of permutations were introduced and thoroughly investigated by Broder, Charikar, Frieze, and Mitzenmacher [3]. They are essential to algorithms for detecting near-duplicate documents, such as used in practice by the AltaVista Web indexing software; see, e.g., [3] for explanation and references. A partial list of subsequent works in this area is [4, 5, 9, 10, 15, 17].

Size of min-wise independent families. Broder et al. [3] showed that the size of any *uniform* min-wise independent family is at least the least common multiple of $2, 3, \ldots, n$, which is of order $e^{n-o(n)}$. They constructed a uniform min-wise independent family of size at most 4^n, and later Takei, Itoh, and Shinozaki [17] provided a construction exactly matching the lower bound just mentioned.

Biased min-wise independent families can be somewhat smaller, but they still require exponential size: The bounds from [3] are $\Omega(\sqrt{n}\,2^n)$ from below and $n2^{n-1}-1$ from above (the latter bound uses the linear programming approach of Koller and Megiddo [13], which we briefly discuss in Section 3).

Thus, min-wise independent families are necessarily exponentially large and thus impractical for some applications. The condition of min-wise independence can be relaxed in various ways.

Approximate min-wise independence. One of them is *approximate min-wise independence* [3], where the probability of $\min \pi(X) = \pi(x)$ is only close to $\frac{1}{|X|}$, rather than equal to it. That is, we require that $\left|\Pr[\min \pi(X) = \pi(x)] - \frac{1}{|X|}\right| \le \frac{\varepsilon}{|X|}$, where $\varepsilon > 0$ is a prescribed error parameter. This concept seems very suitable for applications, and very small approximate min-wise independent families have been constructed. We refer to Broder et al. [3], Indyk [9], and Saks et al. [15] for upper and lower bounds.

Restricted min-wise independence. In this paper we study a different relaxation of min-wise independence, where the min-uniform condition is required to hold exactly, but only for some sets X. Namely, we call a family $\mathcal{F}$ *k-restricted min-wise independent* if the min-uniform condition holds for all $X \subseteq [n]$ with $|X| \le k$.

This perhaps most natural concept of restricted min-wise independence was also introduced by Broder et al. [3]. By a sophisticated method using graph entropy, they proved a lower bound of $\Omega\big(k2^{k/2}\log(n/k)\big)$ for the size of any (possibly biased) k-restricted min-wise independent family. They also noted that for uniform k-restricted min-wise independent families, one obtains a lower bound of $e^{k-o(k)}$ by the method used for uniform min-wise families, and that the upper bound of $\sum_{j=2}^{k} j\binom{n}{j}$ $(= \Theta(n^k)$ for k fixed) for *biased* k-restricted min-wise independent families follows by the linear programming approach. The problem was further investigated by Itoh et al. [10], who constructed an explicit uniform k-restricted min-wise independent family of size $n^{(1+\frac{2}{\lg n})k}$, and proved a lower bound of $n-1$ (for any $k \ge 3$). Recently Norin [14] proved the following stronger lower bound:

Theorem 1 ([14]) *Let $\mathcal{F}$ be a k-restricted min-wise independent family (with an arbitrary, possibly biased, probability distribution), and let $n \ge 2k$. Then*

$$|\mathcal{F}| \ge \binom{n-1}{\lfloor \frac{k-1}{2} \rfloor}.$$

In Section 2 we present an alternative proof of this fact. Although we found it without knowing of Norin's work, it has some similarities to his proof, but it is considerably simpler and shorter. Some of the features of our proof, such as the use of complex random variables, appear to be new in this context and could be useful in other problems as well.

Recently we have learned that Itoh et al. [11] independently proved a lower bound similar to Theorem 1 but slightly stronger, namely, $\sum_{i=0}^{(k-1)/2} \binom{n-1}{i}$ for k odd and $\sum_{i=0}^{k/2-1} \binom{n-1}{i} + \binom{n-2}{k/2-1}$ for k even. Their method differs from both Norin's and ours in several ways.

Non-uniform distributions on S_n. While we do not know whether the lower bound in Theorem 1 can be improved, we show that in a more general setting, the upper bound of order n^k obtained by the linear programming approach is tight.

In that more general setting, we do not want our family $\mathcal{F}$ of permutations to imitate, as far as the minima of at most k-element sets are concerned, the uniform distribution on S_n, but rather distributions where a random permutation is selected according to some given *priorities* of the elements. The priorities are positive real numbers $w_1, w_2, \ldots, w_n$, and a random permutation π is chosen, briefly speaking, by "sampling from the priorities without replacement". More precisely, we select π in n steps as follows: We maintain a current set C of elements, which is initialized to $[n]$ before the first step. In the ith step, we pick the element that becomes the ith element in the ordering given by π (that is, we pick $\pi^{-1}(i)$). We pick it at random from the current set C, where each $a \in C$ is chosen with probability

$$\frac{w_a}{\sum_{b \in C} w_b}.$$

The chosen element is deleted from C and we continue with the next step.

This distribution is known as the stationary distribution of the *Tsetlin library* Markov chain [18], and it arises naturally in computer science (e.g., for self-organizing lists), as well as in many applied contexts; see, for instance, Fill [7].

The above algorithmic description can easily be turned into a formal description of the resulting probability distribution $\mu = \mu(w_1, \ldots, w_n)$ on S_n, but we omit the formulas. The important property for us, whose (straightforward) proof we also omit, is that for $X \subseteq [n]$ and $x \in X$, the probability of $\min \pi(X) = \pi(x)$ for π chosen at random from μ is

$$\frac{w_x}{\sum_{y \in X} w_y}.$$

It should be clear what is meant by saying that a family $\mathcal{F} \subseteq S_n$ (with a probability distribution on it) is k-restricted min-wise independent *with respect to a given probability distribution* μ on S_n: For any at most k-element $X \subseteq [n]$ and any $x \in X$, the probability of $\min \pi(X) = \pi(x)$ is the same for π selected at random from $\mathcal{F}$ and for π selected at random from μ. Now we can state our (exact) result:

Theorem 2

(i) *Let μ be an arbitrary probability distribution on S_n. Then there exists a (biased) family $\mathcal{F} \subset S_n$ of size at most*

$$1 + \sum_{j=2}^{k} (j-1) \binom{n}{j}$$

that is k-restricted min-wise independent with respect to μ.

(ii) *There exist positive priorities $w_1, \ldots, w_n$ such that the upper bound in* (i) *cannot be improved for the distribution $\mu = \mu(w_1, \ldots, w_n)$. In fact, the set of positive vectors $(w_1, \ldots, w_n) \in \mathbb{R}^n$ that do not have this property is contained in the zero set of a non-zero polynomial in $w_1, \ldots, w_n$, and thus it is nowhere dense and of measure zero.*

In particular, for $k = n$ we obtain the bound $(n-2)2^{n-1}$.

The theorem is proved in Section 3. It is analogous to a result of Karloff and Mansour concerning k-wise independent random variables (although it is proved differently); for comparison, we outline the relevant notions and results.

Remark on k-wise independent random variables. Random variables $X_1, X_2, \ldots, X_n$ on some probability space are called *k-wise independent* if every k of them are mutually independent. The investigation of k-wise independence predates the study of min-wise independent permutations (and most of the questions, results, and techniques concerning min-wise independence have their analogs and inspirations there). Here we focus on results which seem most relevant to our topic. For simplicity, we consider k fixed in the following discussion.

Alon et al. [1] constructed a family of n k-wise independent random variables, each attaining values 0 and 1 with probability $\frac{1}{2}$, on a space of size $O(n^{\lfloor k/2 \rfloor})$. They also proved that any family of n k-wise independent random variables require size $\Omega(n^{\lfloor k/2 \rfloor})$, provided that none of the variables attains a single value with probability 1 (a special case of this result was independently obtained by Chor, Goldreich, Håstad, Friedman, Rudich, and Smolensky [6]).

On the other hand, Karloff and Mansour [12] showed that, for suitably chosen probabilities $p_1, \ldots, p_n$ (for example, $p_i = \frac{k}{n}$ will do), a family of k-wise independent 0/1 random variables $X_1, \ldots, X_n$, where X_i attains value 1 with probability p_i, requires space size $\Omega(n^k)$. This is the inspiration for Theorem 2.

Thus, the existence of a space of size $O(n^{\lfloor k/2 \rfloor})$ for the uniform 0/1 case can be considered as a "lucky coincidence" made possibly by symmetry. A very interesting open problem is, whether a "lucky" construction of an k-restricted min-wise independent family of permutations (with respect to the uniform distribution) exists, of size much smaller than n^k.

Restricted min-wise independence for sets of size exactly k. The min-uniform condition for all X of size *exactly* k does not generally imply the min-uniform condition for sets X of smaller size. One way of seeing this is to note that the order of the last $k-1$ elements in the permutations of $\mathcal{F}$ does not affect the min-uniform condition for sets of size k (or larger), while it does affect the min-uniform condition for sets smaller than k.

We thus define, for a set $S \subseteq [n]$, a family of permutations to be S-*restricted min-wise independent* if the min-uniform condition holds for all $X \subseteq [n]$ with $|X| \in S$. We will consider mainly the case with $S = \{k\}$.

The problem of estimating the size of $\{k\}$-restricted min-wise independent families was raised by Emo Welzl, and it seems challenging. We do not know of any substantial improvement of the $O(n^k)$ upper bound, but neither the proof method of Theorem 1 nor some other approaches used for lower bounds in the literature seem applicable here. At first sight, it is not even obvious that such a family cannot have size depending only on k. (On the other hand, if we admit more general distributions on S_n as in Theorem 2, we can obtain an exact bound, $(k-1)\binom{n}{k} + 1$, by following the proof of Theorem 2 almost literally.) Some results in this direction are presented in Section 4.

2 The lower bound for k-restricted min-wise independent families

All known proofs of Theorem 1 use the following somewhat surprising property of k-restricted min-wise independent families. We present a short proof (avoiding some calculations from [14]).

Proposition 3 ([14]) *Let $\mathcal{F}$ be a k-restricted min-wise independent family, let X be a k-element subset of $[n]$, and let $X = A \cup \{x\} \cup B$ be a partition of X. Then, for π chosen from $\mathcal{F}$ at random,*

$$\Pr[\pi(A) < \pi(x) < \pi(B)] = \frac{1}{(a+1)\binom{k}{a+1}}, \tag{1}$$

where $a = |A|$ (this probability is the same as for π chosen uniformly at random from S_n). Here $\pi(A) < \pi(x)$ means that every element of A precedes x in the ordering given by π.

Proof. We proceed by induction on $a = |A|$. The case $a = 0$ follows immediately from the min-uniform condition.

Now we suppose that (1) holds for all X and all partitions $X = A \cup \{x\} \cup B$ with $|A| < a$. Let us consider some X and some partition $X = A \cup \{x\} \cup B$ with $|A| = a$. We have

$$\begin{aligned} \frac{1}{k-a+1} &= \Pr[\pi(x) < \pi(B)\}] = \\ &= \sum_{C \subseteq A} \Pr[\pi(C) < \pi(x) < \pi(B \cup (A \setminus C))]\,. \end{aligned}$$

For every *proper* subset $C \subset A$ probability $\Pr[\pi(C) < \pi(x) < \pi(B \cup (A \setminus C))]$ is, by the inductive assumption, the same as if π were selected uniformly at random from S_n. Therefore, $\Pr[\pi(A) < \pi(x) < \pi(B)]$ must also be the same as for π selected uniformly at random from S_n. The proposition is proved. □

For later use, we note that the proof of the validity of (1) with $|A| \le a$ requires the min-uniform condition only for sets of size $k-a, k-a+1, \ldots, k$.

In our proof of Theorem 1 we are going to use a result on the rank of certain inclusion matrices. Let $L_n(i,j)$, $1 \le j \le i \le n$, be the 0/1 matrix with rows indexed by all i-element subsets $S \subseteq [n]$, columns indexed by all j-element subsets $T \subseteq [n]$, and with the entry at position (S, T) equal to 1 if $T \subseteq S$ and equal to 0 otherwise. This matrix is a representative of a wide class of so-called inclusion matrices; see, e.g., Babai and Frankl [2]. We need the following statement.

Theorem 4 (Gottlieb [8]) *For every i, j, $1 \leq j \leq i \leq n$, the matrix $L_n(i,j)$ has full rank; that is*

$$\operatorname{rank} L_n(i,j) = \min\left\{ \binom{n}{i}, \binom{n}{j} \right\}.$$

A different proof, which establishes much more, can be found in [2].

Proof of Theorem 1. Our basic strategy is inspired by Alon, Babai, and Itai [1]. We suppose that a k-restricted min-wise independent family $\mathcal{F}$ with its probability distribution is given, and we define a suitable family $(X_i : i \in I)$ of random variables on $\mathcal{F}$. We consider the $|I| \times |I|$ matrix $M_I := \Big(\mathbf{E}[X_i X_j]\Big)_{i,j \in I}$ (where $\mathbf{E}[X]$ denotes the expectation of X) and prove that it is non-singular. Then it follows that $|\mathcal{F}| \geq |I|$. Indeed, let us consider each random variable X_i as a vector with $|\mathcal{F}|$ components. We verify that these $|I|$ vectors are linearly independent and, consequently, the vector space of all $|\mathcal{F}|$-component vectors has dimension at least $|I|$.

To check the linear independence of the vectors X_i, let us suppose that $(\alpha_i : i \in I)$ are scalars such that $\sum_{i \in I} \alpha_i X_i = 0$. Then, for every $j \in I$,

$$X_j \sum_{i \in I} \alpha_i X_i = \sum_{i \in I} \alpha_i X_i X_j = 0,$$

and therefore,

$$\mathbf{E}\Bigg[\sum_{i \in I} \alpha_i X_i X_j\Bigg] = \sum_{i \in I} \alpha_i \mathbf{E}[X_i X_j] = 0.$$

Since M_I is non-singular, it follows that $\alpha_i = 0$ for all $i \in I$, and so the X_i are linearly independent.

In the construction of Alon et al. [1], the X_i attain real values, and they are constructed so that M_I is a non-singular diagonal matrix. We are going to use a more general M_I and *complex* random variables.

We let $\ell := \lfloor \frac{k-1}{2} \rfloor$. Our random variables will be indexed by ℓ-element subsets $S \subset [n-1]$; that is, $I = \binom{[n-1]}{\ell}$. They are defined as follows:

$$X_S(\pi) = \begin{cases} 1 & \text{if } \pi(n) < \pi(S), \\ \mathrm{i} & \text{if } \pi(S) < \pi(n), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Here $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. In other words, X_S has value 1 if n appears as the first element of the set $S \cup \{n\}$, value i if n appears as the last element, and value 0 otherwise.

We have

$$X_S X_T = \begin{cases} 1 & \text{if } \pi(n) < \pi(S \cup T), \\ -1 & \text{if } \pi(S \cup T) < \pi(n), \\ \mathrm{i} & \text{if either } \pi(S) < \pi(n) < \pi(T) \text{ or } \pi(T) < \pi(n) < \pi(S), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Using Proposition 3, we can calculate the expectations $\mathbf{E}[X_S X_T]$, i.e., the matrix M_I:

$$\mathbf{E}[X_S X_T] = \begin{cases} 0 & \text{if } S \cap T \neq \emptyset, \text{ and} \\ \mathrm{i}\frac{2}{(\ell+1)\binom{2\ell+1}{\ell+1}} & \text{if } S \cap T = \emptyset. \end{cases}$$

Thus, M_I is a constant multiple of the "intersection matrix" with the entry (S,T) being 1 for $S \cap T = \emptyset$ and 0 otherwise. This, in turn, is the same as the matrix $L_n(n-\ell-1, \ell)$, which is non-singular by Theorem 4. Theorem 1 is proved. □

3 Bounds for more general distributions on S_n

Proof of part (i) of Theorem 2. We follow the method of Koller and Megiddo [13]. Let μ be a given probability distribution on S_n, and for $X \subseteq [n]$ and $x \in X$, let $p_{X,x}$ be the probability of $\min \pi(X) = \pi(x)$ for π chosen at random from μ. (So $p_{X,x} = \frac{1}{|X|}$ if μ is the uniform distribution.)

To each permutation $\pi \in S_n$, we assign a real variable α_π. Under the conditions $\alpha_\pi \geq 0$ for all π and

$$\sum_{\pi \in S_n} \alpha_\pi = 1, \tag{2}$$

these variables specify a probability distribution on S_n. If we, moreover, postulate the conditions

$$\sum_{\pi \in S_n :\, \min \pi(X) = \pi(x)} \alpha_\pi = p_{X,x} \tag{3}$$

for every at most k-element $X \subseteq [n]$ and every $x \in X$, any solution to this system of linear inequalities specifies a probability distribution making S_n into a k-restricted min-wise independent family (with respect to μ).

Next, we note that some of the equations in the above system are redundant. Namely, for every fixed X, we have

$$\sum_{x \in X} \sum_{\pi \in S_n :\, \min \pi(X) = \pi(x)} \alpha_\pi = \sum_{\pi \in S_n} \alpha_\pi = 1,$$

and so one of the $|X|$ equations (3) can be omitted for every X. Hence, the conditions on the α_π can be specified by a system of $m_0 := 1 + \sum_{2 \le |X| \le k} (|X|-1)$ linear equations (m_0 is precisely the number appearing in the upper bound we want to prove).

This system certainly has a solution (namely, the one given by $\alpha_\pi = \mu(\{\pi\})$), and hence there is also a basic solution with at most m_0 non-zero components. By collecting the permutations with $\alpha_\pi \neq 0$ in such a solution, we obtain a (biased) k-restricted min-wise independent family of the desired size. □

It follows from part (ii) of Theorem 2 that no further equations from the above system can be eliminated. This can also be easily shown directly, by checking that the matrix of the system has rank m_0.

For the proof of part (ii) we need some preparations. Let $x = (x_1, x_2, \ldots, x_n)$ be a vector of n variables (indeterminates), and let $\mathbb{R}(x)$ denote the field of rational functions in $x_1, x_2, \ldots, x_n$ with real coefficients. Each element of $\mathbb{R}(x)$ can thus be written as $p(x)/q(x)$ for some n-variate polynomials $p(x)$ and $q(x) \neq 0$.

For each $X \subseteq [n]$ with $2 \le |X| \le k$, we fix one element $a_X \in X$, and we let $I = \{(X, a) : a \in X, a \neq a_X\}$. (We now write a instead of x for the element of X, in order to avoid confusion with the variables x_i.) Note that $|I| = m_0 - 1$, where m_0 is as in the proof of part (i) above.

For every $(X, a) \in I$, we define the rational function

$$f_{X,a}(x) := \frac{x_a}{\sum_{b \in X} x_b}$$

If $w = (w_1, w_2, \ldots, w_n)$ is a vector of priorities specifying a probability distribution $\mu(w) = \mu(w_1, \ldots, w_n)$ on S_n, then $f_{X,a}(w)$ is the probability of $\min \pi(X) = \pi(a)$.

Lemma 5 *The rational functions $f_{X,a}(x)$, $(X,a) \in I$, plus the constant rational function* 1, *are linearly independent when $\mathbb{R}(x)$ is considered as a (real) vector space.*

Proof. Suppose that β_1 and $\beta_{X,a}$, $(X,a) \in I$, are real numbers such that

$$\beta_1 + \sum_{(X,a)\in I} \beta_{X,a} f_{X,a}(x) = 0.$$

(the rational function equal to 0 everywhere). Substituting for $f_{X,a}$ and rearranging, we obtain

$$\beta_1 + \sum_X \frac{N_X(x)}{D_X(x)} = 0, \tag{4}$$

where $D_X(x) = \sum_{b\in X} x_b$ and $N_X(x) = \sum_{a\in X\setminus\{a_X\}} \beta_{X,a} x_a$. We are going to show that $N_X(x)$ is identically 0 for each X; then $\beta_1 = 0$ as well and the desired linear independence follows.

Let us fix some X. We multiply the equation (4) by the polynomial $R(x) := \prod_{Y\neq X} D_Y(x)$ (the product is over all Y of size between 2 and k except for $Y = X$). The left-hand side of the resulting equality is $R(x)N_X(x)/D_X(x)$ plus a polynomial, and so $R(x)N_X(x)/D_X(x)$ is a polynomial, too. Since the (irreducible) polynomial $D_X(x)$ does not divide $R(x)$, it must divide $N_X(x)$, and this is possible only if $N_X(x)$ is the zero polynomial. The lemma is proved. □

Proof of part (ii) of Theorem 2. Suppose that $w = (w_1, \ldots, w_n)$ is a vector of positive priorities, $\mathcal{F} \subset S_n$ is family of $m < m_0$ permutations, and p_π, $\pi \in \mathcal{F}$, are probabilities such that $\mathcal{F}$ with the distribution given by these p_π is k-restricted min-wise independent with respect to the distribution $\mu(w)$ on S_n.

Then the numbers $f_{X,a}(w)$, as well as the number 1, are expressible as a sum of some of the p_π. More explicitly, for $y = (y_\pi : \pi \in \mathcal{F})$, we let $\ell_1(y) = \sum_{\pi\in\mathcal{F}} y_\pi$ and

$$\ell_{X,a}(y) = \sum_{\pi\in\mathcal{F}\colon \pi(a)=\min\pi(X)} y_\pi, \qquad (X,a) \in I.$$

Then we have $1 = \ell_1(p)$ and $f_{X,a}(w) = \ell_{X,a}(p)$.

Now $\ell_1(y)$ and $\ell_{X,a}(y)$ are m_0 homogeneous linear polynomials in $m < m_0$ variables, and so they are linearly dependent. So there exist real numbers β_1 and $\beta_{X,a}$, $(X,a) \in I$, such that $\beta_1 \ell_1(y) + \sum_{(X,a)\in I} \beta_{X,a} \ell_{X,a}(y)$ is the zero polynomial. Then we obtain

$$\beta_1 + \sum_{(X,a)\in I} \beta_{X,a} f_{X,a}(w) = 0. \tag{5}$$

On the other hand, the rational function

$$F(x) = \beta_1 + \sum_{(X,a)\in I} \beta_{X,a} f_{X,a}(x)$$

is non-zero (as an element of $\mathbb{R}(x)$) by Lemma 5. Therefore, the w satisfying (5) are zeros of $F(x)$, and hence zeros of a non-zero polynomial.

So for each fixed family $\mathcal{F} \subset S_n$ of fewer than m_0 permutations, the set of w for which $\mathcal{F}$ can be made k-restricted min-wise independent with respect to $|mu(w)$ by some choice of the probabilities p_π is contained in the zero set of a non-zero polynomial. By taking the product of these polynomials over the finitely many possible choices of $\mathcal{F}$, we get a polynomial $P(x)$ such that whenever there exists a k-restricted min-wise independent family $\mathcal{F}$ with respect to $\mu(w)$ with $|\mathcal{F}| < m_0$, then $P(w) = 0$. Theorem 2 is proved. $\square$

4 Remarks on $\{k\}$-restricted min-wise independent families

Let $N(n,k)$ denote the minimum cardinality of a family $\mathcal{F} \subseteq S_n$ such that for every k-element $X \subseteq [n]$ and every $x \in X$, $\mathcal{F}$ contains a permutation π with $\min \pi(X) = \pi(x)$. Clearly, $N(n,k)$ is a lower bound for the size of any (possibly biased) $\{k\}$-restricted min-wise independent family.

Spencer [16] proved that $N(n,3) \geq \log_2 \log_2 n$ (to see this, consider a family of fewer permutations and, by iterated application of the Erdős–Szekeres lemma, find elements a, b, c such that b is between a and c in each of the permutations). It is also easy to see that $N(n,k) \geq N(n-1,k-1)+1$: Given a family attaining $N(n,k)$, we delete one (arbitrarily chosen) permutation π, as well as the first element under that permutation. We thus get:

Proposition 6 *If $\mathcal{F}$ is a (possibly biased) $\{k\}$-restricted min-wise independent family, $k \geq 3$, then*

$$|\mathcal{F}_k| \geq N(n,k) \geq \lceil \log_2(\log_2(n-k+2)) \rceil + k - 2.$$

For a *uniform* family $\mathcal{F}$, the leading constant can be improved a little. Namely, consider a subfamily $\mathcal{F}' \subseteq \mathcal{F}$ consisting of $N(n,k)-1$ permutations. Then there are X, $|X| = k$, and $x \in X$ such that $\min \pi(X) \neq \pi(x)$ for all $\pi \in \mathcal{F}'$. Then, for π chosen at random from $\mathcal{F}$, we have

$$\frac{1}{k} = \Pr[\min \pi(X) = \pi(x)] \leq \frac{|\mathcal{F}| - |\mathcal{F}'|}{|\mathcal{F}|} = \frac{|\mathcal{F}| - (N(n,k)-1)}{|\mathcal{F}|}.$$

This yields the following:

Proposition 7 *If $\mathcal{F}$ is a uniform $\{k\}$-restricted min-wise independent family, $k \geq 3$, then*

$$|\mathcal{F}| \geq \frac{k}{k-1}\Big(\lceil \log_2(\log_2(n-k+2)) \rceil + (k-3) \Big).$$

As a construction due to Hajnal, presented in [16], shows, we have $N(n,k) = O(\log\log n)$ for every fixed k, and thus, unfortunately, the lower bound in Proposition 6 cannot be further improved by this approach.

The following proposition addresses an "intermediate case" between k-restricted min-wise independent families and $\{k\}$-restricted min-wise independent families.

cases

Proposition 8 *Let $\mathcal{F} \subseteq S_n$ be a (possibly biased) $\{k-1, k\}$-restricted min-wise independent family. Then $|\mathcal{F}| \geq n-k+2$.*

Proof. We proceed as in the proof of Theorem 1, defining a suitable family $(X_i)_{i \in I}$ of random variables. This time we let $I := [n-k+2]$, $S := \{n-k+3, n-k+4, \ldots, n\}$, and

$$X_i = \begin{cases} 1 & \text{if } \pi(i) < \pi(S), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

We have $\mathbf{E}\left[X_i^2\right] = \frac{1}{k-1}$. By the remark following the proof of Proposition 3, we can further calculate $\mathbf{E}[X_i X_j] = \frac{2}{k(k-1)}$ for $i \neq j$. The resulting matrix M_I is easily checked to be non-singular, and the proposition is proved. $\square$

Open problems

1. The most interesting questions seems to be the order of magnitude for k-restricted min-wise independent families of permutations with k fixed and $n \to \infty$. Is the exponent about $\frac{k}{2}$, or about k, or...?

2. Can one improve the lower (or upper) bound for $\{k\}$-restricted min-wise independent families?

Acknowledgment

We would like to thank Emo Welzl for bringing the problems studied in this paper to our attention, formulating one of them, and for numerous discussions and valuable suggestions. We also thank Pavel Pudlák and Jiří Sgall for kindly answering questions and pointing out references. Finally, we thank Yoshinori Takei for kindly sending us the manuscript [11].

References

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

[2] L. Babai and P. Frankl. *Linear algebra methods in combinatorics (Preliminary version 2).* Department of Computer Science, The University of Chicago, 1992.

[3] A.Z. Broder, M. Charikar, A.M. Frieze, M. Mitzenmacher: Min-wise independent permutations, *J. Comput. Syst. Sci.* 60(2000), 630–659. Preliminary version in *Proc. of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, 1998.

[4] A.Z. Broder, M. Charikar, M. Mitzenmacher: A derandomization using min-wise independent permutations. In *Randomization and approximation techniques in computer science (Barcelona, 1998)*, 15–24, Lecture Notes in Comput. Sci., 1518, Springer, Berlin, 1998.

[5] A.Z. Broder, M. Mitzenmacher: Completeness and robustness properties of min-wise independent permutations, *Random Structures and Algorithms*, 18(2001), 18–30.

[6] B. Chor, O. Goldreich, J. Håstad, J. Friedman, S. Rudich, R. Smolensky: The Bit Extraction Problem of t-Resilient Functions (Preliminary Version). *Proc. 26th IEEE Sympos. Foundat. Comput. Sci. (FOCS)*, 1985, pages 396–407.

[7] J. A. Fill: Limits and rates of convergence for the distribution of search cost under the move-to-front rule, *Theor. Comput. Sci.* 164,1-2(1996), 185–206.

[8] D.H. Gottlieb: A certain class of incidence matrices, *Proc. A.M.S.* 17(1966), 1233–1237.

[9] P. Indyk: A small approximately min-wise independent family of hash functions, *J. Algorithms* 38(2001), 84–90.

[10] T. Itoh, Y. Takei, J. Tarui: On permutations with limited-independence, *Proc. 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2000, 137–146.

[11] T. Itoh, Y. Takei, J. Tarui: On the Sample Size of k-Min-Wise Independent Permutations and Other k-Wise Distributions, manuscript, 2002.

[12] H. Karloff and Y. Mansour: On construction of k-wise independent random variables, *Combinatorica* 17,1(1997), 91–107.

[13] D. Koller, N. Megiddo: Constructing small sample spaces satisfying given constraints, *SIAM J. Discrete Math.* 7(1994), 260–274.

[14] S.A. Norin: A polynomial lower bound for the size of any k-min-wise independent set of permutations (in Russian), *Zapiski Nauchnyh Seminarov POMI*, 227(2001), 104–116.

[15] M. Saks, A. Srinivasan, S. Zhou, D. Zuckerman: Low discrepancy sets yield approximate min-wise independent permutation families, *Inform. Process. Lett.* 73(2000), 29–32.

[16] J. Spencer: Minimal scrambling sets of simple orders, *Acta Mathematica Academiae Scientarium Hungaricae* 22(1971), 349–353.

[17] Y. Takei, T. Itoh, T. Shinozaki: Constructing an optimal family of min-wise independent permutations, *IECE Trans. Fundamentals* 83-A,4(2000) 747–755.

[18] M. L. Tsetlin: Finite automata and models of simple forms of behavior, *Russian Math. Surv.* 18,4(1963), 1–27.