Combinatorics, Probability and Computing (2009) 18, 3–15. © 2008 Cambridge University Press
doi:10.1017/S0963548307008917 Printed in the United Kingdom

Perturbed Identity Matrices Have High Rank: Proof and Applications

NOGA ALON[†]

Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
(e-mail: nogaa@tau.ac.il)

Received 15 November 2006; revised 17 November 2007; first published online 16 January 2008

We describe a lower bound for the rank of any real matrix in which all diagonal entries are significantly larger in absolute value than all other entries, and discuss several applications of this result to the study of problems in Geometry, Coding Theory, Extremal Finite Set Theory and Probability. This is partly a survey, containing a unified approach for proving various known results, but it contains several new results as well.

1. Introduction

Let $B = (b_{i,j})$ be an n by n real matrix. It is easy and well known that if, for every i, $|b_{i,i}| > \sum_{j \neq i} |b_{i,j}|$, then B is of full rank. Indeed, assuming this is false, let $c = (c_j)$ be a non-zero column vector such that $Bc = 0$. Let $|c_r| = \max_i |c_i|$ (> 0) and consider the component number r of Bc. The absolute value of this component is

$$\left| \sum_j b_{r,j} c_j \right| \geqslant |b_{r,r} c_r| - \sum_{j \neq r} |b_{r,j} c_j| \geqslant |c_r| \left(|b_{r,r}| - \sum_{j \neq r} |b_{r,j}| \right) > 0,$$

contradicting the assumption $Bc = 0$ and proving that B indeed has full rank. In particular, this implies that if $b_{i,i} = 1$ for all i and $|b_{i,j}| \leqslant \frac{1}{n}$ for all distinct indices i, j, then the rank of B is n.

Suppose we relax the conditions above, and only assume that each diagonal entry is, in absolute value, at least $1/2$ and the absolute value of each other entry is at most ϵ. In this case one can also establish a lower bound for the rank of B, as stated in the following theorem.

Theorem 1.1. *There exists an absolute positive constant c such that the following holds. Let B be an n by n real matrix with $|b_{i,i}| \geqslant 1/2$ for all i and $|b_{i,j}| \leqslant \epsilon$ for all $i \neq j$, where $\frac{1}{2\sqrt{n}} \leqslant \epsilon <$*

[†] Research supported in part by the Israel Science Foundation, by a USA–Israel BSF grant, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

1/4. *Then the rank of B satisfies*

$$\mathrm{rank}(B) \geqslant \frac{c}{\epsilon^2 \log(1/\epsilon)} \log n.$$

This theorem is a slight variation of a result proved in [1]. In this short paper we present the proof of the theorem, and describe several applications in various areas. Some of these applications are known, and some, including a solution of one of the open problems raised in [16], are new.

The rest of the paper is organized as follows. In Section 2 we present the proof of the theorem, in Sections 3, 4, 5, 6 and 7 we describe its applications in Geometry, Coding Theory, Extremal Finite Set Theory, the investigation of pseudo-random sequences, and the study of small sample spaces supporting nearly independent random variables. The final section, Section 8, contains some concluding remarks and open problems.

2. Perturbed identity matrices

It is convenient to first prove the following variant of Theorem 1.1.

Theorem 2.1. *There exists an absolute positive constant c such that the following holds. Let B be an n by n real matrix with $b_{i,i} = 1$ for all i and $|b_{i,j}| \leqslant \epsilon$ for all $i \neq j$. If the rank of B is d, and $\frac{1}{\sqrt{n}} \leqslant \epsilon < 1/2$, then*

$$d \geqslant \frac{c}{\epsilon^2 \log(1/\epsilon)} \log n.$$

This result is proved in [1]. For completeness, we reproduce the proof (omitting the final detailed computation). We need the following well-known lemma proved, among other places, in [8, 1].

Lemma 2.2. *Let $A = (a_{i,j})$ be an n by n real, symmetric matrix with $a_{i,i} = 1$ for all i and $|a_{i,j}| \leqslant \epsilon$ for all $i \neq j$. If the rank of A is d, then*

$$d \geqslant \frac{n}{1 + (n-1)\epsilon^2}.$$

In particular, if $\epsilon \leqslant \frac{1}{\sqrt{n}}$ then $d > n/2$.

Proof. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of A; then their sum is the trace of A, which is n, and at most d of them are non-zero. Thus, by Cauchy–Schwarz, $\sum_{i=1}^{n} \lambda_i^2 \geqslant d(n/d)^2 = n^2/d$. On the other hand, this sum is the trace of $A^t A$, which is precisely $\sum_{i,j} a_{i,j}^2 \leqslant n + n(n-1)\epsilon^2$. Hence $n + n(n-1)\epsilon^2 \geqslant n^2/d$, implying the desired result. □

Lemma 2.3. *Let $B = (b_{i,j})$ be an n by n matrix of rank d, and let $P(x)$ be an arbitrary polynomial of degree k. Then the rank of the n by n matrix $(P(b_{i,j}))$ is at most $\binom{k+d}{k}$. Moreover, if $P(x) = x^k$ then the rank of $(P(b_{i,j}))$ is at most $\binom{k+d-1}{k}$.*

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Proof. Let $\mathbf{v_1} = (v_{1,j})_{j=1}^n$, $\mathbf{v}_2 = (v_{2,j})_{j=1}^n, \ldots, \mathbf{v_d} = (v_{d,j})_{j=1}^n$ be a basis of the row-space of B. Then the vectors $(v_{1,j}^{k_1} \cdot v_{2,j}^{k_2} \cdots v_{d,j}^{k_d})_{j=1}^n$, where $k_1, k_2, \ldots, k_d$ range over all non-negative integers whose sum is at most k, span the rows of the matrix $(P(b_{i,j}))$. For $P(x) = x^k$ it suffices to take all these vectors corresponding to $k_1, k_2, \ldots, k_d$ whose sum is precisely k. □

Remark. It is worth noting that there is no analogue to the last lemma if the entries of the matrix are raised to a fractional power. In fact, for every $n > 1$ there is an n by n real matrix $B = (b_{i,j})$ of rank 2, such that the matrix $(b_{i,j}^{1/2})$ has full rank. Indeed, let $3 = p_1 < p_2 < \cdots < p_n$ be the first n odd primes, and consider the matrix $B = (b_{i,j})$ given by $b_{i,j} = i + p_j - j$. Clearly B has rank 2. We prove, by induction on n, that the matrix $(b_{i,j}^{1/2})$ has rank n. This is trivially true for $n = 1$. Assuming it holds for $n-1$, suppose that the n by n matrix $(b_{i,j}^{1/2})$ does not have full rank. Expanding its determinant according to the last row, we get that $\sqrt{p_n}$ times the determinant of the $(n-1)$ by $(n-1)$ matrix $(b_{i,j}^{1/2})$, $1 \leqslant i, j \leqslant n-1$, which is non-zero by the induction hypothesis, lies in the field $Q[\sqrt{3}, \sqrt{5}, \ldots, \sqrt{p_{n-1}}]$. This implies that $\sqrt{p_n}$ lies in that field, and it is well known that this is false, supplying the desired result.

Proof of Theorem 2.1. We may and will assume that B is symmetric, since otherwise we simply apply the result to $(B + B^t)/2$ whose rank is at most twice the rank of B. If $\epsilon \leqslant 1/n^\delta$ for some fixed $\delta > 0$, the result follows by applying Lemma 2.2 to a $\lfloor \frac{1}{\epsilon^2} \rfloor$ by $\lfloor \frac{1}{\epsilon^2} \rfloor$ submatrix of B. Thus we may assume that $\epsilon \geqslant 1/n^\delta$ for some fixed, small $\delta > 0$. Put $k = \lfloor \frac{\log n}{2\log(1/\epsilon)} \rfloor$, $n' = \lfloor \frac{1}{\epsilon^{2k}} \rfloor$ and note that $n' \leqslant n$ and that $\epsilon^k \leqslant \frac{1}{\sqrt{n'}}$. By Lemma 2.3 the rank of the n' by n' matrix $(b_{i,j}^k)_{i,j \leqslant n'}$ is at most $\binom{d+k}{k} \leqslant (\frac{e(k+d)}{k})^k$. On the other hand, by Lemma 2.2, the rank of this matrix is at least $n'/2$. Therefore

$$\left(\frac{e(k+d)}{k}\right)^k \geqslant \frac{n'}{2} = \frac{1}{2}\left\lfloor \frac{1}{\epsilon^{2k}} \right\rfloor,$$

and the desired result follows by some simple manipulation, that can be found, for example, in [4]. □

Proof of Theorem 1.1. Let $C = (c_{i,j})$ be the n by n 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 2ϵ. The result thus follows from Theorem 2.1. □

3. Distortion in low-dimension embeddings

A well-known lemma of Johnson and Lindenstrauss, proved in [11] (see also [15]), asserts that for any $\epsilon > 0$, any set A of n points in an Euclidean space can be embedded in an Euclidean space of dimension $k = c(\epsilon) \log n$ with distortion at most ϵ. That is, there is a mapping $f : A \mapsto R^k$ such that for any $a, b \in A$, the distance between $f(a)$ and $f(b)$ is at least the distance between a and b, and at most that distance multiplied by $1 + \epsilon$. The proof gives that $c(\epsilon) \leqslant O(\frac{1}{\epsilon^2})$. Theorem 2.1 can be used to show that this is nearly tight: $c(\epsilon)$ must be at least $\Omega\big(\frac{1}{\epsilon^2 \log(1/\epsilon)}\big)$, even for embedding the set of points of a simplex. This is stated in the following proposition, proved in [1].

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Proposition 3.1. *Let $P_0, P_1, \ldots, P_n$ be a set of $n+1$ points in R^k, and suppose that the distance between any two of them is at least 1 and at most $1+\epsilon$, where $\frac{1}{\sqrt{n}} \leqslant \epsilon \leqslant \frac{1}{10}$. Then $k \geqslant \frac{c'}{\epsilon^2 \log(1/\epsilon)} \log n$, where c' is an absolute positive constant.* □

Proof. Put one of the points, say P_0, at the origin, and shift all other points by at most ϵ, making sure that their distance from P_0 is exactly 1. By the triangle inequality, the distance between any pair of the shifted points is still $1 + O(\epsilon)$. Therefore, if v_i is the k-dimensional vector representing the ith point, then the gram matrix $C = (v_i^t \cdot v_j)$ is an n by n matrix in which all diagonal entries are 1, and all other entries are $1/2 + O(\epsilon)$. Moreover, the rank of this matrix is at most k. Therefore, the rank of $B = 2C - J$, where J is the all-1 n by n matrix, is at most $k+1$. By Theorem 2.1 this rank is at least $\Omega\big(\frac{1}{\epsilon^2 \log(1/\epsilon)} \log n\big)$, supplying the required lower bound for the dimension k. □

4. Coding theory

A binary code of length k is a set $C \subset \{0,1\}^k$ of binary vectors with k coordinates. The code is called *ϵ-balanced* if the Hamming distance between any two code-words is at least $\frac{1-\epsilon}{2}k$ and at most $\frac{1+\epsilon}{2}k$. For each vector $v = (v_1, v_2, \ldots, v_k) \in C$, let $x(v)$ denote the vector

$$x(v) = \big((-1)^{v_1}, (-1)^{v_2}, \ldots, (-1)^{v_k}\big) \in \{-1,1\}^k.$$

Note that for any two $u, v \in C$, the inner product between $x(u)$ and $x(v)$ is precisely $k - 2h(u,v)$, where $h(u,v)$ is the Hamming distance between u and v.

It follows that for $\epsilon = 0$, every two vectors $x(u), x(v)$ corresponding to distinct code-words of an ϵ-balanced code are orthogonal, and hence the number of code-words is at most k. Any Hadamard matrix of order k (if one exists) shows that this is tight, hence this is tight for all powers of 2 as well as for many other values of k divisible by 4 (see, *e.g.*, [9] for more information about the existence of Hadamard matrices.)

For positive values of ϵ, the problem of determining or estimating the largest possible cardinality of an ϵ-balanced code of length k is more complicated. Note, first, that ϵ should be at least $1/k$, since otherwise any ϵ-balanced code of length k is, in fact, 0-balanced. A simple probabilistic argument (or an obvious variant of the Gilbert Varshamov bound) shows that there are ϵ-balanced codes of length k with at least $2^{\Omega(\epsilon^2 k)}$ code-words. Theorem 2.1 provides a quick upper bound, as follows.

Proposition 4.1. *There exists an absolute positive constant a such that, for all $\frac{1}{\sqrt{k}} \leqslant \epsilon < 1/2$, the cardinality of any ϵ-balanced code of length k is at most $2^{a\epsilon^2 \log(1/\epsilon)k}$.* □

Proof. Let $C \subset \{0,1\}^k$ be an ϵ-balanced code of length k and maximum cardinality. Put $n = |C|$ and note that we may assume that $n \geqslant k$. Let X be the n by k matrix whose rows are the $|C|$ vectors $\frac{x(v)}{\sqrt{k}}$, $v \in C$. Let B be the n by n matrix defined by $B = (b_{u,v}) = XX^t$. Then each diagonal entry $b_{u,u}$ of B is 1, where every other entry $b_{u,v}$ for $u \neq v$, $u, v \in C$, satisfies $|b_{u,v}| = |\frac{1}{k}(k - 2h(u,v))| \leqslant \epsilon$. Therefore, by Theorem 2.1,

$$k \geqslant \operatorname{rank}(X) \geqslant \operatorname{rank}(B) \geqslant \frac{c}{\epsilon^2 \log(1/\epsilon)} \log n,$$

supplying the desired result. □

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Note that the assertion of the last proposition, at least for fixed ϵ and large n, can be also deduced, in a completely different manner, from the Linear Programming technique of Delsarte and the McEliece–Rodemich–Rumsey–Welch bound (see, *e.g.*, [13, p. 559]). It is also worth noting that, as is well known, the Plotkin bound (see, *e.g.*, [13, pp. 41–43]) implies that any ϵ-balanced code of length k has at most $O(2^{\epsilon k/2}k)$ code-words, and this bound holds even if we do not assume any upper bound on the Hamming weights of the code-words, only a $\frac{1-\epsilon}{2}k$ lower bound. The interesting part in the last proposition is, however, the quadratic dependence on ϵ in the exponent.

When ϵ is smaller than $\frac{1}{\sqrt{k}}$ we can repeat the above proof, but apply Lemma 2.2 instead of Theorem 2.1, as stated in the next proposition.

Proposition 4.2. *Suppose $\epsilon = \frac{1}{w\sqrt{k}}$, where $w > 1$. Then the cardinality of any ϵ-balanced code C of length k is smaller than $k\frac{w^2}{w^2-1}$.* □

Proof. Put $n = |C|$. Applying Lemma 2.2 to the matrix B defined from the code C as in the previous proof, we conclude that

$$k \geqslant \mathrm{rank}(B) \geqslant \frac{n}{1 + (n-1)/(w^2k)},$$

implying the desired bound. □

Thus, in particular, if $w \geqslant \sqrt{2}$ then $n < 2k$, and if w tends to infinity with k, then $n \leqslant (1 + o(1))k$.

5. Cross-intersecting pairs

Extremal Finite Set Theory deals with various instances of the problem of determining or estimating the maximum or minimum possible cardinality of a collection of subsets of a k-element set that satisfies some given conditions. Rank arguments are often useful in obtaining results in this area: see, *e.g.*, [12] for several examples. It is therefore not surprising that one can apply Theorems 1.1 and 2.1 (or Lemma 2.2) in the investigation of problems of this type. Here we only describe one representative example.

Proposition 5.1. *Let c, α be positive constants satisfying $c\alpha > 1$. Let $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ be a collection of n pairs of subsets of a k-element set. Suppose that $X_i \cap Y_i = \emptyset$ for all $i \in [n] = \{1, 2, \ldots, n\}$, and that for all distinct $i, j \in [n]$,*

$$||X_i \cap Y_j| - c(k+1)| < \frac{\sqrt{k+1}}{\alpha}.$$

Then the number of pairs, n, satisfies $n < \frac{c^2\alpha^2}{c^2\alpha^2-1}(k+1)$. □

Proof. Let X be the n by k matrix whose rows are the incidence vectors of the sets X_i, and let Y be the k by n matrix whose columns are the incidence vectors of the sets Y_j. Then the product $Z = XY$ is an n by n matrix in which each diagonal entry is zero, and each other entry deviates

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

from $c(k+1)$ by at most $\frac{\sqrt{k+1}}{\alpha}$. Let J be the n by n matrix in which all entries are 1, and define $B = \frac{1}{c(k+1)}(c(k+1)J - Z)$. Then, each diagonal entry of B is 1, and the absolute value of each other entry is at most

$$\frac{\sqrt{k+1}}{\alpha}\frac{1}{c(k+1)} = \frac{1}{c\alpha\sqrt{k+1}}.$$

Note that the rank of B is at most $k+1$, as the rank of Z does not exceed k. On the other hand, by Lemma 2.2, the rank of B is larger than

$$\frac{n}{1+n/(c^2\alpha^2(k+1))}.$$

It follows that

$$k+1 > \frac{n}{1+n/(c^2\alpha^2(k+1))},$$

implying that $n < \frac{c^2\alpha^2}{c^2\alpha^2-1}(k+1)$, as needed. □

By the above proposition, whenever $c\alpha$ is bounded away from 1, the maximum possible number of pairs is linear in k. The existence of Hadamard matrices shows that, for an appropriate c, this number is at least $(1-o(1))k$ even if α is arbitrarily large, implying that the above estimate is nearly tight.

6. Pseudo-randomness

In a series of papers, Mauduit and Sárközy studied finite pseudo-random binary sequences $E_N = (e_1, \ldots, e_N) \in \{-1,1\}^N$. In particular, they investigated in [14] a certain measure of pseudo-randomness, defined as follows.

Given $k, M \leqslant N$ and $D = \{d_1, \ldots, d_k\}$, where the d_i are integers with $1 \leqslant d_1 < \cdots < d_k \leqslant N - M + 1$, define

$$V(E_N, M, D) = \sum_{0 \leqslant n < M} \prod_{1 \leqslant i \leqslant k} e_{n+d_i} = \sum_{0 \leqslant n < M} \prod_{d \in D} e_{n+d}.$$

The *correlation measure* of order k of E_N is defined as

$$C_k(E_N) = \max\{|V(E_N, M, D)| \quad M \text{ and } D \text{ such that } M - 1 + d_k \leqslant N\}.$$

Improving an estimate of [7], the following is proved in [4] (among other related results).

Theorem 6.1 ([4], Theorem 1.2). *There is an absolute constant $c > 0$ for which the following holds. For any positive integers ℓ and N with $\ell \leqslant N/3$, we have*

$$\max\{C_2(E_N), C_4(E_N), \ldots, C_{2\ell}(E_N)\} \geqslant c\sqrt{\ell N},$$

for all $E_N \in \{-1,1\}^N$.

The proof is a simple consequence of Theorem 2.1. Here is a sketch. Fix a sequence $E_N = (e_1, e_2, \ldots, e_N)$ for which the above maximum is as small as possible, and denote it by T. For

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

every subset A of at most ℓ distinct members of $\{1, 2, \ldots, 2N/3\}$, consider the $\{-1, 1\}$-vector $x(A)$ of length $N/3$ whose ith coordinate, for $1 \leqslant i \leqslant N/3$, is the product $\prod_{a \in A} e_{i+a}$. The set of all vectors $x(A)$ is a set of $\sum_{j=0}^{\ell} \binom{2N/3}{j}$ vectors. The inner product of any two distinct vectors in this set is, in absolute value, at most T. Therefore, the gram matrix of the vectors, divided by $N/3$, has 1 in each diagonal entry, and an element of absolute value at most $3T/N$ in each other entry. It follows, by Theorem 2.1, that its rank is at least

$$\Omega\left(\frac{N^2}{T^2 \log(N/T)} \log\left[\sum_{j=0}^{\ell} \binom{2N/3}{j}\right]\right).$$

However, this rank is at most $2N/3$, implying that $2N/3$ is at least as large as the last expression. This implies the assertion of the theorem by some simple calculation, which is omitted. For more details see [4], where it is also shown that this estimate is sharp up to a logarithmic factor.

7. Derandomization

7.1. Nearly independent random variables

Let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables over a sample space S of size m, and suppose each variable takes values in $\{-1, 1\}$. For every subset $Y \subset [n]$, let X_Y denote the random variable $X_Y = \Pi_{i \in Y} X_i$. The family X is called *ϵ-biased* if, for every non-empty Y,

$$|\text{Prob}[X_Y = 1] - \text{Prob}[X_Y = -1]| \leqslant \epsilon.$$

Note that it is more common to consider random variables attaining values in $\{0, 1\}$, and look at their linear combinations over Z_2, but the above definition is equivalent.

It is known (see [3]) that if S is a *uniform* sample space of size m supporting an ϵ-biased set X as above, where $\epsilon \geqslant 2^{-n/2}$, then $m \geqslant \Omega\left(\frac{n}{\epsilon^2 \log(1/\epsilon)}\right)$. Here we show that the same lower bound applies even without the assumption that S is uniform.

Theorem 7.1. *Let $X = \{X_1, X_2, \ldots, X_n\}$ be an ϵ-biased set of n random variables over a sample space $S = \{s_1, s_2, \ldots, s_m\}$ of size m. If $\epsilon \geqslant 2^{-n/2}$ then $m \geqslant \Omega\left(\frac{n}{\epsilon^2 \log(1/\epsilon)}\right)$. If $\epsilon < 2^{-n/2}$ then $m \geqslant \Omega(2^n)$.*

Proof. Suppose $S = \{s_1, s_2, \ldots, s_m\}$, where $\text{Prob}(s_i) = p_i$. Define a 2^n by m matrix $U = (U_{Y, s_j})$ whose rows are indexed by the family of all subsets Y of $[n]$, and whose columns are indexed by the points of S as follows: $U_{Y, s_j} = X_Y(s_j)\sqrt{p_j}$.

Put $A = UU^T$ and note that for every two subsets Y_1, Y_2 of $[n]$,

$$A_{Y_1, Y_2} = \text{Prob}[X_{Y_1 \oplus Y_2} = 1] - \text{Prob}[X_{Y_1 \oplus Y_2} = -1].$$

Therefore, all diagonal entries of A are 1, whereas all off-diagonal entries are, in absolute value, at most ϵ. By Theorem 2.1, if $\epsilon \geqslant 2^{-n/2}$ then

$$m \geqslant \text{rank}(A) \geqslant \Omega\left(\frac{\log(2^n)}{\epsilon^2 \log(1/\epsilon)}\right),$$

completing the proof for $\epsilon \geqslant 2^{-n/2}$. The result for $\epsilon < 2^{-n/2}$ follows from the case $\epsilon = 2^{-n/2}$. □

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Remark. A similar proof implies that the size m of any (not necessarily uniform) sample space that supports a family of n random variables in which every set of k is ϵ-biased, where $\epsilon \geqslant [\binom{n}{\lfloor k/2 \rfloor}]^{-1/2}$, satisfies

$$m \geqslant \Omega\left(\frac{k \log(n/k)}{\epsilon^2 \log(1/\epsilon)}\right).$$

As is the case with Theorem 7.1, this is tight, up to the $\log(1/\epsilon)$-term. The proof (for the uniform case) appears in [2].

7.2. Nearly min-wise independent permutations

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 non-empty 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 was introduced by Broder, Charikar, Frieze and Mitzenmacher [6], motivated by applications in data mining. It is shown in [6] that there are such families of size at most $O\left(\frac{k^2}{\varepsilon^2} \log\left(\frac{n}{k}\right)\right)$ and that each such family must be of size at least $\Omega\left(k^2(1 - \sqrt{8\varepsilon})\right)$ in the uniform case, and at least

$$\Omega\left(\min\left\{k2^{k/2} \log\left(\frac{n}{k}\right), \frac{\log(1/\varepsilon)(\log n - \log\log(1/\varepsilon))}{\varepsilon^{1/3}}\right\}\right)$$

in the biased case.

The lower estimates are improved in [5], where the following two results are proved. Note that both supply lower bounds for the biased case that improve even the known bounds for the uniform case.

Theorem 7.2. *For any $1/3 > \varepsilon > 0$ and $k \geqslant 3$, 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*

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

Theorem 7.3. *For any $1/3 > \varepsilon > 0$ and $k \geqslant 3$, 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*

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

The proofs are based on Theorem 1.1, together with some additional linear-algebra arguments. Here is the proof of the first result.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Proof of Theorem 7.2. Let $\mathcal{F}$ be an (ε, k)-min-wise independent family of permutations of $[n]$, with respect to the distribution D, where $\varepsilon > 0$, $k \geqslant 3$ 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:

$$u_{ij}^{(h)} = \begin{cases} \sqrt{\text{Prob}_D(\pi_j)} & \text{if } \min(\pi_j(X_0 \cup X_i)) = \pi_j(h), \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

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 $\Omega\left(\frac{1}{\varepsilon^2 \log(1/\varepsilon)} \log m\right)$. 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 $6s$, we get a matrix in which each diagonal entry is at least $\frac{1}{2}$, and each off-diagonal entry is in absolute value at most 2ε. As the above subtraction and multiplication can change the rank by at most 1, the assertion of the claim follows from Theorem 1.1. 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 $|\mathcal{F}| = d \geqslant \text{rank}(V)$, the assertion of the theorem follows. □

The proof of Theorem 7.3 is similar, with an extra combinatorial argument. The idea is to replace the family of sets $\{X_1, X_2, \ldots, X_{L-1}\}$ in the proof above by a larger family of s-subsets of $[n] - X_0$, so that the intersection of every two of them is at most ϵs. The full details appear in [5].

7.3. Min-wise independence for sets of size exactly k

Call a family of permutations $\mathcal{F}$ of $[n]$, with a distribution D, *exactly k min-wise independent* if, for every subset X of exactly k elements of $[n]$ and every $x \in X$, when a random permutation π is chosen according to the distribution D, then $\text{Prob}_D(\min \pi(X) = \pi(x)) = \frac{1}{|X|}$.

In case the above holds for every subset X of at most k elements of $[n]$, $\mathcal{F}$ is called *at most k min-wise independent*. Note that this last notion coincides with the notion of (ϵ, k)-min-wise independent family considered in Section 7.2, for the special case $\epsilon = 0$. There are several papers dealing with the minimum possible cardinality of a family of at most k min-wise independent permutations. In [10] it is shown, slightly improving estimates of [17] and [16], that any such family must be of size at least $\sum_{i=0}^{(k-1)/2} \binom{n-1}{i}$ for odd k, and of size at least $\sum_{i=0}^{k/2-1} \binom{n-1}{i} + \binom{n-2}{k/2-1}$ for even k. On the other hand, it is known (see [16], slightly improving [6]), that there are such families of size at most $1 + \sum_{j=2}^{k} (j-1)\binom{n}{j}$.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

Much less is known about the minimum possible size of exactly k min-wise independent families of permutations of $[n]$. The best-known upper bound is $1+(k-1)\binom{n}{k}$, proved in [16], which is similar to the bound for at most k min-wise independent families, whereas the best-known lower bound is only $\lceil \log_2 \log_2(n-k+2)\rceil + k - 2$. Indeed, one of the two open problems mentioned in [16] is the problem of improving this lower bound. This is done in the following theorem.

Theorem 7.4. *For any n and $k \geqslant 3$, and for any exactly k min-wise independent family of permutations of $[n]$ $\mathcal{F}$ with respect to a distribution D,*

$$|\mathcal{F}| \geqslant \frac{(k-2)(n-2k+3)}{k-1}.$$

While this bound is still significantly smaller than the upper bound, it is far better than the double logarithmic known bound, and for every fixed k, it grows linearly with n. The proof, given below, does not apply Theorems 1.1 and 2.1, but we include it here as it is similar to the proof of Theorem 7.2, combining the basic approach therein with a simple probabilistic argument.

Proof of Theorem 7.4. Let $\mathcal{F}$ be an exactly k min-wise independent family of permutations with respect to a distribution D, where $k \geqslant 3$. Clearly, for every subset Y of $k-1$ elements of $[n]$ there is at least one element $y \in Y$ such that

$$\mathrm{Prob}_D(\min \pi(Y) = \pi(y)) \geqslant \frac{1}{k-1} \left(> \frac{1}{k}\right).$$

It follows that if one chooses, randomly and uniformly, a subset $Y \subset [n]$ of cardinality $k-1$, and a member $y \in Y$, then with probability at least $1/(k-1)$, the probability (with respect to D) that $\min \pi(Y) = \pi(y)$ exceeds $1/k$.

Let X be a random subset of cardinality $k-2$ of $[n]$, and define, for each $x \in X$, a set Y_x as follows:

$$Y_x = \left\{ y \in [n] - X : \ \mathrm{Prob}_D(\min \pi(X \cup \{y\}) = \pi(x)) > \frac{1}{k} \right\}.$$

By linearity of expectation, the expected value of $\sum_{x\in X} |Y_x|$ is at least $\frac{(k-2)(n-k+2)}{k-1}$, and thus there exists an X for which the size of $\sum_{x\in X} |Y_x|$ is at least this fraction. Fix such a set X, suppose $\mathcal{F} = \{\pi_1, \ldots, \pi_d\}$ and define, for each $x \in X$, a $|Y_x|$ by d matrix $U^{(x)} = (u^{(x)}_{y,j})$ where $y \in Y_x$ and $1 \leqslant j \leqslant d$ as follows:

$$u^{(x)}_{y,j} = \begin{cases} \sqrt{\mathrm{Prob}_D(\pi_j)} & \text{if } \min(\pi_j(X \cup y)) = \pi_j(x), \\ 0 & \text{otherwise.} \end{cases} \tag{7.2}$$

As in the proof of Theorem 7.2, define $V^{(x)} = (v^{(x)}_{y,y'}) = U^{(x)}(U^{(x)})^T$ and observe that $v^{(x)}_{y,y}$ is precisely the probability that x is the minimum element of $X \cup \{y\}$ (according to the distribution D on $\mathcal{F}$), whereas for $y \neq y'$, $v^{(x)}_{y,y'}$ is the probability that x is the minimum element of $X \cup \{y, y'\}$ according to the same distribution. By the definition of Y_x, and the assumption on $\mathcal{F}$ and D, each diagonal entry of $V^{(x)}$ is strictly greater than $1/k$, whereas each other entry is exactly $1/k$. Therefore, the rank of $V^{(x)}$ is at least $|Y_x| - 1$ (as subtracting $1/k$ from each of its entries creates

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

a matrix of full rank). In addition, note that the definition of the matrices $U^{(x)}$ implies that for any distinct $x, x' \in X$, $U^{(x)}(U^{(x')})^T = 0$.

Put $p = \sum_{x \in X} |Y_x|$ and let U be the p by d matrix obtained by putting all matrices $U^{(x)}$, $(x \in X)$ together, one on top of the other. Then $V = UU^T$ is a block-diagonal matrix whose blocks are the matrices $V^{(x)}$, implying that its rank is the sum of ranks of the matrices $V^{(x)}$.

Since the rank of each $V^{(x)}$ is at least $|Y_x| - 1$, and

$$\sum_{x \in X} |Y_x| \geqslant \frac{(k-2)(n-k+2)}{k-1},$$

it follows that

$$|\mathcal{F}| = d \geqslant \mathrm{rank}(V) \geqslant \frac{(k-2)(n-k+2)}{k-1} - (k-2) = \frac{(k-2)(n-2k+3)}{k-1},$$

completing the proof. □

Remark. Note that k is a trivial lower bound for the size of any exactly k min-wise independent family of permutations. Indeed, fix an arbitrary set X of k elements, and observe that each $x \in X$ has to appear first among the elements of X in at least one of the permutations. Therefore, by the last theorem, $\Omega(n)$ is a lower bound for the size of any exacly k min-wise independent family of permutations of $[n]$, for all $n \geqslant k \geqslant 3$. (For $k = 2$ and any n the two permutations $1, 2, \ldots, n$ and $n, n-1, \ldots, 1$ suffice, of course.)

8. Concluding remarks

The proof of Theorems 1.1 and 2.1 can be easily modified to supply a more general result, as follows.

Theorem 8.1. *Let $B = (b_{i,j})$ be an n by n real, symmetric matrix of rank d, and let $P(z)$ be an arbitrary polynomial of degree k. Then the following inequality holds:*

$$\binom{d+k}{k} \geqslant \frac{[\sum_{i=1}^{n} P(b_{i,i})]^2}{\sum_{i,j=1}^{n} P^2(b_{i,j})}.$$

Indeed, this follows by noticing that the proof of Lemma 2.2 implies the known fact that the rank of any real, symmetric matrix is at least the ratio between the square of its trace, and the trace of its square, and by applying this fact, together with the assertion of Lemma 2.3, to the matrix $P(b_{i,j})$. As mentioned in Section 2, here too the symmetry assumption is not very crucial, as any matrix can be made symmetric by averaging it with its transpose, a process that does not change the rank by more than a factor of 2, maintains the trace, and does not increase the trace of the square.

The main open problem concerning the assertion of Theorems 1.1 and 2.1 is whether it is possible to remove the $\log(1/\epsilon)$-term in their statement when n is sufficiently large as a function of ϵ. If possible, this would be tight up to a constant factor, as shown by many of the applications described throughout the paper, where the gap between the upper and lower bounds is $\Theta(\log(1/\epsilon))$. Note that when $\epsilon = \frac{1}{\sqrt{n}}$, the $\log(1/\epsilon)$-term cannot be omitted.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

In most of the proofs throughout the paper, and in particular, in the proof of Theorem 2.1, we made no attempt to optimize the absolute constants involved. In some cases these constants may be of interest, and it is thus worthwhile to note that the estimates can be improved by replacing the polynomial $P(z) = z^k$ used in the proof of Theorem 2.1 by an appropriate Chebyshev polynomial. Indeed, the proof suggests that the best choice of a polynomial P of degree k for which we consider the matrix $P(b_{i,j})$ is the polynomial of degree P for which the maximum value of $|P(z)|$ over $z \in [-\epsilon, \epsilon]$ is minimum, among all polynomials P satisfying $P(1) = 1$. It is known (see [18]) that the optimal polynomial P for this problem can be obtained as follows.

The Chebyshev polynomials of the first kind, $T_k(z)$, can be defined by $T_0(Z) = 1$, $T_1(z) = z$ and $T_{k+1}(z) = 2zT_k(z) - T_{k-1}(z)$ for all $k \geqslant 1$. Equivalently, $T_k(z) = \cosh(k \cosh^{-1}(z))$, where $\cosh(z) = \frac{e^z+e^{-z}}{2}$. It is known that if $[a, b]$ is a real interval where $b > a > 0$, then among all polynomials t of degree k that satisfy $t(0) = 1$, the one for which the maximum of the absolute value in $[a, b]$ is minimal, is the polynomial

$$t_k(z) = \frac{T_k(\frac{a+b-2z}{b-a})}{T_k(\frac{a+b}{b-a})}.$$

For this polynomial,

$$\max_{z\in[a,b]} |t_k(z)| = t_k(a) = \frac{1}{T_k(\frac{a+b}{b-a})}.$$

It follows that, for our purpose, the best polynomial of degree k is obtained by taking $a = 1 - \epsilon$, $b = 1 + \epsilon$ and $P_k(z) = t_k(1 - z)$ for t_k as above. Therefore, $P_k(1) = 1$, and the maximum value of $|P_k(z)|$ in $[-\epsilon, \epsilon]$ is $T_k(1/\epsilon)^{-1}$. Since $\cosh^{-1}(z) = \ln(z + \sqrt{z^2 - 1})$ and $T_k(z) = \cosh(k \cosh^{-1}(z))$, it is not difficult to check that, for small ϵ, $T_k(1/\epsilon)^{-1}$ is roughly $\frac{\epsilon^k}{2^{k-1}}$ for all $k \geqslant 1$. It follows that by using this polynomial instead of the polynomial z^k in the proof of Theorem 2.1, if ϵ is small and k is large, one can roughly replace ϵ by $\epsilon/2$ in the conclusion of the theorem, improving its estimate by roughly a factor of 4. This does not shed any light on the problem of deciding whether or not the $\log(1/\epsilon)$-term in the statement of the theorem can be removed for sufficiently large n.

Acknowledgement

This work was partly performed during a visit to the IHES in Bures sur Yvette, France. I thank my hosts in Bures for their hospitality, and I also thank an anonymous referee for helpful comments.

References

[1] Alon, N. (2003) Problems and results in extremal combinatorics I. *Discrete Math.* **273** 31–53.
[2] Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R. and Xie, N. (2007) Testing k-wise and almost k-wise independence. In *Proc. 39th ACM Symposium on Theory of Computing*, pp. 496–505.
[3] Alon, N., Goldreich, O., Hastad, J. and Peralta, R. (1990) Simple constructions of almost k-wise independent random variables. In *Proc. 31st IEEE Symposium on Foundations of Computer Science*, pp. 544–553. Also *Random Struct. Alg.* **3** (1992) 289–304.
[4] Alon, N., Kohayakawa, Y., Mauduit, C., Moreira, C. G. and Rödl, V. (2006) Measures of pseudorandomness for finite sequences: Minimal values. *Combin. Probab. Comput.* **15** 1–29.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press

[5] Alon, N., Itoh, T. and Nagatani, T. (2007) On (ε,k)-min-wise independent permutations. *Random Struct. Alg.* **31** 384–389.
[6] Broder, A., Charikar, M., Frieze, A. and Mitzenmacher, M. (2000) Min-wise independent permutations. *J. Comput. System Sci.* **60** 630–659. A preliminary version appeared in *Proc. 30th Annual ACM Symposium on Theory of Computing* (1998), pp. 327–336.
[7] Cassaigne, J., Mauduit, C. and Sárközy, A. (2002) On finite pseudorandom binary sequences VII: The measures of pseudorandomness. *Acta Arith.* **103** 97–118.
[8] Codenotti, B., Pudlák, P. and Resta, G. (2000) Some structural properties of low-rank matrices related to computational complexity. *Theoret. Comput. Sci.* **235** 89–107.
[9] Hall, M. (1986) *Combinatorial Theory*, 2nd edn, Wiley.
[10] Itoh, T., Takei, Y. and Tarui, J. (2000) On permutations with limited independence. In *Proc. 11th Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 137–146.
[11] Johnson, W. B. and Lindenstrauss, J. (1984) Extensions of Lipschitz mappings into a Hilbert space. In Vol. 26 of *Contemporary Mathematics*, AMS, Providence, RI, pp. 189–206.
[12] Jukna, S. (2001) *Extremal Combinatorics*, Springer, Berlin.
[13] MacWilliams, F. J. and Sloane, N. J. A. (1977) *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam.
[14] Mauduit, C. and Sárközy, A. (1997) On finite pseudorandom binary sequences I: Measure of pseudorandomness, the Legendre symbol. *Acta Arith.* **82** 365–377.
[15] Matoušek, J. (2002) *Lectures on Discrete Geometry*, Springer.
[16] Matoušek, J. and Stojaković, M. (2003) On restricted min-wise independence of permutations. *Random Struct. Alg.* **23** 397–408.
[17] Norin, S. (2001) A polynomial lower bound for the size of any k-min-wise independent set of permutation. *Zapiski Nauchnyh Seminarov* (POMI) **277** 104–116 (in Russian). Available at `http://www.pdmi.ras.ru/znsl/`
[18] Rivin, T. J. (1990) *The Chebyshev Polynomials*, Wiley, New York.

https://doi.org/10.1017/S0963548307008917 Published online by Cambridge University Press