Sampling from Large Matrices: An Approach through Geometric Functional Analysis

MARK RUDELSON

University of Missouri, Columbia, Missouri

AND

ROMAN VERSHYNIN

University of California, Davis, California

Abstract. We study random submatrices of a large matrix A. We show how to approximately compute A from its random submatrix of the smallest possible size $O(r \log r)$ with a small error in the spectral norm, where $r = \|A\|_F^2/\|A\|_2^2$ is the numerical rank of A. The numerical rank is always bounded by, and is a stable relaxation of, the rank of A. This yields an asymptotically optimal guarantee in an algorithm for computing *low-rank approximations* of A. We also prove asymptotically optimal estimates on the *spectral norm* and the *cut-norm* of random submatrices of A. The result for the cut-norm yields a slight improvement on the best-known sample complexity for an approximation algorithm for MAX-2CSP problems. We use methods of Probability in Banach spaces, in particular the law of large numbers for operator-valued random variables.

Categories and Subject Descriptors: F.2.1 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems; G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Randomized algorithms, Monte-Carlo methods, massive data sets, singular-value decompositions, low-rank approximations

The work of M. Rudelson was partially supported by NSF grant DMS-0245380.

R. Vershynin is an Alfred P. Sloan Research Fellow. His work was also partially supported by NSF grant DMS-0401032.

This project started when the authors participated in the PIMS Thematic Programme on Asymptotic Geometric Analysis at the University of British Columbia in Summer 2002. M. Rudelson was a PIMS postdoctoral fellow at that time. We are grateful to PIMS for its hospitality.

The final part of this research was done when M. Rudelson visited University of California, Davis.

Authors' addresses: M. Rudelson, Department of Mathematics, University of Missouri, Mathematical Sciences Building, Columbia, MO 65211, e-mail: rudelson@math.missouri.edu; R. Vershynin, University of California—Davis, 2218 Math Science Building, Davis, CA 95616, e-mail: vershynin@math.ucdavis.edu.

ACM Reference Format:

Rudelson, M. and Vershynin, R. 2007. Sampling from large matrices: an approach through geometric functional analysis. J. ACM 54, 4, Article 21 (July 2007), 19 pages. DOI = 10.1145/1255443.1255449
http://doi.acm.org/10.1145/1255443.1255449

1. *Introduction*

This article studies random submatrices of a large matrix A. The study of random submatrices spans several decades and is related to diverse areas of mathematics and computer science. Two main reasons for the interest in random submatrices are:

(1) one can learn properties of A from the properties of its random submatrices;

(2) properties of A may improve by passing to its random submatrices.

We address both aspects of random submatrices in this article. We show how to approximate A by its random submatrix in the spectral norm, and we compute the asymptotics of the spectral and the cut norms of random submatrices. This yields improvements upon known algorithms for computing low rank approximations, Singular Value Decompositions, and approximations to MAX-2CSP problems.

1.1. THE SPECTRAL NORM: LOW RANK APPROXIMATIONS AND SVD. Can one approximate A by only knowing a random submatrix of A of a fixed size? If so, what is the sample complexity, the minimal size of a submatrix which yields a good approximation with a small error in some natural norm, and with high probability?

This problem belongs to a class of problems the Statistical Learning Theory is concerned with. These problems inevitably bear the assumption that the the object to be learned belongs to a relatively small "target" class. To be able to learn A from a matrix of small size thus of small rank, we have to assume that A itself has *small rank*–or can be approximated by an (unknown) matrix of a small rank. We thus strive to find a low rank approximation to a matrix A, whenever such an approximation exists, from only knowing a small random submatrix of A.

Solving this problem is essential for development of fast Monte-Carlo algorithms for computations on large matrices. An extremely large matrix A – say, of the order of $10^5 \times 10^5$ – is impossible to upload into the Random Access Memory (RAM) of a computer; it is instead stored in an external memory. On the other hand, sampling a small submatrix of A, storing it in RAM and computing its small rank approximation is feasible.

The crucial assumption that A is essentially a low rank matrix holds in many applications. For example, this is a model hypothesis in the Latent Semantic Indexing (see Jerry and Linoff [1997], Papadimitriou et al. [1998], Berry et al. [1995, 1999], Deerwester et al. [1990], and Azar et al. [2001]). There A is the "document-term matrix", which is formed of the frequencies of occurrence of various terms in the documents of a large collection. The hypothesis that the documents are related to a small number of (unknown) topics translates into the assumption that A can be approximated by an (unknown) low rank matrix. Finding such an approximation would determine the "best" topics the collection is really about. Other examples where this problem arises include clustering of graphs [Drineas et al. 2004], DNA microarray data, facial recognition, web search (see Drineas et al. [2006a]), lossy data compression and cryptography (see Berry et al. [1999]).

The best fixed rank approximation to A is obviously given by the partial sums of the Singular Value Decomposition (SVD)

$$A = \sum_j \sigma_j(A)\, u_j \otimes v_j,$$

where $\sigma_j(A)$ is the nonincreasing and nonnegative sequence of the singular values of A, and u_j and v_j are left and right singular vectors of A, respectively. The best rank k approximation to A in both the spectral and Frobenius norms is thus AP_k, where P_k is the orthogonal projection onto the top k left singular vectors of A. In particular, for the spectral norm, we have

$$\min_{B:\, \mathrm{rank} B \le k} \|A - B\|_2 = \|A - AP_k\|_2 = \sigma_{k+1}(A). \tag{1}$$

However, computing P_k, which gives the first elements of the SVD of a $m \times n$ matrix A, is often impossible in practice because (1) it would take many passes through A, which is prohibitively slow for a matrix stored in an external memory; (2) this would take superlinear time in $m + n$. Instead, it was proposed in Frieze et al. [2004], Drineas and Kannan [2003], and Drineas et al. [2006a, 2006b] to use the Monte-Carlo methodology: namely, approximate the kth partial sum of the SVD of A by the kth partial sum of the SVD of a random submatrix of A. In this article, we show that this can be done:

(1) with almost linear sample complexity $O(r \log r)$, that is, by sampling only $O(r \log r)$ random rows of A, if A is approximable by a rank r matrix;

(2) in one pass through A if the matrix is stored row-by-row, and in two passes if its entries are stored in arbitrary order;

(3) using RAM space and time $O(n + m)$ (and polynomial in r and k).

THEOREM 1.1. *Let A be an $m \times n$ matrix with numerical rank $r = \|A\|_F^2 \,/\, \|A\|_2^2$. Let $\varepsilon, \delta \in (0, 1)$, and let $d \le m$ be an integer such that*

$$d \ge C\Big(\frac{r}{\varepsilon^4 \delta}\Big) \log\Big(\frac{r}{\varepsilon^4 \delta}\Big). \tag{2}$$

Consider a $d \times n$ matrix $\tilde{A}$, which consists of d normalized rows of A picked independently with replacement, with probabilities proportional to the squares of their Euclidean lengths. Then with probability at least $1 - 2\exp(-c/\delta)$ the following holds. For a positive integer k, let P_k be the orthogonal projection onto the top k left singular vectors of $\tilde{A}$. Then

$$\|A - AP_k\|_2 \le \sigma_{k+1}(A) + \varepsilon\, \|A\|_2\,. \tag{3}$$

Here and in the sequel, $C, c, C_1, \ldots$ denote positive absolute constants.

Comparing (3) with the best approximation (1) given by the SVD, we see an additional error $\varepsilon\, \|A\|_2$ which can be made small by increasing the size d of the sample.

Remark 1.2 (*Optimality*). The almost linear sample complexity $d = O(r \log r)$ achieved in Theorem 1.1 is optimal, see Proposition 3.9 below. The best known previous result, due to Drineas, Kannan and Mahoney, had with the quadratic sample complexity $d = O(r^2)$ [Drineas et al. 2006a, Theorem 4] see also Drineas et al.

[2006b, Theorem 3]. The approximation scheme in Theorem 1.1 was developed in Frieze et al. [2004], Drineas and Kannan [2003], and Drineas et al. [2006a, 2006b].

Remark 1.3 (*Numerical Rank*). The numerical rank $r = r(A) = \|A\|_F^2 \,/\, \|A\|_2^2$ in Theorem 1.1 is a relaxation of the exact notion of rank. Indeed, one always has $r(A) \leq \operatorname{rank}(A)$. But as opposed to the exact rank, the numerical rank is stable under small perturbations of the matrix A. In particular, the numerical rank of A tends to be low when A is close to a low rank matrix, or when A is sufficiently sparse. So results like Theorem 1.1, which depend on the numerical rather than exact rank, should be useful in many applications, such as the Principal Component Analysis.

Remark 1.4 (*Law of Large Numbers for Operator-Valued Random Variables*). The new feature in our proof of Theorem 1.1 is a use of the first author's argument about random vectors in the isotropic position [Rudelson 1999]. It yields a law of large numbers for operator-valued random variables. We apply it for independent copies of a rank one random operator, which is given by a random row of the matrix $A^T A$.

1.2. The Cut-Norm: Decay, MAX-CSP Problems.

Alon et al. [2002, 2003] reduced the problem of additive approximation of the MAX-CSP problems (which are NP-hard) to computing the cut-norm of random submatrices. The cut norm of an $n \times n$ matrix A is the maximum sum of the entries of its submatrix,

$$\|A\|_C = \max_{I,J} \Big| \sum_{i \in I, j \in J} A_{ij} \Big|,$$

and it is equivalent to the $\ell_\infty \to \ell_1$ operator norm.

The problem is to understand how the cut norm of A decreases when we pass to its random submatrix. Let Q be a a random subset of $\{1, \dots, n\}$ of expected cardinality q. This means that each element of $\{1, \dots, n\}$ is included into Q independently with probability q/n. We form a $Q \times Q$ random submatrix $A|_{Q \times Q} = (A_{ij})_{i,j \in Q}$.

Intuitively, $A|_{Q\times Q}$ is $(q/n)^2$ times smaller than A if A is diagonal-free, but only (q/n) times smaller than A if A is a diagonal matrix. We prove a general estimate of the cut-norm of random submatrices, which combines both of these types of decay:

THEOREM 1.5. *Let A be an $n \times n$ matrix. Let Q be a random subset of $\{1, \dots, n\}$ of expected cardinality q. Then*

$$\mathbb{E}\|A|_{Q\times Q}\|_C \leq O\Big(\Big(\frac{q}{n}\Big)^2 \|A - D(A)\|_C + \Big(\frac{q}{n}\Big) \|D(A)\|_C + \Big(\frac{q}{n}\Big)^{3/2} (\|A\|_{\mathrm{Col}} + \|A^T\|_{\mathrm{Col}})\Big),$$

where $\|A\|_{\mathrm{Col}}$ is the sum of the Euclidean lengths of the columns of A, and $D(A)$ is the diagonal part of A.

Remark 1.6 (*Optimality*). The estimate in this theorem is optimal, see Section 4.2.

We now state a partial case of Theorem 1.5 in the form useful for MAX-CSP problems. Note that $\|A\|_{\mathrm{Col}} \leq \sqrt{n}\, \|A\|_F$. Then we have:

COROLLARY 1.7. *Under the hypotheses of Theorem* 1.5*, let* $q = \Omega(\varepsilon^{-2})$*. Assume that* $\|A\|_C = O(\varepsilon n^2)$*, and* $\|A\|_F = O(n)$*,* $\|A\|_\infty = O(\varepsilon^{-1})$*, where* $\|A\|_\infty$ *denotes the maximum of the absolute values of the entries of* A*. Then*

$$\mathbb{E}\|A|_{Q\times Q}\|_C = O(\varepsilon q^2).$$

In solving MAX-2-CSP problems, one approximates the edge-weight matrix W of the graph on n vertices by a cut approximation W', and checks that the the error matrix $A = W - W'$ satisfies the assumptions of Corollary 1.7, see Fernandez de la Vega [1996]; Alon et al. [2002, 2003]. Hence, the Corollary says that for a random induced graph on $q = \Omega(\varepsilon^{-2})$ vertices, the same cut-approximation (induced on the q vertices of the random subgraph) works. Namely, the error in cut-norm is at most εq^2.

A weaker form of Corollary 1.7 was proved by Alon et al. [2003, Theorem 6]. Their result has a bigger sample complexity $q = \Omega(\varepsilon^{-4}\log(1/\varepsilon))$ and an extra assumption $n = \exp(\Omega(\varepsilon^{-2}))$, but it works for multidimensional arrays rather than for matrices (= 2-dimensional arrays).

Using Corollary 1.7 instead of Alon et al. [2003, Theorem 6] slightly improves the best known sample complexity for the approximation algorithm for MAX-2-CSP problems due to Alon et al. [2003]. The solution to a MAX-2SCP problem on n variables can be approximated within the additive error εn^2 by the solution of the problem induced by a randomly chosen q variables. The best known sample complexity, due to Alon et al. [2003], is $q = O(\varepsilon^{-4}\log(1/\varepsilon))$. Using Corollary 1.7 in the argument of Alon et al. [2003, Theorem 6] improves the sample complexity to $q = O(\varepsilon^{-4})$.

Our proof of Theorem 1.5 uses the technique of probability in Banach spaces, and includes decoupling, symmetrization, and application of a version of Slepian's lemma for Bernoulli random variables due to Talagrand.

1.3. THE SPECTRAL NORM: DECAY. Perhaps the most important matrix norm is the spectral norm. Nevertheless, its decay under passing to submatrices has not been sufficiently understood.

Let A be an $n \times n$ matrix, and Q be a random subset of $\{1, \dots, n\}$ of expected cardinality q (as above). We consider a random row-submatrix $A|_Q = (A_{ij})_{i\in Q, j\le n}$, which consists of the rows of A in Q.

When one orthogonally projects a vector $x \in \mathbb{R}^n$ onto $\mathbb{R}^Q$, its Euclidean length reduces in average by the factor of $\sqrt{\frac{q}{n}}$. So, one should expect a similar type of decay for the spectral norm—something like $\mathbb{E}\|A|_Q\|_2 \le \sqrt{\frac{q}{n}}\|A\|_2$.

However, similarly to the previous section, the diagonal matrices exhibit a different type of decay. For example, there is no decay at all for the identity matrix. One can check that the correct order of decay for diagonal matrices is

$$\|A\|_{(k)} = \text{the average of } k \text{ biggest Euclidean lengths of the columns of } A,$$

where $k = n/q$. General matrices again combine both types of decay:

THEOREM 1.8. *Let* A *be an* $n\times n$ *matrix. Let* Q *be a random subset of* $\{1, \dots, n\}$ *of expected cardinality* q*. Then*

$$\mathbb{E}\|A|_Q\|_2 \le O\left(\sqrt{\frac{q}{n}}\|A\|_2 + \sqrt{\log q}\,\|A\|_{(n/q)}\right).$$

Remark 1.9 (*Optimality*). The estimate in this theorem is optimal. The example considered in the proof of Proposition 3.9 below shows that the coefficient $\sqrt{\log q}$ is necessary.

Generalizing an earlier result of Lunin [1975], Kashin and Tzafriri (unpublished notes) (see Vershynin [2001]) essentially proved the *existence* of a subset Q of cardinality q and such that

$$\|A|_Q\|_2 \le O\left(\sqrt{\frac{q}{n}}\|A\|_2 + \frac{\|A\|_F}{\sqrt{n}}\right).$$

Note that $\frac{\|A\|_F}{\sqrt{n}} = \big(\frac{1}{n}\sum_{i=1}^n |A_i|^2\big)^{1/2}$ is the average of the lengths of *all* columns of A. As the example of diagonal operators shows, for random subsets Q this term has to be replaced by the average of the few biggest columns. Talagrand [1995] proved deep results on the more general operator norms $\ell_2 \to X$, where X is a 2-smooth Banach space. However, the decay on $\frac{q}{n}$ in his results is logarithmic rather than polynomial.

1.4. Stable Matrix Results. Many problems on random submatrices, of both theoretical and practical importance, have functional-analytic rather than linear-algebraic nature. These problems, like those this article considers, are about estimating operator norms. We thus see a matrix A as a linear operator A between finite dimensional normed spaces—say, between ℓ_2^n and ℓ_2^n for the spectral norm, and between ℓ_∞^n and ℓ_1^n for the cut norm.

From this perspective, the dimension n of the ambient normed space should play a minor role, while the real control of the picture should be held by (hopefully few) quantities tied to the operator rather than the space. As a trivial example, if A is not of full rank then the dimension n is useless compared to the rank of A. Further, we are looking for stable results, those not ruined by small perturbations of the linear operators. This is a natural demand in applications, and this differs our analytic perspective from the linear algebraic one. It would thus be natural to look for *stable* quantities tied to linear operators, which govern the picture. For example, operator norms are stable quantities, while the rank is not.

This article advances such approach to matrices. The low rank approximations in Theorem 1.1 are only controlled by the numerical rank $r(A) = \|A\|_F^2/\|A\|_2^2$ of the matrix, which is a stable relaxation of the rank. The norms of random matrices in Theorems 1.5 and 1.8 are essentially controlled by the norms of the original matrix (and naturally by the sampling factor, the ratio of the size of the submatrix to the size of the original matrix). The dimension n of the matrix does not play a separate role in these results (although the matrix norms may grow with the dimension).

2. *Notation*

For $p \le \infty$, the finite dimensional ℓ_p spaces are denoted by ℓ_p^n. Thus, ℓ_p^n is the Banach space $(\mathbb{R}^n, \|\cdot\|_p)$, where $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \le \infty$, and $\|x\|_\infty = \max_i |x_i|$. The closed unit ball of ℓ_p is denoted by $B_p^n := \{x \mid \|x\|_p \le 1\}$.

The canonical basis of $\mathbb{R}^n$ is denoted by $(e_1, \dots, e_n)$. Let $x, y \in \mathbb{R}^n$. The canonical inner product is denoted by $\langle x, y\rangle := x^T y$. The tensor product is defined as $x \otimes y := y x^T$; thus $(x \otimes y)z = \langle x, z\rangle\, y$ for all $z \in \mathbb{R}^n$.

Let $A = (A_{ij})_{ij}$ be an $m \times n$ real matrix. The spectral norm of A is the operator norm $\ell_2 \to \ell_2$, defined as

$$\|A\|_2 := \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1(A),$$

where $\sigma_1(A)$ is the largest singular value of A. The Frobenius norm $\|A\|_F$ of A is defined as

$$\|A\|_F^2 := \sum_{i,j} A_{ij}^2 = \sum_j \sigma_j(A)^2,$$

where $\sigma_j(A)$ are the singular values of A.

Finally, $C, C_1, c, c_1, \dots$ denote positive absolute constants. The $a = O(b)$ notation means that $a \le Cb$ for some absolute constant C.

3. *Low-Rank Approximations*

In this section, we prove Theorem 1.1, discuss the algorithm for finding low rank approximations, and show that the sample complexity in Theorem 1.1 is optimal. Our argument will be based on the law of large numbers for operator-valued random variables.

3.1. Law of Large Numbers for Operator-Valued Random Variables.

Theorem 1.1 is about random independent sampling the rows of the matrix A. Such sampling can be viewed as an empirical process taking values in the set of rows. If we sample enough rows, then the matrix constructed from them would nicely approximate the original matrix A in the spectral norm. For the scalar random variables, this effect is the classical Law of Large Numbers. For example, let X be a bounded random variable and let $X_1 \cdots X_d$ be independent copies of X. Then

$$\mathbb{E}\left|\frac{1}{d}\sum_{j=1}^{d} X_j - \mathbb{E}X\right| = O\left(\frac{1}{\sqrt{d}}\right). \tag{4}$$

Furthermore, the large deviation theory allows one to estimate the probability that the empirical mean $\frac{1}{d}\sum_{j=1}^{d} X_j$ stays close to the true mean $\mathbb{E}X$.

Operator-valued versions of this inequality are harder to prove. The absolute value must be replaced by the operator norm. So, instead of proving a large deviation estimate for a single random variable, we have to estimate the supremum of a random process. This requires deeper probabilistic techniques. The following Theorem generalizes the main result of Rudelson [1999].

Theorem 3.1. *Let y be a random vector in $\mathbb{R}^n$, which is uniformly bounded almost everywhere: $\|y\|_2 \le M$. Assume for normalization that $\|\mathbb{E}y \otimes y\|_2 \le 1$. Let $y_1 \cdots y_d$ be independent copies of y. Let*

$$a := C\sqrt{\frac{\log d}{d}} \cdot M.$$

Then

(i) *If* $a < 1$, *then*

$$\mathbb{E}\left\|\frac{1}{d}\sum_{i=1}^{d} y_i \otimes y_i - \mathbb{E}\, y \otimes y\right\|_2 \le a.$$

(ii) *For every* $t \in (0, 1)$,

$$\mathbb{P}\left\{\left\|\frac{1}{d}\sum_{i=1}^{d} y_i \otimes y_i - \mathbb{E} y \otimes y\right\|_2 > t\right\} \le 2\exp(-ct^2/a^2).$$

Remark 3.2. Part (i) is a law of large numbers, and part (ii) is a large deviation estimate for operator-valued random variables. Comparing this result with its scalar prototype (4), we see an additional logarithmic factor. This factor is essential, as we show in Remark 3.4 below.

Remark 3.3. The boundedness assumption $\|y\|_2 \le M$ can be too strong for some applications. The proof of Theorem 3.1 shows that, in part (i), the boundedness almost everywhere can be relaxed to the moment assumption $\mathbb{E}\|y\|_2^q \le M^q$, where $q = \log d$. Part (ii) also holds under an assumption that the moments of $\|y\|_2$ have a nice decay. However, we do not need these improvements here.

Remark 3.4. The estimate in Theorem 3.1 is in general optimal. Indeed, consider the random vector y taking values $\sqrt{n}e_1, \ldots, \sqrt{n}e_n$ each with probability $1/n$, where (e_i) is the canonical basis of $\mathbb{R}^n$. Then $\mathbb{E}y \otimes y = I$. Then

$$\mathbb{E}\left\|\frac{1}{d}\sum_{j=1}^{d} y_j \otimes y_j - I\right\|_2 = \mathbb{E}\max_{i=1\ldots n}\left|\frac{n}{d}\,|\{j \mid y_j = \sqrt{n}e_i\}| - 1\right|.$$

If we want this quantity to be $O(1)$, then it is not hard to check that d should be of order at least $n\log n$. Therefore, the coefficient $\sqrt{\log(d)/d}$ in Theorem 3.1 is optimal.

3.2. PROOF OF THEOREM 3.1. The proof consists of two steps. First, we use the standard symmetrization technique for random variables in Banach spaces (see for example, Ledoux and Talagrand [1991, Section 6]). Then we adapt the technique of Rudelson [1999] to obtain a bound on a symmetric random process. To obtain the probability estimate in part (ii), we shall estimate the high moments rather than the first moment in part (i).

Let $\varepsilon_1 \cdots \varepsilon_d$ denote independent Bernoulli variables taking values $1, -1$ with probability $1/2$. Let $y_1 \cdots y_d$, $\bar{y}_1 \cdots \bar{y}_d$ be independent copies of y. We shall denote by $\mathbb{E}_y$, $\mathbb{E}_{\bar{y}}$ and $\mathbb{E}_\varepsilon$ the expectations according to (y_i), $(\bar{y}_i)$ and (ε_i) respectively.

Let $p \ge 1$. We shall estimate

$$E_p := \left(\mathbb{E}\left\|\frac{1}{d}\sum_{i=1}^{d} y_i \otimes y_i - \mathbb{E}y \otimes y\right\|_2^p\right)^{1/p}. \tag{5}$$

Note that $\mathbb{E}_y\, y \otimes y = \mathbb{E}_{\bar{y}}\, \bar{y} \otimes \bar{y} = \mathbb{E}_{\bar{y}}\left(\frac{1}{d}\sum_{i=1}^{d} \bar{y}_i \otimes \bar{y}_i\right)$. We put this into (5). Since

$x \mapsto \|x\|_2^p$ is a convex function on $\mathbb{R}^n$, Jensen's inequality implies that

$$E_p \le \left(\mathbb{E}_y \mathbb{E}_{\bar{y}} \left\| \frac{1}{d} \sum_{i=1}^{d} y_i \otimes y_i - \frac{1}{d} \sum_{i=1}^{d} \bar{y}_i \otimes \bar{y}_i \right\|_2^p \right)^{1/p}.$$

Since $y_i \otimes y_i - \bar{y}_i \otimes \bar{y}_i$ is a symmetric random variable, it is distributed identically with $\varepsilon_i (y_i \otimes y_i - \bar{y}_i \otimes \bar{y}_i)$. Thus

$$E_p \le \left(\mathbb{E}_y \mathbb{E}_{\bar{y}} \mathbb{E}_\varepsilon \left\| \frac{1}{d} \sum_{i=1}^{d} \varepsilon_i (y_i \otimes y_i - \bar{y}_i \otimes \bar{y}_i) \right\|_2^p \right)^{1/p}.$$

Denote $Y = \frac{1}{d} \sum_{i=1}^{d} \varepsilon_i y_i \otimes y_i$ and $\bar{Y} = \frac{1}{d} \sum_{i=1}^{d} \varepsilon_i \bar{y}_i \otimes \bar{y}_i$. Then $\|Y - \bar{Y}\|_2^p \le (\|Y\|_2 + \|\bar{Y}\|_2)^p \le 2^p (\|Y\|_2^p + \|\bar{Y}\|_2^p)$, and $\mathbb{E}\|Y\|_2^p = \mathbb{E}\|\bar{Y}\|_2^p$. Thus, we obtain

$$E_p \le 2 \left(\mathbb{E}_y \mathbb{E}_\varepsilon \left\| \frac{1}{d} \sum_{i=1}^{d} \varepsilon_i y_i \otimes y_i \right\|_2^p \right)^{1/p}.$$

We shall estimate the last expectation using a lemma from Rudelson [1999].

LEMMA 3.5. *Let $y_1 \cdots y_d$ be vectors in R^k and let $\varepsilon_1 \cdots \varepsilon_d$ be independent Bernoulli variables taking values* $1, -1$ *with probability* $1/2$. *Then*

$$\left(\mathbb{E} \left\| \sum_{i=1}^{d} \varepsilon_i y_i \otimes y_i \right\|_2^p \right)^{1/p} \le C_0 (p + \log k)^{1/2} \cdot \max_{i=1\cdots d} \|y_i\|_2 \cdot \left\| \sum_{i=1}^{d} y_i \otimes y_i \right\|_2^{1/2}.$$

Remark 3.6. We can consider the vectors $y_1 \cdots y_d$ as vectors in their linear span, so we can always choose the dimension k of the ambient space at most d.

Combining Lemma 3.5 with Remark 3.6 and using Hölder's inequality, we obtain

$$E_p \le 2C_0 \frac{(p + \log d)^{1/2}}{d} \cdot M \cdot \left(\mathbb{E} \left\| \sum_{i=1}^{d} y_i \otimes y_i \right\|_2^p \right)^{1/2p}. \tag{6}$$

By Minkowski's inequality, we have

$$\left(\mathbb{E} \left\| \sum_{i=1}^{d} y_i \otimes y_i \right\|_2^p \right)^{1/p} \le d \left[\left(\mathbb{E} \left\| \frac{1}{d} \sum_{i=1}^{d} y_i \otimes y_i - \mathbb{E}\, y \otimes y \right\|_2^p \right)^{1/p} + \|\mathbb{E}\, y \otimes y\|_2 \right]$$
$$\le d(E_p + 1).$$

So we obtain

$$E_p \le \frac{ap^{1/2}}{2} (E_p + 1), \qquad \text{where} \quad a = 4C_0 \left(\frac{\log d}{d} \right)^{1/2} M.$$

It follows that

$$\min(E_p, 1) \le ap^{1/2}. \tag{7}$$

To prove part (i) of the theorem, note that $a \le 1$ by the assumption. It thus follows that $E_1 \le a$. This proves part (i).

To prove part (ii), we can $E_p = (\mathbb{E}\, Z^p)^{1/p}$, where

$$Z = \Big\| \frac{1}{d} \sum_{i=1}^{d} y_i \otimes y_i - \mathbb{E} y \otimes y \Big\|_2.$$

So (7) implies that

$$\big(\mathbb{E} \min(Z, 1)^p\big)^{1/p} \le \min(E_p, 1) \le a p^{1/2}. \tag{8}$$

This moment bound can be expressed as a tail probability estimate using the following standard lemma (see, e.g., Ledoux and Talagrand [1991, Lemmas 3.7 and 4.10]).

LEMMA 3.6. *Let Z be a nonnegative random variable. Assume that there exists a constant $K > 0$ such that $(\mathbb{E}\, Z^p)^{1/p} \le K p^{1/2}$ for all $p \ge 1$. Then*

$$\mathbb{P}\{Z > t\} \le 2 \exp(-c_1 t^2 / K^2) \qquad \textit{for all } t > 0.$$

It thus follows this and from (8) that

$$\mathbb{P}\{\min(Z, 1) > t\} \le 2 \exp(-c_1 t^2 / a^2) \qquad \text{for all } t > 0.$$

This completes the proof of the theorem. □

3.3. PROOF OF THEOREM 1.1. By the homogeneity, we can assume $\|A\|_2 = 1$. The following lemma of Drineas and Kannan [2003] (see also [Drineas et al. 2006a]) reduces Theorem 1.1 to a comparison of A and a sample $\tilde{A}$ in the spectral norm.

LEMMA 3.8 (DRINEAS, KANNAN).

$$\|A - AP_k\|_2^2 \le \sigma_{k+1}(A)^2 + 2\|A^T A - \tilde{A}^T \tilde{A}\|_2.$$

PROOF. We have

$$\begin{aligned} \|A - AP_k\|_2^2 &= \sup_{x \in \ker P_k,\ \|x\|_2 = 1} \|Ax\|_2^2 = \sup_{x \in \ker P_k,\ \|x\|_2 = 1} \langle A^T A x, x \rangle \\ &\le \sup_{x \in \ker P_k,\ \|x\|_2 = 1} \langle (A^T A - \tilde{A}^T \tilde{A}) x, x \rangle + \sup_{x \in \ker P_k,\ \|x\|_2 = 1} \langle \tilde{A}^T \tilde{A} x, x \rangle \\ &= \|A^T A - \tilde{A}^T \tilde{A}\|_2 + \sigma_{k+1}(\tilde{A}^T \tilde{A}). \end{aligned}$$

By a result of perturbation theory, $|\sigma_{k+1}(A^T A) - \sigma_{k+1}(\tilde{A}^T \tilde{A})| \le \|A^T A - \tilde{A}^T \tilde{A}\|_2$. This proves Lemma 3.8. □

Let $x_1 \cdots x_m$ denote the rows of the matrix A. Then

$$A^T A = \sum_{j=1}^{m} x_j \otimes x_j.$$

We shall view the matrix $A^T A$ as the *true mean* of a bounded operator valued random variable, whereas $\tilde{A}^T \tilde{A}$ will be its *empirical mean*; then we shall apply the Law of Large Numbers for operator-valued random variables—Theorem 3.1. To this end, define a random vector $y \in \mathbb{R}^m$ as

$$\mathbb{P}\left(y = \frac{\|A\|_F}{\|x_j\|_2} x_j\right) = \frac{\|x_j\|_2^2}{\|A\|_F^2}.$$

Let $y_1 \cdots y_d$ be independent copies of y. Let the matrix $\tilde{A}$ consist of rows $\frac{1}{\sqrt{d}} y_1 \cdots \frac{1}{\sqrt{d}} y_d$. (The normalization of $\tilde{A}$ here is different than in the statement of Theorem 1.1: in the proof, it is convenient to multiply $\tilde{A}$ by the factor $\frac{1}{\sqrt{d}}\|A\|_F$. However, note that the singular vectors of $\tilde{A}$ and thus P_k do not change.) Then

$$A^T A = \mathbb{E} y \otimes y, \qquad \tilde{A}^T \tilde{A} = \frac{1}{d} \sum_{i=1}^{d} y_j \otimes y_j, \qquad M := \|y\|_2 = \|A\|_F = \sqrt{r}.$$

We can thus apply Theorem 3.1. Due to our assumption on d, we have

$$a := 4C_0 \Big(\frac{\log d}{d} \cdot r\Big)^{1/2} \le \frac{\varepsilon^2 \delta^{1/2}}{2} < 1.$$

Thus, Theorem 3.1 yields (with $t = \varepsilon^2/2$) that, with probability at least $1 - 2\exp(-c/\delta)$, we have

$$\|\tilde{A}^T \tilde{A} - A^T A\|_2 \le \frac{\varepsilon^2}{2}.$$

Whenever this event holds, we can conclude by Lemma 3.8 that

$$\|A - AP_k\|_2 \le \sigma_{k+1}(A) + \sqrt{2}\|A^T A - \tilde{A}^T \tilde{A}\|_2^{1/2} \le \sigma_{k+1}(A) + \varepsilon.$$

This proves Theorem 1.1.

3.4. ALGORITHMIC ASPECTS OF THEOREM 1.1. Finding a good low rank approximation to a matrix A amounts, due to Theorem 1.1, to sampling a random submatrix $\tilde{A}$ and computing its SVD (actually, only left singular vectors are needed). The algorithm works well if the numerical rank $r = r(A) = \|A\|_F^2/\|A\|_2^2$ of the matrix A is small. This is the case, in particular, when A is essentially a low-rank matrix, because $r(A) \le \operatorname{rank}(A)$.

First, the algorithm samples $d = O(r \log r)$ random rows of A. Namely, it takes d independent samples of the random vector y whose law is

$$\mathbb{P}\left(y = \frac{A_j}{\|A_j\|_2}\right) = \frac{\|A_j\|_2^2}{\|A\|_F^2}$$

where A_j is the jth row of A. This sampling can be done in *one pass* through A if the matrix is stored row-by-row, and in *two passes* if its entries are stored in arbitrary order [Drineas et al. 2004, Section 5.1].

Then, the algorithm computes the SVD of the $d \times n$ matrix $\tilde{A}$, which consists of the normalized sampled rows. This can be done in time $O(dn)+$ the time needed to compute the SVD of a $d \times d$ matrix. The latter can be done by one of the known methods. This takes significantly less time than computing SVD of the original $m \times n$ matrix A. In particular, the running time of this algorithm is linear in the dimensions of the matrix (and polynomial in d).

3.5. Optimality of the Sample Complexity. The sample complexity $d = O(r \log r)$ in Theorem 1.1 is best possible:

Proposition 3.9. *There exist matrices A with arbitrarily big numerical rank $r = \|A\|_F^2/\|A\|_2^2$ and such that whenever*

$$d < \frac{1}{10} r \log r,$$

the conclusion (3) *of Theorem* 1.1 *fails for $k = n$ and for all $\varepsilon \in (0, 1)$.*

Proof. Let $n, m \in \mathbb{N}$ be arbitrary numbers such that $n < m$. We define the $m \times n$ matrix by its entries as follows:

$$A_{ij} = \sqrt{\frac{n}{m}}\, \delta_{\lceil \frac{n}{m} i \rceil, j},$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

Then each row of A contains exactly one entry of value $\sqrt{\frac{n}{m}}$, and each row is repeated m/n times. The jth column of A contains exactly one block of values $\sqrt{\frac{n}{m}}$ in positions $i \in (\frac{m}{n}(j-1), \frac{m}{n} j] =: I_j$. In particular, the columns are orthonormal. Also, $\|A\|_2 = 1$, $\|A\|_F = \sqrt{n}$, thus $r = n$.

Now we form a submatrix $\tilde{A}$ as described in Theorem 1.1 – by picking d rows of A independently and with uniform distribution. If $d < \frac{1}{10} n \log n$, then with high probability there exists at least one block I_j from which no rows i are picked. Call this block I_{j_0}. It follows that j_0-th column of $\tilde{A}$ is zero. Consider the coordinate vector $e_{j_0} = (0, \dots, 0, 1, 0, \dots, 0)$ of n positions, with 1 at position j_0. Then, $e_{j_0} \in \ker \tilde{A} \subseteq \ker P_k \subseteq \ker(AP_k)$. Thus, $\|(A - AP_k)e_{j_0}\|_2 = \|Ae_{j_0}\|_2 = 1$. Hence,

$$\|A - AP_k\|_2 \ge 1, \quad \text{while} \quad \sigma_{n+1}(A) = 0, \ \ \|A\|_2 = 1.$$

Hence, (3) fails for $k = n$ and for all $\varepsilon \in (0, 1)$. □

4. *The Decay of the Cut Norm*

In this section, we prove Theorem 1.5 on the cut norm of random submatrices and show that it is optimal. Our argument will be based on the tools of probability in Banach spaces: decoupling, symmetrization, and Slepian's Lemma (more precisely, its version for the Rademacher random variables due to M.Talagrand).

4.1. Proof of Theorem 1.5. It is known and easy to check that

$$\frac{1}{4}\|A\|_{\infty\to 1} \le \|A\|_C \le \|A\|_{\infty\to 1},$$

where $\|A\|_{\infty\to 1}$ denotes the operator norm of A from ℓ_∞^n into ℓ_∞^n:

$$\|A\|_{\infty\to 1} := \sup_{x\in\mathbb{R}^n} \frac{\|Ax\|_1}{\|x\|_\infty} = \sup_{x\in B_\infty^n} \|Ax\|_1$$

(recall that B_∞^n denotes the unit ball of ℓ_∞^n). Note also that both these norms are self-dual:

$$\|A^T\|_C = \|A\|_C, \qquad \|A^T\|_{\infty\to 1} = \|A\|_{\infty\to 1}.$$

So we can prove Theorem 1.5 for the norm $\|\cdot\|_{\infty\to 1}$ instead of the cut norm.

We shall use the following decoupling lemma due to Bourgain and Tzafriri [1987].

LEMMA 4.1. *Let (ξ_i) be a finite sequence of bounded i.i.d. random variables, and (ξ_i') be its independent copy. Then for any sequence of vectors (x_{ij}) in a Banach space with $x_{ii} = 0$,*

$$\mathbb{E}\Big\| \sum_{i,j} \xi_i \xi_j x_{ij} \Big\| \le 20\mathbb{E}\Big\| \sum_{i,j} \xi_i \xi_j' x_{ij} \Big\|.$$

Let $\delta_1 \cdots \delta_n$ be independent Bernoulli random variables, which take value 1 with probability $\delta := q/n$. Let P_Δ denote the coordinate projection on the random set of coordinates $\{j \mid \delta_j = 1\}$.

Denote by $D(A)$ the diagonal part of A. Then

$$P_\Delta A P_\Delta = P_\Delta(A - D(A))P_\Delta + P_\Delta D(A) P_\Delta = \sum_{i \ne j} \delta_i \delta_j A_{ij} e_i \otimes e_j + \sum_{i=1}^n \delta_i A_{ii} e_i \otimes e_i.$$

We can use Lemma 4.1 to estimate the first summand, taking $x_{ij} = A_{ij} e_i \otimes e_j$ if $i \ne j$ and $x_{ij} = 0$ if $i = j$. To this end, let (δ_j') be an independent copy of (δ_j), and let $P_{\Delta'}$ denote the coordinate projection on the random set of coordinates $\{j \mid \delta_j' = 1\}$. Then, by Lemma 4.1, and by the triangle inequality, we obtain

$$\mathbb{E}\|P_\Delta A P_\Delta\|_{\infty\to 1} \le 20\mathbb{E}\|P_\Delta(A - D(A))P_{\Delta'}\|_{\infty\to 1} + \delta \sum_{i=1}^n |A_{ii}|.$$

Clearly, $\sum_{i=1}^n |A_{ii}| = \|D(A)\|_{\infty\to 1}$. Thus, to complete the proof, we can assume that the diagonal of A is zero, and prove the inequality as stated in the theorem for $\mathbb{E}\|P_\Delta A P_{\Delta'}\|_{\infty\to 1}$, that is,

$$\mathbb{E}\|P_\Delta A P_{\Delta'}\|_{\infty\to 1} \le C\delta^2 \|A\|_{\infty\to 1} + C\delta^{3/2}(\|A\|_{\mathrm{Col}} + \|A^T\|_{\mathrm{Col}}). \tag{9}$$

Note that

$$\mathbb{E}\|A P_{\Delta'}\|_{\infty\to 1} = \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n |\langle A P_{\Delta'} x, e_i\rangle|,$$

hence

$$\begin{aligned}\mathbb{E}\|P_\Delta A P_{\Delta'}\|_{\infty\to 1} &= \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \delta_i |\langle A P_{\Delta'} x, e_i\rangle| \\ &= \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n (\delta_i - \delta) |\langle A P_{\Delta'} x, e_i\rangle| + \delta \cdot \mathbb{E}\|A P_{\Delta'}\|_{\infty\to 1}. \end{aligned} \tag{10}$$

We proceed with a known symmetrization argument, which we used in the beginning of Section 3.2. Since $\delta_i - \delta$ are mean zero, we can replace δ by δ_i'', an independent copy of δ_i, which can only increase the quantity in (10). Then, the first term in (10) does not exceed

$$\mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n (\delta_i - \delta_i'') |\langle A P_{\Delta'} x, e_i\rangle|. \tag{11}$$

The random variable $\delta_i - \delta_i''$ is symmetric: hence, it is distributed identically with $\varepsilon_i(\delta_i - \delta_i'')$, where ε_i are -1, 1-valued symmetric random variables independent of all other random variables. Therefore, the expression in (11) bounded by

$$\mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \delta_i |\langle A P_{\Delta'} x, e_i \rangle| + \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \delta_i'' |\langle A P_{\Delta'} x, e_i \rangle| \leq 2\mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \delta_i |\langle A P_{\Delta'} x, e_i \rangle|. \quad (12)$$

To estimate this, we use Slepian's inequality for Rademacher random variables proved by Talagrand. This estimate allows us to remove the absolute values in (12). Precisely, a partial case of Slepian's inequality due to Talagrand (see Ledoux and Talagrand [1991, Eq. (4.20)]) states that, for arbitrary $y_1, \dots, y_n \in \mathbb{R}^n$, one has

$$\mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i |\langle x, y_i \rangle| \leq \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \langle x, y_i \rangle = \mathbb{E} \Big\| \sum_{i=1}^n \varepsilon_i y_i \Big\|_1.$$

Therefore

$$\begin{aligned}
\mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \delta_i |\langle A P_{\Delta'} x, e_i \rangle| &= \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \varepsilon_i \Big| \langle x, P_{\Delta'} A^T \delta_i e_i \rangle \Big| \\
&\leq \mathbb{E} \Big\| P_{\Delta'} A^T \Big(\sum_{i=1}^n \varepsilon_i \delta_i e_i \Big) \Big\|_1 \\
&= \mathbb{E} \sum_{j=1}^n \delta_j' \Big| \Big\langle A^T \Big(\sum_{i=1}^n \varepsilon_i \delta_i e_i \Big), e_j \Big\rangle \Big| \\
&= \delta \cdot \mathbb{E} \sum_{j=1}^n \Big| \sum_{i=1}^n \varepsilon_i \delta_i A_{ij} \Big| \\
&\leq \delta \cdot \sum_{j=1}^n \Big(\mathbb{E} \Big| \sum_{i=1}^n \varepsilon_i \delta_i A_{ij} \Big|^2 \Big)^{1/2} \\
&= \delta \cdot \sum_{j=1}^n \Big(\mathbb{E} \sum_{i=1}^n |\delta_i A_{ij}|^2 \Big)^{1/2} \qquad \text{(averaging over } (\varepsilon_i)) \\
&= \delta^{3/2} \cdot \sum_{j=1}^n \Big(\sum_{i=1}^n |A_{ij}|^2 \Big)^{1/2} = \delta^{3/2} \|A\|_{\mathrm{Col}}.
\end{aligned}$$

We have proved that the first term in (10) does not exceed $\delta^{3/2}\|A\|_{\mathrm{Col}}$. To estimate the second term, note that

$$\mathbb{E}\|A P_{\Delta'}\|_{\infty \to 1} = \mathbb{E}\|P_\Delta A^T\|_{\infty \to 1} = \mathbb{E} \sup_{x \in B_\infty^n} \sum_{i=1}^n \delta_i |\langle A^T x, e_i \rangle|.$$

So we can essentially repeat the argument above to bound this expression by

$$\leq \delta^{1/2}\|A^T\|_{\mathrm{Col}} + \delta\|A^T\|_{\infty \to 1} = \delta^{1/2}\|A^T\|_{\mathrm{Col}} + \delta\|A\|_{\infty \to 1}.$$

Putting this together, we can estimate (10) as

$$\begin{aligned}\mathbb{E}\|P_\Delta A P_{\Delta'}\|_{\infty\to 1} &\le \delta^{3/2}\|A\|_{\mathrm{Col}} + \delta(\delta^{1/2}\|A^T\|_{\mathrm{Col}} + \delta\|A\|_{\infty\to 1})\\ &\le \delta^{3/2}\|A\|_{\mathrm{Col}} + \delta^{3/2}\|A^T\|_{\mathrm{Col}} + \delta^2\|A\|_{\infty\to 1},\end{aligned}$$

as desired. This completes the proof of Theorem 1.5.

4.2. OPTIMALITY. All terms appearing in Theorem 1.8 are necessary. Their optimality can be witnessed on different types of matrices. To see that the first term is necessary, consider a matrix A, all whose entries are equal 1. For this matrix $\|A\|_C = n^2$, and for any $Q \subset \{1, \dots n\}$, $\|A_{Q\times Q}\|_C = |Q|^2$.

The optimality of the second term can be seen in the case when A is the identity matrix. In this case $\|A\|_C = n$, while $\|A_{Q\times Q}\| = |Q|$.

To prove that the third term is also necessary, assume that $A = (\varepsilon_{i,j})$ is a random ± 1 matrix. Then $\|D(A)\|_C = n$, and $\|A\|_{\mathrm{Col}} = \|A^T\|_{\mathrm{Col}} = n^{3/2}$. It is easy to show that $\mathbb{E}_\varepsilon \|A\|_C \le Cn^{3/2}$, so for $q < n$ the third term in Theorem 1.8 is dominant. Indeed, by Azuma's inequality, for any $x, y \in \{0, 1\}^n$

$$\mathbb{P}_\varepsilon\left(\left|\frac{1}{n}\sum_{i,j=1}^{n} \varepsilon_{ij} x_i y_j\right| > t\right) \le C\exp(-t^2/2).$$

Hence,

$$\mathbb{P}_\varepsilon\big(\|A\|_C > sn^{3/2}\big) \le 4^n \cdot C\exp(-s^2 n),$$

which implies the desired bound for the expectation.

Now fix a ± 1 matrix A such that $\|A\|_C \le Cn^{3/2}$. Let Q be any subset of $\{1, \dots, n\}$. Recall that the norms $\|A\|_C$ and $\|A\|_{\infty\to 1}$ are equivalent. We claim that

$$\|A|_{Q\times Q}\|_{\infty\to 1} \ge \frac{1}{\sqrt{2}}|Q|^{3/2}.$$

Indeed, let $\delta_i,\ i \in Q$ be independent ± 1 random variables. Then, by Khinchin's inequality

$$\sum_{j\in Q}\mathbb{E}_\delta\left|\sum_{i\in Q}\varepsilon_{ij}\delta_i\right| \ge \frac{1}{\sqrt{2}}|Q|^{3/2}.$$

Choose $x \in \{-1, 1\}^Q$ such that $\sum_{j\in Q}\left|\sum_{i\in Q}\varepsilon_{ij}x_i\right| \ge \frac{1}{\sqrt{2}}|Q|^{3/2}$. For $j \in Q$ set

$$y_j = \operatorname{sign}\left(\sum_{i\in Q}\varepsilon_{ij}x_i\right).$$

Then

$$\|A|_{Q\times Q}\|_{\infty\to 1} \ge \left|\sum_{i,j\in Q}\varepsilon_{ij}x_i y_j\right| \ge \frac{1}{\sqrt{2}}|Q|^{3/2}.$$

Therefore,

$$\mathbb{E}_Q\|A|_{Q\times Q}\|_C \ge \frac{1}{4\sqrt{2}}\left(\frac{q}{n}\right)^{3/2}\cdot\big(\|A\|_{\mathrm{Col}} + \|A^T\|_{\mathrm{Col}}\big).$$

5. *The Decay of the Spectral Norm*

In this section, we prove Theorem 1.8 on the spectral norm of random submatrices.

By homogeneity, we can assume that $\|A\|_2 = 1$. Let $\delta_1, \dots, \delta_n$ be $\{0, 1\}$-valued independent random variables with $\mathbb{E}\delta_j = \delta = \frac{q}{n}$. So our random set is $Q = \{j \mid \delta_j = 1\}$.

Let $x_1 \cdots x_n$ denote the columns of A. Then

$$A = \sum_{j=1}^{n} e_j \otimes x_j, \qquad A|_Q = \sum_{j=1}^{n} \delta_j e_j \otimes x_j.$$

The spectral norm can be computed as

$$\|A\|_2 = \|A^T A\|_2^{1/2} = \left\| \sum_{j=1}^{n} x_j \otimes x_j \right\|_2^{1/2},$$

and similarly

$$\|A|_Q\|_2 = \left\| \sum_{j=1}^{n} \delta_j x_j \otimes x_j \right\|_2^{1/2}.$$

To estimate the latter norm, we shall first apply the standard symmetrization argument (see Ledoux and Talagrand [1991, Lemma 6.3]), like we did in the beginning of Section 3.2 and in Section 4. Then, we will apply Lemma 3.5. Set

$$E = \mathbb{E}\|A|_Q\|_2.$$

The symmetrization argument yields

$$E \le \mathbb{E} \left\| \sum_{j=1}^{n} (\delta_j - \delta) x_j \otimes x_j \right\|_2^{1/2} + \sqrt{\delta}\, \|A\|_2^{1/2} \le 2\mathbb{E}_\delta \left(\mathbb{E}_\varepsilon \left\| \sum_{j=1}^{n} \varepsilon_j \delta_j x_j \otimes x_j \right\|_2 \right)^{1/2} + \sqrt{\delta}.$$

Now we apply Lemma 3.5 with $p = 1$ to bound $\mathbb{E}_\varepsilon \| \sum_{j=1}^{n} \varepsilon_j \delta_j x_j \otimes x_j \|_2$ for fixed (δ_j). By Remark 3.6, we can assume k in this Lemma equal

$$n(\delta) := e + \sum_{j \le n} \delta_j.$$

Then, using Cauchy–Schwartz inequality, we obtain

$$\begin{aligned} E &\le \mathbb{E}_\delta \left(C\sqrt{\log n(\delta)} \cdot \max_{j=1\cdots n} \delta_j \, \|x_j\|_2 \cdot \left\| \sum_{j=1}^{n} \delta_j x_j \otimes x_j \right\|_2^{1/2} \right)^{1/2} + \sqrt{\delta} \\ &\le C \Big(\mathbb{E}_\delta \Big(\sqrt{\log n(\delta)} \cdot \max_{j=1\cdots n} \delta_j \, \|x_j\|_2 \Big) \Big)^{1/2} \left(\mathbb{E}_\delta \left\| \sum_{j=1}^{n} \delta_j x_j \otimes x_j \right\|_2^{1/2} \right)^{1/2} + \sqrt{\delta}. \end{aligned} \tag{13}$$

To estimate the first term in the product here, we use the following

LEMMA 5.1. *Let $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$ and let $\delta_1 \cdots \delta_n$ be independent Bernoulli random variables taking value* 1 *with probability $\delta > 2/n$. Then*

$$\frac{\delta}{4e}\sqrt{\log \delta n} \cdot \sum_{j=1}^{1/\delta} a_j \leq \mathbb{E}\left(\sqrt{\log n(\delta)} \cdot \max_{j=1\cdots n} \delta_j a_j\right) \leq 4\delta\sqrt{\log \delta n} \cdot \sum_{j=1}^{1/\delta} a_j.$$

PROOF. To prove the upper estimate note that

$$\max_{j=1\cdots n} \delta_j a_j \leq \sum_{j=1}^{1/\delta} \delta_j a_j + a_{1/\delta}.$$

Hence,

$$\mathbb{E}\left(\sqrt{\log n(\delta)} \cdot \max_{j=1\cdots n} \delta_j a_j\right) \leq \mathbb{E}\left(\sqrt{\log n(\delta)} \cdot \sum_{j=1}^{1/\delta} \delta_j a_j\right) + a_{1/\delta} \cdot \mathbb{E}\sqrt{\log n(\delta)}. \tag{14}$$

Jensen's inequality yields

$$\mathbb{E}\sqrt{\log n(\delta)} \leq \sqrt{\log\left(\mathbb{E}\sum_{i=1}^{n} \delta_i + e\right)} \leq 2\sqrt{\log \delta n}. \tag{15}$$

By the linearity of expectation, the first term in the right hand side of (14) equals

$$\sum_{j=1}^{1/\delta} a_j \mathbb{E}\left(\delta_j \sqrt{\log n(\delta)}\right) \leq \sum_{j=1}^{1/\delta} a_j \mathbb{E}\left(\delta_j \sqrt{\log\left(\sum_{i \neq j} \delta_i + 1 + e\right)}\right),$$

where we estimated $n(\delta)$ replacing δ_j by 1. Taking the expectation first with respect to δ_j and then with respect to the other δ_i, and using Jensen's inequality, we bound the last expression by

$$\delta \sum_{j=1}^{1/\delta} a_j \cdot \sqrt{\log(\delta n + 1 + e)} \leq 2\delta \sum_{j=1}^{1/\delta} a_j \cdot \sqrt{\log \delta n}. \tag{16}$$

Finally, substituting (15) and (16) into (14), we obtain

$$\mathbb{E}\left(\sqrt{\log n(\delta)} \cdot \max_{j=1\cdots n} \delta_j a_j\right) \leq \left(2\delta \sum_{j=1}^{1/\delta} a_j + 2a_{1/\delta}\right)$$
$$\times \sqrt{\log \delta n} \leq 4\delta \sum_{j=1}^{1/\delta} a_j \cdot \sqrt{\log \delta n}.$$

To prove the lower bound, we estimate the product in Lemma 5.1 from below to make the terms independent. We have

$$\mathbb{E}\left(\sqrt{\log n(\delta)}\cdot \max_{j=1\cdots n}\delta_j a_j\right) \ge \mathbb{E}\left(\sqrt{\log\left(\sum_{i=1/\delta+1}^{n}\delta_i + e\right)}\cdot \max_{j=1\cdots 1/\delta}\delta_j a_j\right)$$
$$= \mathbb{E}\sqrt{\log\left(\sum_{i=1/\delta+1}^{n}\delta_i + e\right)}\cdot \mathbb{E}\max_{j=1\cdots 1/\delta}\delta_j a_j. \quad (17)$$

These terms will be estimated separately. Since $\mathbb{P}\left(\sum_{i=1/\delta+1}^{n}\delta_i \ge \delta n/2\right) \ge 1/2$,

$$\mathbb{E}\sqrt{\log\left(\sum_{i=1/\delta+1}^{n}\delta_i + e\right)} \ge \frac{1}{2}\sqrt{\log\frac{\delta n}{2}}.$$

Let $1 \le k \le 1/\delta$. Denote by A_k the event $\{\delta_k = 1, \delta_j = 0 \text{ for } 1 \le j \le 1/\delta,\ j \ne k\}$. Then

$$\mathbb{P}(A_k) = \delta\cdot(1-\delta)^{1/\delta-1} \ge \delta/e.$$

Since the events $A_1, \cdots, A_{1/\delta}$ are disjoint,

$$\mathbb{E}\max_{j=1\cdots 1/\delta}\delta_j a_j \ge \sum_{k=1}^{1/\delta} a_k\mathbb{P}(A_k) \ge \frac{\delta}{e}\sum_{j=1}^{1/\delta} a_j.$$

Substituting this estimate into (17) finishes the proof of Lemma 5.1. □

Now we can complete the proof of Theorem 1.8. Combining Lemma 5.1 and (13), we get

$$E \le C\left(4\delta\sqrt{\log\delta n}\sum_{j=1}^{1/\delta}\|x_j\|_2\right)^{1/2} E^{1/2} + \sqrt{\delta} = 2C\left(\sqrt{\log\delta n}\|A\|_{(1/\delta)}\right)^{1/2} E^{1/2} + \sqrt{\delta}.$$

It can be easily checked that $E \le aE^{1/2} + b$ implies $E \le 4a^2 + 2b$. Hence, recalling that $\delta = q/n$, we conclude that

$$E \le 16C^2\sqrt{\log q}\cdot\|A\|_{(n/q)} + 2\sqrt{q/n}.$$

This completes the proof of Theorem 1.8.

ACKNOWLEDGMENTS. We are grateful for R. Kannan for his comments on the initial version of this paper, and to M. Karpinski for explaining what was the correct consequence of Corollary 1.7 for MAX-2-CSP problems. Finally, we thank the referees for their valuable comments and suggestions.

REFERENCES

ALON, N., FERNANDEZ DE LA VEGA, W., KANNAN, R., AND KARPINSKI, M. 2002. Random sampling and approximation of MAX-CSPs. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, ACM, New York, 232–239.

ALON, N., FERNANDEZ DE LA VEGA, W., KANNAN, R., AND KARPINSKI, M. 2003. Random Sampling and approximation of MAX-CSPs. *J. Comput. Syst. Sci. 67*, 212–243.

AZAR, Y., FIAT, A., KARLIN, A., MCSCHERRY, F., AND SAIA, J. 2001. Spectral analysis for data mining. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*, ACM, New York, 619–626.

BERRY, M. W., DRMAC, Z., AND JESSUP, E. R. 1999. Matrices, vector spaces and information retrieval. *SIAM Rev. 41*, 335–362.

BERRY, M. W., DUMAIS, S. T., AND O'BRIAN, S. T. 1995. Using linear algebra for intelligent information retrieval. *SIAM Rev. 37*, 573–595.

BOURGAIN, J., AND TZAFRIRI, L. 1987. Invertibility of "large" sumatricies with applications to the geometry of Banach spaces and harmonic analysis. Israel *Journal of Mathematics 57*, 137–223.

DEERWESTER, S. T., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R.H. 1990. Indexing by latent semantic analysis. *J. Amer. Soci. Inf. Sci. 41*, 391–407.

DRINEAS, P., FRIEZE, A., KANNAN, R., VEMPALA, S., AND VINAY, V. 2004. Clustering large graphs via Singular Value Decomposition. *Mach. Learn. 56*, 9–33.

DRINEAS, P., AND KANNAN, R. 2003. Pass efficient algorithms for approximating large matrices. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms* (Baltimore, MD), ACM, New York, 223–232.

DRINEAS, P., KANNAN, R., AND MAHONEY, M. 2006a. Fast Monte-Carlo algorithms for Matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput. 36*, 158–183.

DRINEAS, P., MAHONEY, M. P., AND KANNAN, R. 2006b. Fast Monte-Carlo algorithms for matrices III: Computing an efficient approximate decomposition of a matrix. *SIAM J. Comput. 36*, 184–206.

FERNANDEZ DE LA VEGA, W. 1996. MAX-CUT has a randomized approximation scheme in dense graphs. *Rand. Struct. Algorithms 8*, 187–199.

FRIEZE, A., KANNAN, R., AND VEMPALA, S. 2004. Fast Monte-Carlo algorithms for finding low-rank approximations. *J. ACM 51*, 1025–1041.

JERRY, M. J., AND LINOFF, G. 1997. Data mining techniques. Wiley, New York.

KASHIN, B., AND TZAFRIRI, L. Some remarks on the restrictions of operators to coordinate subspaces. Unpublished notes.

LEDOUX, M., AND TALAGRAND, M. 1991. *Probability in Banach spaces*, Springer-Verlag, New York.

LUNIN, A. A. 1975. On operator norms of submatrices. *Math. USSR Sbornik 27*, 481–502.

PAPADIMITRIOU, C. H., RAGHVAN, P., TAMAKI, H., AND VEMPALA, S. 1998. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci. 61*, 217–235.

RUDELSON, M. 1999. Random vectors in isotropipc position. *J. Funct. Anal. 164*, 60–72.

TALAGRAND, M. 1995. Sections of smooth convex bodies via majorizing measures. *Acta Math. 175*, 273–300

VERSHYNIN, R. 2001. John's decompositions: Selecting a large part. *Isr. J. Math. 122*, 253–277.

RECEIVED MARCH 2005; REVISED DECEMBER 2006; ACCEPTED MAY 2007