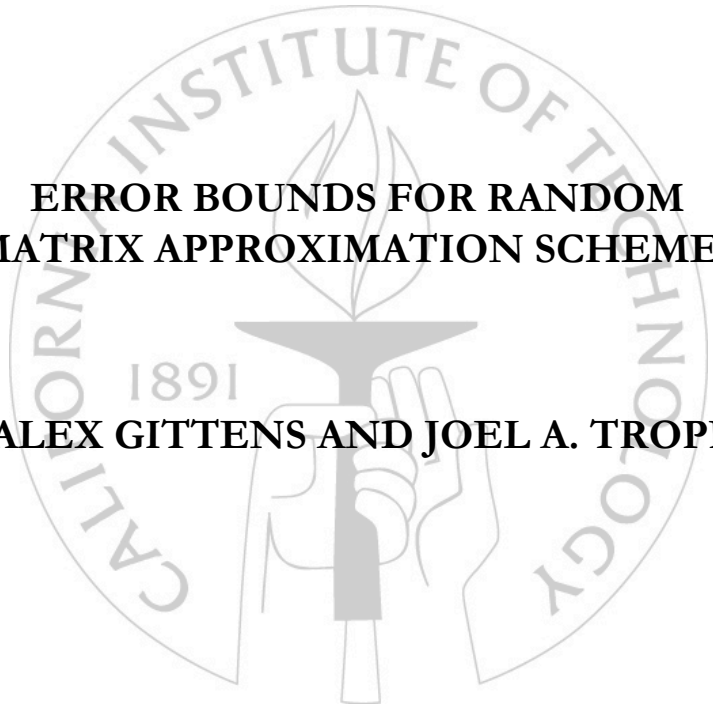

ERROR BOUNDS FOR RANDOM MATRIX APPROXIMATION SCHEMES

ALEX GITTENS AND JOEL A. TROPP

Technical Report No. 2014-01
August 2014

APPLIED & COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
mail code 9-94 · pasadena, ca 91125

ERROR BOUNDS FOR RANDOM MATRIX APPROXIMATION SCHEMES

A. GITTENS AND J. A. TROPP

Abstract. Randomized matrix sparsification has proven to be a fruitful technique for producing faster algorithms in applications ranging from graph partitioning to semidefinite programming. In the decade or so of research into this technique, the focus has been—with few exceptions—on ensuring the quality of approximation in the spectral and Frobenius norms. For certain graph algorithms, however, the $\infty \to 1$ norm may be a more natural measure of performance.

This paper addresses the problem of approximating a real matrix $\boldsymbol{A}$ by a sparse random matrix $\boldsymbol{X}$ with respect to several norms. It provides the first results on approximation error in the $\infty \to 1$ and $\infty \to 2$ norms, and it uses a result of Latała to study approximation error in the spectral norm. These bounds hold for a reasonable family of random sparsification schemes, those which ensure that the entries of $\boldsymbol{X}$ are independent and average to the corresponding entries of $\boldsymbol{A}$. Optimality of the $\infty \to 1$ and $\infty \to 2$ error estimates is established. Concentration results for the three norms hold when the entries of $\boldsymbol{X}$ are uniformly bounded. The spectral error bound is used to predict the performance of several sparsification and quantization schemes that have appeared in the literature; the results are competitive with the performance guarantees given by earlier scheme-specific analyses.

1. Introduction

Massive datasets are ubiquitous in modern data processing. Classical dense matrix algorithms are poorly suited to such problems because their running times scale superlinearly with the size of the matrix. When the dataset is sparse, one prefers to use sparse matrix algorithms, whose running times depend more on the sparsity of the matrix than on the size of the matrix. Of course, in many applications the matrix is *not* sparse. Accordingly, one may wonder whether it is possible to approximate a computation on a large dense matrix with a related computation on a sparse approximant to the matrix.

Let $\|\cdot\|$ be a norm on matrices. We may phrase the following question: Given a matrix $\boldsymbol{A}$, how can one efficiently generate a sparse matrix $\boldsymbol{X}$ for which the approximation error $\|\boldsymbol{A} - \boldsymbol{X}\|$ is small?

In the seminal papers [AM07, AM01], Achlioptas and McSherry demonstrate that one can bound, *a priori*, the spectral and Frobenius norm errors incurred when using a particular random approximation scheme. They use this scheme as the basis of an efficient algorithm for calculating near optimal low-rank approximations to large matrices. In a related work [AHK06], Arora, Hazan, and Kale present another randomized scheme for computing sparse approximants with controlled Frobenius and spectral norm errors.

The literature has concentrated on the behavior of the approximation error in the spectral and Frobenius norms; however, these norms are not always the most natural choice. For instance, the problem of graph sparsification is naturally posed as a question of preserving the cut-norm of a graph Laplacian. The strong equivalency of the cut-norm and the $\infty \to 1$ norm suggests that, for graph-theoretic applications, it may be fruitful to consider the behavior of the $\infty \to 1$ norm under sparsification. In other applications, e.g., the column subset selection algorithm in [Tro09], the $\infty \to 2$ norm is the norm of interest.

This paper investigates the errors incurred by approximating a fixed real matrix with a random matrix. Our results apply to any scheme in which the entries of the approximating matrix are independent and average to the corresponding entries of the fixed matrix. Our main contribution is a bound on the expected $\infty \to p$ norm error, which we specialize to the case of the $\infty \to 1$ and $\infty \to 2$

norms. We also use a result of Latała [Lat05] to find an optimal bound on the expected spectral approximation error, and we establish the subgaussianity of the spectral approximation error.

1.1. **Summary of results.** Consider the matrix $\boldsymbol{A}$ as an operator from ℓ_∞^n to ℓ_p^m, where $1 \le p \le \infty$. The operator norm of $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\|_{\infty\to p} = \max_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{u}\|_p}{\|\boldsymbol{u}\|_\infty}.$$

The major novel contributions of this paper are estimates of the $\infty \to 1$ and $\infty \to 2$ norm approximation errors induced by random matrix sparsification schemes. Let $\boldsymbol{A}$ be a target matrix. Suppose that $\boldsymbol{X}$ is a random matrix with independent entries that satisfies $\mathbb{E}\boldsymbol{X} = \boldsymbol{A}$. Then

$$\mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to 1} \le 2\left[\sum\nolimits_k \left(\sum\nolimits_j \operatorname{Var}(X_{jk})\right)^{1/2} + \sum\nolimits_j \left(\sum\nolimits_k \operatorname{Var}(X_{jk})\right)^{1/2}\right]$$

and

$$\mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to 2} \le 2\left(\sum\nolimits_{jk} \operatorname{Var}(X_{jk})\right)^{1/2} + 2\sqrt{m}\min_{\boldsymbol{D}}\max_j \left(\sum\nolimits_k \frac{\operatorname{Var}(X_{jk})}{d_k^2}\right)^{1/2}$$

where $\boldsymbol{D}$ is a positive diagonal matrix with $\operatorname{trace}(\boldsymbol{D}^2) = 1$. We complement these estimates with tail bounds for the approximation error in each norm.

We also note that

$$\mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\| \le \mathrm{C}\left[\max_j \left(\sum\nolimits_k \operatorname{Var}(X_{jk})\right)^{1/2} + \max_k \left(\sum\nolimits_j \operatorname{Var}(X_{jk})\right)^{1/2} + \left(\sum\nolimits_{jk} \mathbb{E}(X_{jk}-a_{jk})^4\right)^{1/4}\right]$$

where C is a universal constant. This estimate follows from Latała's work [Lat05]. We use our spectral norm results to analyze the sparsification and quantization schemes in [AHK06] and [AM07], and we show that our analysis yields error estimates competitive with those established in the respective works.

While on the path to proving the $\infty\to1$ and $\infty\to2$ norm approximation error estimates, we derive analogous relations of independent interest that bound the expected norm of a random matrix $\boldsymbol{Z}$ with independent, zero-mean entries:

$$\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 1} \le 2\mathbb{E}\left(\|\boldsymbol{Z}\|_{\mathrm{col}} + \left\|\boldsymbol{Z}^T\right\|_{\mathrm{col}}\right),$$

where $\|\boldsymbol{A}\|_{\mathrm{col}}$ is the sum of the ℓ_2 norms of the columns of $\boldsymbol{A}$, and

$$\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 2} \le 2\mathbb{E}\,\|\boldsymbol{Z}\|_{\mathrm{F}} + 2\min_{\boldsymbol{D}} \mathbb{E}\left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2\to\infty},$$

where $\boldsymbol{D}$ is a diagonal positive matrix with $\operatorname{trace}(\boldsymbol{D}^2) = 1$. More generally,

$$\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to p} \le 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{z}_k\right\|_p + 2\max_{\|\boldsymbol{u}\|_q=1} \mathbb{E}\sum\nolimits_k \left|\sum\nolimits_j \varepsilon_j Z_{jk} u_j\right|,$$

where q is the conjugate exponent to p and $\boldsymbol{z}_k$ is the kth column of $\boldsymbol{Z}$. Here, $\{\varepsilon_k\}$ is a sequence of independent random variables uniform on $\{\pm1\}$.

1.2. **Outline.** Section 2 offers an overview of the related strands of research, and Section 3 introduces the reader to our notations. In Section 4, we establish the foundations for our $\infty\to1$ and $\infty\to2$ error estimates: an estimate of the expected $\infty\to p$ norm of a random matrix with independent, zero-mean entries and a tail bound for this quantity. In Sections 5 and 6, these estimates are specialized to the cases $p = 1$ and $p = 2$, respectively, and we establish the optimality of the resulting expressions. The bounds are provided in both their most generic forms and ones more suitable for applications. In Section 7, we use a result due to Latała [Lat05] to find an optimal estimate of the spectral approximation error. A deviation bound is provided for the spectral norm which captures the correct (subgaussian) tail behavior. We show that our estimates applied to the

sparsification schemes in [AHK06] and [AM07] recover performance guarantees comparable with those obtained from the original analyses, under slightly weaker hypotheses.

2. Related Work

In recent years, much attention has been paid to the problem of using sampling methods to approximate linear algebra computations efficiently. Such algorithms find applications in areas like data mining, computational biology, and other areas where the relevant matrices may be too large to fit in RAM, or large enough that the computational requirements of standard algorithms become prohibitive. Here we review the streams of literature that have motivated and influenced this work.

2.1. **Randomized sparsification.** The seminal research of Frieze, Kannan, and Vempala [FKV98, FKV04] on using random sampling to decrease the running time of linear algebra algorithms focuses on approximations to $\boldsymbol{A}$ constructed from random subsets of its rows. The subsequent influential works of Drineas, Kannan, and Mahoney [DKM06a, DKM06b, DKM06c] also analyze the performance of Monte Carlo algorithms which sample from the columns or rows of $\boldsymbol{A}$.

In [AM01], Achlioptas and McSherry advance a different approach: instead of using low-rank approximants, they sample the entries of $\boldsymbol{A}$ to produce a sparse matrix $\boldsymbol{X}$ that has a spectral decomposition close to that of $\boldsymbol{A}$. This transition from row/column sampling to independent sampling of the entries allows them to bring to bear powerful techniques from random matrix theory. In [AM01], their main tool is a result on the concentration of the spectral norm of a random matrix with independent, zero-mean, bounded entries. In a follow-up paper, [AM07], they obtain better estimates by using a sharper concentration result derived from Talagrand's inequality. They point out the importance of using schemes which sparsify a given entry in $\boldsymbol{A}$ with a probability proportional to its magnitude. Such adaptive sparsification schemes keep the variance of the individual entries small, which tends to keep the error in approximating $\boldsymbol{A}$ with $\boldsymbol{X}$ small. Their scheme requires either two passes through the matrix or prior knowledge of an upper bound for the largest magnitude in $\boldsymbol{A}$.

In [AHK06], Arora, Hazan, and Kale describe a random sparsification algorithm which partially quantizes its inputs and requires only one pass through the matrix. They use an epsilon-net argument and Chernoff bounds to establish that with high probability the resulting approximant has small error and high sparsity.

2.2. **Probability in Banach spaces.** In [RV07], Rudelson and Vershynin take a different approach to the Monte Carlo methodology for low-rank approximation. They consider $\boldsymbol{A}$ as a linear operator between finite-dimensional Banach spaces and apply techniques of probability in Banach spaces: decoupling, symmetrization, Slepian's lemma for Rademacher random variables, and a law of large numbers for operator-valued random variables. They show that, if $\boldsymbol{A}$ can be approximated by any rank-r matrix, then it is possible to obtain an accurate rank-k approximation to $\boldsymbol{A}$ by sampling $\mathrm{O}(r \log r)$ rows of $\boldsymbol{A}$. Additionally, they quantify the behavior of the $\infty \to 1$ and $\infty \to 2$ norms of random submatrices.

Our methods are similar to those of Rudelson and Vershynin in [RV07] in that we consider $\boldsymbol{A}$ as a linear operator between finite-dimensional Banach spaces and use some of the same tools of probability in Banach spaces. Whereas Rudelson and Vershynin consider the behavior of the norms of random submatrices of $\boldsymbol{A}$, we consider the behavior of the norms of matrices formed by randomly sparsifying (or quantizing) the entries of $\boldsymbol{A}$. This yields error bounds applicable to schemes that sparsify or quantize matrices entrywise. Since some graph algorithms depend more on the number of edges in the graph than the number of vertices, such schemes may be useful in developing algorithms for handling large graphs.

2.3. **Random sampling of graphs.** The bulk of the literature has focused on the behavior of the spectral and Frobenius norms under randomized sparsification, but the $\infty \to 1$ norm occurs naturally in connection with graph theory. Let us review the relevant ideas.

Consider a weighted simple graph $G = (V, E, \omega)$ with adjacency matrix $\boldsymbol{A}$ given by

$$a_{jk} = \begin{cases} \omega_{jk}, & (j,k) \in E \\ 0, & \text{otherwise.} \end{cases}$$

A *cut* is a partition of the vertices into two blocks: $V = S \cup \overline{S}$. The *cost* of a cut is the sum of the weights of all edges in E which have one vertex in S and one vertex in $\overline{S}$. Several problems relating to cuts are of considerable practical interest. In particular, the MAXCUT problem, to determine the cut of maximum cost in a graph, is common in computer science applications. The cuts of maximum cost are exactly those which realize the *cut-norm* of the adjacency matrix, which is defined as

$$\|\boldsymbol{A}\|_{\mathrm{C}} = \max_{S \subset E} \left| \sum\nolimits_{(j,k) \in E} \omega_{jk} (\mathbb{1}_S)_j (\mathbb{1}_{\overline{S}})_k \right|,$$

where $\mathbb{1}_S$ is the indicator vector for S. Finding the cut-norm of a general matrix is NP-hard, but in [AN04], the authors offer a randomized polynomial-time algorithm which finds a submatrix $\tilde{\boldsymbol{A}}$ of $\boldsymbol{A}$ such that $|\sum_{jk} \tilde{a}_{jk}| \geq 0.56 \|\boldsymbol{A}\|_{\mathrm{C}}$. One crucial point in the derivation of the algorithm is the fact that the $\infty \to 1$ norm is strongly equivalent with the cut-norm:

$$\|\boldsymbol{A}\|_{\mathrm{C}} \leq \|\boldsymbol{A}\|_{\infty \to 1} \leq 4 \|\boldsymbol{A}\|_{\mathrm{C}}.$$

In his thesis [Kar95] and the sequence of papers [Kar94a, Kar94b, Kar96], Karger introduces the idea of random sampling to increase the efficiency of calculations with graphs, with a focus on cuts. In [Kar96], he shows that by picking each edge of the graph with a probability inversely proportional to the density of edges in a neighborhood of that edge, one can construct a *sparsifier*, i.e., a graph with the same vertex set and significantly fewer edges that preserves the value of each cut to within a factor of $(1 \pm \epsilon)$.

In [SS08], Spielman and Srivastava improve upon this sampling scheme, instead keeping an edge with probability proportional to its *effective resistance*—a measure of how likely it is to appear in a random spanning tree of the graph. They provide an algorithm which produces a sparsifier with $\mathrm{O}\big((n \log n)/\epsilon^2\big)$ edges, where n is the number of vertices in the graph. They obtain this result by reducing the problem to the behavior of projection matrices Π_G and $\Pi_{G'}$ associated with the original graph and the sparsifier, and appealing to a spectral norm concentration result.

The $\log n$ factor in [SS08] seems to be an unavoidable consequence of using spectral norm concentration. In [BSS09], Batson et. al. prove that the $\log n$ factor is not intrinsic: they establish that every graph has a sparsifier that has $\mathrm{O}(n)$ edges. The proof is constructive and provides a deterministic algorithm for constructing such optimal sparsifiers in $\mathrm{O}\big(n^3 m\big)$ time, where m is the number of edges in the original graph.

The algorithm of [BSS09] is clearly not suitable for sparsifying graphs with a large number of vertices. Part of our motivation for investigating the $\infty \to 1$ approximation error is the belief that the equivalence of the cut-norm with the $\infty \to 1$ norm means that matrix sparsification in the $\infty \to 1$ norm might be useful for efficiently constructing optimal sparsifiers for such graphs.

3. Background

We establish the notation used in the sequel and introduce our key technical tools.

All quantities are real. The kth column of the matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{a}_k$ and the entries are denoted a_{jk}.

For $1 \leq p \leq \infty$, the ℓ_p norm of $\boldsymbol{x}$ is written as $\|\boldsymbol{x}\|_p$. We treat $\boldsymbol{A}$ as an operator from ℓ_p^n to ℓ_q^m, and the $p \to q$ operator norm of $\boldsymbol{A}$ is written as $\|\boldsymbol{A}\|_{p \to q}$. The $p \to q$ and $q' \to p'$ norms are dual

in the sense that

$$\|\boldsymbol{A}\|_{p\to q} = \left\|\boldsymbol{A}^T\right\|_{q'\to p'}.$$

This paper is concerned primarily with the spectral norm and the $\infty\to 1$ and $\infty\to 2$ norms. The spectral norm $\|\boldsymbol{A}\|$ is the largest singular value of $\boldsymbol{A}$. The $\infty\to 1$ and $\infty\to 2$ norms do not have such nice interpretations, and they are NP-hard to compute for general matrices [Roh00]. We remark that $\|\boldsymbol{A}\|_{\infty\to 1} = \|\boldsymbol{A}\boldsymbol{x}\|_1$ and $\|\boldsymbol{A}\|_{\infty\to 2} = \|\boldsymbol{A}\boldsymbol{y}\|_2$ for certain vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ whose components take values ± 1.

An additional operator norm, the $2\to\infty$ norm, is also of interest: it is the largest ℓ_2 norm achieved by a row of $\boldsymbol{A}$. We encounter two norms in the sequel that are not operator norms: the Frobenius norm, denoted by $\|\boldsymbol{A}\|_{\mathrm{F}}$, and the column norm

$$\|\boldsymbol{A}\|_{\mathrm{col}} = \sum\nolimits_k \|\boldsymbol{a}_k\|_2.$$

The expectation of a random variable X is written $\mathbb{E}X$ and its variance is written $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$. The expectation taken with respect to one variable X, with all others fixed, is written $\mathbb{E}_X$. The L_q norm of X is denoted by $\mathbb{E}^q(X) = (\mathbb{E}|X|^q)^{1/q}$.

The expression $X \sim Y$ indicates the random variables X and Y are identically distributed. Given a random variable X, the symbol X' denotes a random variable independent of X such that $X' \sim X$. The indicator variable of the event $X > Y$ is written $\mathbb{1}_{X>Y}$.

The Bernoulli distribution with expectation p is written $\mathrm{Bern}(p)$ and the Binomial distribution of n independent trials each with success probability p is written $\mathrm{Bin}(n, p)$. We write $X \sim \mathrm{Bern}(p)$ to indicate X is Bernoulli.

A Rademacher random variable takes on the values ± 1 with equal probability. A vector whose components are independent Rademacher variables is called a Rademacher vector. A real sum whose terms are weighted by independent Rademacher variables is called a Rademacher sum. The Khintchine inequality [Sza76] gives information on the moments of a Rademacher sum; in particular, it tells us the expected value of the sum is equivalent with the ℓ_2 norm of the vector $\boldsymbol{x}$:

Proposition 1 (Khintchine inequality)**.** *Let $\boldsymbol{x}$ be a real vector, and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Then*

$$\frac{1}{\sqrt{2}}\|\boldsymbol{x}\|_2 \le \mathbb{E}\left|\sum\nolimits_k \varepsilon_k x_k\right| \le \|\boldsymbol{x}\|_2.$$

4. The $\infty\to p$ norm of a Random Matrix

We are interested in schemes that approximate a given matrix $\boldsymbol{A}$ by means of a random matrix $\boldsymbol{X}$ in such a way that the entries of $\boldsymbol{X}$ are independent and $\mathbb{E}\boldsymbol{X} = \boldsymbol{A}$. It follows that the error matrix $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$ has independent, zero-mean entries. Our intellectual concern is the class of sparse random matrices, but this property does not play a role at this stage of the analysis.

In this section, we derive a bound on the expected value of the $\infty\to p$ norm of a random matrix with independent, zero-mean entries. We also study the tails of this error. In the next two sections, we use the results of this section to reach more detailed conclusions on the $\infty\to 1$ and $\infty\to 2$ norms of $\boldsymbol{Z}$.

4.1. Expected $\infty\to p$ norm.

The main tool used to derive the bound on the expected norm of $\boldsymbol{Z}$ is the following symmetrization argument [vW96, Lemma 2.3.1 et seq.].

Proposition 2. *Let $Z_1, \dots, Z_n, Z_1', \dots, Z_n'$ be independent random variables satisfying $Z_i \sim Z_i'$, and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Let $\mathcal{F}$ be a family of functions such that*

$$\sup_{f\in\mathcal{F}} \sum\nolimits_{k=1}^n (f(Z_k) - f(Z_k'))$$

is measurable. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum\nolimits_{k=1}^{n} (f(Z_k) - f(Z_k')) = \mathbb{E} \sup_{f \in \mathcal{F}} \sum\nolimits_{k=1}^{n} \varepsilon_k (f(Z_k) - f(Z_k')).$$

Since we work with finite-dimensional probability models and linear functions, measurability concerns can be ignored.

The other critical tool is a version of Talagrand's Rademacher comparison theorem [LT91, Theorem 4.12 et seq.].

Proposition 3. *Fix finite-dimensional vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Then*

$$\mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_{k=1}^{n} \varepsilon_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| \leq \mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_{k=1}^{n} \varepsilon_k \langle \boldsymbol{z}_k, \boldsymbol{u} \rangle.$$

Now we state and prove the bound on the expected norm of $\boldsymbol{Z}$.

Theorem 1. *Let $\boldsymbol{Z}$ be a random matrix with independent, zero-mean entries and let $\boldsymbol{\varepsilon}$ be a Rademacher vector independent of $\boldsymbol{Z}$. Then*

$$\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to p} \leq 2\mathbb{E} \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{z}_k \right\|_p + 2 \max_{\|\boldsymbol{u}\|_q = 1} \mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j \varepsilon_j Z_{jk} u_j \right|$$

where q is the conjugate exponent of p.

Proof of Theorem 1. By duality,

$$\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to p} = \mathbb{E} \left\| \boldsymbol{Z}^T \right\|_{q \to 1} = \mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle|.$$

Center the terms in the sum and apply subadditivity of the maximum to get

$$\begin{aligned} \mathbb{E} \|\boldsymbol{Z}\|_{\infty \to p} &\leq \mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k (|\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| - \mathbb{E}' |\langle \boldsymbol{z}_k', \boldsymbol{u} \rangle|) + \max_{\|\boldsymbol{u}\|_q = 1} \mathbb{E} \sum\nolimits_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| \\ &=: F + S. \end{aligned} \tag{1}$$

Begin with the first term in (1). Use Jensen's inequality to draw the expectation outside of the maximum:

$$F \leq \mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k (|\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| - |\langle \boldsymbol{z}_k', \boldsymbol{u} \rangle|).$$

Now apply Proposition 2 to symmetrize the random variable:

$$F \leq \mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k \varepsilon_k (|\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| - |\langle \boldsymbol{z}_k', \boldsymbol{u} \rangle|).$$

By the subadditivity of the maximum,

$$F \leq \mathbb{E} \left(\max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k \varepsilon_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| + \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k -\varepsilon_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle| \right) = 2\mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k \varepsilon_k |\langle \boldsymbol{z}_k, \boldsymbol{u} \rangle|,$$

where we have invoked the fact that $-\varepsilon_k$ has the Rademacher distribution. Apply Proposition 3 to get the final estimate of F:

$$F \leq 2\mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k \varepsilon_k \langle \boldsymbol{z}_k, \boldsymbol{u} \rangle = 2\mathbb{E} \max_{\|\boldsymbol{u}\|_q = 1} \left\langle \sum\nolimits_k \varepsilon_k \boldsymbol{z}_k, \boldsymbol{u} \right\rangle = 2\mathbb{E} \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{z}_k \right\|_p.$$

Now consider the last term in (1). Use Jensen's inequality to prepare for symmetrization:

$$\begin{aligned} S &= \max_{\|\boldsymbol{u}\|_q = 1} \mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j Z_{jk} u_j \right| = \max_{\|\boldsymbol{u}\|_q = 1} \mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j (Z_{jk} - \mathbb{E}' Z_{jk}') u_j \right| \\ &\leq \max_{\|\boldsymbol{u}\|_q = 1} \sum\nolimits_k \mathbb{E} \left| \sum\nolimits_j (Z_{jk} - Z_{jk}') u_j \right|. \end{aligned}$$

Apply Proposition 2 to the expectation of the inner sum to see

$$S \le \max_{\|\boldsymbol{u}\|_q=1} \sum\nolimits_k \mathbb{E}\left|\sum\nolimits_j \varepsilon_j (Z_{jk} - Z'_{jk}) u_j\right|.$$

The triangle inequality gives us the final expression:

$$S \le \max_{\|\boldsymbol{u}\|_q=1} 2\mathbb{E} \sum\nolimits_k \left|\sum\nolimits_j \varepsilon_j Z_{jk} u_j\right|.$$

Introduce the bounds for F and S into (1) to complete the proof. □

4.2. **Tail bound for $\infty \to p$ norm.** We now develop a deviation bound for the $\infty \to p$ approximation error. The argument is based on a bounded differences inequality.

First we establish some notation. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a measurable function of n random variables. Let $X_1, \dots, X_n$ be independent random variables, and write $W = g(X_1, \dots, X_n)$. Let W_i denote the random variable obtained by replacing the ith argument of g with an independent copy: $W_i = g(X_1, \dots, X'_i, \dots, X_n)$.

The following bounded differences inequality states that if g is insensitive to changes of a single argument, then W does not deviate much from its mean.

Proposition 4 ([BLM03])**.** *Let W and $\{W_i\}$ be random variables defined as above. Assume that there exists a positive number C such that, almost surely,*

$$\sum\nolimits_{i=1}^n (W - W_i)^2 \mathbb{1}_{W > W_i} \le C.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(W > \mathbb{E}W + t\right) \le \mathrm{e}^{-t^2/(4C)}.$$

To apply Proposition 4, we let $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$ be our error matrix, $W = \|\boldsymbol{Z}\|_{\infty \to p}$, and $W^{jk} = \left\|\boldsymbol{Z}^{jk}\right\|_{\infty \to p}$, where $\boldsymbol{Z}^{jk}$ is a matrix obtained by replacing $a_{jk} - X_{jk}$ with an identically distributed variable $a_{jk} - X'_{jk}$ while keeping all other variables fixed. The $\infty \to p$ norms are sufficiently insensitive to each entry of the matrix that Proposition 4 gives us a useful deviation bound.

Theorem 2. *Fix an $m \times n$ matrix $\boldsymbol{A}$, and let $\boldsymbol{X}$ be a random matrix with independent entries for which $\mathbb{E}X = \boldsymbol{A}$. Assume $|X_{jk}| \le \frac{D}{2}$ almost surely for all j, k. Then, for all $t > 0$,*

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty \to p} > \mathbb{E}\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty \to p} + t\right) \le \mathrm{e}^{-t^2/(4D^2 n m^s)}$$

where $s = \max\{0, 1 - 2/q\}$ and q is the conjugate exponent to p.

Proof. Let q be the conjugate exponent of p, and choose $\boldsymbol{u}, \boldsymbol{v}$ such that $W = \boldsymbol{u}^T \boldsymbol{Z} \boldsymbol{v}$ and $\|\boldsymbol{u}\|_q = 1$ and $\|\boldsymbol{v}\|_\infty = 1$. Then

$$(W - W^{jk}) \mathbb{1}_{W > W^{jk}} \le \boldsymbol{u}^T \left(\boldsymbol{Z} - \boldsymbol{Z}^{jk}\right) \boldsymbol{v}\, \mathbb{1}_{W > W^{jk}} = (X'_{jk} - X_{jk}) u_j v_k \mathbb{1}_{W > W^{jk}} \le D|u_j v_k|.$$

This implies

$$\sum\nolimits_{j,k} (W - W^{jk})^2 \mathbb{1}_{W > W^{jk}} \le D^2 \sum\nolimits_{j,k} |u_j v_k|^2 \le n D^2 \|\boldsymbol{u}\|_2^2,$$

so we can apply Proposition 4 if we have an estimate for $\|\boldsymbol{u}\|_2^2$. We have the bounds $\|\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_q$ for $q \in [1, 2]$ and $\|\boldsymbol{u}\|_2 \le m^{1/2 - 1/q} \|\boldsymbol{u}\|_q$ for $q \in [2, \infty]$. Therefore,

$$\sum\nolimits_{j,k} (W - W^{jk})^2 \mathbb{1}_{W > W^{jk}} \le D^2 \begin{cases} n m^{1-2/q}, & q \in [2, \infty] \\ n, & q \in [1, 2]. \end{cases}$$

It follows from Proposition 4 that

$$\mathbb{P}\left(\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p} > \mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p} + t\right) = \mathbb{P}\left(W > \mathbb{E}W + t\right) \leq \mathrm{e}^{-t^2/(4D^2nm^s)}$$

where $s = \max\{0, 1-2/q\}$. □

It is often convenient to measure deviations on the scale of the mean. Taking $t = \delta\mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p}$ in Theorem 2 gives the following result.

Corollary 1. *Under the conditions of Theorem 2, for all $\delta > 0$,*

$$\mathbb{P}\left(\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p} > (1+\delta)\mathbb{E}\,\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p}\right) \leq \mathrm{e}^{-\delta^2\left(\mathbb{E}\|\boldsymbol{A}-\boldsymbol{X}\|_{\infty\to p}\right)^2/(4D^2nm^s)}.$$

5. Approximation in the $\infty\to 1$ norm

In this section, we develop the $\infty\to 1$ error bound as a consequence of Theorem 1. We then prove that one form of the error bound is optimal, and we describe an example of its application to matrix sparsification.

5.1. Expected $\infty\to 1$ norm.

To derive the $\infty\to 1$ error bound, we first apply Theorem 1 with $p = 1$.

Theorem 3. *Suppose that $\boldsymbol{Z}$ is a random matrix with independent, zero-mean entries. Then*

$$\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 1} \leq 2\mathbb{E}(\|\boldsymbol{Z}\|_{\mathrm{col}} + \left\|\boldsymbol{Z}^T\right\|_{\mathrm{col}}).$$

Proof. Apply Theorem 1 to get

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 1} &\leq 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{z}_k\right\|_1 + 2\max_{\|\boldsymbol{u}\|_\infty=1}\mathbb{E}\sum\nolimits_k\left|\sum\nolimits_j \varepsilon_j Z_{jk}u_j\right| \\ &=: F + S.\end{aligned} \tag{2}$$

Use Hölder's inequality to bound the first term in (2) with a sum of squares:

$$\begin{aligned}F &= 2\mathbb{E}\sum\nolimits_j\left|\sum\nolimits_k \varepsilon_k Z_{jk}\right| = 2\mathbb{E}_{\boldsymbol{Z}}\sum\nolimits_j \mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum\nolimits_k \varepsilon_k Z_{jk}\right| \\ &\leq 2\mathbb{E}_{\boldsymbol{Z}}\sum\nolimits_j\left(\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum\nolimits_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2}.\end{aligned}$$

The inner expectation can be computed exactly by expanding the square and using the independence of the Rademacher variables:

$$F \leq 2\mathbb{E}\sum\nolimits_j\left(\sum\nolimits_k Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\left\|\boldsymbol{Z}^T\right\|_{\mathrm{col}}.$$

We treat the second term in the same manner. Use Hölder's inequality to replace the sum with a sum of squares and invoke the independence of the Rademacher variables to eliminate cross terms:

$$S \leq 2\max_{\|\boldsymbol{u}\|_\infty=1}\mathbb{E}_{\boldsymbol{Z}}\sum\nolimits_k\left(\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum\nolimits_j \varepsilon_j Z_{jk}u_j\right|^2\right)^{1/2} = 2\max_{\|\boldsymbol{u}\|_\infty=1}\mathbb{E}\sum\nolimits_k\left(\sum\nolimits_j Z_{jk}^2u_j^2\right)^{1/2}.$$

Since $\|\boldsymbol{u}\|_\infty = 1$, it follows that $u_j^2 \leq 1$ for all j, and

$$S \leq 2\mathbb{E}\sum\nolimits_k\left(\sum\nolimits_j Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\,\|\boldsymbol{Z}\|_{\mathrm{col}}.$$

Introduce these estimates for F and S into (2) to complete the proof. □

Taking $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$ in Theorem 3, we find

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1} \le 2\mathbb{E}\left[\sum\nolimits_k \left(\sum\nolimits_j (a_{jk} - X_{jk})^2\right)^{1/2} + \sum\nolimits_j \left(\sum\nolimits_k (a_{jk} - X_{jk})^2\right)^{1/2}\right].$$

A simple application of Jensen's inequality gives an error bound in terms of the variances of the entries of $\boldsymbol{X}$.

Corollary 2. *Fix the matrix $\boldsymbol{A}$, and let $\boldsymbol{X}$ be a random matrix with independent entries for which $\mathbb{E}X_{jk} = a_{jk}$. Then*

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1} \le 2\left[\sum\nolimits_k \left(\sum\nolimits_j \operatorname{Var}(X_{jk})\right)^{1/2} + \sum\nolimits_j \left(\sum\nolimits_k \operatorname{Var}(X_{jk})\right)^{1/2}\right].$$

5.2. **Optimality.** The bound in Corollary 2 is optimal in the sense that there are families of matrices $\boldsymbol{A}$ and random approximants $\boldsymbol{X}$ for which $\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1}$ grows like one of the terms in the bound and dominates the other term in the bound. To show this, we construct specific examples.

Let $\boldsymbol{A}$ be a tall $m \times \sqrt{m}$ matrix of ones and choose the approximant $X_{jk} \sim 2\,\text{Bern}\left(\frac{1}{2}\right)$. With this choice, $\operatorname{Var}(X_{jk}) = 1$, so the first term in the bound is m and the second term is $m^{5/4}$. The following argument from [RV07, Sec. 4.2] establishes that $\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1}$ grows like $m^{5/4}$.

Observe that the matrix $\boldsymbol{A} - \boldsymbol{X} = [\varepsilon_{jk}]$, where ε_{jk} are i.i.d. Rademacher variables. Its $\infty \to 1$ norm is

$$\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1} = \max_{\substack{\|\boldsymbol{y}\|_\infty = 1 \\ \|\boldsymbol{x}\|_\infty = 1}} \sum\nolimits_{j,k} \varepsilon_{jk} x_j y_k.$$

Let δ_k be a sequence of i.i.d. Rachemacher variables. By the scalar Khintchine inequality,

$$\mathbb{E}_{\boldsymbol{\delta}} \sum\nolimits_j \left|\sum\nolimits_k \varepsilon_{jk}\delta_k\right| \ge \sum\nolimits_j \frac{1}{\sqrt{2}} \|\boldsymbol{\varepsilon}_{j\cdot}\|_2 = \frac{1}{\sqrt{2}} m^{5/4}.$$

The probabilistic method shows that there is a sign vector $\boldsymbol{x}$ for which

$$\sum\nolimits_j \left|\sum\nolimits_k \varepsilon_{jk} x_k\right| \ge \frac{1}{\sqrt{2}} m^{5/4}.$$

Choose the vector $\boldsymbol{y}$ with components $y_j = \operatorname{sgn}(\sum_k \varepsilon_{jk} x_k)$. Then

$$\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1} \ge \sum\nolimits_j \sum\nolimits_k \varepsilon_{jk} y_j x_k = \sum\nolimits_j \left|\sum\nolimits_k \varepsilon_{jk} x_k\right| \ge \frac{1}{\sqrt{2}} m^{5/4}.$$

This shows $\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1}$ grows like the second term in the error bound, so this term of the bound cannot be ignored.

These arguments, applied to a fat $\sqrt{n} \times n$ matrix of ones, also establish the necessity of the first term.

5.3. **Example application.** In this section we provide an example illustrating the application of Corollary 2 to matrix sparsification.

From Corollary 2 we infer that a good scheme for sparsifying a matrix $\boldsymbol{A}$ while minimizing the expected relative $\infty \to 1$ error is one which drastically increases the sparsity of $\boldsymbol{X}$ while keeping the relative error

$$\frac{\sum_k \left(\sum_j \operatorname{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \operatorname{Var}(X_{jk})\right)^{1/2}}{\|\boldsymbol{A}\|_{\infty\to 1}}$$

small. Once a sparsification scheme is chosen, the hardest part of estimating this quantity is probably estimating the $\infty \to 1$ norm of $\boldsymbol{A}$. The example shows, for a simple family of approximation schemes, what kind of sparsification results can be obtained using Corollary 2 when we have a very good handle on this quantity.

Consider the case where $\boldsymbol{A}$ is an $n \times n$ matrix whose entries all lie within an interval bounded away from zero; for definiteness, take them to be positive. Let γ be a desired bound on the expected relative $\infty \to 1$ norm error. We choose the randomization strategy $X_{jk} \sim \frac{a_{jk}}{p} \operatorname{Bern}(p)$ and ask how small can p be without violating our bound on the expected error.

In this case,

$$\|\boldsymbol{A}\|_{\infty\to 1} = \sum\nolimits_{j,k} a_{jk} = \mathrm{O}\big(n^2\big)\,,$$

and $\operatorname{Var}(X_{jk}) = \frac{a_{jk}^2}{p} - a_{jk}^2$. Consequently, the first term in Corollary 2 satisfies

$$\begin{aligned}\sum\nolimits_k \Big(\sum\nolimits_j \operatorname{Var}(X_{jk})\Big)^{1/2} &= \sum\nolimits_k \left(\frac{1}{p}\,\|\boldsymbol{a}_k\|_2^2 - \|\boldsymbol{a}_k\|_2^2\right)^{1/2} = \left(\frac{1-p}{p}\right)^{1/2} \|\boldsymbol{A}\|_{\mathrm{col}} \\ &= \mathrm{O}\left(\left(\frac{1-p}{p}\right)^{1/2} n\sqrt{n}\right)\end{aligned}$$

and likewise the second term satisfies

$$\sum\nolimits_j \Big(\sum\nolimits_k \operatorname{Var}(X_{jk})\Big)^{1/2} = \mathrm{O}\left(\left(\frac{1-p}{p}\right)^{1/2} n\sqrt{n}\right).$$

Therefore the relative $\infty \to 1$ norm error satisfies

$$\frac{\sum_k \Big(\sum_j \operatorname{Var}(X_{jk})\Big)^{1/2} + \sum_j \Big(\sum_k \operatorname{Var}(X_{jk})\Big)^{1/2}}{\|\boldsymbol{A}\|_{\infty\to 1}} = \mathrm{O}\left(\left(\frac{1-p}{pn}\right)^{1/2}\right).$$

It follows that $\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 1} < \gamma$ for p on the order of $(1 + n\gamma^2)^{-1}$ or larger. The expected number of nonzero entries in $\boldsymbol{X}$ is pn^2, so for matrices with this structure, we can sparsify with a relative $\infty \to 1$ norm error smaller than γ while reducing the number of expected nonzero entries to as few as $\mathrm{O}(\frac{n^2}{1+n\gamma^2}) = \mathrm{O}(\frac{n}{\gamma^2})$. Intuitively, this sparsification result is optimal in the dimension: it seems we must keep on average at least one entry per row and column if we are to faithfully approximate $\boldsymbol{A}$.

6. Approximation in the $\infty \to 2$ norm

In this section, we develop the $\infty \to 2$ error bound stated in the introduction, establish the optimality of a related bound, and provide examples of its application to matrix sparsification. To derive the error bound, we first specialize Theorem 1 to the case of $p = 2$.

Theorem 4. *Suppose that $\boldsymbol{Z}$ is a random matrix with independent, zero-mean entries. Then*

$$\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 2} \le 2\mathbb{E}\,\|\boldsymbol{Z}\|_{\mathrm{F}} + 2\min_{\boldsymbol{D}} \mathbb{E}\,\big\|\boldsymbol{Z}\boldsymbol{D}^{-1}\big\|_{2\to\infty}$$

where $\boldsymbol{D}$ is a positive diagonal matrix that satisfies $\operatorname{trace}(\boldsymbol{D}^2) = 1$.

Proof. Apply Theorem 1 to get

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\|_{\infty\to 2} &\le 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{z}_k\right\|_2 + 2\max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\sum\nolimits_k \left|\sum\nolimits_j \varepsilon_j Z_{jk} u_j\right| \\ &=: F + S.\end{aligned} \tag{3}$$

Expand the first term, and use Jensen's inequality to move the expectation with respect to the Rademacher variables inside the square root:

$$F = 2\mathbb{E}\left(\sum\nolimits_j \left|\sum\nolimits_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2} \le 2\mathbb{E}_{\boldsymbol{Z}}\left(\sum\nolimits_j \mathbb{E}_{\boldsymbol{\varepsilon}} \left|\sum\nolimits_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2}.$$

The independence of the Rademacher variables implies that the cross terms cancel, so

$$F \le 2\mathbb{E}\left(\sum\nolimits_j \sum\nolimits_k Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\,\|\boldsymbol{Z}\|_{\mathrm{F}}\,.$$

We use the Cauchy–Schwarz inequality to replace the ℓ_1 norm with an ℓ_2 norm in the second term of (3). A direct application would introduce a possibly suboptimal factor of $\sqrt{n}$ (where n is the number of columns in $\boldsymbol{Z}$), so instead we choose $d_k > 0$ such that $\sum_k d_k^2 = 1$ and use the corresponding weighted ℓ_2 norm:

$$S = 2\max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\sum\nolimits_k \frac{\left|\sum_j \varepsilon_j Z_{jk} u_j\right|}{d_k} d_k \le 2\max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\left(\sum\nolimits_k \frac{\left|\sum_j \varepsilon_j Z_{jk} u_j\right|^2}{d_k^2}\right)^{1/2}.$$

Move the expectation with respect to the Rademacher variables inside the square root and observe that the cross terms cancel:

$$S \le 2\max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}_{\boldsymbol{Z}}\left(\sum\nolimits_k \frac{\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum_j \varepsilon_j Z_{jk} u_j\right|^2}{d_k^2}\right)^{1/2} = 2\max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\left(\sum\nolimits_{j,k} \frac{Z_{jk}^2 u_j^2}{d_k^2}\right)^{1/2}.$$

Use Jensen's inequality to pass the maximum through the expectation, and note that if $\|\boldsymbol{u}\|_2 = 1$ then the vector formed by elementwise squaring $\boldsymbol{u}$ lies on the ℓ_1 unit ball, thus

$$S \le 2\mathbb{E}\left(\max_{\|\boldsymbol{u}\|_1=1} \sum\nolimits_{j,k} \left(\frac{Z_{jk}}{d_k}\right)^2 u_j\right)^{1/2}.$$

Clearly this maximum is achieved when $\boldsymbol{u}$ is chosen so $u_j = 1$ at an index j for which $\left(\sum_k \left(\frac{Z_{jk}}{d_k}\right)^2\right)^{1/2}$ is maximal and $u_j = 0$ otherwise. Consequently, the maximum is the largest of the ℓ_2 norms of the rows of $\boldsymbol{Z}\boldsymbol{D}^{-1}$, where $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_n)$. Recall that this quantity is, by definition, $\left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2\to\infty}$. Therefore $S \le 2\mathbb{E}\left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2\to\infty}$. The theorem follows by optimizing our choice of $\boldsymbol{D}$ and introducing our estimates for F and S into (3). □

Taking $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$ in Theorem 4, we have

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 2} \le 2\mathbb{E}\left(\sum\nolimits_{j,k} (X_{jk} - a_{jk})^2\right)^{1/2} + 2\min_{\boldsymbol{D}} \mathbb{E}\max_j \left(\sum\nolimits_k \frac{(X_{jk} - a_{jk})^2}{d_k^2}\right)^{1/2}. \tag{4}$$

We now derive a bound which depends only on the variances of the X_{jk}.

Corollary 3. *Fix the $m \times n$ matrix $\boldsymbol{A}$ and let $\boldsymbol{X}$ be a random matrix with independent entries so that $\mathbb{E}X = \boldsymbol{A}$. Then*

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|_{\infty\to 2} \le 2\left(\sum\nolimits_{j,k} \operatorname{Var}(X_{jk})\right)^{1/2} + 2\sqrt{m}\min_{\boldsymbol{D}} \max_j \left(\sum\nolimits_k \frac{\operatorname{Var}(X_{jk})}{d_k^2}\right)^{1/2}$$

where $\boldsymbol{D}$ is a positive diagonal matrix with $\operatorname{trace}(\boldsymbol{D}^2) = 1$.

Proof. Let F and S denote, respectively, the first and second term of (4). An application of Jensen's inequality shows that $F \le 2\left(\sum_{j,k} \operatorname{Var}(X_{jk})\right)^{1/2}$. A second application shows that

$$S \le 2\min_{\boldsymbol{D}} \left(\mathbb{E}\max_j \sum\nolimits_k \frac{(X_{jk} - a_{jk})^2}{d_k^2}\right)^{1/2}.$$

Bound the maximum with a sum:

$$S \leq 2 \min_{\boldsymbol{D}} \left(\sum\nolimits_j \mathbb{E} \sum\nolimits_k \frac{(X_{jk} - a_{jk})^2}{d_k^2} \right)^{1/2}.$$

The sum is controlled by a multiple of its largest term, so

$$S \leq 2\sqrt{m} \min_{\boldsymbol{D}} \left(\max_j \sum\nolimits_k \frac{\operatorname{Var}(X_{jk})}{d_k^2} \right)^{1/2},$$

where m is the number of rows of $\boldsymbol{A}$. □

6.1. **Optimality.** We now show that Theorem 4 gives an optimal bound, in the sense that each of its terms is necessary. In the following, we reserve the letter $\boldsymbol{D}$ for a positive diagonal matrix with $\operatorname{trace}(\boldsymbol{D}^2) = 1$.

First, we establish the necessity of the Frobenius term by identifying a class of random matrices whose $\infty \to 2$ norms are larger than their weighted $2 \to \infty$ norms but comparable to their Frobenius norms. Let $\boldsymbol{Z}$ be a random $m \times \sqrt{m}$ matrix such that the entries in the first column of $\boldsymbol{Z}$ are equally likely to be positive or negative ones, and all other entries are zero. With this choice, $\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to 2} = \mathbb{E} \|\boldsymbol{Z}\|_{\mathrm{F}} = \sqrt{m}$. Meanwhile, $\mathbb{E} \left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2 \to \infty} = \frac{1}{d_{11}}$, so $\min_{\boldsymbol{D}} \mathbb{E} \left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2 \to \infty} = 1$, which is much smaller than $\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to 2}$. Clearly, the Frobenius term is necessary.

Similarly, to establish the necessity of the weighted $2 \to \infty$ norm term, we consider a class of matrices whose $\infty \to 2$ norms are larger than their Frobenius norms but comparable to their weighted $2 \to \infty$ norms. Consider a $\sqrt{n} \times n$ matrix $\boldsymbol{Z}$ whose entries are all equally likely to be positive or negative ones. It is a simple task to confirm that $\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to 2} \geq n$ and $\mathbb{E} \|\boldsymbol{Z}\|_{\mathrm{F}} = n^{3/4}$; it follows that the weighted $2 \to \infty$ norm term is necessary. In fact,

$$\min_{\boldsymbol{D}} \mathbb{E} \left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2 \to \infty} = \min_{\boldsymbol{D}} \mathbb{E} \max_{j = 1, \dots, \sqrt{n}} \left(\sum\nolimits_{k=1}^{n} \frac{Z_{jk}^2}{d_{kk}^2} \right)^{1/2} = \min_{\boldsymbol{D}} \left(\sum\nolimits_{k=1}^{n} \frac{1}{d_{kk}^2} \right)^{1/2} = n,$$

so we see that $\mathbb{E} \|\boldsymbol{Z}\|_{\infty \to 2}$ and the weighted $2 \to \infty$ norm term are comparable.

6.2. **Example application.** From Theorem 4 we infer that a good scheme for sparsifying a matrix $\boldsymbol{A}$ while minimizing the expected relative $\infty \to 2$ norm error is one which drastically increases the sparsity of $\boldsymbol{X}$ while keeping the relative error

$$\frac{\mathbb{E} \|\boldsymbol{Z}\|_{\mathrm{F}} + \min_{\boldsymbol{D}} \mathbb{E} \left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2 \to \infty}}{\|\boldsymbol{A}\|_{\infty \to 2}}$$

small, where $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$.

As before, consider the case where $\boldsymbol{A}$ is an $n \times n$ matrix all of whose entries are positive and in an interval bounded away from zero. Let γ be a desired bound on the expected relative $\infty \to 2$ norm error. We choose the randomization strategy $X_{jk} \sim \frac{a_{jk}}{p} \operatorname{Bern}(p)$ and ask how much can we sparsify while respecting our bound on the relative error. That is, how small can p be? We appeal to Theorem 4. In this case,

$$\|\boldsymbol{A}\|_{\infty \to 2} = \left(\sum\nolimits_j \sum\nolimits_k a_{jk}^2 + 2 \sum\nolimits_j \sum\nolimits_{\ell < m} a_{j\ell} a_{jm} \right)^{\frac{1}{2}} = \mathrm{O}\Big(\big(n^2 + n^2(n-1) \big)^{\frac{1}{2}} \Big).$$

By Jensen's inequality,

$$\mathbb{E} \|\boldsymbol{Z}\|_{\mathrm{F}} \leq \mathbb{E} \|\boldsymbol{A}\|_{\mathrm{F}} + \mathbb{E} \|\boldsymbol{X}\|_{\mathrm{F}} \leq \left(1 + \frac{1}{\sqrt{p}} \right) \|\boldsymbol{A}\|_{\mathrm{F}} = \mathrm{O}\left(n \left(1 + \frac{1}{\sqrt{p}} \right) \right).$$

We bound the other term in the numerator, also using Jensen's inequality:

$$\min_{\boldsymbol{D}} \mathbb{E} \left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2 \to \infty} \leq \sqrt{n} \mathbb{E} \|\boldsymbol{Z}\|_{2 \to \infty} \leq \sqrt{n} \left(1 + \frac{1}{\sqrt{p}} \right) \|\boldsymbol{A}\|_{2 \to \infty} = \mathrm{O}\left(n \left(1 + \frac{1}{\sqrt{p}} \right) \right)$$

to get

$$\frac{\mathbb{E}\,\|\boldsymbol{Z}\|_{\mathrm{F}} + \min_{\boldsymbol{D}} \mathbb{E}\,\left\|\boldsymbol{Z}\boldsymbol{D}^{-1}\right\|_{2\to\infty}}{\|\boldsymbol{A}\|_{\infty\to 2}} = \mathrm{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{pn}}\right) = \mathrm{O}\left(\frac{1}{\sqrt{pn}}\right)$$

We conclude that, for this class of matrices and this family of sparsification schemes, we can reduce the number of expected nonzero terms to $\mathrm{O}\left(\frac{n}{\gamma^2}\right)$ while maintaining an expected $\infty \to 2$ norm relative error of γ.

7. Spectral error bound

In this section we establish a bound on $\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|$ as an immediate consequence of Latała's result [Lat05]. We then derive a deviation inequality for the spectral approximation error using a log-Sobolev inequality from [BLM03], and use it to compare our results to those of Achlioptas and McSherry [AM07] and Arora, Hazan, and Kale [AHK06].

Theorem 5. *Suppose $\boldsymbol{A}$ is a fixed matrix, and let $\boldsymbol{X}$ be a random matrix with independent entries for which $\mathbb{E}X = \boldsymbol{A}$. Then*

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\| \le \mathrm{C}\left[\max_j \left(\sum\nolimits_k \mathrm{Var}(X_{jk})\right)^{1/2} + \max_k \left(\sum\nolimits_j \mathrm{Var}(X_{jk})\right)^{1/2} + \left(\sum\nolimits_{jk} \mathbb{E}(X_{jk} - a_{jk})^4\right)^{1/4}\right]$$

where C *is a universal constant.*

In [Lat05], Latała considered the spectral norm of random matrices with independent, zero-mean entries, and he showed that, for any such matrix $\boldsymbol{Z}$,

$$\mathbb{E}\|\boldsymbol{Z}\| \le \mathrm{C}\left[\max_j \left(\sum\nolimits_k \mathbb{E}Z_{jk}^2\right)^{1/2} + \max_k \left(\sum\nolimits_j \mathbb{E}Z_{jk}^2\right)^{1/2} + \left(\sum\nolimits_{jk} \mathbb{E}Z_{jk}^4\right)^{1/4}\right],$$

where C is some universal constant. Unfortunately, no estimate for C is available. Theorem 5 follows from Latała's result, by taking $\boldsymbol{Z} = \boldsymbol{A} - \boldsymbol{X}$.

The bounded differences argument from Section 4 establishes the correct (subgaussian) tail behavior of $\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|$.

Theorem 6. *Fix the matrix $\boldsymbol{A}$, and let $\boldsymbol{X}$ be a random matrix with independent entries for which $\mathbb{E}X = \boldsymbol{A}$. Assume $|X_{jk}| \le D/2$ almost surely for all j, k. Then, for all $t > 0$,*

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{X}\| > \mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\| + t\right) \le \mathrm{e}^{-t^2/(4D^2)}.$$

Proof. The proof is exactly that of Theorem 2, except now $\boldsymbol{u}$ and $\boldsymbol{v}$ are both in the ℓ_2 unit sphere. □

We find it convenient to measure deviations on the scale of the mean.

Corollary 4. *Under the conditions of Theorem 6, for all $\delta > 0$,*

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{X}\| > (1+\delta)\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\|\right) \le \mathrm{e}^{-\delta^2(\mathbb{E}\|\boldsymbol{A}-\boldsymbol{X}\|)^2/(4D^2)}.$$

7.1. Comparison with previous results.

To demonstrate the applicability of our bound on the spectral norm error, we consider the sparsification and quantization schemes used by Achlioptas and McSherry [AM07], and the quantization scheme proposed by Arora, Hazan, and Kale [AHK06]. We show that our spectral norm error bound and the associated concentration result give results of the same order, with less effort. Throughout these comparisons, we take $\boldsymbol{A}$ to be a $m \times n$ matrix, with $m < n$, and we define $b = \max_{jk} |a_{jk}|$.

7.1.1. *A matrix quantization scheme.* First we consider the scheme proposed by Achlioptas and McSherry for quantization of the matrix entries:

$$X_{jk} = \begin{cases} b & \text{with probability } \frac{1}{2} + \frac{a_{jk}}{2b} \\ -b & \text{with probability } \frac{1}{2} - \frac{a_{jk}}{2b} \end{cases}.$$

With this choice $\operatorname{Var}(X_{jk}) = b^2 - a_{jk}^2 \leq b^2$, and $\mathbb{E}(X_{jk} - a_{jk})^4 = b^2 - 3a^4 + 2a^2b^2 \leq 3b^4$, so the expected spectral error satisfies

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{X}\| \leq \mathrm{C}(\sqrt{n}b + \sqrt{m}b + b\sqrt[4]{3mn}) \leq 4\mathrm{C}b\sqrt{n}.$$

Applying Corollary 4, we find that the error satisfies

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{X}\| > 4\mathrm{C}b\sqrt{n}(1+\delta)\right) \leq \mathrm{e}^{-\delta^2\mathrm{C}^2 n}.$$

In particular, with probability at least $1 - \exp(-\mathrm{C}^2 n)$,

$$\|\boldsymbol{A} - \boldsymbol{X}\| \leq 8\mathrm{C}b\sqrt{n}.$$

Achlioptas and McSherry proved that for $n \geq n_0$, where n_0 is on the order of 10^9, with probability at least $1 - \exp(-19(\log n)^4)$,

$$\|\boldsymbol{A} - \boldsymbol{X}\| < 4b\sqrt{n}.$$

Thus, Theorem 6 provides a bound of the same order in n which holds with higher probability and over a larger range of n.

7.1.2. *A nonuniform sparsification scheme.* Next we consider an analog to the nonuniform sparsification scheme proposed in the same paper. Fix a number p in the range $(0, 1)$ and sparsify entries with probabilities proportional to their magnitudes:

$$X_{jk} \sim \frac{a_{jk}}{p_{jk}} \operatorname{Bern}(p_{jk}), \text{ where } p_{jk} = \max\left\{ p\left(\frac{a_{jk}}{b}\right)^2, \sqrt{p\left(\frac{a_{jk}}{b}\right)^2 \times (8\log n)^4/n} \right\}.$$

Achlioptas and McSherry determine that, with probability at least $1 - \exp(-19(\log n)^4)$,

$$\|\boldsymbol{A} - \boldsymbol{X}\| < 4b\sqrt{n/p}.$$

Further, the expected number of nonzero entries in $\boldsymbol{X}$ is less than

$$pmn \times \operatorname{Avg}[(a_{jk}/b)^2] + m(8\log n)^4. \tag{5}$$

Their choice of p_{jk}, in particular the insertion of the $(8\log n)^4/n$ factor, is an artifact of their method of proof. Instead, we consider a scheme which compares the magnitudes of a_{jk} and b to determine p_{jk}. Introduce the quantity $R = \max_{a_{jk} \neq 0} b/|a_{jk}|$ to measure the spread of the entries in $\boldsymbol{A}$, and take

$$X_{jk} \sim \begin{cases} \frac{a_{jk}}{p_{jk}} \operatorname{Bern}(p_{jk}), \text{ where } p_{jk} = \frac{pa_{jk}^2}{pa_{jk}^2 + b^2}, & a_{jk} \neq 0 \\ 0, & a_{jk} = 0. \end{cases}$$

With this scheme, $\operatorname{Var}(X_{jk}) = 0$ when $a_{jk} = 0$, otherwise $\operatorname{Var}(X_{jk}) = b^2/p$. Likewise, $\mathbb{E}(X_{jk} - a_{jk})^4 = 0$ if $a_{jk} = 0$, otherwise

$$\mathbb{E}(X_{jk} - a_{jk})^4 \leq \operatorname{Var}(X_{jk}) \|X_{jk} - a_{jk}\|_\infty^2 = \frac{b^2}{p} \max\left\{ |a_{jk}|, |a_{jk}| \left(\frac{pa_{jk}^2 + b^2}{pa_{jk}^2} - 1 \right) \right\}^2 \leq \frac{b^4}{p^2} R^2,$$

so

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\| \leq \mathrm{C}\left(b\sqrt{\frac{n}{p}} + b\sqrt{\frac{m}{p}} + b\sqrt{\frac{R}{p}}\sqrt[4]{mn} \right) \leq \mathrm{C}(2 + \sqrt{R})b\sqrt{\frac{n}{p}}.$$

Applying Corollary 4, we find that with probability at least $1 - \exp(-\mathrm{C}^2(2+\sqrt{R})^2 pn/16)$,

$$\|\boldsymbol{A} - \boldsymbol{X}\| \leq 2\mathrm{C}(2+\sqrt{R})b\sqrt{\frac{n}{p}}.$$

Thus, Theorem 5 and Achlioptas and McSherry's scheme-specific analysis yield results of the same order in n and p. As before, we see that our bound holds with higher probability and over a larger range of n. Furthermore, since the expected number of nonzero entries in $\boldsymbol{X}$ satisfies

$$\sum\nolimits_{jk} p_{jk} = \sum\nolimits_{jk} \frac{pa_{jk}^2}{pa_{jk}^2 + b^2} \leq pnm \times \operatorname{Avg}\left[\left(\frac{a_{jk}}{b}\right)^2\right],$$

we have established a smaller limit on the expected number of nonzero entries.

7.1.3. *A scheme which simultaneously sparsifies and quantizes.* Finally, we use Theorem 6 to estimate the error in using the scheme from [AHK06] which simultaneously quantizes and sparsifies. Fix $\delta > 0$ and consider

$$X_{jk} = \begin{cases} \operatorname{sgn}(a_{jk})\frac{\delta}{\sqrt{n}} \operatorname{Bern}\left(\frac{|a_{jk}|\sqrt{n}}{\delta}\right), & |a_{jk}| \leq \frac{\delta}{\sqrt{n}} \\ a_{jk}, & \text{otherwise.} \end{cases}$$

Then $\operatorname{Var}(X_{jk}) = 0$ if $|a_{jk}| \geq \delta/\sqrt{n}$, otherwise

$$\operatorname{Var}(X_{jk}) = |a_{jk}|^3\frac{\sqrt{n}}{\delta} - 2a_{jk}^2 + |a_{jk}|\frac{\delta}{\sqrt{n}} \leq \frac{\delta^2}{n}.$$

Also the fourth moment term is zero for large enough a_{jk}, otherwise

$$\mathbb{E}(X_{jk} - a_{jk})^4 = |a_{jk}|^5\frac{\sqrt{n}}{\delta} - 4a_{jk}^4 + 6|a_{jk}|^3\frac{\delta}{\sqrt{n}} - 4a_{jk}^2\frac{\delta^2}{n} + |a_{jk}|\left(\frac{\delta}{\sqrt{n}}\right)^3 \leq 8\frac{\delta^4}{n^2}.$$

This gives the estimates

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{X}\| \leq C\left(\sqrt{n}\frac{\delta}{\sqrt{n}} + \sqrt{m}\frac{\delta}{\sqrt{n}} + 2\frac{\delta}{\sqrt{n}}\sqrt[4]{mn}\right) \leq 4C\delta$$

and

$$\mathbb{P}\left(\|\boldsymbol{A} - \boldsymbol{X}\| > 4C\delta(\gamma+1)\right) \leq \mathrm{e}^{-\gamma^2 C^2 n}.$$

Taking $\gamma = 1$, we see that with probability at least $1 - \exp(-C^2 n)$,

$$\|\boldsymbol{A} - \boldsymbol{X}\| \leq 8C\delta.$$

Let $S = \sum_{j,k} |A_{jk}|$, then appealing to Lemma 1 in [AHK06], we find that $\boldsymbol{X}$ has $\mathrm{O}\left(\frac{\sqrt{n}S}{\gamma}\right)$ nonzero entries with probability at least $1 - \exp\left(-\Omega\left(\frac{\sqrt{n}S}{\gamma}\right)\right)$.

Arora, Hazan, and Kale establish that this scheme guarantees $\|\boldsymbol{A} - \boldsymbol{X}\| \leq \mathrm{O}(\delta)$ with probability at least $1 - \exp(-\Omega(n))$, so we see that our general bound recovers a bound of the same order.

In conclusion, we see that the bound on expected spectral error in Theorem 6 in conjunction with the deviation result in Corollary 4 provide guarantees comparable to those derived with scheme-specific analyses. We anticipate that the flexibility demonstrated here will make these useful tools for analyzing and guiding the design of novel sparsification schemes.

References

[AHK06] Sanjeev Arora, Elad Hazan, and Satyen Kale, *A Fast Random Sampling Algorithm for Sparsifying Matrices*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer Berlin, 2006, pp. 272–279.

[AM01] Dimitris Achlioptas and Frank McSherry, *Fast Computation of Low Rank Matrix Approximations*, Proceedings of the 33rd annual ACM symposium on Theory of Computing, 2001, pp. 611–618.

[AM07] ———, *Fast Computation of Low Rank Matrix Approximations*, Journal of the ACM **54** (2007), no. 2.

[AN04] Noga Alon and Assaf Naor, *Approximating the Cut-Norm via Grothendieck's inequality*, Proceedings of the 36th Annual ACM symposium on Theory of Computing, 2004, pp. 72–80.

[BLM03] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities using the entropy method*, The Annals of Probability **31** (2003), 1583–1614.

[BSS09] Joshua Batson, Daniel Spielman, and Nikhil Srivastava, *Twice-Ramanujan Sparsifiers*, Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009, pp. 255–262.

[DKM06a] Petros Drineas, Ravi Kannan, and Michael W. Mahoney, *Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication*, SIAM Journal of Computing **36** (2006), no. 1, 132–157.

[DKM06b] ———, *Fast Monte Carlo Algorithms for Matrices II: Computing Low-Rank Approximations to a Matrix*, SIAM Journal of Computing **36** (2006), no. 1, 158–183.

[DKM06c] ———, *Fast Monte Carlo Algorithms for Matrices III: Computing an Efficient Approximate Decomposition of a Matrix*, SIAM Journal of Computing **36** (2006), no. 1, 184–206.

[FKV98] Alan Frieze, Ravi Kannan, and Santosh Vempala, *Fast Monte-Carlo Algorithms for finding low-rank approximations*, Proceedings of the 39th Annual Symposium on Foundations of Computer Science, 1998, pp. 378–390.

[FKV04] ———, *Fast monte-carlo algorithms for finding low-rank approximations*, Journal of the ACM **51** (2004), 1025–1041.

[Kar94a] David R. Karger, *Random sampling in cut, flow, and network design problems*, Proceedings of the 26th Annual ACM Symposium on Theory of Computing, May 1994, pp. 648–657.

[Kar94b] ———, *Using randomized sparsification to approximate minimum cuts*, Proceedings of the 5th Annual ACM-SIAM Symposium on Discrete Algorithms, January 1994, pp. 424–432.

[Kar95] ———, *Random Sampling in Graph Optimization Problems*, Ph.D. thesis, Stanford University, 1995.

[Kar96] ———, *Approximating s–t minimum cuts in $\tilde{O}(n^2)$ time*, Proceedings of the 28th Annual ACM Symposium on Theory of Computing, May 1996, pp. 47–55.

[Lat05] Rafal Latała, *Some estimates of norms of random matrices*, Proceedings of the American Mathematical Society **133** (2005), 1273–1282.

[LT91] Michel Ledoux and Michel Talagrand, *Probability in Banach Spaces*, Springer-Verlag, 1991.

[Roh00] J. Rohn, *Computing the Norm $\|A\|_{\infty\to 1}$ is NP-Hard*, Linear and Multilinear Algebra **47** (2000), 195–204.

[RV07] Mark Rudelson and Roman Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, Journal of the ACM **54** (2007), no. 4.

[Seg00] Yoav Seginer, *The Expected Norm of Random Matrices*, Combinatorics, Probability and Computing **9** (2000), no. 2, 149–166.

[SS08] Daniel A. Spielman and Nikhil Srivastava, *Graph sparsification by effective resistances*, Proceedings of the 40th annual ACM symposium on Theory of computing, 2008, pp. 563–568.

[Sza76] S. J. Szarek, *On the best constants in the Khintchin-inequality*, Studia Math **58** (1976), 197–208.

[Tro09] Joel Tropp, *Column subset selection, matrix factorization, and eigenvalue optimization*, Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 978–986.

[vW96] Aad W. van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, 1996.