Chapter 3

Randomized sparsification in NP-hard norms

Massive matrices 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. Here is one way to frame this challenge mathematically: Given a matrix $\mathbf{A}$, how can one efficiently generate a sparse matrix $\mathbf{X}$ for which the approximation error $\|\mathbf{A}-\mathbf{X}\|$ is small?

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. Sometimes it is more appropriate to consider the matrix as an operator from a finite-dimensional ℓ_p space to a finite-dimensional ℓ_q space, and investigate the behavior of the approximation error in the associated $p \to q$ operator norm. As an example, the problem of graph sparsification is naturally posed as a question of preserving the so-called *cut norm* of a matrix associated with the graph.

The strong equivalency of the cut norm and the $\infty\rightarrow1$ norm suggests that, for graph-theoretic applications, it may be fruitful to consider the behavior of the $\infty\rightarrow1$ norm under sparsification. In other applications, e.g., the column subset selection algorithm in [Tro09], the $\infty\rightarrow2$ norm is the norm of interest.

This chapter investigates the errors incurred by approximating a fixed real matrix with a random matrix[1]. 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 $\infty\rightarrow p$ norm error, which we specialize to the case of the $\infty\rightarrow1$ and $\infty\rightarrow2$ norms. We also use a result of Latała [Lat05] to bound the expected spectral approximation error, and we establish the subgaussianity of the spectral approximation error.

Our methods are similar to those of Rudelson and Vershynin in [RV07] in that we treat $\mathbf{A}$ 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 $\mathbf{A}$, we consider the behavior of the norms of matrices formed by randomly sparsifying (or quantizing) the entries of $\mathbf{A}$. 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. In particular, the algorithm of [BSS09] is not suitable for sparsifying graphs with a large number of vertices. Part of our motivation for investigating the $\infty\rightarrow1$ approximation error is the belief that the equivalence of the cut norm with the $\infty\rightarrow1$ norm means that matrix sparsification in the $\infty\rightarrow1$ norm might be useful for efficiently constructing optimal sparsifiers for such graphs.

[1] The content of this chapter is adapted from the technical report [GT11] co-authored with Joel Tropp.

3.1 Notation

We establish the notation particular to this chapter.

All quantities are real. For $1 \le p \le \infty$, the ℓ_p^n norm of $\mathbf{x} \in \mathbb{R}^n$ is written as $\|\mathbf{x}\|_p$. Each space ℓ_p^n has an associated dual space $\ell_{p'}^n$, where p', the *conjugate exponent* to p, is determined by the relation $p^{-1} + (p')^{-1} = 1$. The dual space of ℓ_1^n (respectively, ℓ_∞^n) is ℓ_∞^n (respectively, ℓ_1^n).

The kth column of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{(k)}$, and the (j,k)th element is denoted by a_{jk}. We treat $\mathbf{A}$ as an operator from ℓ_p^n to ℓ_q^m, and the $p \to q$ operator norm of $\mathbf{A}$ is written as $\|\mathbf{A}\|_{p\to q}$. The spectral norm, i.e. the $2 \to 2$ operator norm, is written $\|\mathbf{A}\|_2$. Recall that given an operator $\mathbf{A} : \ell_p^n \to \ell_q^m$, the associated adjoint operator ($\mathbf{A}^T$, in the case of a matrix) maps from $\ell_{q'}^m$ to $\ell_{p'}^n$. Further, the $p \to q$ and $q' \to p'$ norms are dual in the sense that

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

This chapter is concerned primarily with the spectral norm and the $\infty\to1$ and $\infty\to2$ norms. The $\infty\to1$ and $\infty\to2$ norms are not unitarily invariant, so do not have simple interpretations in terms of singular values; in fact, they are NP-hard to compute for general matrices [Roh00]. We remark that $\|\mathbf{A}\|_{\infty\to1} = \|\mathbf{A}\mathbf{x}\|_1$ and $\|\mathbf{A}\|_{\infty\to2} = \left\|\mathbf{A}\mathbf{y}\right\|_2$ for certain vectors $\mathbf{x}$ and $\mathbf{y}$ whose components take values ±1. An additional operator norm, the $2\to\infty$ norm, is of interest: it is the largest ℓ_2 norm achieved by a row of $\mathbf{A}$. In the sequel we also encounter the column norm

$$\|\mathbf{A}\|_{\text{col}} = \sum\nolimits_k \|\mathbf{A}_{(k)}\|_2.$$

The variance of X is written $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)^2$. The expectation taken with respect to one variable X, with all others fixed, is written $\mathbb{E}_X$. The 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 with mean p.

3.1.0.1 Graph sparsification

Graphs are often represented and fruitfully manipulated in terms of matrices, so the problems of graph sparsification and matrix sparsification are strongly related. We now introduce the relevant notation before surveying the literature.

Let $G = (V, E, \omega)$ be a weighted simple undirected graph with n vertices, m edges, and adjacency matrix $\mathbf{A}$ given by

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

Orient the edges of G in an arbitrary manner. Then define the corresponding $2m \times n$ *oriented incidence matrix* $\mathbf{B}$ in the following manner: $b_{2i-1,j} = b_{2i,k} = \omega_i$ and $b_{2i-1,k} = b_{2i,j} = -\omega_i$ if edge i is oriented from vertex j to vertex k, and all other entries of B are identically zero.

A *cut* is a partition of the vertices of G 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 that correspond to the *cut-norm* of the oriented incidence matrix $\mathbf{B}$, which is defined as

$$\|\mathbf{B}\|_{\mathrm{C}} = \max_{I\subset\{1,\ldots,2m\}, J\subset\{1,\ldots,n\}} \left|\sum\nolimits_{i\in I, j\in J} b_{ij}\right|.$$

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 $\tilde{\mathbf{B}}$ of $\mathbf{B}$ for which $|\sum_{jk}\tilde{b}_{jk}| \geq 0.56\,\|\mathbf{B}\|_{\mathrm{C}}$. This algorithm thereby gives a feasible means of approximating the MAXCUT value for arbitrary graphs. A crucial point in the derivation of the algorithm is the fact that for general matrices the $\infty\rightarrow 1$ operator norm is strongly equivalent with the cut-norm:

$$\|\mathbf{A}\|_{\mathrm{C}} \leq \|\mathbf{A}\|_{\infty\rightarrow 1} \leq 4\,\|\mathbf{A}\|_{\mathrm{C}}\,;$$

in fact, in the particular case of oriented incidence matrices, $\|\mathbf{B}\|_{\mathrm{C}} = \|\mathbf{B}\|_{\infty\rightarrow 1}$.

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 $\mathbf{\Pi}_G$ and $\mathbf{\Pi}_{G'}$ associated with the original graph and the sparsifier, and then 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 $\Omega(n)$ edges. The existence proof is constructive and provides a deterministic algorithm for constructing such sparsifiers in $\mathrm{O}(n^3 m)$ time, where m is the number of edges in the original graph.

3.2 Preliminaries

Bounded differences inequalities are useful tools for establishing measure concentration for functions of independent random variables that are insensitive to changes in a single argument. In this chapter, we use a bounded differences inequality to show that the norms of the random matrices that we encounter exhibit measure concentration.

Before stating the inequality of interest to us, 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.

Lemma 3.1 ([BLM03, Corollary 3]). *Let W and $\{W_i\}$ be random variables defined as above. Assume that there exists a positive number C such that, almost surely,*

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

Then, for all $t > 0$,

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

Rademacher random variables take the values ±1 with equal probability. Rademacher vectors are vectors of i.i.d. Rademacher random variables. We make use of Rademacher variables in this chapter to simplify our analyses through the technique of Rademacher symmetrization. Essentially, given a random variable Z, Rademacher symmetrization allows us to estimate the behavior of Z in terms of that of the random variable $Z_{\text{sym}} = \varepsilon(Z - Z')$, where Z' is an i.i.d. copy of Z and ε is a Rademacher random variable that is independent of the pair (Z, Z'). The variable Z_{sym} is often easier to manipulate than Z, since it is guaranteed to be symmetric (i.e., Z_{sym} and $-Z_{\text{sym}}$ are identically distributed); in particular, $\mathbb{E}Z_{\text{sym}} = 0$. The following basic symmetrization result is drawn from [vW96, Lemma 2.3.1 et seq.].

Lemma 3.2. *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* $\mathscr{F}$ *be a family of functions such that*

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

is measurable. Then

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

Since we work with finite-dimensional probability models and linear functions, measurability concerns can be ignored in our applications of Lemma 3.2.

While Lemma 3.2 allows us to replace certain random processes with Rademacher processes, Talagrand's Rademacher comparison theorem [LT91, Theorem 4.12 et seq.] shows that certain complicated Rademacher processes are bounded by simpler Rademacher processes. Together, these two results often allow us to reduce the analysis of complicated random processes to the analysis of simpler Rademacher processes.

Lemma 3.3. *Fix finite-dimensional vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$ and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Then*

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

Lemma 3.3 involves Rademacher sums, i.e. sums of the form $\sum_k \varepsilon_k x_k$ where $\boldsymbol{\varepsilon}$ is a Rademacher vector and $\mathbf{x}$ is a fixed vector. One of the most basic tools for understanding Rademacher sums is the Khintchine inequality [Sza76], which 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 $\mathbf{x}$.

Lemma 3.4 (Khintchine Inequality)**.** *Let $\mathbf{x}$ be a real vector, and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Then*

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

In its more general form, which we do not use in this thesis, the Khintchine inequality implies that Rademacher sums are subguassian random variables.

3.3 The $\infty \to p$ norm of a random matrix

We are interested in schemes that approximate a given matrix $\mathbf{A}$ by means of a random matrix $\mathbf{X}$ in such a way that the entries of $\mathbf{X}$ are independent and $\mathbb{E}\mathbf{X} = \mathbf{A}$. It follows that the error matrix $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ has independent, zero-mean entries. Ultimately we aim to construct $\mathbf{X}$ so that it has the property that, with high probability, many of its entries are identically zero, 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 $\mathbf{Z}$.

3.3.1 The expected $\infty \to p$ norm

The main tools used to derive the bound on the expected norm of $\mathbf{Z}$ are Lemma 3.2, a result on the Rademacher symmetrization of random processes, and Lemma 3.3, Talagrand's Rademacher comparison theorem.

We now state and prove the bound on the expected norm of $\mathbf{Z}$.

Theorem 3.5. *Let* $\mathbf{Z}$ *be a random matrix with independent, zero-mean entries and let* $\boldsymbol{\varepsilon}$ *be a Rademacher vector independent of* $\mathbf{Z}$. *Then*

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

where q *is the conjugate exponent of* p.

Proof of Theorem 3.5. By duality,

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

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

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to p} &\le \mathbb{E}\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k(|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| - \mathbb{E}'|\langle \mathbf{Z}'_{(k)}, \mathbf{u}\rangle|) + \max_{\|\mathbf{u}\|_q=1}\mathbb{E}\sum\nolimits_k |\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| \\ &= F + S.\end{aligned} \tag{3.3.1}$$

Begin with the first term on the right-hand side of (3.3.1). Use Jensen's inequality to draw the expectation outside of the maximum:

$$F \le \mathbb{E}\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k(|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| - |\langle \mathbf{Z}'_{(k)}, \mathbf{u}\rangle|).$$

Now apply Lemma 3.2 to symmetrize the random variable:

$$F \le \mathbb{E}\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k \varepsilon_k(|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| - |\langle \mathbf{Z}'_{(k)}, \mathbf{u}\rangle|).$$

By the subadditivity of the maximum,

$$F \le \mathbb{E}\left(\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k \varepsilon_k|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| + \max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k -\varepsilon_k|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle|\right) = 2\mathbb{E}\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k \varepsilon_k|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle|,$$

where we have invoked the fact that $-\varepsilon_k$ has the Rademacher distribution. Apply Lemma 3.3 to get the final estimate of F:

$$F \le 2\mathbb{E}\max_{\|\mathbf{u}\|_q=1}\sum\nolimits_k \varepsilon_k\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle = 2\mathbb{E}\max_{\|\mathbf{u}\|_q=1}\left\langle \sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}, \mathbf{u}\right\rangle = 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}\right\|_p.$$

Now consider the last term on the right-hand side of (3.3.1). Use Jensen's inequality to prepare for symmetrization:

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

Apply Lemma 3.2 to the expectation of the inner sum to see

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

The triangle inequality gives us the final expression:

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

Introduce the bounds for F and S into (3.3.1) to complete the proof. □

3.3.2 A tail bound for the $\infty \to p$ norm

We now develop a deviation bound for the $\infty \to p$ approximation error. The argument is based on Lemma 3.1, a bounded differences inequality.

To apply Lemma 3.1, we let $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ be our error matrix, $W = \|\mathbf{Z}\|_{\infty \to p}$, and $W^{jk} = \left\|\mathbf{Z}^{jk}\right\|_{\infty \to p}$, where $\mathbf{Z}^{jk}$ 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 Lemma 3.1 gives us a useful deviation bound.

Theorem 3.6. *Fix an $m \times n$ matrix $\mathbf{A}$, and let $\mathbf{X}$ be a random matrix with independent entries for* *which* $\mathbb{E}X = \mathbf{A}$*. Assume* $\left|X_{jk}\right| \leq \frac{D}{2}$ *almost surely for all* j,k*. Then, for all* $t > 0$*,*

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

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 $\mathbf{u}, \mathbf{v}$ such that $W = \mathbf{u}^T\mathbf{Z}\mathbf{v}$ and $\|\mathbf{u}\|_q = 1$ and $\|\mathbf{v}\|_\infty = 1$. Then

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

This implies

$$\sum\nolimits_{j,k}(W - W^{jk})^2\mathbb{1}_{W>W^{jk}} \leq D^2\sum\nolimits_{j,k}|u_jv_k|^2 \leq nD^2\|\mathbf{u}\|_2^2,$$

so we can apply Lemma 3.1 if we have an estimate for $\|\mathbf{u}\|_2^2$. We have the bounds $\|\mathbf{u}\|_2 \leq \|\mathbf{u}\|_q$ for $q \in [1,2]$ and $\|\mathbf{u}\|_2 \leq m^{1/2-1/q}\|\mathbf{u}\|_q$ for $q \in [2,\infty]$. Therefore,

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

It follows from Lemma 3.1 that

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

where $s = \max\left\{0, 1-2/q\right\}$. □

It is often convenient to measure deviations on the scale of the mean. Taking $t = \delta\mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p}$ in Theorem 3.6 gives the following result.

Corollary 3.7. *Under the conditions of Theorem 3.6, for all $\delta > 0$,*

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

3.4 Approximation in the $\infty\to 1$ norm

In this section, we develop the $\infty\to 1$ error bound as a consequence of Theorem 3.5. We then prove that one form of the error bound is optimal, and we describe an example of its application to matrix sparsification.

3.4.1 The expected $\infty\to 1$ norm

To derive the $\infty\to 1$ error bound, we first apply Theorem 3.5 with $p = 1$.

Theorem 3.8. *Suppose that $\mathbf{Z}$ is a random matrix with independent, zero-mean entries. Then*

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

Proof. Apply Theorem 3.5 to get

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} &\leq 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}\right\|_1 + 2\max_{\|\mathbf{u}\|_\infty=1}\mathbb{E}\sum\nolimits_k\left|\sum\nolimits_j \varepsilon_j Z_{jk}u_j\right| \\ &= F + S.\end{aligned} \tag{3.4.1}$$

Use Hölder's inequality to bound the first term in (3.4.1) with a sum of squares:

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

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

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

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:

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

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

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

Introduce these estimates for F and S into (3.4.1) to complete the proof. □

Taking $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ in Theorem 3.8, we find

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

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

Corollary 3.9. *Fix the matrix* $\mathbf{A}$, *and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}X_{jk} = a_{jk}$. *Then*

$$\mathbb{E}\left\|\mathbf{A}-\mathbf{X}\right\|_{\infty\to1} \leq 2\left[\sum\nolimits_k \left(\sum\nolimits_j \mathrm{Var}(X_{jk})\right)^{1/2} + \sum\nolimits_j \left(\sum\nolimits_k \mathrm{Var}(X_{jk})\right)^{1/2}\right].$$

3.4.2 Optimality

A simple estimate using the Khintchine inequality shows that the bound on the expected value of the ∞→1 norm given in Theorem 3.8 is in fact optimal up to constants.

Corollary 3.10. *Suppose that* $\mathbf{Z}$ *is a random matrix with independent, zero-mean entries. Then*

$$\frac{1}{2\sqrt{2}}\mathbb{E}(\|\mathbf{Z}\|_{\mathrm{col}} + \left\|\mathbf{Z}^T\right\|_{\mathrm{col}}) \leq \mathbb{E}\left\|\mathbf{Z}\right\|_{\infty\to1} \leq 2\mathbb{E}(\|\mathbf{Z}\|_{\mathrm{col}} + \left\|\mathbf{Z}^T\right\|_{\mathrm{col}}).$$

Proof. First we establish the inequality

$$\|\mathbf{Z}\|_{\mathrm{col}} \leq \sqrt{2}\,\|\mathbf{Z}\|_{\infty\to1} \tag{3.4.2}$$

as a consequence of the Khintchine inequality, Lemma 3.4. Indeed, since

$$\|\mathbf{Z}\|_{\mathrm{col}} = \sum\nolimits_j \left\|\mathbf{z}_{(j)}\right\|_2$$

and the Khintchine inequality gives the estimate

$$\left\|\mathbf{Z}_{(j)}\right\|_2 \leq \sqrt{2}\mathbb{E}\left|\sum\nolimits_i \varepsilon_i Z_{ij}\right|,$$

we see that

$$\begin{aligned}\|\mathbf{Z}\|_{\text{col}} &\leq \sqrt{2}\mathbb{E}\sum\nolimits_j\left|\sum\nolimits_i \varepsilon_i Z_{ij}\right| \\ &= \sqrt{2}\mathbb{E}\left\|\mathbf{Z}^T\varepsilon\right\|_1 \leq \sqrt{2}\sup_{\|\mathbf{x}\|_\infty=1}\left\|\mathbf{Z}^T\mathbf{x}\right\|_1 \\ &= \sqrt{2}\left\|\mathbf{Z}^T\right\|_{\infty\to1} = \sqrt{2}\|\mathbf{Z}\|_{\infty\to1}.\end{aligned}$$

Since the $\infty\to1$ norms of $\mathbf{Z}$ and $\mathbf{Z}^T$ are equal, it also follows that

$$\left\|\mathbf{Z}^T\right\|_{\text{col}} \leq \sqrt{2}\left\|\mathbf{Z}^T\right\|_{\infty\to1} = \sqrt{2}\|\mathbf{Z}\|_{\infty\to1}.$$

The lower bound on $\mathbb{E}\|\mathbf{Z}\|_{\infty\to1}$ is now a consequence of (3.4.2),

$$\frac{1}{2\sqrt{2}}\mathbb{E}(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}) \leq \mathbb{E}\|\mathbf{Z}\|_{\infty\to1},$$

while the upper bound is given by Theorem 3.8. □

Remark 3.11. Using standard arguments, one can establish that the deterministic bounds

$$\frac{1}{2\sqrt{2}}\left(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}\right) \leq \|\mathbf{Z}\|_{\infty\to1} \leq \frac{\sqrt{n}}{2}\left(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}\right)$$

hold for *any* square $n \times n$ matrix $\mathbf{Z}$. Corollary 3.10 is a refinement of this equivalence relation that holds when $\mathbf{Z}$ is a random, zero-mean matrix. In particular, the corollary tells us that when we assume this model for **Z**, the equivalence relation does not depend on the dimensions of **Z**, and thus if what we care about is the expected $\infty \to 1$ norm of **Z**, we can work with the expected column norm of **Z** without losing any sharpness.

3.4.3 An example application

In this section we provide an example illustrating the application of Corollary 3.9 to matrix sparsification.

From Corollary 3.9 we infer that a good scheme for sparsifying a matrix **A** while minimizing the expected relative $\infty \to 1$ error is one which drastically increases the sparsity of **X** while keeping the relative error

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

small. Once a sparsification scheme is chosen, the hardest part of estimating this quantity is probably estimating the $\infty \to 1$ norm of **A**. The example shows, for a simple family of approximation schemes, what kind of sparsification results can be obtained using Corollary 3.9 when we have a very good handle on this quantity.

Consider the case where **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 $X_{jk} \sim \frac{a_{jk}}{p}\,\mathrm{Bern}(p)$ and ask how small can p be without violating our bound on the expected error.

In this case,

$$\|\mathbf{A}\|_{\infty\to 1} = \sum_{j,k} a_{jk} = \mathrm{O}(n^2),$$

and $\mathrm{Var}(X_{jk}) = \frac{a_{jk}^2}{p} - a_{jk}^2$. Consequently, the first term in Corollary 3.9 satisfies

$$\sum\nolimits_k \left(\sum\nolimits_j \mathrm{Var}(X_{jk})\right)^{1/2} = \sum\nolimits_k \left(\frac{1}{p}\left\|\mathbf{a}_k\right\|_2^2 - \left\|\mathbf{a}_k\right\|_2^2\right)^{1/2} = \left(\frac{1-p}{p}\right)^{1/2} \|\mathbf{A}\|_{\mathrm{col}}$$
$$= \mathrm{O}\left(\left(\frac{1-p}{p}\right)^{1/2} n\sqrt{n}\right)$$

and likewise the second term satisfies

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

Therefore the relative ∞→1 norm error satisfies

$$\frac{\sum_k \left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2}}{\|\mathbf{A}\|_{\infty\to1}} = \mathrm{O}\left(\left(\frac{1-p}{pn}\right)^{1/2}\right).$$

It follows that $\mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to1} < \gamma$ for p on the order of $(1+n\gamma^2)^{-1}$ or larger. The expected number of nonzero entries in $\mathbf{X}$ is pn^2, so for matrices with this structure, we can sparsify with a relative ∞→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 $\mathbf{A}$.

3.5 Approximation in the ∞→2 norm

In this section, we develop the ∞→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 3.5 to the case of $p = 2$.

Theorem 3.12. *Suppose that* $\mathbf{Z}$ *is a random matrix with independent, zero-mean entries. Then*

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

where $\mathbf{D}$ *is a positive diagonal matrix that satisfies* $\mathrm{Tr}(\mathbf{D}^2) = 1$.

Proof. Apply Theorem 3.5 to get

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} &\leq 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}\right\|_2 + 2\max_{\|\mathbf{u}\|_2=1}\mathbb{E}\sum\nolimits_k\left|\sum\nolimits_j \varepsilon_j Z_{jk} u_j\right| \\ &=: F + S.\end{aligned} \tag{3.5.1}$$

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

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

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

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

We use the Cauchy–Schwarz inequality to replace the ℓ_1 norm with an ℓ_2 norm in the second term of (3.5.1). A direct application would introduce a possibly suboptimal factor of $\sqrt{n}$ (where n is the number of columns in $\mathbf{Z}$), so instead we choose $d_k > 0$ such that $\sum_k d_k^2 = 1$ and use the corresponding weighted ℓ_2 norm:

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

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

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

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

$$S \leq 2\mathbb{E} \left(\max_{\|\mathbf{u}\|_1=1} \sum\nolimits_j u_j \cdot \left(\sum_k (Z_{jk}/d_k)^2 \right) \right)^{1/2}.$$

Clearly this maximum is achieved when $\mathbf{u}$ is chosen so $u_j = 1$ at an index j for which $\sum_k (Z_{jk}/d_k)^2$ is maximal and $u_j = 0$ otherwise. Consequently, the maximum is the largest of the ℓ_2 norms of the rows of $\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. Recall that this quantity is, by definition, $\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty}$. Therefore $S \leq 2\mathbb{E}\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty}$. The theorem follows by optimizing our choice of $\mathbf{D}$ and introducing our estimates for F and S into (3.5.1). □

Taking $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ in Theorem 3.12, we have

$$\mathbb{E}\,\|\mathbf{A} - \mathbf{X}\|_{\infty\to 2} \leq 2\mathbb{E} \left(\sum\nolimits_{j,k} (X_{jk} - a_{jk})^2 \right)^{1/2} + 2 \min_{\mathbf{D}} \mathbb{E} \max_j \left(\sum\nolimits_k \frac{(X_{jk} - a_{jk})^2}{d_k^2} \right)^{1/2}. \quad (3.5.2)$$

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

Corollary 3.13. *Fix the $m \times n$ matrix $\mathbf{A}$ and let $\mathbf{X}$ be a random matrix with independent entries so that $\mathbb{E}X = \mathbf{A}$. Then*

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

where $\mathbf{D}$ *is a positive diagonal matrix with* $\mathrm{Tr}(\mathbf{D}^2) = 1$.

Proof. Let F and S denote, respectively, the first and second term of (3.5.2). An application of Jensen's inequality shows that $F \le 2\left(\sum_{j,k} \mathrm{Var}(X_{jk})\right)^{1/2}$. A second application shows that

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

Bound the maximum with a sum:

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

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

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

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

3.5.1 Optimality

We now show that Theorem 3.12 gives an optimal bound, in the sense that each of its terms is necessary. In the following, we reserve the letter $\mathbf{D}$ for a positive diagonal matrix with $\mathrm{Tr}(\mathbf{D}^2) = 1$.

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

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

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

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

3.5.2 An example application

From Theorem 3.12 we infer that a good scheme for sparsifying a matrix $\mathbf{A}$ while minimizing the expected relative ∞→2 norm error is one which drastically increases the sparsity of $\mathbf{X}$ while keeping the relative error

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

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

As before, consider the case where $\mathbf{A}$ 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 ∞→2 norm error. We choose the randomization strategy $X_{jk} \sim \frac{a_{jk}}{p}\,\mathrm{Bern}(p)$ 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 3.12. In this case,

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

By Jensen's inequality,

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

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

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

to get

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

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.

3.6 A spectral error bound

In this section we establish a bound on $\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|$ 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 3.14. *Suppose* $\mathbf{A}$ *is a fixed matrix, and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}\mathbf{X} = \mathbf{A}$*. Then*

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

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 $\mathbf{Z}$,

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

where C is some universal constant. Unfortunately, no estimate for C is available. Theorem 3.14 follows from Latała's result, by taking $\mathbf{Z} = \mathbf{A} - \mathbf{X}$.

The bounded differences argument from Section 3.3 establishes the correct (subgaussian) tail behavior of $\mathbb{E}\|\mathbf{A} - \mathbf{X}\|$.

Theorem 3.15. *Fix the matrix* $\mathbf{A}$*, and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}X = \mathbf{A}$*. Assume* $|X_{jk}| \leq D/2$ *almost surely for all* j, k*. Then, for all* $t > 0$*,*

$$\mathbb{P}\{\|\mathbf{A} - \mathbf{X}\| > \mathbb{E}\|\mathbf{A} - \mathbf{X}\| + t\} \leq \mathrm{e}^{-t^2/(4D^2)}.$$

Proof. The proof is exactly that of Theorem 3.6, except now $\mathbf{u}$ and $\mathbf{v}$ are both in the ℓ_2 unit sphere. □

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

Corollary 3.16. *Under the conditions of Theorem 3.15, for all* $\delta > 0$*,*

$$\mathbb{P}\{\|\mathbf{A} - \mathbf{X}\| > (1+\delta)\mathbb{E}\|\mathbf{A} - \mathbf{X}\|\} \leq \mathrm{e}^{-\delta^2(\mathbb{E}\|\mathbf{A}-\mathbf{X}\|)^2/(4D^2)}.$$

3.6.1 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 **A** to be a $m \times n$ matrix, with $m < n$, and we define $b = \max_{jk} |a_{jk}|$.

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

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

Applying Corollary 3.16, we find that the error satisfies

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

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

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

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

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

Thus, Theorem 3.15 provides a bound of the same order in n which holds with higher probability and over a larger range of n.

3.6.1.2 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:

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

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

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

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

$$pmn \times \operatorname{Avg}[(a_{jk}/b)^2] + m(8\log n)^4, \tag{3.6.1}$$

where the notation Avg$(\cdot)$ indicates the average of a quantity over all the entries of $\mathbf{A}$.

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 $\mathbf{A}$, and take

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

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

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

so

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

Applying Corollary 3.16, we find that the error satisfies

$$\mathbb{P}\left\{ \|\mathbf{A} - \mathbf{X}\| > \mathrm{C}(2 + \sqrt{R}) b \sqrt{\frac{n}{p}} (\epsilon + 1) \right\} \leq e^{-\epsilon^2 \mathrm{C}^2 (2+\sqrt{R})^2 pn/16},$$

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

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

Thus, Theorem 3.14 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 $\mathbf{X}$ 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.

3.6.1.3 A scheme which simultaneously sparsifies and quantizes

Finally, we use Theorem 3.15 to estimate the error of 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

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

The fourth moment term is zero when $|a_{jk}| \geq \delta/\sqrt{n}$, and when $|a_{jk}| < \delta/\sqrt{n}$,

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

This gives the estimates

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

and

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

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

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

Let $S = \sum_{j,k} |A_{jk}|$, then appealing to Lemma 1 in [AHK06], we find that $\mathbf{X}$ 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 $\|\mathbf{A}-\mathbf{X}\| = \mathrm{O}(\delta)$ with probability at least $1 - \exp(-\Omega(n))$, so we see that our general bound recovers a bound of the same order.

3.7 Comparison with later bounds

The papers [NDT10, DZ11, AKL13], written after the results in this chapter were obtained, present alternative schemes for sparsification and quantization.

The scheme presented in [NDT10] sparsifies a matrix by zeroing out all sufficiently small entries of $\mathbf{A}$, keeping all sufficiently large entries, and randomly sampling the remaining entries of the matrix with a probability depending on their magnitudes. More precisely, given a parameter $s > 0$, it generates an approximation whose entries are distributed as

$$X_{jk} = \begin{cases} 0, & a_{jk}^2 \leq (\log^2(n)/n)\|\mathbf{A}\|_{\mathrm{F}}^2/s \\ a_{jk} & a_{jk}^2 \geq \|\mathbf{A}\|_{\mathrm{F}}^2/s \\ (a_{jk}/p_{jk})\mathrm{Bern}(p_{jk}), & \text{otherwise, where } p_{jk} = s a_{jk}^2/\|\mathbf{A}\|_{\mathrm{F}}^2. \end{cases}$$

The analysis offered guarantees that if $s = \Omega(\epsilon^{-2} n \log^3 n)$, then with probability at least $1 - n^{-1}$, $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$ and, in expectation, $\mathbf{X}$ has less than $2s$ nonzero entries. It is not clear whether or not this scheme can be analyzed using Theorem 3.14. It is straightforward to establish that $\mathrm{Var}(X_{jk}) \leq \epsilon^2/(n \log^3 n)$ for this scheme, but obtaining a sufficiently small upper bound on the fourth moment $\mathbb{E}(X_{jk} - a_{jk})^4$ is challenging. In particular, the estimate

$$\mathbb{E}(X_{jk} - a_{jk})^4 \leq \mathrm{Var}(X_{jk}) \|X_{jk} - a_{jk}\|_\infty$$

gives an upper bound on the order of $\epsilon a_{jk}^2 n / \log^5 n$, which is sufficient only to establish a much weaker guarantee on the error $\mathbb{E}\|\mathbf{A} - \mathbf{X}\|_2$ than the guarantee given in [NDT10].

The scheme introduced in [DZ11] first zeroes out all entries of $\mathbf{A} \in \mathbb{R}^{n \times n}$ of sufficiently small magnitude, then samples elements from $\mathbf{A}$ in s i.i.d. trials with replacement. The elements are selected with probabilities proportional to their squared magnitudes. Thus, the approximant can be written in the form

$$\mathbf{X} = \frac{1}{s} \sum\nolimits_{t=1}^{s} \frac{a_{j_t k_t}}{p_{j_t k_t}} \boldsymbol{e}_{j_t} \boldsymbol{e}_{k_t}^T,$$

where (j_t, k_t) is the index of the element of $\mathbf{A}$ selected in the tth trial, $p_{jk} = a_{jk}^2 / \|\mathbf{A}\|_\mathrm{F}^2$ is the probability that the entry a_{jk} is selected, and $\boldsymbol{e}_j$ denotes the jth standard basis vector in n. Clearly $\mathbf{X}$ has at most s nonzero entries. Let $s = \Omega(\epsilon^{-2} n \log(n) \|\mathbf{A}\|_\mathrm{F}^2)$. Then the authors show that, with probability at least $1 - n^{-1}$, the error of the approximation satisfies $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$. This scheme is not easily analyzable using our Theorem 3.14. Since the approximant $\mathbf{X}$ is a sum of rank-one matrices, it is most natural to analyze its approximation error using tail bounds for sums of independent random matrices. Indeed, the authors of [DZ11] use a matrix Bernstein inequality to provide their results.

Finally, the scheme presented in [AKL13] computes an approximation of the same form as the scheme introduced in [DZ11], but samples entries of $\mathbf{A}$ with probabilities proportional their absolute values. That is,

$$\mathbf{X} = \frac{1}{s}\sum\nolimits_{t=1}^{s} \frac{a_{j_t k_t}}{p_{j_t k_t}} \boldsymbol{e}_{j_t} \boldsymbol{e}_{k_t}^T,$$

where $p_{jk} = |a_{jk}| / \sum_{pq} |a_{pq}|$. Again, this scheme is not amenable to analysis using Theorem 3.14. Recall that $\mathbf{A}^{(k)}$ denotes the kth row of $\mathbf{A}$. The authors establish that, when

$$s = \Omega\left(\epsilon^{-2}\log(n/\delta)\left(\sum\nolimits_{jk}|A_{jk}|\right)\max_k \|\mathbf{A}^{(k)}\|_1\right),$$

the error bound $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$ is satisfied with probability at least $1 - \delta$. The approximant $\mathbf{X}$ has, in expectation, at most $2s$ nonzero entries.

Comparing the extents to which we were able to reproduce the guarantees of the sparsification schemes introduced in [AM01, AHK06, AM07, NDT10, DZ11, AKL13], we see that Theorem 3.14 sometimes can recover competitive guarantees on the approximation errors of element-wise sparsification schemes in which X_{jk} is directly related to a_{jk} through a simple expression. When $\mathbf{X}$ is more naturally represented as a sum of rank-1 matrices, Theorem 3.14 is not easily applicable.