Randomized Algorithms for Low-Rank Matrix Factorizations: Sharp Performance Bounds

Rafi Witten* and Emmanuel Candès†

August 2013; Revised April 2014

Abstract

The development of randomized algorithms for numerical linear algebra, e.g. for computing approximate QR and SVD factorizations, has recently become an intense area of research. This paper studies one of the most frequently discussed algorithms in the literature for dimensionality reduction—specifically for approximating an input matrix with a low-rank element. We introduce a novel and rather intuitive analysis of the algorithm in [6], which allows us to derive sharp estimates and give new insights about its performance. This analysis yields theoretical guarantees about the approximation error and at the same time, ultimate limits of performance (lower bounds) showing that our upper bounds are tight. Numerical experiments complement our study and show the tightness of our predictions compared with empirical observations.

1 Introduction

Almost any method one can think of in data analysis and scientific computing relies on matrix algorithms. In the era of 'big data', we must now routinely deal with matrices of enormous sizes and reliable algorithmic solutions for computing solutions to least-squares problems, for computing approximate QR and SVD factorizations and other such fundamental decompositions are urgently needed. Fortunately, the development of randomized algorithms for numerical linear algebra has seen a new surge in recent years and we begin to see computational tools of a probabilistic nature with the potential to address some of the great challenges posed by big data. In this paper, we study one of the most frequently discussed algorithms in the literature for dimensionality reduction, and provide a novel analysis which gives sharp performance bounds.

1.1 Approximate low-rank matrix factorization

We are concerned with the fundamental problem of approximately factorizing an arbitrary $m \times n$ matrix A as

$$\underset{m \times n}{A} \quad \approx \quad \underset{m \times \ell}{B} \quad \underset{\ell \times n}{C} \tag{1.1}$$

where $\ell \leq \min(m, n) = m \wedge n$ is the desired rank. The goal is to compute B and C such that $A - BC$ is as small as possible. Typically, one measures the quality of the approximation by taking either

*Bit Body, Inc., Cambridge, MA

†Departments of Mathematics and of Statistics, Stanford University, Stanford CA

the spectral norm $\|\cdot\|$ (the largest singular value, also known as the 2 norm) or the Frobenius norm $\|\cdot\|_F$ (the root-sum of squares of the singular values) of the residual $A - BC$. It is well-known that the best rank-ℓ approximation, measured either in the spectral or Frobenius norm, is obtained by truncating the singular value decomposition (SVD), but this can be prohibitively expensive when dealing with large matrix dimensions.

Recent work [6] introduced a randomized algorithm for matrix factorization with lower computational complexity. This algorithm is the same as a version presented by Sarlós in [9, Section 4].

Algorithm 1 Randomized algorithm for matrix approximation

Require: Input: $m \times n$ matrix A and desired rank ℓ.
Sample an $n \times \ell$ test matrix G with independent mean-zero, unit-variance Gaussian entries.
Compute $H = AG$.
Construct $Q \in \mathbb{R}^{m\times\ell}$ with columns forming an orthonormal basis for the range of H.
return the approximation $B = Q$, $C = Q^*A$.

The algorithm is simple to understand: $H = AG$ is an approximation of the range of A; we therefore project the columns of A onto this approximate range by means of the orthogonal projector QQ^* and hope that $A \approx BC = QQ^*A$. Of natural interest is the accuracy of this procedure: how large is the size of the residual? Specifically, how large is $\|(I - QQ^*)A\|$?

The subject of the beautiful survey [3] as well as [6] is to study this problem and provide an analysis of performance. Before we state the sharpest results known to date, we first recall that if $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m\wedge n}$ are the ordered singular values of A, then the best rank-ℓ approximation obeys

$$\min\{\|A - B\| : \mathrm{rank}(B) \leq \ell\} = \sigma_{\ell+1}.$$

It is known that there are choices of A such that $\mathbb{E}\|(I-QQ^*)A\|$ is greater than $\sigma_{\ell+1}$ by an arbitrary multiplicative factor, see e.g. [3]. For example, setting

$$A = \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix}$$

and $\ell = 1$, direct computation shows that $\lim_{t\to\infty} \mathbb{E}\|(I - QQ^*)A\| = \infty$. Thus, we write $\ell = k + p$ (where $p > 0$) and seek $\underline{b}$ and $\overline{b}$ such that

$$\underline{b}(m, n, k, p) \leq \sup_{A\in\mathbb{R}^{m\times n}} \mathbb{E}\|(I - QQ^*)A\|/\sigma_{k+1} \leq \overline{b}(m, n, k, p).$$

We are now ready to state the best results concerning the performance of Algorithm 1 we are aware of.

Theorem 1.1 ([3])**.** *Let A be an $m \times n$ matrix and run Algorithm 1 with $\ell = k + p$, then*

$$\mathbb{E}\|(I - QQ^*)A\| \leq \left[1 + \frac{4\sqrt{k+p}}{p-1}\sqrt{m \wedge n}\right] \sigma_{k+1}. \tag{1.2}$$

It is a priori unclear whether this upper bound correctly predicts the expected behavior or not. That is to say, is the dependence upon the problem parameters in the right-hand side of the right order of magnitude? Is the bound tight or can it be substantially improved? Are there lower bounds which would provide ultimate limits of performance? The aim of this paper is merely to provide some definite answers to such questions.

1.2 Sharp bounds

It is convenient to write the residual in Algorithm 1 as

$$f(A, G) := (I - QQ^*)A \tag{1.3}$$

as to make the dependence on the random test matrix G explicit. Our main result states that there is an explicit random variable whose size completely determines the accuracy of the algorithm. This statement uses a natural notion of stochastic ordering; below we write $X \overset{d}{\geq} Y$ if and only if the random variables X and Y obey $\mathbb{P}(X \geq t) \geq \mathbb{P}(Y \geq t)$ for all $t \in \mathbb{R}$.

Theorem 1.2. *Suppose without loss of generality that $m \geq n$. Then in the setup of Theorem 1.1, for each matrix $A \in \mathbb{R}^{m\times n}$,*

$$\|(I - QQ^*)A\| \overset{d}{\leq} \sigma_{k+1}\, W,$$

where W is the random variable

$$W = \|f(I_{n-k}, X_2) \begin{bmatrix} X_1\Sigma^{-1} & I_{n-k} \end{bmatrix}\|; \tag{1.4}$$

here, X_1 and X_2 are respectively $(n-k)\times k$ and $(n-k)\times p$ matrices with i.i.d. $\mathcal{N}(0,1)$ entries, Σ is a $k\times k$ diagonal matrix with the singular values of a $(k+p)\times k$ Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and I_{n-k} is the $(n-k)$-dimensional identity matrix. Furthermore, X_1, X_2 and Σ are all independent (and independent from G). In the other direction, for any $\epsilon > 0$, there is a matrix A with the property

$$\|(I - QQ^*)A\| \overset{d}{\geq} (1-\epsilon)\sigma_{k+1}\, W.$$

In particular, this gives

$$\sup_A\ \mathbb{E}\|(I - QQ^*)A\|/\sigma_{k+1} = \mathbb{E}W.$$

The proof of the theorem is in Section 2.3. To obtain very concrete bounds from this theorem, one can imagine using Monte Carlo simulations by sampling from W to estimate the worst error the algorithm commits. Alternatively, one could derive upper and lower bounds about W by analytic means. The corollary below is established in Section 2.4.

Corollary 1.3 (Nonasymptotic bounds). *With W as in (1.4) (recall $n \leq m$),*

$$\sqrt{n - (k+p+2)}\,\mathbb{E}\|\Sigma^{-1}\| \leq \mathbb{E}W \leq 1 + \left(\sqrt{n-k} + \sqrt{k}\right)\,\mathbb{E}\|\Sigma^{-1}\|. \tag{1.5}$$

In the regime of interest where n is very large and $k + p \ll n$, the ratio between the upper and lower bound is practically equal to 1 so our analysis is very tight. Furthermore, Corollary 1.3 clearly emphasizes why we would want to take $p > 0$. Indeed, when $p = 0$, Σ is square and nearly singular so that both $\mathbb{E}\|\Sigma^{-1}\|$ and the lower bound become very large. In contrast, increasing p yields a sharp decrease in $\mathbb{E}\|\Sigma^{-1}\|$ and, thus, improved performance.

It is further possible to derive explicit bounds by noting (Lemma A.2) that

$$\frac{1}{\sqrt{p+1}} \leq \mathbb{E}\|\Sigma^{-1}\| \leq e\frac{\sqrt{k+p}}{p}. \tag{1.6}$$

Plugging the right inequality into (1.5) improves upon (1.2) from [3]. In the regime where $k+p \ll n$ (we assume throughout this section that $n \leq m$), taking $p = k$, for instance, yields an upper bound roughly equal to $e\sqrt{2n/k} \approx 3.84\sqrt{n/k}$ and a lower bound roughly equal to $\sqrt{n/k}$, see Figure 1.

When k and p are reasonably large, it is well-known (see Lemma A.3) that

$$\sigma_{\min}(\Sigma) \approx \sqrt{k+p} - \sqrt{k}$$

so that in the regime of interest where $k + p \ll n$, both the lower and upper bounds in (1.5) are about equal to

$$\frac{\sqrt{n}}{\sqrt{k+p} - \sqrt{k}}. \tag{1.7}$$

We can formalize this as follows: in the limit of large dimensions where $n \to \infty$, $k, p \to \infty$ with $p/k \to \rho > 0$ (in such a way that $\limsup\ (k+p)/n < 1$), we have almost surely

$$\limsup\ \frac{W}{\overline{b}(n,k,p)} \leq 1, \qquad \overline{b}(n,k,p) = \frac{\sqrt{n-k} + \sqrt{k}}{\sqrt{k+p} - \sqrt{k}}. \tag{1.8}$$

Conversely, it holds almost surely that

$$\liminf\ \frac{W}{\underline{b}(n,k,p)} \geq 1, \qquad \underline{b}(n,k,p) = \frac{\sqrt{n-k-p}}{\sqrt{k+p} - \sqrt{k}}. \tag{1.9}$$

A short justification of this limit behavior may also be found in Section 2.4.

1.3 Innovations

Whereas the analysis in [3] uses sophisticated concepts and tools from matrix analysis and from perturbation analysis, our method is different and only uses elementary ideas (for instance, it should be understandable by an undergraduate student with no special training). In a nutshell, the authors in [3] control the error of Algorithm 1 by establishing an upper bound about $\|f(A,X)\|$ holding for all matrices X (the bound depends on X). From this, they deduce bounds about $\|f(A,G)\|$ in expectation and in probability by integrating with respect to G. A limitation of this approach is that it does not provide any estimate of how close the upper bound is to being tight.

In contrast, we perform a sequence of reductions, which ultimately identifies the worst-case input matrix. The crux of this reduction is a monotonicity property, which roughly says that if the spectrum of a matrix A is larger than that of another matrix B, then the singular values of the residual $f(A,G)$ are stochastically greater than those of $f(B,G)$, see Lemma 2.1 in Section 2.1 for details. Hence, applying the algorithm to A results in a larger error than when the algorithm is applied to B. In turn, this monotonicity property allows us to write the worst-case residual in a very concrete form. With this representation, we can recover the deterministic bound from [3] and immediately see the extent to which it is sub-optimal. Most importantly, our analysis admits matching lower and upper bounds as discussed earlier.

Our analysis of Algorithm 1, presented in Section 2.3, shows that the approximation error is heavily affected by the spectrum of the matrix A past its first $k+1$ singular values.[1] In fact, suppose $m \geq n$ and let D_{n-k} be the diagonal matrix of dimension $n-k$ equal to $\mathrm{diag}(\sigma_{k+1}, \sigma_{k+2}, \ldots, \sigma_n)$.

[1] To accommodate this, previous works also provide bounds in terms of the singular values of A past σ_{k+1}.

Then our method show that the worst case error for matrices with this tail spectrum is equal to the random variable

$$W(D_{n-k}) = \|f(D_{n-k}, X_2) \begin{bmatrix} X_1\Sigma^{-1} & I_{n-k} \end{bmatrix}\|.$$

In turn, a very short argument gives the expected upper bound below:

Theorem 1.4. *Take the setup of Theorem 1.1 and let σ_i be the ith singular value of A. Then*

$$\mathbb{E}\|(I - QQ^*)A\| \leq \Big(1 + \sqrt{\frac{k}{p-1}}\Big)\sigma_{k+1} + \mathbb{E}\|\Sigma^{-1}\| \sqrt{\sum_{i>k} \sigma_i^2}. \tag{1.10}$$

Substituting $\mathbb{E}\|\Sigma^{-1}\|$ with the upper bound in (1.6) *recovers Theorem 10.6 from [3].*

This bound is tight in the sense that setting $\sigma_{k+1} = \sigma_{k+2} = \ldots = \sigma_n = 1$ essentially yields the upper bound from Corollary 1.3, which as we have seen, cannot be improved.

Similarly, we can derive bounds in the Frobenius norm, in the style of Sarlós [9, Theorem 14].

Theorem 1.5. *Take the setup of Theorem 1.1 and let σ_i be the ith singular value of A. Then*

$$\mathbb{E}\|(I - QQ^*)A\|_F^2 \leq \left(1 + \frac{k}{p-1}\right) \sum_{i>k} \sigma_i^2. \tag{1.11}$$

Therefore, there is a constant approximation error in the Frobenius sense as k and p grow in proportion. A proof of this inequality is in Section 2.6. Again, this bound cannot really be improved.

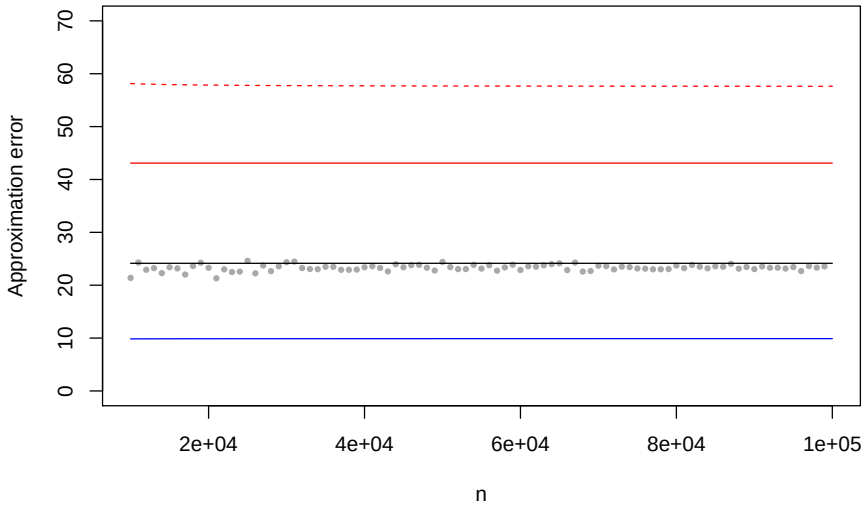
1.4 Experimental results

To examine the tightness of our analysis of performance, we apply Algorithm 1 to the 'worst-case' input matrix and compute the spectral norm of the residual, performing such computations for fixed values of m, n, k and p. We wish to compare the sampled errors with our deterministic upper and lower bounds, as well as with the previous upper bound from Theorem 1.1 and our error proxy (1.7). Because of Lemma 2.2, the worst-case behavior of the algorithm does not depend on m and n separately but only on $\min(m, n)$. Hence, we set $m = n$ in this section.

Figure 1 reveals that the new upper and lower bounds are tight up to a small multiplicative factor, and that the previous upper bound is also fairly tight. Further, the plots also demonstrate the effect of concentration in measure—the outcomes of different samples each lie remarkably close to the yellow rule of thumb, especially for larger n, suggesting that for practical purposes the algorithm is deterministic. Hence, these experimental results reinforce the practical accuracy of the error proxy (1.7) in the regime $k + p \ll n$ since we can see that the worst-case error is just about (1.7).

Figure 2 gives us a sense of the variability of the algorithm for two fixed values of the triple (n, k, p). As expected, as k and p grow the variability of the algorithm decreases, demonstrating the effect of concentration of measure.

(a) $k = p = 0.01\,n$



(b) $k = p = 100$

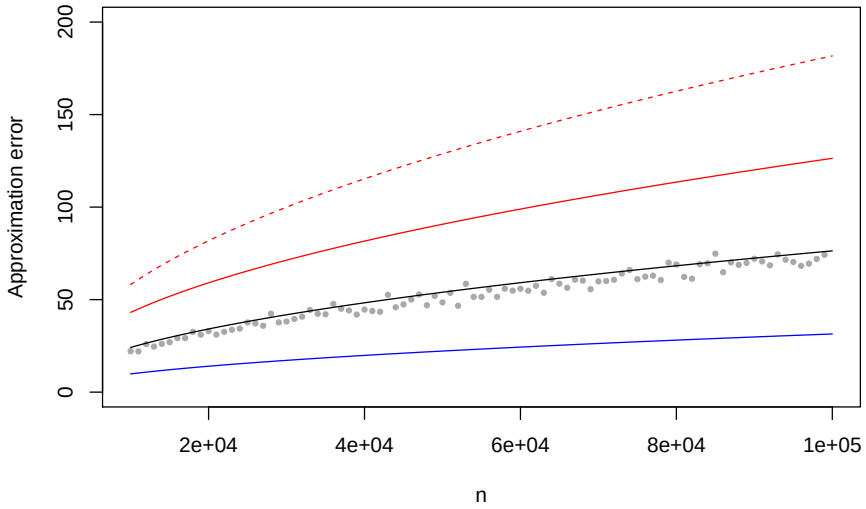


Figure 1: *Spectral norm* $\|(I - QQ^*)A\|$ *of the residual with worst-case matrix of dimension* $n \times n$ *as input with* n *varying between* 10^4 *and* 10^5*. Each grey dot represents the error of one run of Algorithm 1. The lines are bounds on the spectral norm: the red dashed line plots the previous upper bound* (1.2)*. The red (resp. blue) solid line is the upper (resp. lower) bound combining Corollary 1.3 and* (1.6)*. The black line is the error proxy* (1.7)*. In the top plot,* $k = p = 0.01\,n$*. Keeping fixed ratios results in constant error. Holding* k *and* p *fixed in the bottom plot while increasing* n *results in approximations of increasing error.*

1.5 Complements

In order to make the paper a little more self-contained, we briefly review some of the literature as to explain the centrality of the range finding problem (Algorithm 1). For instance, suppose we

(a) $m = n = 10^5, k = p = 10^2$ **(b)** $m = n = 10^5, k = p = 10^3$

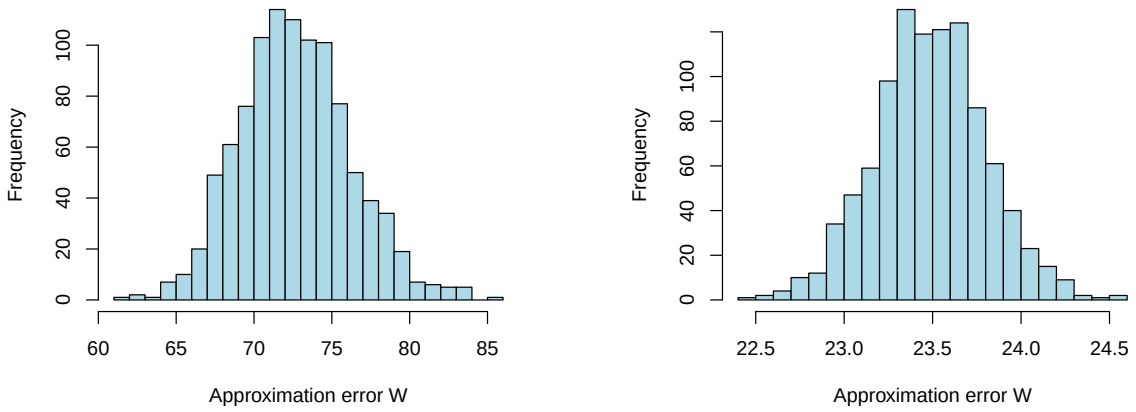


Figure 2: *We fix n (recall that $m = n$), k and p and plot the approximation error W across* 1000 *independent runs of the algorithm. For the larger value of k and p, we see reduced variability—the approximation errors range from about* 61 *to* 85 *for $k = p = 10^2$ and* 22.5 *to* 24.5 *for $k = p = 10^3$, much less variability in both absolute and percentage terms. Similarly, the empirical standard deviation with $k = p = 10^2$ is approximately* 3.6 *while with $k = p = 10^3$ it falls to* .32.

wish to construct an approximate SVD of a very large matrix $m \times n$ matrix A. Then this can be achieved by running Algorithm 1 and then computing the SVD of the 'small' matrix $C = Q^*A$, check Algorithm 2 below, which returns an approximate SVD $A \approx U\Sigma V^*$. Assuming the extra steps in Algorithm 2 are exact, we have

$$\|A - U\Sigma V^*\| = \|A - Q\hat{U}\Sigma V^*\| = \|A - QC\| = \|A - QQ^*A\| \tag{1.12}$$

so that the approximation error is that of Algorithm 1 whose study is the subject of this paper.

Algorithm 2 Randomized algorithm for approximate SVD computation

Require: Input: $m \times n$ matrix A and desired rank ℓ.
Run Algorithm 1.
Compute $C = Q^*A$.
Compute the SVD of $C = \hat{U}\Sigma V^*$.
return the approximation $U = Q\hat{U}$, Σ, V.

The point here of course is that since the matrix $C = Q^*A$ is $\ell \times n$—we typically have $\ell \ll n$—the computational cost of forming its SVD is on the order of $O(\ell^2 n)$ flops and fairly minimal. (For reference, we note that there is an even more effective single-pass variant of Algorithm 2 in which we do not need to re-access the input matrix A once Q is available, please see [3] and references therein for details.)

Naturally, we may wish to compute other types of approximate matrix factorizations of A such as eigenvalue decompositions, QR factorizations, interpolative decompositions (where one searches

for an approximation $A \approx BC$ in which B is a subset of the columns of A), and so on. All such computations would follow the same pattern: (1) apply Algorithm 1 to find an approximate range, and (2) perform classical matrix factorizations on a matrix of reduced size. This general strategy, namely, approximation followed by standard matrix computations, is of course hardly anything new. To be sure, the classical Businger-Golub algorithms for computing partial QR decompositions follows this pattern. Again, we refer the interested reader to the survey [3].

We have seen that when the input matrix A does not have a rapidly decaying spectrum, as this may be the case in a number of data analysis applications, the error of approximation Algorithm 1 commits—the random variable W in Theorem 1.2—may be quite large. In fact, when the singular values hardly decay at all, it typically is on the order of the error proxy (1.7). This results in poor performance. On the other hand, when the singular values decay rapidly, we have seen that the algorithm is provably accurate, compare Theorem 1.4. This suggests using a power iteration, similar to the block power method, or the subspace iteration in numerical linear algebra: Algorithm 3 was proposed in [7].

Algorithm 3 Randomized algorithm with power trick for matrix approximation

Require: Input: $m \times n$ matrix A and desired rank ℓ.
Sample an $n \times \ell$ test matrix G with independent mean-zero, unit-variance Gaussian entries.
Compute $H = (AA^*)^q AG$.
Construct $Q \in \mathbb{R}^{m \times \ell}$ with columns forming an orthonormal basis for the range of H.
return the approximation $B = Q$, $C = Q^*A$.

The idea in Algorithm 3 is of course to turn a slowly decaying spectrum into a rapidly decaying one at the cost of more computations: we need $2q + 1$ matrix-matrix multiplies instead of just one. The benefit is improved accuracy. Letting P be any orthogonal projector, then a sort of Jensen inequality states that for any matrix A,

$$\|PA\| \leq \|P(AA^*)^q A\|^{1/(2q+1)},$$

see [3] for a proof. Therefore, if Q is computed via the power trick (Algorithm 3), then

$$\|(I - QQ^*)A\| \leq \|(I - QQ^*)(AA^*)^q A\|^{1/(2q+1)}.$$

This immediately gives a corollary to Theorem 1.2.

Corollary 1.6. *Let A and W be as in Theorem 1.2. Applying Algorithm 3 yields*

$$\|(I - QQ^*)A\| \overset{d}{\leq} \sigma_{k+1} W^{1/(2q+1)}. \tag{1.13}$$

It is amusing to note that with, say, $n = 10^9$, $k = p = 200$, and the error proxy (1.7) for W, the size of the error factor $W^{1/(2q+1)}$ in (1.13) is about 3.41 when $q = 3$.

Further, we would like to note that our analysis exhibits a sequence of matrices (2.1) that have approximation errors that limit to the worst case approximation error when $q = 0$. However, when $q = 1$, this same sequence of matrices limits to having an approximation error exactly equal to 1, which is the best possible since this is the error achieved by truncating the SVD. It would be interesting to study the tightness of the upper bound (1.13) and we leave this to future research.

1.6 Notation

In Section 3, we shall see that among all possible test matrices, Gaussian inputs are in some sense optimal. Otherwise, the rest of the paper is mainly devoted to proving the novel results we have just presented. Before we do this, however, we pause to introduce some notation that shall be used throughout. We reserve I_n to denote the $n \times n$ identity matrix.

We use the partial ordering of n-dimensional vectors and write $x \geq y$ if $x - y$ has nonnegative entries. Similarly, we use the semidefinite (partial) ordering and write $A \succeq B$ if $A - B$ is positive semidefinite. We also introduce a notion of stochastic ordering in $\mathbb{R}^n$: given two n-dimensional random vectors z_1 and z_2, we say that $z_1 \overset{d}{\geq} z_2$ if $\mathbb{P}(z_1 \geq x) \geq \mathbb{P}(z_2 \geq x)$ for all $x \in \mathbb{R}^n$. If instead $\mathbb{P}(z_1 \geq x) = \mathbb{P}(z_2 \geq x)$ then we say that z_1 and z_2 are equal in distribution and write $z_1 \overset{d}{=} z_2$. A function $g : \mathbb{R}^n \to \mathbb{R}$ is monotone non-decreasing if $x \geq y$ implies $g(x) \geq g(y)$. Note that if g is monotone non-decreasing and $z_1 \overset{d}{\geq} z_2$, then $\mathbb{E}g(z_1) \geq \mathbb{E}g(z_2)$.

2 Proofs

2.1 Monotonicity

The key insight underlying our analysis is this:

Lemma 2.1 (**Monotonicity**). *Let $A, B \in \mathbb{R}^{m \times n}$ be fixed matrices, $G \in \mathbb{R}^{n \times \ell}$ be a Gaussian test matrix and $f(A, G)$ and $f(B, G)$ be the residuals as defined in* (1.3)*. If $\sigma(B) \leq \sigma(A)$, then*

$$\sigma\big(f(B, G)\big) \overset{d}{\leq} \sigma\big(f(A, G)\big).$$

The implications of this lemma are two-fold. First, to derive error estimates, we can work with diagonal matrices without any loss of generality. Second and more importantly, it follows from the monotonicity property that

$$\sup_{A \in \mathbb{R}^{m \times n}} \mathbb{E}\|f(A, G)\|/\sigma_{k+1} = \lim_{t \to \infty} \mathbb{E}\|f(M(t), G)\|, \quad M(t) = \begin{bmatrix} tI_k & 0 \\ 0 & I_{n-k} \end{bmatrix}. \tag{2.1}$$

The rest of this section is thus organized as follows:

- The proof of the monotonicity lemma is in Section 2.2.
- To prove Theorem 1.2, it suffices to consider a matrix input as $M(t)$ in (2.1) with $t \to \infty$. This is the object of Section 2.3.
- Bounds on the worst case error (the proof of Corollary 1.3) are given in Section 2.4.
- The mixed norm Theorem 1.4 is proved in Section 2.5.

Additional supporting materials are found in the Appendix.

2.2 Proof of the monotonicity property

We prove the monotonicity property in two steps. First, we establish an invariance property with respect to multiplication by unitary matrices. This step, which uses the distributional assumption about the test matrix, allows to reduce matters to diagonal inputs. Second, we show that the monotinicity property holds for diagonal matrices regardless of the test matrix that is applied.

Lemma 2.2 (Rotational invariance). *Let $A \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ be arbitrary matrices with U and V orthogonal. Then if $G \in \mathbb{R}^{n\times \ell}$ is a Gaussian test matrix,*

$$f(UAV, G) \stackrel{d}{=} Uf(A, G)V.$$

In particular, if $A = U\Sigma V^*$ is a singular value decomposition of A, then $\sigma(f(A, G)) \stackrel{d}{=} \sigma(f(\Sigma, G))$.

Proof. Letting P_C be the orthogonal projector onto the column space of C, we have $P_{UAVG} = UP_{AVG}U^*$. Hence,

$$f(UAV, G) = UAV - P_{UAVG}UAV = UAV - UP_{AVG}AV = U(I_m - P_{AVG})V.$$

The claim follows from the fact that $VG \stackrel{d}{=} G$ so that $I_m - P_{AVG} \stackrel{d}{=} I_m - P_{AG}$. □

We now establish a *deterministic* monotonicity property for pairs of diagonal matrices obeying $\Sigma_1 \succeq \Sigma_2$.

Lemma 2.3. *Let Σ_1 and Σ_2 be $n \times n$ diagonal and positive semidefinite matrices obeying $\Sigma_1 \succeq \Sigma_2$. Let $G \in \mathbb{R}^{n\times\ell}$ be an arbitrary test matrix. Then for all x,*

$$\|f(\Sigma_1, G)x\|_2 \geq \|f(\Sigma_2, G)x\|_2.$$

This implies that $\sigma\left(f(\Sigma_1, G)\right) \geq \sigma\left(f(\Sigma_2, G)\right)$. Therefore, if G is a random variable,

$$\sigma\left(f(\Sigma_1, G)\right) \stackrel{d}{\geq} \sigma\left(f(\Sigma_2, G)\right).$$

Proof. We first prove the property in the case where $\ell = 1$ so that $g := G \in \mathbb{R}^n$. Introduce

$$\begin{array}{rccl} h: & \mathbb{R}^n_+ & \to & \mathbb{R} \\ & \sigma & \mapsto & \|f(\mathrm{diag}(\sigma), g)x\|_2^2 \end{array}$$

in which $\mathrm{diag}(\sigma)$ is the diagonal matrix with σ_i on the diagonal. Clearly, it suffices to show that $\partial_i h(\sigma) \geq 0$ to prove the first claim. Putting $\Sigma = \mathrm{diag}(\sigma)$, $h(\sigma)$ is given by

$$h(\sigma) = x^*\left(\Sigma\left(I - \frac{(\Sigma g)(\Sigma g)^*}{\|\Sigma g\|_2^2}\right)\Sigma\right)x = \sum_k \sigma_k^2 x_k^2 - \frac{\left(\sum_{k=1}^n \sigma_k^2 g_k x_k\right)^2}{\sum_{k=1}^n \sigma_k^2 g_k^2}.$$

Taking derivatives gives

$$\frac{\partial}{\partial\sigma_i}h(\sigma) = 2\sigma_i x_i^2 - 4\sigma_i x_i t_i + 2\sigma_i t_i^2, \qquad t_i = g_i\frac{\sum_{k=1}^n \sigma_k^2 g_k x_k}{\sum_{k=1}^n \sigma_k^2 g_k^2}.$$

Hence, $\partial_i h(\sigma) = 2\sigma_i(x_i - t_i)^2 \geq 0$.

We now consider the case where $\ell \geq 1$. Fix $x \in \mathbb{R}^n$ and put $z = \Sigma_1 x$. The vector z uniquely decomposes as the sum of $z^\perp$ and $z^\parallel$, where $z^\parallel$ belongs to the range of $\Sigma_1 G$ and $z^\perp$ is the orthogonal component. Now, since $z^\parallel$ is in the range of $\Sigma_1 G$, there exists g such that $\Sigma_1 g = z^\parallel$. We have

$$f(\Sigma_1, G)x = z^\perp = f(\Sigma_1, g)x.$$

Since we know that $\|f(\Sigma_1, g)x\|_2 \geq \|f(\Sigma_2, g)x\|_2$, it follows

$$\|f(\Sigma_1, G)x\|_2 = \|f(\Sigma_1, g)x\|_2 \geq \|f(\Sigma_2, g)x\|_2 \geq \|f(\Sigma_2, G)x\|_2.$$

The last inequality holds because g is in the range of G. In details, let $P_{\Sigma_2 G}$ (resp. $P_{\Sigma_2 g}$) be the orthogonal projector onto the range of $\Sigma_2 G$ (resp. $\Sigma_2 g$). Then Pythagoras' theorem gives

$$\begin{aligned}\|f(\Sigma_2, g)x\|_2^2 = \|(I - P_{\Sigma_2 g})\Sigma_2 x\|_2^2 &= \|(I - P_{\Sigma_2 G})\Sigma_2 x\|_2^2 + \|(P_{\Sigma_2 G} - P_{\Sigma_2 g})\Sigma_2 x\|_2^2 \\ &= \|f(\Sigma_2, G)x\|_2^2 + \|(P_{\Sigma_2 G} - P_{\Sigma_2 g})\Sigma_2 x\|_2^2.\end{aligned}$$

The second statement, namely, $\sigma\left(f(\Sigma_1, G)\right) \geq \sigma\left(f(\Sigma_2, G)\right)$, follows from Lemma 2.4 below, whose result is a consequence of Corollary 4.3.3 in [4] and the following fact: $\|Bx\|_2^2 \leq \|Ax\|_2^2$ for all x if and only if $B^*B \preceq A^*A$. □

Lemma 2.4. *If $\|Bx\|_2^2 \leq \|Ax\|_2^2$ for all x, then $\sigma(B) \leq \sigma(A)$.*

2.3 Proof of Theorem 1.2

We let D_{n-k} be an $(n-k)$-dimensional diagonal matrix and work with

$$M(t) = \begin{bmatrix} tI_k & 0 \\ 0 & D_{n-k} \end{bmatrix}. \tag{2.2}$$

Set $G \in \mathbb{R}^{n \times (k+p)}$ and partition the rows as

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \qquad \begin{matrix} G_1 \in \mathbb{R}^{k \times (k+p)} \\ G_2 \in \mathbb{R}^{(n-k) \times (k+p)} \end{matrix}.$$

Next, introduce an SVD for G_1

$$G_1 = U \begin{bmatrix} \Sigma & 0 \end{bmatrix} V^*, \quad U \in \mathbb{R}^{k \times k}, \quad V \in \mathbb{R}^{(k+p) \times (k+p)}, \quad \Sigma \in \mathbb{R}^{k \times k},$$

and partition $G_2 V$ as

$$G_2 V = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \quad X_1 \in \mathbb{R}^{(n-k) \times k}, \quad X_2 \in \mathbb{R}^{(n-k) \times p}.$$

A simple calculation shows that

$$H = M(t)G = \begin{bmatrix} tG_1 \\ D_{n-k} G_2 \end{bmatrix} = \begin{bmatrix} U & 0 \\ D_{n-k} X_1 \Sigma^{-1}/t & D_{n-k} X_2 \end{bmatrix} \begin{bmatrix} t\Sigma & 0 \\ 0 & I_p \end{bmatrix} V^*.$$

Hence, H and $\begin{bmatrix} U & 0 \\ D_{n-k} X_1 \Sigma^{-1}/t & D_{n-k} X_2 \end{bmatrix}$ have the same column space. If Q_2 is an orthonormal basis for the range of $D_{n-k} X_2$, we conclude that

$$\begin{bmatrix} U & 0 \\ D_{n-k} X_1 \Sigma^{-1}/t & Q_2 \end{bmatrix} \text{ and, therefore, } \tilde{H} = \begin{bmatrix} U & 0 \\ (I - Q_2 Q_2^*) D_{n-k} X_1 \Sigma^{-1}/t & Q_2 \end{bmatrix}$$

have the same column space as H. Note that the last p columns of $\tilde{H}$ are orthonormal.

Continuing, we let B be the first k columns of $\tilde{H}$. Then Lemma 2.5 below allows us to decompose B as

$$B = Q + E(t),$$

where Q is orthogonal with the same range space as B and $E(t)$ has a spectral norm at most $O(1/t^2)$. Further since the first k columns of $\tilde{H}$ are orthogonal to the last p columns, we have

$$\tilde{H} = \tilde{Q} + \tilde{E}(t),$$

where $\tilde{Q}$ is orthogonal with the same range space as $\tilde{H}$ and $\tilde{E}(t)$ has a spectral norm also at most $O(1/t^2)$. This gives

$$\begin{aligned}
\lim_{t\to\infty} M(t) - QQ^*M(t) &= \lim_{t\to\infty} M(t) - (\tilde{H} - \tilde{E}(t))(\tilde{H} - \tilde{E}(t))^* M(t) \\
&= \lim_{t\to\infty} M(t) - \tilde{H}\tilde{H}^* M(t) \\
&= \begin{bmatrix} 0 & 0 \\ -(I - Q_2Q_2^*)D_{n-k}X_1\Sigma^{-1}U^* & (I - Q_2Q_2^*)D_{n-k} \end{bmatrix}.
\end{aligned}$$

We have reached the conclusion

$$\lim_{t\to\infty} \|f(M(t), G)\| = \|f(D_{n-k}, X_2) \begin{bmatrix} X_1\Sigma^{-1} & I_{n-k} \end{bmatrix} \|. \tag{2.3}$$

When $D_{n-k} = I_{n-k}$, this gives our theorem.

Lemma 2.5. *Let $A = \begin{bmatrix} I_k \\ t^{-1}B \end{bmatrix} \in \mathbb{R}^{n\times k}$. Then A is $O(t^{-2}\|B^*B\|_F)$ in Frobenius norm away from a matrix with the same range and orthonormal columns.*

Proof. We need to construct a matrix $E \in \mathbb{R}^{(n+k)\times k}$ obeying $\|E\|_F = O(\|B^*B\|/t^2)$ and such that (1) $Q = A + E$ is orthogonal and (2) Q and A have the same range. Let $U\Sigma V^*$ be a reduced SVD decomposition for A,

$$A = U\Sigma V^* = UV^* + U(\Sigma - I)V^* := Q + E.$$

Clearly, Q is an orthonormal matrix with the same range as A. Next, the ith singular value of A obeys

$$\sigma_i(A) = \sqrt{\lambda_i(A^*A)} = \sqrt{\lambda_i(I_k + t^{-2}B^*B)} = \sqrt{1 + t^{-2}\lambda_i(B^*B)} \approx 1 + \frac{1}{2}t^{-2}\lambda_i(B^*B).$$

Hence,

$$\|E\|_F^2 = \sum_i (\sigma_i(A) - 1)^2 \approx \left(\frac{1}{2}t^{-2}\right)^2 \sum_i \lambda_i^2(B^*B) = \left(\frac{1}{2}t^{-2}\|B^*B\|_F\right)^2,$$

which proves the claim. □

2.4 Proof of Corollary 1.3

Put $L = f(I_{n-k}, X_2)X_1\Sigma^{-1}$. Since $f(I_{n-k}, X_2)$ is an orthogonal projector, $\|f(I_{n-k}, X_2)\| \le 1$ and (2.3) gives

$$\|L\| \le W \le \|L\| + \|f(I_{n-k}, X_2)\| \le \|L\| + 1. \tag{2.4}$$

Also,

$$\|L\| \le \|f(I_{n-k}, X_2)\|\|X_1\|\|\Sigma^{-1}\| \le \|X_1\|\|\Sigma^{-1}\|. \tag{2.5}$$

Standard estimates in random matrix theory (Lemma A.1) give

$$\sqrt{n-k} - \sqrt{k} \le \mathbb{E}\sigma_{\min}(X_1) \le \mathbb{E}\sigma_{\max}(X_1) \le \sqrt{n-k} + \sqrt{k}.$$

Hence,

$$\mathbb{E}\|L\| \le \mathbb{E}\|X_1\|\,\mathbb{E}\|\Sigma^{-1}\| \le (\sqrt{n-k} + \sqrt{k})\,\mathbb{E}\|\Sigma^{-1}\|.$$

Conversely, letting i be the index corresponding to the largest entry of Σ^{-1}, we have $\|L\| \ge \|Le_i\|$. Therefore,

$$\|L\| \ge \|\Sigma^{-1}\|\,\|f(I_{n-k}, X_2)z\|,$$

where z is the ith column of X_1. Now $f(I_{n-k}, X_2)z$ is the projection of a Gaussian vector onto a plane of dimension $n-k-p$ drawn independently and uniformly at random. This says that $\|f(I_{n-k}, X_2)z\| \stackrel{d}{=} \sqrt{Y}$, where Y is a chi-square random variable with $d = n - (k+p)$ degrees of freedom. If g is a nonnegative random variable, then[2]

$$\mathbb{E}g \ge \sqrt{\frac{(\mathbb{E}g^2)^3}{\mathbb{E}g^4}}. \tag{2.6}$$

Since $\mathbb{E}Y = d$ and $\mathbb{E}Y^2 = d^2 + 2d$, we have

$$\mathbb{E}\sqrt{Y} \ge \sqrt{\frac{(\mathbb{E}Y)^3}{\mathbb{E}Y^2}} = \sqrt{d}\sqrt{\frac{1}{1+2/d}} \ge \sqrt{d}\sqrt{1-\frac{2}{d}}.$$

Hence,

$$\mathbb{E}\|L\| \ge \mathbb{E}\|X_1\|\,\mathbb{E}\|\Sigma^{-1}\| \ge \sqrt{n-(k+p+2)}\,\mathbb{E}\|\Sigma^{-1}\|,$$

which establishes Corollary 1.3.

The limit bounds (1.8) and (1.9) are established in a similar manner. The upper estimate is a consequence of the bound $W \le 1 + \|X_1\|\|\Sigma^{-1}\|$ together with Lemma A.3. The lower estimate follows from $W \ge Y^{1/2}\|\Sigma^{-1}\|$, where Y is a chi-square as before, together with Lemma A.3. We forgo the details.

2.5 Proof of Theorem 1.4

Take $M(t)$ as in (2.2) with $\sigma_{k+1}, \sigma_{k+2}, \ldots, \sigma_n$ on the diagonal of D_{n-k}. Applying (2.3) gives

$$\|f(D_{n-k}, X_2)X_1\Sigma^{-1}\| \le \lim_{t\to\infty} \|f(M(t), G)\| \le \|f(D_{n-k}, X_2)X_1\Sigma^{-1}\| + \|f(D_{n-k}, X_2)\|. \tag{2.7}$$

[2]This follows from Hölder's inequality $\mathbb{E}|XY| \le (\mathbb{E}|X|^{3/2})^{2/3}(\mathbb{E}|Y|^3)^{1/3}$ with $X = g^{2/3}$, $Y = g^{4/3}$.

We follow [3, Proof of Theorem 10.6] and use an inequality of Gordon to establish

$$\mathbb{E}_{X_1}\|f(D_{n-k}, X_2)X_1\Sigma^{-1}\| \le \|f(D_{n-k}, X_2)\|\|\Sigma^{-1}\|_F + \|f(D_{n-k}, X_2)\|_F\|\Sigma^{-1}\|,$$

where $\mathbb{E}_{X_1}$ is expectation over X_1. Further, it is well-known that

$$\mathbb{E}\|\Sigma^{-1}\|_F^2 = \frac{k}{p-1}.$$

The reason is that $\|\Sigma^{-1}\|_F^2 = \operatorname{trace}(M^{-1})$, where M is a Wishart matrix $M \sim \mathcal{W}(I_k, k+p)$. The identity follows from $\mathbb{E}M^{-1} = (p-1)^{-1}I_k$ [5, Exercise 3.4.13]. In summary, the expectation of the right-hand side in (2.7) is bounded above by

$$\left(1 + \sqrt{\frac{k}{p-1}}\right)\mathbb{E}\|f(D_{n-k}, X_2)\| + \mathbb{E}\|\Sigma^{-1}\|\, E\|f(D_{n-k}, X_2)\|_F.$$

Since $\sigma(f(D_{n-k}, X_2)) \le \sigma(D_{n-k})$, the conclusion of the theorem follows.

2.6 Proof of Theorem 1.5

With $W(D_{n-k}) = f(D_{n-k}, X_2)[X_1\Sigma^{-1} \quad I_{n-k}]$, we have

$$\|W(D_{n-k})\|_F^2 \le \|D_{n-k}[X_1\Sigma^{-1} \quad I_{n-k}]\|_F^2 = \|D_{n-k}X_1\Sigma^{-1}\|_F^2 + \|D_{n-k}]\|_F^2;$$

the inequality follows from the fact that for any orthogonal projector P, $\|PA\|_F \le \|A\|_F$, and the second is Pythagoras' identity. Now a simple calculation we omit gives

$$\mathbb{E}\|D_{n-k}X_1\Sigma^{-1}\|_F^2 = \|D_{n-k}\|_F^2\mathbb{E}\|\Sigma^{-1}\|_F^2.$$

The theorem then follows from the previous identity $\mathbb{E}\|\Sigma^{-1}\|_F^2 = k/(p-1)$.

3 Discussion

We have developed a new method for characterizing the performance of a well-studied algorithm in randomized numerical linear algebra and used it to prove sharp performance bounds. A natural question to ask when using Algorithm 1 is if one should draw G from some distribution other than the Gaussian. It turns out that for all values m, n, k and p, choosing G with Gaussian entries minimizes

$$\sup_A \mathbb{E}\|A - QQ^*A\|/\sigma_{k+1}.$$

This is formalized as follows:

Lemma 3.1. *Choosing $G \in \mathbb{R}^{m\times\ell}$ with i.i.d. Gaussian entries minimizes the supremum of $\mathbb{E}\|(I - QQ^*)A\|/\sigma_{k+1}$ across all choices of A.*

Proof. Fix A with $\sigma_{k+1}(A) = 1$ (this is no loss of generality) and suppose F is sampled from an arbitrary measure with probability 1 of being rank ℓ (since being lower rank can only increase the error). The expected error is $\mathbb{E}_F\|\mathcal{P}_{AF}(A)\|$, where for arbitrary matrices, $\mathcal{P}_A(B) = (I - P)B$ in which P is the orthogonal projection onto the range of A. Suppose further that U is drawn

uniformly at random from the space of orthonormal matrices. Then if G is sampled from the Gaussian distribution,

$$\mathbb{E}_G \|\mathcal{P}_{AG}(A)\| = \mathbb{E}_{F,U} \|\mathcal{P}_{AUF}(A)\|$$

since UF chooses a subspace uniformly at random as does G. Therefore, there exists U_0 with the property $\mathbb{E}_F \|\mathcal{P}_{AU_0F}(A)\| \geq \mathbb{E}_G \|\mathcal{P}_{AG}(A)\|$, whence

$$\mathbb{E}_G \|\mathcal{P}_{AG}(A)\| \leq \mathbb{E}_F \|\mathcal{P}_{AU_0F}(A)\| = \mathbb{E}_F \|\mathcal{P}_{AU_0F}(AU_0)\|.$$

Hence, the expected error using a test matrix drawn from the Gaussian distribution on A is smaller or equal to that when using a test matrix drawn from another distribution on AU_0. Since the singular values of A and AU_0 are identical since U_0 is orthogonal, the Gaussian measure (or any measure that results in a rotationally invariant choice of rank k subspaces) is worst-case optimal for the spectral norm. □

The analysis presented in this paper does not generalize to a test matrix G drawn from the subsampled random Fourier transform (SRFT) distribution as suggested in [11]. Despite their inferior performance in the sense of Lemma 3.1, SRFT test matrices are computationally attractive since they come with fast algorithms for matrix-matrix multiply.

A Appendix

We use well-known bounds to control the expectation of the extremal singular values of a Gaussian matrix. These bounds are recalled in [8], though known earlier.

Lemma A.1. *If $m > n$ and A is a $m \times n$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries, then*

$$\sqrt{m} - \sqrt{n} \leq \mathbb{E}\sigma_{\min}(A) \leq \mathbb{E}\sigma_{\max}(A) \leq \sqrt{m} + \sqrt{n}.$$

Next, we control the expectation of the norm of the pseudo-inverse $A^\dagger$ of a Gaussian matrix A.

Lemma A.2. *In the setup of Lemma A.1, we have*

$$\frac{1}{\sqrt{m-n}} \leq \mathbb{E}\|A^\dagger\| \leq e\frac{\sqrt{m}}{m-n}.$$

Proof. The upper bound is the same as is used in [3] and follows from the work of [1]. For the lower bound, set $B = (A^*A)^{-1}$ which has an inverse Wishart distribution, and observe that

$$\|A^\dagger\|^2 = \|B\| \geq B_{11},$$

where B_{11} is the entry in the $(1,1)$ position. It is known that $B_{11} \stackrel{d}{=} 1/Y$, where Y is distributed as a chi-square variable with $d = m - n + 1$ degrees of freedom [5, Page 72]. Hence,

$$\mathbb{E}\|A^\dagger\| \geq \mathbb{E}\frac{1}{\sqrt{Y}} \geq \frac{1}{\sqrt{\mathbb{E}Y}} = \frac{1}{\sqrt{m-n+1}}.$$

□

The limit laws below are taken from [10] and [2].

Lemma A.3. *Let $A_{m,n}$ be a sequence of $m \times n$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries such that $\lim_{n\to\infty} m/n = c \geq 1$. Then*

$$\frac{1}{\sqrt{n}}\sigma_{\min}(A_{m,n}) \overset{a.s.}{\to} \sqrt{c} - 1$$
$$\frac{1}{\sqrt{n}}\sigma_{\max}(A_{m,n}) \overset{a.s.}{\to} \sqrt{c} + 1.$$

Acknowledgements

E. C. is partially supported by NSF via grant CCF-0963835 and by a gift from the Broadcom Foundation. We thank Carlos Sing-Long for useful feedback about an earlier version of the manuscript. These results were presented in July 2013 at the European Meeting of Statisticians. We would like to thank the reviewers for useful comments.

References

[1] Z. Chen and J. J. Dongarra, "Condition numbers of Gaussian random matrices". *SIAM J. Matrix Anal. Appl.*, Vol. 27, pp. 603–620, 2005.

[2] S. Geman, "A limit theorem for the norm of random matrices". *Ann. Prob.*, Vol. 8, pp. 252–261, 1980.

[3] N. Halko, P-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions". *SIAM Review*, Vol. 53, pp. 217–288, 2011.

[4] R. A. Horn, and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[5] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[6] P-G. Martinsson, V. Rokhlin, and M. Tygert, "A Fast Randomized Algorithm for the Approximation of Matrices". *Appl. Comp. Harm. Anal.*, Vol. 30, pp. 47–68, 2011. (Early version published as YALEU/DCS/TR-1361, 2006.)

[7] V. Rokhlin, A. Szlam, and M. Tygert, "A randomized algorithm for principal component analysis". *SIAM J. Matrix Anal. Appl.*, Vol. 31, pp. 1100–1124, 2009.

[8] M. Rudelson, and R. Vershynin, "Non-asymptotic theory of random matrices: extreme singular values". In *Proceedings of the International Congress of Mathematicians*, Vol. III, pp. 1576–1602, Hindustan Book Agency, New Delhi, 2010.

[9] T. Sarlós, "Improved approximation algorithms for large matrices via random projections". *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 143–152, 2006.

[10] J. W. Silverstein, "The smallest eigenvalue of a large dimensional Wishart matrix". *Ann. Prob.*, Vol. 13, pp. 1364-1368, 1985.

[11] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert "A fast randomized algorithm for the approximation of matrices" *Appl. Comp. Harmon. Anal.*, Vol. 25, pp. 335–366, 2008.