RESEARCH SUMMARY

ALEX GITTENS

1. Overview

Modern scientific computing demands efficient algorithms for dealing with large datasets. Often these datasets can be fruitfully represented and manipulated as matrices; in this case, fast low-error methods for making basic linear algebra computations are key to efficient algorithms. Examples of such foundational computational tools are low-rank approximations, matrix sparsification, and randomized column subset selection.

My research focuses on ways in which randomness can be turned to our advantage in the development of methods for dealing with these massive datasets. The underlying intuition is that when dealing with an excess of structured data (e.g., a large matrix which has low numerical rank), then one can toss away a large portion of this data, thereby reducing the computational load, without introducing much additional error into the computation. I am interested in the analysis of the performance of randomized algorithms based upon this idea of information reduction, and the development of tools to facilitate this analysis.

2. Matrix sparsification

One approach to computing with large matrices is to use classical algorithms for dealing with sparse matrices. These algorithms have the advantage that their complexity depends more on the sparsity of the matrix (the number of nonzero entries) than the dimension of the matrix. For instance, the classical dense SVD scales like $\mathrm{O}(n^3)$ on a square $n \times n$ matrix while sparse SVD computations take $\mathrm{O}(ns + nk^2)$, where $s = \#\mathrm{nnz}(\mathbf{A})$ is the sparsity of the matrix. Since one may be dealing with a dense matrix, one should sparsify it before using these algorithms. This leads to the question: given a sparsity level s, find $\arg\min_{\hat{\mathbf{A}}:\#\mathrm{nnz}(\hat{\mathbf{A}})=s} \left\|\mathbf{A} - \hat{\mathbf{A}}\right\|_2$. Unfortunately, this does not lead to a workable algorithm, so other approaches to sparsification must be considered.

A related problem is that of quantization, or reducing the number of bits required to store the entries in a matrix—rather than reducing the cost of computations, the aim of quantization is to reduce the cost of storage and transmission of datasets. Random quantization schemes are attractive because they can be designed to allow one to trade off accuracy of the approximation with the severity of the quantization.

In joint work with Joel Tropp [GTa], I consider random matrix approximation schemes that approximate a fixed $m \times n$ matrix $\mathbf{A}$ with a random matrix $\mathbf{X}$ having the properties that the entries of $\mathbf{X}$ are independent and average to the corresponding entries of $\mathbf{A}$. This investigation was initiated by the observation that several algorithms for random matrix quantization and sparsification are based on approximations that have these properties [AM07, AM01, AHK06]. One of our goals in [GTa] was to establish a generic framework for the analysis of such approximation schemes that would recapitulate the known guarantees for these particular algorithms and provide a means of quickly deriving guarantees for *any* schemes respecting these conditions.

We show that the spectral norm approximation error can be controlled in terms of the variances and fourth moments of the entries of $\mathbf{X}$ as follows:

$$(1)\quad \mathbb{E}\left\|\mathbf{A} - \mathbf{X}\right\|_2 \leq \mathrm{C}\left[\max_j \left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2} + \max_k \left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2} + \left(\sum_{jk} \mathbb{E}(X_{jk} - a_{jk})^4\right)^{1/4}\right],$$

where C is a universal constant. This expectation bound was obtained by leveraging work done by Latała on the spectral norm of random matrices with zero mean entries [Lat05]. The results hold without any restrictions on the size of $\mathbf{A}$. The constant C is inherited from Latała; extracting a numerical estimate of C from his proof would require considerable effort. We leave this for the future.

Date: May 25, 2012.

When the entries of $\mathbf{A}$ are bounded (so that the variances of the entries of $\mathbf{X}$ are small), an argument based on a bounded difference inequality shows that the approximation error does not exceed this expectation by much: if $|X_{jk}| \leq D$ almost surely, then

$$\mathbb{P}\left\{\left\|\mathbf{A}-\mathbf{X}\right\|_2 \geq (1+\delta)\mathbb{E}\left\|\mathbf{A}-\mathbf{X}\right\|_2\right\} \leq \exp\left(-\frac{\delta^2(\mathbb{E}\left\|\mathbf{A}-\mathbf{X}\right\|_2)^2}{D^2}\right) \tag{2}$$

for any $\delta > 0$. Equation (1) identifies properties desirable in randomized approximation schemes: namely that they minimize the maximum column and row norms of the variances of the entries, as well as the fourth moments of all entries. Thus we have guidance in the design of future approximation schemes. As I now demonstrate, our bounds also yield sharper analyses of current quantization and sparsification schemes.

In the seminal work [AM07, AM01], Achlioptas and McSherry propose two approximation schemes. Their first scheme approximates $\mathbf{A}$ with a 1-bit quantized random matrix $\mathbf{X}$ whose entries satisfy

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

Here b is the largest modulus of the entries in $\mathbf{A}$.

After some algebra, (1) and (2) give the guarantee

$$\left\|\mathbf{A}-\mathbf{X}\right\|_2 \leq 8Cb\sqrt{n}$$

for some universal constant C (the same as above) with a success rate of at least $1-\exp(-C^2 n)$. This error guarantee is on the same order as that given in the original analysis, but our sucess rate is much higher than the estimate $1-\exp(-19(\log n)^4)$ given in [AM07, AM01], and we eliminate a technical restriction on the size of $\mathbf{A}$ that was present in the original analysis.

The second scheme proposed in the same paper sparsifies $\mathbf{A}$ by zeroing entries with probabilities proportional to their magnitudes and a base probability $p \in (0,1)$. Essentially, the smaller the entry, the more likely it is to be sparsified, but if a small entry *is not* sparsified, then it is multiplied by a large constant; this ensures that $\mathbb{E}X_{jk} = a_{jk}$. Achlioptas and McSherry determine that, with probability at least $1-\exp(-19(\log n)^4)$, the matrix $\mathbf{X}$ satisfies

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

and that the expected number of nonzero entries in $\mathbf{X}$ is smaller than

$$pmn \times \mathrm{Avg}[(a_{jk}/b)^2] + m(8\log n)^4. \tag{3}$$

We propose a variation of this scheme with the same type of error guarantee that holds with higher probability, and has a tighter bound on the number of nonzero entries in $\mathbf{X}$. The key idea is that, in addition to the information that b provides on the absolute spread of the entries of the matrix, we should exploit information on the relative spread of the entries of the matrix as measured by $R = \max_{a_{jk}\neq 0} b/|a_{jk}|$. Specifically, we take

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

Then the expected number of nonzeros in $\mathbf{X}$ is smaller than

$$pnm \times \mathrm{Avg}[(a_{jk}/b)^2]$$

and

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

with probability at least $1-\exp(-\mathrm{C}^2(2+\sqrt{R})^2 pn/16)$. Thus this scheme performs better than the one proposed in [AM07, AM01] if the spread R is not too large and the base probability p is not too small.

Arora, Hazan, and Kale consider a scheme in [AHK06] that simultaneously quantizes and sparsifies $\mathbf{A}$. We are able to recover comparable error bounds using (1) and (2).

Our second goal in [GTa] was to analyze the performance of these randomized matrix approximation schemes as measured using non-unitary invariant norms. The literature on randomized matrix approximation has, with few exceptions, focused on the behavior of the spectral and Frobenius norms. However, depending on the application, other norms are of more interest; for instance, the $p \to q$ norms naturally arise when one considers $\mathbf{A}$ as a map from $\ell_p(\mathbb{R}^n)$ to $\ell_q(\mathbb{R}^m)$:

$$\|\mathbf{A}\|_{p\to q} = \max_{\mathbf{x}\,:\,\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_q.$$

Consider, in particular, the $\infty \to 1$ and $\infty \to 2$ norms, both of which are NP-hard to compute. The $\infty \to 1$ norm has applications in graph theory and combinatorics. One oft-invoked connection involves the MAXCUT problem, that of determining a cut of maximum cost in a graph. This problem is known to be NP-Hard [Roh00]. In [AN04] the authors construct a matrix $\mathbf{\Pi}_G$, a simple extension of the classical edge-vertex incidence matrix, that has the property that the value of the maximum cut in G is given by $1/4\,\|\mathbf{\Pi}_G\|_{\infty\to 1}$. They use this observation, in connection with Grothendieck's inequality, to provide an approximation algorithm for the MAXCUT problem.

The $\infty \to 2$ norm has applications in numerical linear algebra. In particular, it is a useful tool in the column subset selection problem: that of, given a matrix $\mathbf{A}$ with unit norm columns, choosing a large subset of the columns of $\mathbf{A}$ so that the resulting submatrix has a norm smaller than some fixed constant (larger than one). Kashin and Tzafriri established that for some constant C such a submatrix exists, but did not offer a procedure for obtaining this submatrix. In [Tro09], Tropp introduces a randomized algorithm for finding such a submatrix. A key point used in this algorithm is that the spectral norm of a matrix is no greater than a small multiple of the $\infty \to 2$ norm of a sufficiently large submatrix.

In a similar way that sparsification can assist in applications where the spectral norm is relevant, we believe it can be of assistance in applications such as these where the norm of interest is a $p \to q$ norm. Our main result is a bound on the expected $\infty \to p$ norm of random matrices whose entries are independent and have mean zero:

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

Here ε is a vector of i.i.d. uniformly random signs, $\mathbf{z}_k$ is the kth column of $\mathbf{Z}$, and q is the conjugate exponent of p. This implies the following bounds on the $\infty \to 1$ and $\infty \to 2$ norms:

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} &\leq 2\mathbb{E}(\|\mathbf{Z}\|_{\text{col}} + \|\mathbf{Z}\|_{\text{row}})\\ \mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} &\leq 2\mathbb{E}\big\|\mathbf{Z}\big\|_{\text{F}} + 2\min_{\mathbf{D}}\mathbb{E}\big\|\mathbf{Z}\mathbf{D}^{-1}\big\|_{2\to\infty},\end{aligned}$$

where the minimization is taken over the set of positive diagonal matrices satisfying $\text{trace}(\mathbf{D}^2) = 1$. The norm $\big\|\mathbf{A}\big\|_{2\to\infty}$ is the largest of the Euclidean norms of the rows of the matrix, $\big\|\mathbf{A}\big\|_{\text{F}}$ is the Frobenius norm, and the column norm $\|\mathbf{A}\|_{\text{col}}$ is the sum of the Euclidean norms of the columns of the matrix. Likewise, $\|\mathbf{A}\|_{\text{row}}$, the row norm of $\mathbf{A}$, is the sum of the Euclidean norms of the rows of the matrix. As in the case of the spectral norm, a bounded differences inequality guarantees us that if the entries of $\mathbf{Z}$ are bounded, then the errors concentrate about these expectations. Thus we have bounds on quantities which are NP-hard to compute, in terms of much simpler quantities.

Both these bounds are optimal in the sense that each term can be shown to be necessary: e.g., there are classes of random matrices whose expected $\infty \to 1$ norms are not comparable to their expected column norms but are comparable to their expected row norms, and vice versa. In the case of the $\infty \to 1$, there is a matching lower bound that lends our bound an interesting interpretation. Littlewood established that, as a consequence of Khintchine's inequality, the relation

$$\max\{\|\mathbf{A}\|_{\text{col}}\,, \|\mathbf{A}\|_{\text{row}}\} \leq \sqrt{2}\,\|\mathbf{A}\|_{\infty\to 1}$$

holds for any matrix $\mathbf{A}$. A standard argument establishes an inequality in the opposite direction; thus for $n \times n$ matrices we have the equivalence

$$\frac{2}{\sqrt{n}}\,\|\mathbf{A}\|_{\infty\to 1} \leq \|\mathbf{A}\|_{\text{col}} + \|\mathbf{A}\|_{\text{row}} \leq 2\sqrt{2}\,\|\mathbf{A}\|_{\infty\to 1}\,.$$

Call this Littlewood's equivalence. Note that there are $\mathbf{A}$ for which the leftmost inequality is an equality—for instance, the matrix of all ones. Thus, in general, as n increases, the gap between $\|\mathbf{A}\|_{\infty\to 1}$ and $\|\mathbf{A}\|_{\text{col}} + \|\mathbf{A}\|_{\text{row}}$ increases. Our result can be interpreted as a *dimensionless* Littlewood's equivalence on the vector space of random zero-mean matrices with independent entries:

$$\frac{1}{2}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} \leq \mathbb{E}\,\|\mathbf{Z}\|_{\text{col}} + \mathbb{E}\,\|\mathbf{Z}\|_{\text{row}} \leq 2\sqrt{2}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1}\,.$$

3. Random matrix theory

Random matrices are employed in an increasing number of applications; to mention just a few: machine learning [Ach04], robust convex optimization [So09b, Nem07], approximation algorithms for NP hard optimization problems [Sar06, So09a, RFP10], and fast computational linear algebra [HMT11, BD09]. In each of these areas, the feasibility and efficacy of randomized algorithms is determined by the spectra of the random matrices used. In other cases, such as covariance estimation, we investigate random matrices in their own right. I am interested in the development of simple but versatile tools for probing the properties of random matrices.

3.1. **Eigenvalue bounds.** Classical asymptotic random matrix theory tools—e.g. the method of moments and the Stieltjes transforms—can be used to obtain results which describe, for certain ensembles of random matrices, the limit of the spectral distributions as the matrix size approaches infinity. Unfortunately, these results address the convergence of the empirical spectral distributions and the extreme eigenvalues, not that of the interior eigenvalues. Furthermore, the métier of these techniques is the determination of convergence, rather than the development of tail bounds that hold at a fixed dimension. To develop such bounds, we turn to the complementary field of nonasymptotic random matrix theory.

In addition to providing quantitative bounds on the extreme eigenvalues of random matrices of fixed dimensions, the tools of nonasymptotic random matrix theory apply to a wider class of matrices than those of asymptotic random matrix theory. Perhaps the most generally applicable tool in the arsenal of nonasymptotic random matrix theory is the matrix Laplace transform technique pioneered by Ahlswede and Winter, which applies to sums of independent random matrices [AW02, Tro].

However, the Laplace transform technique yields bounds on only the extremal eigenvalues. In [GTb], Joel Tropp and I introduce a simple technique, based upon the variational characterization of the eigenvalues of self-adjoint matrices and the Laplace transform machinery, for bounding *all* eigenvalues of sums of independent random self-adjoint matrices.

The Laplace transform technique allows one to develop matrix corollaries of the classical exponential probability inequalities for sums of independent random variables. The scalar Laplace transform technique uses Markov's inequality, the monotonicity of the scalar exponential mapping, and the fact that the moment-generating function of a sum of independent random variables is the product of the individual moment-generating functions. In the matrix case, we still have Markov's inequality and the monotonicity of the scalar exponential mapping; the role of the moment generating function in the scalar case is played by the trace exponential in the matrix case. This gives the basic bound

$$\mathbb{P}\left\{\lambda_{\max}\left(\mathbf{Y}\right) \geq t\right\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{tr}\exp\left(\mathrm{e}^{\theta\mathbf{Y}}\right)$$

For sums of independent random matrices $\mathbf{Y} = \sum_i \mathbf{X}_i$, we do not retain the property that the moment generating function is the product of the summands' moment-generating functions. Ahlswede and Winter's seminal approach is to use the Golden-Thompson inequality iteratively to get a weaker form of this separation. We follow instead the approach in [Tro], which uses the fact that, when $\mathbf{H}$ is a fixed self-adjoint matrix, the function

$$\mathbf{A} \mapsto \mathrm{tr}\exp\left(\mathbf{H} + \log\mathbf{A}\right)$$

is concave on the positive-definite cone to establish the relation

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_i \theta\mathbf{X}_i\right) \leq \mathrm{tr}\exp\left(\sum\nolimits_i \log\mathbb{E}\mathrm{e}^{\theta\mathbf{X}_i}\right).$$

When one has sufficiently strong semidefinite bounds on the matrix cumulant generating functions $\log\mathbb{E}\mathrm{e}^{\theta\mathbf{X}_i}$ of the summands, the Laplace transform technique yields exponential probability bounds on the extreme eigenvalues of $\mathbf{Y} = \sum_i \mathbf{X}_i$.

The minimax Laplace transform we introduce in [GTb] takes advantage of the Courant–Fischer variational characterization of eigenvalues of self-adjoint matrices. For integers d and n satisfying $1 \leq d \leq n$, the complex Stiefel manifold

$$V_d(\mathbb{C}^n) = \{\mathbf{V} \in \mathbb{C}^{n\times d} \,:\, \mathbf{V}^*\mathbf{V} = \mathbf{I}\}$$

is the collection of orthonormal bases for the d-dimensional subspaces of $\mathbb{C}^n$, or, equivalently, the collection of all isometric embeddings of $\mathbb{C}^d$ into $\mathbb{C}^n$. Let $\mathbf{A}$ be a self-adjoint matrix with dimension n. Then we have the representation

$$\lambda_k(\mathbf{A}) = \min_{\mathbf{V} \in V_{n-k+1}(\mathbb{C}^n)} \lambda_{\max}\left(\mathbf{V}^*\mathbf{A}\mathbf{V}\right).$$

This allows us to relate the behavior of the k-th eigenvalue of a random self-adjoint matrix to the behavior of its compressions to subspaces:

$$\mathbb{P}\left\{\lambda_k(\mathbf{Y}) \geq t\right\} \leq \inf_{\theta > 0} \min_{\mathbf{V} \in V_{n-k+1}(\mathbb{C}^n)} \left\{ \mathrm{e}^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \exp\left(\mathrm{e}^{\theta \mathbf{V}^* \mathbf{Y} \mathbf{V}}\right) \right\}$$

We then exploit the fact that, for all $\mathbf{V}$ with orthonormal columns, the function

$$\mathbf{A} \mapsto \operatorname{tr} \exp\left(\mathbf{H} + \log(\mathbf{V}^*\mathbf{A}\mathbf{V})\right)$$

is concave on the positive-definite cone to establish the relation

$$\mathbb{E} \operatorname{tr} \exp\left(\sum\nolimits_i \theta \mathbf{V}^* \mathbf{X}_i \mathbf{V}\right) \leq \operatorname{tr} \exp\left(\sum\nolimits_i \log \mathbb{E} \mathrm{e}^{\theta \mathbf{V}^* \mathbf{X}_i \mathbf{V}}\right).$$

Thus, when one has sufficiently strong semidefinite bounds on the matrix cumulant generating functions $\log \mathbb{E}\mathrm{e}^{\theta \mathbf{V}^* \mathbf{X}_i \mathbf{V}}$ of the compressions of the summands $\mathbf{X}_i$, the minimax Laplace transform technique yields exponential probability bounds for all the eigenvalues of $\mathbf{Y} = \sum_i \mathbf{X}_i$.

We employ the minimax Laplace transform to produce eigenvalue Chernoff, Bennett, and Bernstein bounds. As an example of the efficacy of this technique, we use the Chernoff bounds to find new bounds on the interior eigenvalues of matrices formed by sampling columns from matrices with orthonormal rows. We also demonstrate that our Bernstein bounds are powerful enough to recover known estimates on the number of samples needed to accurately estimate the eigenvalues of the covariance matrix of a Gaussian process by the eigenvalues of the sample covariance matrix.

4. Low-rank approximation

It is a classical result that the spectral norm distance of any matrix $\mathbf{A}$ from the set of rank-k matrices is exactly $\sigma_k(\mathbf{A})$, and that a rank-k matrix $\mathbf{A}_k$ that achieves this error can be obtained from the truncated SVD of $\mathbf{A}$. Likewise the Frobenius norm distance of $\mathbf{A}$ from the set of rank-k matrices is exactly $(\sum_{i>k} \sigma_i(\mathbf{A})^2)^{1/2}$, and $\mathbf{A}_k$ also achieves this error. Such low-rank approximations are ubiquitious in scientific computation, but classical approaches involve using a truncated SVD, which can be expensive to compute and difficult to parallelize. Modern randomized algorithms for low-rank approximation, such as those espoused in [HMT11, NDT09, WLRT06], can achieve comparable errors to the classical approach with low failure rates, and are often amenable to parallelization.

4.1. Subsampled orthogonal transformations for faster low-rank approximation.

In [BG], along with my collaborator Christos Boutsidis, I offer a new analysis of the Subsampled Randomized Hadamard Transform (SRHT) approach to low-rank approximation. This is a specific instance of a class of low-rank approximation algorithms based on random projections. The intuition behind these methods is essentially the same as that behind the classical subspace iteration algorithms: if $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be approximated well by a rank-k matrix (i.e. if $\mathbf{A}$ has sufficient spectral decay) then one can capture the top k-dimensional singular spaces of $\mathbf{A}$ by applying $\mathbf{A}$ to a collection of more than k random vectors. The corresponding low-rank approximation is then just the projection of $\mathbf{A}$ onto the span of these vectors. The use of *more* than k random vectors allows these projection methods to find good low-rank approximations without iteration; as the oversampling increases, the probability that the approximation returned is at least as accurate as the optimal rank-k approximation increases.

Let $\ell > k$ be a positive integer and let $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ be a matrix whose columns are random vectors, then projection methods approximate $\mathbf{A}$ with $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$, which has rank at most ℓ (The notation $\mathbf{P}_{\mathbf{M}}$ denotes the projection onto the range of $\mathbf{M}$). The selection of the distribution of $\mathbf{S}$ crucial: if the entries of $\mathbf{S}$ are i.i.d. standard Gaussians, then the error decreases sharply as a function of $\ell - k$. However, one can reduce the cost of the algorithm by using random matrices whose structure allows for fast multiplication—specifically, one can reduce the cost of forming the product $\mathbf{AS}$ from $\mathrm{O}(mn\ell)$ to $\mathrm{O}(mn \log \ell)$. One choice of a structured random matrix is the (transpose of the) subsampled randomized Hadamard transform (SRHT),

$$\mathbf{S} = \sqrt{\frac{n}{\ell}} \cdot \mathbf{D}\mathbf{H}^T\mathbf{R}^T.$$

Here, $\mathbf{D}$ is a diagonal matrix whose entries are independent random uniformly distributed signs, $\mathbf{H}$ is a normalized Walsh-Hadamard matrix (a particular kind of orthogonal matrix, each of whose entries has modulus $n^{-1/2}$), and $\mathbf{R}$ is a matrix that restricts an n-dimensional vector to a random size ℓ subset of its coordinates. It is not necessary that $\mathbf{H}$ be a normalized Walsh-Hadamard matrix; other orthogonal transforms whose entries are on the order of $n^{-1/2}$ can be used as well, such as the discrete cosine transform or the discrete Hartley transform.

The question of how much the approximation error increases when we switch from the Gaussian distribution to the SRHT model has been studied by several authors [Tro11, NDT09, HMT11]. The previous tightest result for the spectral norm is given in [HMT11], where it is shown that

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2 \leq \left(1 + \sqrt{\frac{7n}{\ell}}\right) \|\mathbf{A} - \mathbf{A}_k\|_2$$

with probability at least $1 - \mathrm{O}(1/k)$ when ℓ is at least on the order of $k \log k$. If $\mathbf{A}$ is close to full rank and has no significant spectral decay, this result is perhaps optimal. But in the situations where low-rank approximation makes sense—namely, when $\mathbf{A}$ is rank-deficient or has fast spectral decay—, this result does not reflect the relevant spectral properties. In fact, this error is of the same order as one would get by naïvely sampling the same number of columns from $\mathbf{A}$ and approximating $\mathbf{A}$ with its projection onto their span [Git]. In [BG] we show that in fact

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2 \leq \mathrm{O}\left(\sqrt{\frac{\log(kn)}{\log(k^2)}}\right) \|\mathbf{A} - \mathbf{A}_k\|_2 + \mathrm{O}\left(\sqrt{\frac{1}{\log(k^2)}}\right) \|\mathbf{A} - \mathbf{A}_k\|_\mathrm{F}$$

with the same probability of failure for the same number of samples. The factor in front of the optimal error has been reduced logarithmically at the cost of the introduction of a Frobenius term. This Frobenius term is small when $\mathbf{A}$ is nicely low-rank approximable. In fact, there is reason to expect that this is the correct form for the bound: an analogous Frobenius term arises naturally in the analysis of the Gaussian case in [HMT11] as a consequence of a Gaussian large-deviations result. We also recover a slightly improved guarantee on the Frobenius error, that

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_\mathrm{F} \leq (1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\mathrm{F}$$

with probability at least $1 - \delta$ when ℓ is on the order of $k \log(k/\delta)$.

The key consideration in bounding the errors of SRHT approximation is the interaction of $\mathbf{S}$ with the singular spaces of $\mathbf{A}$. Let $\mathbf{U}_1$ be the matrix of (normalized) right singular vectors corresponding to the top k-dimensional right singular subspace of $\mathbf{A}$,$\mathbf{U}_2$ span the bottom $(n-k)$-dimensional right singular subspace of $\mathbf{A}$, and $\mathbf{\Sigma}_2$ be the corresponding diagonal matrix of $n-k$ singular values. Several researchers have shown deterministic results relating the approximation errors to the behavior of the matrix $\mathbf{U}_1^T\mathbf{S}$. Specifically, if $\mathbf{S}$ has enough samples that this matrix has full row-rank, then

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_\xi^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \|\mathbf{\Sigma}_2\mathbf{U}_2^T\mathbf{S}(\mathbf{U}_1^T\mathbf{S})^\dagger\|_\xi^2,$$

where $\xi = 2$ or $\xi = F$ and $\dagger$ denotes the pseudoinversion operation. The condition that $\mathbf{U}_1^T\mathbf{S}$ have full row-rank is quite natural: geometrically, it reflects the fact that $\mathbf{AS}$ can only capture the top k-dimensional singular spaces of $\mathbf{A}$ if $\mathbf{S}$ has nonzero projections onto all the relevant right singular vectors of $\mathbf{A}$. This is analogous to the fact that subspace iteration can only succeed if the starting matrix has components in the direction of the eigenspace we are trying to recover.

In the case of $\xi = 2$, previous analyses have proceeded by using submultiplicativity to estimate the second term in this bound with $\|\mathbf{\Sigma}_2\|_2^2\|\mathbf{U}_2^T\mathbf{S}\|_2^2\|(\mathbf{U}_1^T\mathbf{S})^\dagger\|_2^2$. This throws out all possibility of accounting for spectral decay. We avoid this trap by bounding the second term with $\|\mathbf{\Sigma}_2\mathbf{U}_2^T\mathbf{S}\|_\xi^2 \ \|(\mathbf{U}_1^T\mathbf{S})^\dagger\|_2^2$. We then need to answer the question of how the singular values of products $\mathbf{MS}$ behave, where $\mathbf{M}$ is a general rectangular matrix.

In [Tro11], Tropp considers the same question, but in the special case where $\mathbf{M}$ has orthogonal rows. There he shows that the maximum column norm of a matrix with orthonormal rows to which an SRHT matrix has been applied is, with high probability, not much larger than the root-mean squared average of the column norms of the original matrix. We extend this result to the case where $\mathbf{M}$ is a general matrix. A bound on the spectral norm of $\mathbf{MS}$ then follows from a matrix Chernoff bound. This allows us to obtain our stated result on the spectral norm error of SRHT low-rank approximation.

4.2. **Positivity-preserving low-rank approximations of positive matrices.** The randomized projection method just described does not preserve positivity, so in the case where $\mathbf{A}$ is positive and one desires to approximate it with a positive low-rank matrix, another approach must be adopted. In [Git], I consider the simple Nyström extension, which approximates $\mathbf{A}$ in terms of linear combinations of a small subset of its columns and rows. Nyström extensions are popular in machine learning, where they are often used to simplify computations with large dense PSD matrices [FBCM04, WDT$^+$09, BF].

A simple Nyström extension of $\mathbf{A}$ is formed by sampling uniformly at random (with or without replacement) ℓ columns of $\mathbf{A}$ to form a matrix $\mathbf{C}$. Let $\mathbf{W}$ denote the $\ell \times \ell$ 'coupling' matrix formed by the intersection of the columns in $\mathbf{C}$ and the corresponding rows in $\mathbf{A}$. The matrix $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ is then a simple Nyström extension of $\mathbf{A}$. Since $\mathbf{W}^\dagger$ is a principal submatrix of $\mathbf{A}$, it is PSD, so the Nyström extension is also PSD. Because of the pseudoinversion operation, the cost of forming a simple Nyström extension is $\mathrm{O}(n\ell^2 + \ell^3)$.

As in the case of projection-based low-rank approximations, the goal in forming a Nyström extension is to achieve (spectral and Frobenius) errors comparable to the errors of the optimal rank-k approximation while using as few column samples ℓ as possible. There are no available relative-error Frobenius norm bounds. [Git] provides the first relative-error spectral norm bound (the preprint [CD] released at the same time as [Git] contains a similar relative-error bound).

Since it is based on uniform column-sampling, the simple Nyström extension performs best when the information in the top k-dimensional eigenspace is distributed evenly throughout the columns of $\mathbf{A}$. One way to quantify this idea uses the concept of *coherence*, taken from the matrix completion literature [CR09]. Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^n$. The coherence of $\mathcal{S}$ is

$$\mu(\mathcal{S}) = \frac{n}{k} \max\nolimits_i (\mathbf{P}_{\mathcal{S}})_{ii}.$$

The coherence of the dominant k-dimensional eigenspace of $\mathbf{A}$ is a measure of how much comparative influence the individual columns of $\mathbf{A}$ have on this subspace: if μ is small, then all columns have essentially the same influence; if μ is large, then it is possible that there is a single column in $\mathbf{A}$ which alone determines one of the top k eigenvectors of $\mathbf{A}$. We mention that if $\mathbf{A}$ is rank-k, then the quantities $(\mathbf{P}_{\mathcal{S}})_{ii}$ are known to statisticians as the leverage scores of the columns of $\mathbf{A}$ [DM10].

Talwalkar and Rostamizadeh were the first to use coherence in the analysis of Nyström extensions. Let $\mathbf{A}$ be exactly rank-k and μ denote the coherence of its top k-dimensional eigenspace. In [TR10], they show that if one samples on the order of $\mu k \log(k/\delta)$ columns to form a simple Nyström extension, then with probability at least $1 - \delta$

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^t\right\|_2 = 0.$$

That is, one achieves exact recovery.

My first result can be viewed as a generalization of Talwalkar's, as I show that for a general PSD matrix $\mathbf{A}$, if one samples the same number of columns, the error of the Nyström extension satisfies

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^t\right\|_2 \leq \lambda_{k+1}(\mathbf{A})\left(1 + \frac{2n}{\ell}\right)$$

with probability at least $1 - \delta$. This result shows that not only is exact recovery achieved if $\mathbf{A}$ is exactly rank-k, but also that the approximation error is small if $\lambda_{k+1}(\mathbf{A})$ is small. The effect of spectral decay can also be taken into account: I also show that

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^t\right\|_2 \leq \lambda_{k+1}(\mathbf{A}) + \frac{2}{\delta}\sum\nolimits_{i>k} \lambda_i(\mathbf{A})$$

with probability at least $1 - 2\delta$ when the same number of samples are taken. These two results show that the performance of the simple Nyström extension is dependent on the coherence of the top k-dimensional eigenspace of matrix it is applied to: if this coherence is small, then the simple Nyström extension will be effective with $\ell = \mathrm{O}(k \log k)$, if it is large, then the simple Nyström extension will not be effective. Experiments bear out this prediction.

These bounds follow from a relation between the simple Nyström extension of $\mathbf{A}$ and the column subset selection problem for $\mathbf{A}^{1/2}$. Let $\mathbf{S}$ be a random matrix distributed as the first ℓ columns of a permutation matrix chosen uniformly at random. Then we can take $\mathbf{C} = \mathbf{A}\mathbf{S}$ in the definition of the simple Nyström extension. I show that

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 = \left\|\mathbf{A}^{1/2} - \mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}\mathbf{A}^{1/2}\right\|_2^2.$$

The quantity on the right hand side is related to the informativity of the columns selected by $\mathbf{S}$: if they are highly informative, then the projection of $\mathbf{A}^{1/2}$ onto $\mathbf{A}^{1/2}\mathbf{S}$ should be close to $\mathbf{A}^{1/2}$ and thus this quantity should be close to the optimal rank-k approximation error $\lambda_k(\mathbf{A})$. Note that this quantity is of the same form considered in the last section. Thus we need to consider the behavior of the matrix $\mathbf{U}_1^T\mathbf{S}$, where the columns of $\mathbf{U}_1$ are orthonormal and span the top k-dimensional eigenspace of $\mathbf{A}^{1/2}$. Specifically, how many samples are required so that this matrix has full row-rank? A matrix Chernoff argument shows that the correct number of samples is $\mathrm{O}(\mu k \log k)$, and that with this many samples, the norm of the pseudo-inverse of $\mathbf{U}_1^T\mathbf{S}$ is less than $\frac{n}{\ell}$. This leads to my results.

5. Current and future projects

5.1. **Nystrom approximation: other norms, and algorithms for larger scale problems.** I am currently working on a manuscript proving relative error bounds for the simple Nyström extension in the Frobenius and trace norms. The Frobenius norm is of interest to the machine learning community. The motivation for considering the trace norm is that it seems that no truly $(1+\varepsilon)$ approximations can be obtained in the spectral or Frobenius norms (instead, the multiplicative factor in front of the optimal error is on the order of $\frac{n}{\ell}$), but such bounds can be obtained for the trace norm.

As part of this manuscript, I am preparing an empirical comparison of the performance of Nyström extensions based on several different schemes using realistic datasets. The goal of this comparison is to provide practitioners with a set of principles for determining which Nyström extension is appropriate for their dataset. For instance, I have observed empirically that Nyström extensions formed by sampling columns according to their leverage scores converge to the optimal rank-k error faster (with fewer samples) than the other Nyström extensions, but once this error level is achieved, the error does not decrease as more columns are used in the extension. This is because leverage score sampling is biased towards selecting columns which contain more information on the top k-dimensional eigenspace than the bottom $(n-k)$-dimensional eigenspace. By way of comparison, the simple Nyström extension does not have such a bias, so its error converges to the optimal rank-k error slower, but the error continues to decrease as more columns are used.

Prior researchers have suggested using ensemble Nyström extensions—formed by taking linear combinations of several base Nyström extensions—when increased accuracy is desired, in lieu of a single large Nyström extension. This is because forming p simple Nyström extensions each using ℓ columns costs $\mathrm{O}(p\ell^3 + p\ell^2 n)$ while forming a single large extension from $p\ell$ columns costs $\mathrm{O}((p\ell)^3 + (p\ell)^2 n)$. Another approach is to use $p\ell$ columns directly, but replace the pseudoinverse of $\mathbf{W}$ with the pseudoinverse of a low-rank approximation to $\mathbf{W}$. Specifically, a low-rank approximation of the form $\mathbf{P}_{\mathbf{WS}}\mathbf{W}\mathbf{P}_{\mathbf{WS}}$, where $\mathbf{S}$ is a random matrix, can be pseudoinverted in time $\mathrm{O}((p\ell)^2 r + r^3)$. [LKL10] provides an estimate of the error incurred by this approximation of $\mathbf{W}^\dagger$, but there is a large gap between their bounds and the observed behavior of this algorithm. I am working on providing a sharper analysis of this algorithm.

5.2. **Determinant approximation.** It takes cubic time to compute the determinant of a matrix $\mathbf{A}$. I am interested in finding faster approximation algorithms for computing determinants.

One possible route to such an algorithm involves the identity

$$\det(\mathrm{e}^{\mathbf{A}}) = \mathrm{e}^{\operatorname{tr}\mathbf{A}}.$$

From this identity, we see that the problem of estimating the determinant of $\mathbf{A}$ is equivalent to estimating the trace of a matrix logarithm of $\mathbf{A}$. Approaching it from this angle seems feasible because we may be able to exploit results on the estimation of the trace of matrices using random sampling techniques. This would require being able to quickly calculate the application of a logarithm of $\mathbf{A}$ to a given vector. In [OSV], the authors show how to quickly calculate matrix-vector products for the matrix exponential of $\mathbf{A}$; perhaps their results can be extended to the matrix logarithm.

5.3. **Screening methods.** Consider the lasso problem, defined for a given matrix $\mathbf{A}$, vector $\mathbf{b}$, and scalar parameter $\lambda > 0$ as follows:

$$\mathbf{x}_* = \arg\min_{\mathbf{x}} \left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2 + \lambda\|\mathbf{x}\|_1.$$

Because of the ℓ_1 term, the optimal solution vector $\mathbf{x}_*$ is typically sparse. Thus only a few columns of $\mathbf{A}$ are actually relevant to this problem, but an algorithm which solves this problem does not know in advance which columns are relevant. This suggests that the computational cost of solving the lasso problem can be reduced by first running a fast procedure that winnows out the columns that do not contribute to the solution. A few such procedures, called screening tests, are already known for the lasso problem.

There is no reason that only the lasso problem should admit screening tests. I would like to understand the available screening tests, and see if the reasoning behind them can be extended to other problems of interest, such as PCA, linear regression, or K-means clustering.

5.4. **Approximating principal angles between subspaces, and convergence of subspaces.** Consider two matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathbb{R}^{n\times k}$. The principal angles between the subspaces spanned by $\mathbf{A}$ and $\mathbf{B}$ are defined by their cosine values, which are the singular values of $\mathbf{Q}_{\mathbf{A}}^T\mathbf{Q}_{\mathbf{B}}$, or equivalently, the singular values of $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}$. The principal angles are of interest in machine learning.

As is clear from the definition, there is a naïve QR-based algorithm for computing the principal angles between the two subspaces that runs in $\mathrm{O}(nk^2)$ time. If $\mathbf{A}$ and $\mathbf{B}$ are tall-and-thin matrices, then a faster algorithm can provide additive approximations to the cosines of the principal angles. The idea is that one can left-multiply $\mathbf{A}$ and $\mathbf{B}$ by a randomized projection matrix (e.g. an SRHT matrix) to produce $k\log k\times k$ sketches $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, then use the principal angles between these sketches as approximations of the principal angles between $\mathbf{A}$ and $\mathbf{B}$. This algorithm runs in $\mathrm{O}(nk\log k)$ time. This result is due to Avron, Boutsidis, Toledo, and Zouzias, and has been submitted for publication to NIPS 2012.

The question I'm interested in pursuing is: what about almost square or square matrices? Is there an efficient approximation algorithm for this case that has provable theoretical guarantees? An algorithm based on left-multiplication would not give significant gains in computational speed. Instead, I propose dimension-reduction using right multiplication: form the sketches $\tilde{\mathbf{A}} = \mathbf{P}_{\mathbf{AS}}\mathbf{A}$ and $\tilde{\mathbf{B}} = \mathbf{P}_{\mathbf{BS}}\mathbf{B}$ and approximate the (top $r < k$) angles between $\mathbf{A}$ and $\mathbf{B}$ using the angles between these sketches. One could also use a two step procedure, by further applying the algorithm for the tall-and-thin matrices to approximate the angles between $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$.

The proposed algorithm is based on our intuition that the range space of $\mathbf{P}_{\mathbf{AS}}$ is well-aligned with the top r-dimensional invariant subspace of $\mathbf{A}$. This is the same intuition behind the random projection-based low-rank approximation algorithms, but the proofs that such low-rank approximations are accurate *do not* verify this intuition. It seems that the route to take toward showing the effectiveness of this algorithm would involve quantifying this intuition. One way to do so is to consider the square of the sine of the angle between $\mathbf{Q}_{\tilde{\mathbf{A}}}$ and $\mathbf{Q}_{\mathbf{A}_r}$, defined as

$$1 - \lambda_{\min}\left(\mathbf{Q}_{\mathbf{A}_r}^T\mathbf{Q}_{\tilde{\mathbf{A}}}\mathbf{Q}_{\tilde{\mathbf{A}}}^T\mathbf{Q}_{\mathbf{A}_r}\right).$$

Clearly if this quantity is close to zero, the range of $\tilde{\mathbf{A}}$ almost contains that of $\mathbf{A}_r$. This seems a prerequisite for the proposed algorithm to work. I have a deterministic expression for this angle,

$$1 - \lambda_r(\mathbf{A}_r\mathbf{S}(\mathbf{S}^T\mathbf{A}\mathbf{A}^T\mathbf{S})^\dagger\mathbf{S}^T\mathbf{A}_r),$$

that holds when $\mathbf{A}_r\mathbf{S}$ has rank r. Perhaps this result can be of use.

5.5. **Spectral clustering.** Consider a dataset consisting of m objects which we would like to divide into k clusters of similar objects. Suppose that these objects are represented as rows of a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, where n is the number of features used to represent each object. One approach to clustering the objects (for the case $k = 2$) is to compute an $m\times m$ similarity matrix $\mathbf{S}$ that measures the similarity of the objects: s_{ij} measures the similarity of object i with object j. One popular similarity matrix, the Gaussian radial basis function kernel, is given by $s_{ij} = \exp(-\beta\left\|\mathbf{A}^i - \mathbf{A}^j\right\|_2)$ for some $\beta > 0$. Here $\mathbf{A}^i$ denotes the ith row of $\mathbf{A}$.

The eigenvectors of $\mathbf{S}$ contain information which can be used to cluster the objects. Specifically, the Fiedler vector, the eigenvector corresponding to the smallest nonzero eigenvalue of $\mathbf{S}$, can be used to partition the objects. Objects whose corresponding elements of the Fiedler vector have the same sign are assigned to the same partition. A straightforward implementation of this algorithm takes time cubic in m, but algorithms are available which run in less time than this. Their error bounds tend to be unsatisfactory.

The question I would like to consider is: can I provide better approximation guarantees for an existing algorithm or design an algorithm which has better approximation guarantees? In practice the spectral clustering algorithm proposed in [FBCM04], which uses the Nyström extension technique and a similarity kernel based on χ^2 distances, performs well. I would like to analyze the theoretical performance of this algorithm.

References

[Ach04] D. Achlioptas, *Random Matrices in Data Analysis*, Machine Learning: ECML 2004, Springer Berlin, 2004, pp. 1–7.

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

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

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

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

[AW02] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579.

[BD09] C. Boutsidis and P. Drineas, *Random projections for the nonnegative least-squares problem*, Linear Algebra Appl. **431** (2009), 760–771.

[BF] A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, submitted 2011.

[BG] C. Boutsidis and A. Gittens, *Improved matrix algorithms via the Subsampled Randomized Hadamard Transform*, Preprint, arXiv:1204.0062, 2012.

[CD] J. Chiu and L. Demanet, *Sublinear randomized algorithms for skeleton decompositions*, Preprint, arXiv:1110.4193, October 2011.

[CR09] E. Candés and B. Recht, *Exact Matrix Completion via Convex Optimization*, Found. Comput. Math. **9** (2009), 717–772.

[DM10] P. Drineas and M. W. Mahoney, *Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving*, Preprint, arXiv:1005.3097, 2010.

[FBCM04] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, *Spectral Grouping Using the Nyström Method*, IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), 214–225.

[Git] A. Gittens, *The spectral error of a simple CUR decomposition for positive semidefinite matrices*, Preprint, arXiv:1110.5305, 2011.

[GTa] A. Gittens and J. A. Tropp, *Error bounds for random matrix approximation schemes*, Preprint, arXiv:0911.4108, 2009.

[GTb] ______, *Tail bounds for all eigenvalues of a sum of random matrices*, Preprint, arXiv:1104.4513, 2011.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev. **53** (2011), no. 2, 217–288.

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

[LKL10] M. Li, J. T. Kwok, and B. Lu, *Making Large-Scale Nyström Approximation Possible*, Proc. 27th International Conference on Machine Learning (ICML 2010), 2010, pp. 1097–1104.

[NDT09] Nam H. Nguyen, Thong T. Do, and Trac D. Tran, *A fast and efficient algorithm for low-rank approximation of a matrix*, Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC 2009), 2009, pp. 215–224.

[Nem07] A. Nemirovski, *Sums of random symmetric matrices and quadratic optimization under orthogonality constraints*, Mathem. Program. **109** (2007), 283–317.

[OSV] L. Orecchia, S. Sachdeva, and N. Vishnoi, *Approximating the exponential, the Lanczos method, and an $\tilde{\mathrm{O}}(m)$-time spectral algorithm for balanced separator*, Preprint, arXiv:1111.1491, 2011.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization*, SIAM Rev. **52** (2010), no. 3, 471–501.

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

[Sar06] T. Sarlos, *Improved Approximation Algorithms for Large Matrices via Random Projections*, Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS 2006), IEEE Computer Society, 2006, pp. 143–152.

[So09a] A. M. So, *Improved approximation bound for quadratic optimization problems with orthogonality constraints*, Proc. 20th ACM-SIAM Symposium on Discrete Algorithms (SODA 2009), SIAM, 2009, pp. 1201–1209.

[So09b] ______, *Moment inequalities for sums of random matrices and their applications in optimization*, Math. Program. (2009), 1–27.

[TR10] A. Talwalkar and A. Rostamizadeh, *Matrix Coherence and the Nyström Method*, Proc. 26th Conference on Uncertainty in Artifical Intelligence (UAI 2010), 2010.

[Tro] J. A. Tropp, *User-Friendly Tail Bounds for Sums of Random Matrices*, Preprint, arXiv:1004.4389, 2011. Submitted, April 2010.

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

[Tro11] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal. **3** (2011), 115–126.

[WDT$^+$09] J. Wang, Y. Dong, X. Tong, Z. Lin, and B. Guo, *Kernel Nyström method for light transport*, ACM Trans. Graph. **28** (2009), 29:1–29:10.

[WLRT06] Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert, *A fast randomized algorithm for the approximation of matrices, preliminary report*, Technical Report 1380, Yale University, Department of Computer Science, April 2006.