An Explicit Sampling Dependent Spectral Error Bound for Column Subset Selection

Tianbao Yang TIANBAO-YANG@UIOWA.EDU
Department of Computer Science, the University of Iowa, Iowa City, USA

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Rong Jin RONGJIN@CSE.MSU.EDU
Department of Computer Science and Engineering, Michigan State University, East Lansing, USA
Institute of Data Science and Technologies at Alibaba Group, Seattle, USA

Shenghuo Zhu SHENGHUO@GMAIL.COM
Institute of Data Science and Technologies at Alibaba Group, Seattle, USA

Abstract

In this paper, we consider the problem of column subset selection. We present a novel analysis of the spectral norm reconstruction for a simple randomized algorithm and establish a new bound that depends explicitly on the sampling probabilities. The sampling dependent error bound (i) allows us to better understand the tradeoff in the reconstruction error due to sampling probabilities, (ii) exhibits more insights than existing error bounds that exploit specific probability distributions, and (iii) implies better sampling distributions. In particular, we show that a sampling distribution with probabilities proportional to the square root of the statistical leverage scores is better than uniform sampling, and is better than leverage-based sampling when the statistical leverage scores are very nonuniform. And by solving a constrained optimization problem related to the error bound with an efficient bisection search we are able to achieve better performance than using either the leverage-based distribution or that proportional to the square root of the statistical leverage scores. Numerical simulations demonstrate the benefits of the new sampling distributions for low-rank matrix approximation and least square approximation compared to state-of-the art algorithms.

1. Introduction

Give a data matrix $A \in \mathbb{R}^{m\times n}$, **column subset selection (CSS)** is an important technique for constructing a compressed representation and a low rank approximation of A by selecting a small number of columns. Compared with conventional singular value decomposition (SVD), CSS could yield more interpretable output while maintaining performance close to SVD (Mahoney, 2011). Recently, CSS has been applied successfully to problems of interest to geneticists such as genotype reconstruction, identifying substructure in heterogeneous populations, etc. (Mahoney, 2011).

Let $C \in \mathbb{R}^{m\times \ell}$ be the matrix formed by ℓ selected columns of A. The key question to CSS is how to select the columns to minimize the reconstruction error:

$$\|A - P_C A\|_\xi,$$

where $P_C = CC^\dagger$ denotes the projection onto the column space of C with $C^\dagger$ being the pseudo-inverse of C and $\xi = 2$ or F corresponds to the spectral norm or the Frobenius norm (Meyer, 2000). In this paper, we are particularly interested in the spectral norm reconstruction with respect to a target rank k.

Our analysis is developed for a randomized algorithm that selects $\ell > k$ columns from A according to sampling probabilities $\mathbf{s} = (s_1, \ldots, s_n)$. Building on advanced matrix concentration inequalities (e.g., matrix Chernoff bound and Bernstein inequality), we develop a novel analysis of the spectral norm reconstruction and establish a sampling dependent relative spectral error bound with a high probabil-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

ity as following:

$$\|A - P_C A\|_2 \leq (1 + \epsilon(\mathbf{s}))\|A - A_k\|_2,$$

where A_k is the best rank-k approximation of A based on SVD and $\epsilon(\mathbf{s})$ is a quantity dependent on the sampling probabilities $\mathbf{s}$ besides the scalars n, k, ℓ. As revealed in our main theorem (Theorem 1), the quantity $\epsilon(\mathbf{s})$ also depends on the **statistical leverage scores (SLS)** inherent to the data (c.f. Section 3), based on which are several important randomized algorithms for CSS.

To the best of our knowledge, this is the first such kind of error bound for CSS. Compared with existing error bounds, the sampling dependent error bound brings us several benefits: (i) it allows us to better understand the tradeoff in the spectral error of reconstruction due to sampling probabilities, complementary to a recent result on the tradeoff from a statistical perspective (Ma et al., 2014) for least square regression; (ii) it implies that a distribution with sampling probabilities proportional to the square root of the SLS is better than the uniform sampling, and is potentially better than that proportional to the SLS when they are skewed; (iii) it motivates an optimization approach by solving a constrained optimization problem related to the error bound to attain better performance. In addition to the theoretical analysis, we also develop an efficient bisection search algorithm to solve the constrained optimization problem for finding better sampling probabilities.

2. Related Work

In this section, we review some previous work on CSS, low-rank matrix approximation, and other closely related work on randomized algorithms for matrices. We focus our discussion on the spectral norm reconstruction.

Depending on whether the columns are selected deterministically or randomly, the algorithms for CSS can be categorized into deterministic algorithms and randomized algorithms. Deterministic algorithms select $\ell \geq k$ columns with some deterministic selection criteria. Representative algorithms in this category are rank revealing QR factorization and its variants from the filed of numerical linear algebra (Gu & Eisenstat, 1996; Pan, 2000; Pan & Tang, 1999). A recent work (Boutsidis et al., 2011) based on the dual set spectral sparsification also falls into this category which will be discussed shortly. Randomized algorithms usually define sampling probabilities $\mathbf{s} \in \mathbb{R}^n$ and then select $\ell \geq k$ columns based on these sampling probabilities. Representative sampling probabilities include ones that depend the squared Euclidean norm of columns (better for Frobenius norm reconstruction) (Frieze et al., 2004), the squared volume of simplices defined by the selected subsets of columns (known as volume sampling) (Deshpande & Rademacher, 2010), and the SLS (known as **leverage-based sampling** or subspace sampling) (Drineas et al., 2008; Boutsidis et al., 2009).

Depending on whether $\ell > k$ is allowed, the error bounds for CSS are different. If exactly k columns are selected to form C, the best bound was achieved by the rank revealing QR factorization (Gu & Eisenstat, 1996) with the error bound given by:

$$\|A - P_C A\|_2 \leq \sqrt{1 + k(n-k)}\|A - A_k\|_2. \quad (1)$$

with a running time $O(mnk \log n)$. The same error bound was also achieved by using volume sampling (Deshpande & Rademacher, 2010). The running time of volume sampling based algorithms can be made close to linear to the size of the target matrix.

If more than k columns are allowed to be selected, i.e., $\ell > k$, better error bounds can be achieved. In the most recent work by Boutsidis et al. (2011), nearly optimal error bounds were shown by selecting $\ell > k$ columns with a deterministic selection criterion based on the dual set spectral sparsification. In particular, a deterministic polynomial-time algorithm [1] was proposed that achieves the following error bound:

$$\|A - P_C A\|_2 \leq \left(1 + \frac{1 + \sqrt{n/\ell}}{1 - \sqrt{k/\ell}}\right)\|A - A_k\|_2 \quad (2)$$

in $T_{V_k} + O(n\ell k^2)$ time where T_{V_k} is the time needed to compute the top k right singular vectors of A and $O(n\ell k^2)$ is the time needed to compute the selection scores. This bound is close to the lower bound $\Omega\left(\sqrt{\frac{n+\alpha^2}{\ell+\alpha^2}}\right), \alpha > 0$ established in their work. It is worth mentioning that the selection scores in (Boutsidis et al., 2011) computed based on the dual set spectral sparsification is difficult to understand than the SLS.

Although our sampling dependent error bound is not directly comparable to these results, our analysis exhibits that the derived error bound could be better than that in (2) when the SLS are nonuniform. Most importantly, the sampling probabilities in our algorithm are only related to the SLS and that can be computed more efficiently (e.g., exactly in $O(T_{V_k})$ or approximately in $O(mn \log n)$ (Drineas et al., 2012)). In simulations, we observe that the new sampling distributions could yield even better spectral norm reconstruction than the deterministic selection criterion in (Boutsidis et al., 2011), especially when the SLS are nonuniform.

For low rank matrix approximation, several other randomized algorithms have been recently developed. For example, Halko et al. (2011) used a random Gaussian matrix $\Omega \in \mathbb{R}^{n \times \ell}$ or a subsampled random Fourier transform to construct a matrix Ω and then project A into the column

[1] A slower deterministic algorithm with a time complexity $T_{\text{SVD}} + O(\ell n(k^2 + (\rho - k)^2))$ was also presented with an error bound $O(\sqrt{\rho/\ell})\|A - A_k\|_2$, where ρ is the rank of A.

space of $Y = A\Omega$, and they established numerous spectral error bounds. Among them is a comparable error bound $O(\sqrt{n/\ell})\|A - A_k\|_2$ to (2) using the subsampled random Fourier transform. Other randomized algorithm for low rank approximation include CUR decomposition (Drineas et al., 2008; Wang & Zhang, 2012; 2013) and the Nyström based approximation for PSD matrices (Drineas & Mahoney, 2005; Gittens & Mahoney, 2013).

Besides low rank matrix approximation and column selection, CSS has also been successfully applied to least square approximation, leading to faster and interpretable algorithms for over-constrained least square regression. In particular, if let $\Omega \in \mathbb{R}^{\ell\times m}$ denote a scaled sampling matrix corresponding to selecting $\ell < m$ rows from A, the least square problem $\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2$ can be approximately solved by $\min_{\mathbf{x}\in\mathbb{R}^n} \|\Omega A\mathbf{x} - \Omega\mathbf{b}\|_2^2$ (Drineas et al., 2008; 2006b; 2011). Ma et al. (2014) studied CSS for least square approximation from a statistical perspective. They exhibited the expectation and variance of the solution to the approximated least square with uniform sampling and leverage-based sampling. They found that leveraging based estimator could suffer from a large variance when the SLS are very nonuniform while uniform sampling is less vulnerable to very small SLS. This tradeoff is complementary to our observation. However, our observation follows directly from the spectral norm error bound. Moreover, our analysis reveals that the sampling distribution with probabilities proportional to the square root of the SLS is always better than uniform sampling, suggesting that intermediate sampling probabilities between SLS and their square roots by solving a constrained optimization problem could yield better performance than the mixing strategy that linearly combines the SLS and uniform probabilities as suggested in (Ma et al., 2014).

There are much more work on studying the Frobenius norm reconstruction of CSS (Drineas et al., 2006a; Guruswami & Sinop, 2012; Boutsidis et al., 2011; Drineas et al., 2008; Boutsidis et al., 2009). For more references, we refer the reader to the survey (Mahoney, 2011). It remains an interesting question to establish sampling dependent error bounds for other randomized matrix algorithms.

3. Preliminaries

Let $A \in \mathbb{R}^{m\times n}$ be a matrix of size $m\times n$ and have a rank of $\rho \le \min(m, n)$. Let $k < \rho$ be a target rank to approximate A. We write the SVD decomposition of A as

$$A = U \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix}$$

where $\Sigma_1 \in \mathbb{R}^{k\times k}$, $\Sigma_2 \in \mathbb{R}^{(\rho-k)\times(\rho-k)}$, $V_1 \in \mathbb{R}^{n\times k}$ and $V_2 \in \mathbb{R}^{n\times(\rho-k)}$. We use $\sigma_1, \sigma_2, \ldots$ to denote the singular values of A in the descending order, and $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ to denote the maximum and minimum eigenvalues of a PSD matrix X. For any orthogonal matrix $U \in \mathbb{R}^{n\times\ell}$, let $U^\perp \in \mathbb{R}^{n\times(n-\ell)}$ denote an orthogonal matrix whose columns are an orthonormal basis spanning the subspace of $\mathbb{R}^n$ that is orthogonal to the column space of U.

Let $\mathbf{s} = (s_1, \ldots, s_n)$ be a set of scores such that $\sum_{i=1}^n s_i = k$ [2], one for each column of A. We will drawn ℓ independent samples with replacement from the set $[n] = \{1, \ldots, n\}$ using a multinomial distribution where the probability of choosing the ith column is $p_i = s_i/\sum_{j=1}^n s_j$. Let $i_1, \ldots, i_\ell$ be the indices of $\ell > k$ selected columns [3], and $S \in \mathbb{R}^{n\times\ell}$ be the corresponding sampling matrix, i.e,

$$S_{i,j} = \begin{cases} 1, & \text{if } i = i_j \\ 0, & \text{otherwise,} \end{cases}$$

and $D \in \mathbb{R}^{\ell\times\ell}$ be a diagonal rescaling matrix with $D_{jj} = \dfrac{1}{\sqrt{s_{i_j}}}$. Given S, we construct the C matrix as

$$C = AS = (A_{i_1}, \ldots, A_{i_\ell}). \tag{3}$$

Our interest is to bound the spectral norm error between A and $P_C A$ for a given sampling matrix S, i.e., $\|A - P_C A\|_2$, where $P_C A$ projects A onto the column space of C. For the benefit of presentation, we define $\Omega = SD \in \mathbb{R}^{n\times\ell}$ to denote the sampling-and-rescaling matrix, and

$$Y = A\Omega, \quad \Omega_1 = V_1^\top\Omega, \quad \Omega_2 = V_2^\top\Omega, \tag{4}$$

where $\Omega_1 \in \mathbb{R}^{k\times\ell}$ and $\Omega_2 \in \mathbb{R}^{(\rho-k)\times\ell}$. Since the column space of Y is the same to that of C, therefore

$$\|A - P_C A\|_2 = \|A - P_Y A\|_2$$

and we will bound $\|A - P_Y A\|_2$ in our analysis. Let $V_1^\top = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^{k\times n}$ and $V_2^\top = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{R}^{(\rho-k)\times n}$. It is easy to verify that

$$\Omega_1 = (\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_\ell})D, \quad \Omega_2 = (\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_\ell})D$$

Finally, we let $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ denote the SLS of A relative to the best rank-k approximation to A (Mahoney, 2011), i.e., $s_i^* = \|\mathbf{v}_i\|_2^2$. It is not difficult to show that $\sum_{i=1}^n s_i^* = k$.

4. Main Result

Before presenting our main result, we first characterize scores in $\mathbf{s}$ by two quantities as follows:

$$c(\mathbf{s}) = \max_{1\le i\le n} \frac{s_i^*}{s_i}, \quad q(\mathbf{s}) = \max_{1\le i\le n} \frac{\sqrt{s_i^*}}{s_i} \tag{5}$$

Both quantities compare $\mathbf{s}$ to the SLS $\mathbf{s}^*$. With $c(\mathbf{s})$ and $q(\mathbf{s})$, we are ready to present our main theorem regarding the spectral error bound.

Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$ have rank ρ and $C \in \mathbb{R}^{m\times\ell}$*

[2] For the sake of discussion, we are not restricting the sum of these scores to be one but to be k, which does not affect our conclusions.

[3] Note that some of the selected columns could be duplicate.

contain the selected columns according to sampling scores in $\mathbf{s}$*. With a probability* $1-\delta-2k\exp(-\ell/[8kc(\mathbf{s})])$*, we have*

$$\|A-P_CA\|_2 \leq \sigma_{k+1}(1+\epsilon(\mathbf{s}))$$

where $\epsilon(\mathbf{s})$ *is*

$$\epsilon(\mathbf{s}) = 3\left[\sqrt{c(\mathbf{s})\frac{k(\rho+1-k)\log\left[\frac{\rho}{\delta}\right]}{\ell}} + q(\mathbf{s})\frac{k\log\left[\frac{\rho}{\delta}\right]}{\ell}\right]$$

where $\sigma_{k+1} = \|A-A_k\|_2$ *is the* $(k+1)$*th singular value of* A*.*

Remark: The proof is deferred to Section 6. Clearly, the spectral error bound and the successful probability in Theorem 1 depend on the quantities $c(\mathbf{s})$ and $q(\mathbf{s})$. In the subsection below, we study the two quantities to facilitate the understanding of the result in Theorem 1.

4.1. More about the two quantities and their tradeoffs

The result in Theorem 1 implies that the smaller the quantities $c(\mathbf{s})$ and $q(\mathbf{s})$, the better the error bound. Therefore, we first study when $c(\mathbf{s})$ and $q(\mathbf{s})$ achieve their minimum values. The key results are presented in the following two lemmas with their proofs deferred to the supplement.

Lemma 1. *The set of scores in* $\mathbf{s}$ *that minimize* $q(\mathbf{s})$ *is given by* $s_i \propto \sqrt{s_i^*}$*, i.e.,* $s_i = \frac{k\sqrt{s_i^*}}{\sum_{i=1}^n \sqrt{s_i^*}}$*.*

Remark: The sampling distribution with probabilities that are proportional to the square root of $s_i^*, i\in[n]$ falls in between the uniform sampling and the leverage-based sampling.

Lemma 2. $c(\mathbf{s}) \geq 1, \forall \mathbf{s}$ *such that* $\sum_{i=1}^m s_i = k$*. The set of scores in* $\mathbf{s}$ *that minimize* $c(\mathbf{s})$ *is given by* $s_i = s_i^*$*, and the minimum value of* $c(\mathbf{s})$ *is* 1*.*

Next, we discuss three special samplings with $\mathbf{s}$ (i) proportional to the square root of the SLS, i.e., $s_i \propto \sqrt{s_i^*}$ (referred to as square-root leverage-based sampling or **sqL-sampling** for short), (ii) equal to the SLS, i.e., $s_i = s_i^*$ (referred to as leverage-based sampling or **L-sampling** for short), and (iii) equal to uniform scalars $s_i = k/n$ (referred to as uniform sampling or **U-sampling** for short).

sqL-sampling. Firstly, if $s_i \propto \sqrt{s_i^*}$, $q(\mathbf{s})$ achieves its minimum value and we have the two quantities written as

$$\begin{aligned} q_{sqL} &= \frac{1}{k}\sum_{i=1}^n \sqrt{s_i^*} \\ c_{sqL} &= \max_i \frac{s_i^*\sum_i\sqrt{s_i^*}}{k\sqrt{s_i^*}} = q_{sqL}\max_i\sqrt{s_i^*} \end{aligned} \tag{6}$$

In this case, when $\mathbf{s}^*$ is flat (all SLS are equal), then $q_{sqL} = \sqrt{\frac{n}{k}}$ and $c^{sqL} = 1$. The bound becomes $\widetilde{O}(\sqrt{(\rho+1-k)k/\ell}+\sqrt{nk/\ell^2})\sigma_{k+1}$ that suppresses logarithmic terms. To analyze q_{sqL} and c_{sqL} for skewed SLS, we consider a power-law distributed SLS, i.e., there exists a small constant a and power index $p>2$, such that $s_{[i]}^*, i=1,\ldots,n$ ranked in descending order satisfy

$$s_{[i]}^* \leq a^2 i^{-p}, \quad i=1,\ldots,n \tag{7}$$

Then it is not difficult to show that

$$\frac{1}{k}\sum_{i=1}^n\sqrt{s_i^*} \leq \frac{a}{k}\left(1+\frac{2}{p-2}\right)$$

which is independent of n. Then the error bound in Theorem 1 becomes $O\left(\sqrt{\frac{\rho+1-k}{\ell}}+\frac{1}{\ell}\right)\sigma_{k+1}$, which is better than that in (2). This result is summarized in Corollary 3 at the end of this subsection.

L-sampling. Secondly, if $s_i = s_i^*$, then $c(\mathbf{s})$ achieves its minimum value and we have the two quantities written as

$$q_L = \max_i \frac{1}{\sqrt{s_i^*}}, \quad c_L = 1 \tag{8}$$

In this case, when $\mathbf{s}^*$ is flat, we have $q_L = \sqrt{\frac{n}{k}}$ and $c_L = 1$ and the same bound $\widetilde{O}(\sqrt{(\rho+1-k)k/\ell}+\sqrt{nk/\ell^2})\sigma_{k+1}$ follows. However, when $\mathbf{s}^*$ is skewed, i.e., there exist very small SLS, then q_L could be very large. As a comparison, the $q(\mathbf{s})$ for sqL-sampling is always smaller than that for L-sampling due to the following inequality

$$\begin{aligned} q_{sqL} &= \frac{1}{k}\sum_{i=1}^n\sqrt{s_i^*} = \frac{1}{k}\sum_{i=1}^n\frac{s_i^*}{\sqrt{s_i^*}} < \max_i\frac{1}{\sqrt{s_i^*}}\frac{\sum_{i=1}^n s_i^*}{k} \\ &= \max_i\frac{1}{\sqrt{s_i^*}} = q_L \end{aligned}$$

U-sampling. Lastly, we consider the uniform sampling $s_i = \frac{k}{n}$. Then the two quantities become

$$q_U = \max_i\frac{n\sqrt{s_i^*}}{k}, \quad c_U = \max_i\frac{ns_i^*}{k} \tag{9}$$

Similarly, if $\mathbf{s}_*$ is flat, $q_U = \sqrt{\frac{n}{k}}$ and $c_U = 1$. Moreover, it is interesting to compare the two quantities for the sqL-sampling in (6) and for the uniform sampling in (9).

$$q_{sqL} = \frac{1}{k}\sum_{i=1}^n\sqrt{s_i^*} \leq \max_i\frac{n\sqrt{s_i^*}}{k} = q_U$$

$$c_{sqL} = \max_i\frac{1}{k}\sqrt{s_i^*}\sum_{i=1}^n\sqrt{s_i^*} \leq \max_i\frac{ns_i^*}{k} = c_U$$

From the above discussions, we can see that when $\mathbf{s}_*$ is a flat vector, there is no difference between the three sampling scores for $\mathbf{s}$. The difference comes from when $\mathbf{s}_*$ tends to be skewed. In this case, $s_i \propto \sqrt{s_i^*}$ works almost for sure better than uniform distribution and could also be potentially better than $s_i = s_i^*$ according to the sampling dependent error bound in Theorem 1. A similar tradeoff between the L-sampling and U-sampling but with a different taste was observed in (Ma et al., 2014), where they showed that for least square approximation the CSS leveraging-based least square estimator could have a large variance when there exist very small SLS. Nonetheless, our bound here exhibits more insights, especially on the sqL-sampling. More importantly, the sampling dependent bound renders the flexibility in choosing the sampling scores by adjusting them according to the distribution of the SLS. In next subsection, we present an optimization approach to find better sampling scores. In Figure 1, we give a quick view of different sampling strategies.

Corollary 2. *Let $A \in \mathbb{R}^{m\times n}$ have rank ρ and $C \in \mathbb{R}^{m\times \ell}$ contain the selected columns according to sampling scores in $\sqrt{\mathbf{s}_*}$. Assume the scores in $\mathbf{s}_*$ follow a power-law distribution as in (7) with $p > 2$ and $\ell \ge 8a^2(1+2/(p-2))\log(k/\delta)$. With a probability at least $1-3\delta$, we have*

$$\|A - P_C A\|_2 \le \sigma_{k+1} O\left[\sqrt{\frac{(\rho+1-k)\log[\frac{\rho}{\delta}]}{\ell}} + \frac{\log[\frac{\rho}{\delta}]}{\ell}\right]$$

where the big $O(\cdot)$ suppresses the dependence on a and p.

4.2. Optimizing the error bound

As indicated by the result in Theorem 1, in order to achieve a good performance, we need to make a balance between $c(\mathbf{s})$ an $q(\mathbf{s})$, where $c(\mathbf{s})$ affects not only the error bound but also the successful probability. To address this issue, we propose a constrained optimization approach. More specifically, to ensure that the failure probability is no more than 3δ, we impose the following constraint on $c(\mathbf{s})$

$$\frac{\ell}{8kc(\mathbf{s})} \ge \log\left(\frac{k}{\delta}\right), \ i.e., \ \max_i \frac{s_i^*}{s_i} \le \frac{\ell}{8k\log\left(\frac{k}{\delta}\right)} := \gamma \qquad (10)$$

Then we cast the problem into minimizing $q(\mathbf{s})$ under the constraint in (10), i.e.,

$$\min_{\mathbf{s}\in\mathbb{R}^n_+} \max_{1\le i\le n} \frac{\sqrt{s_i^*}}{s_i}$$
$$\text{s.t.} \quad \mathbf{s}^\top \mathbf{1} = k, \ s_i^* \le \gamma s_i, i = 1,\ldots,n \qquad (11)$$

It is easy to verify that the optimization problem in (11) is convex. Next, we develop an efficient bisection search algorithm to solve the above problem with a linear convergence rate. To this end, we introduce a slack variable t and rewrite the optimization problem in (11) as

$$\min_{\mathbf{s}\in\mathbb{R}^n_+, t\ge 0} \quad t, \quad \text{s.t.} \quad \mathbf{s}^\top\mathbf{1} = k$$
$$\text{and} \quad \frac{s_i^*}{s_i} \le \min\left(\gamma, t\sqrt{s_i^*}\right), i = 1,\ldots,n \qquad (12)$$

We now find the optimal solution by performing bisection search on t. Let $t_{\max}$ and $t_{\min}$ be the upper and lower bounds for t. We set $t = (t_{\min}+t_{\max})/2$ and decide the feasibility of t by simply computing the quantity

$$f(t) = \sum_{i=1}^n \frac{s_i^*}{\min\left(\gamma, t\sqrt{s_i^*}\right)}$$

Evidently, t is a feasible solution if $f(t) \le k$ and is not if $f(t) > k$. Hence, we will update $t_{\max} = t$ if $f(t) \le k$ and $t_{\min} = t$ if $f(t) > k$. To run the bisection algorithm, we need to decide initial $t_{\min}$ and $t_{\max}$. We can set $t_{\min} = 0$. To compute $t_{\max}$, we make an explicit construction of $\mathbf{s}$ by distributing the $(1-\gamma^{-1})$ share of the largest element of $\mathbf{s}_*$ to the rest of the list. More specifically, let j be the index for the largest entry in $\mathbf{s}^*$. We set $s_j = \|\mathbf{s}^*\|_\infty \gamma^{-1}$ and $s_i = s_i^* + (1-\gamma^{-1})\|\mathbf{s}^*\|_\infty/(n-1)$ for $i \neq j$. Evidently, this solution satisfies the constraints $s_i^* \le \gamma s_i, i \in [n]$ for $\gamma \ge 1$. With this construction, we can show that

$$q(\mathbf{s}) \le \max\left(\frac{\gamma}{\sqrt{\|\mathbf{s}^*\|_\infty}}, \frac{n-1}{\sqrt{\|\mathbf{s}^*\|_\infty}(1-\gamma^{-1})}\right)$$

Therefore, we set initial $t_{\max}$ to the value in R.H.S of the above inequality. Given the optimal value of $t = t_*$ we compute the optimal value of s_i by $s_i = \frac{s_i^*}{\min(\gamma, t_*\sqrt{s_i^*})}$. The corresponding sampling distribution clearly lies between L-sampling and sqL-sampling. In particular, when $\gamma = 1$ the resulting sampling distribution is L-sampling due to Lemma 2 and when $\gamma \to \infty$ the resulting sampling distribution approaches sqL-sampling.

Finally, we comment on the value of ℓ. In order to make the constraint in (10) feasible, we need to ensure $\gamma \ge 1$. Therefore, we need $\ell \ge \Omega(k\log\left(\frac{k}{\delta}\right))$.

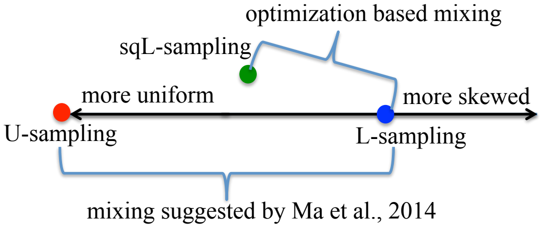
optimization based mixing
sqL-sampling
more uniform
more skewed
U-sampling
L-sampling
mixing suggested by Ma et al., 2014

Figure 1. An illustration of different sampling strategies. The mixing strategy suggested by (Ma et al., 2014) is a convex combination of U-sampling and L-sampling. Our optimization approach gives an intermediate sampling between the sqL-sampling and the L-sampling.

4.3. Subsequent Applications

Next, we discuss two subsequent applications of CSS, one for low rank approximation and one for least square approximation.

Rank-k approximation. If a rank-k approximation is desired, we need to do some postprocessing since $P_C A$ might have rank larger than k. We can use the same algorithm as presented in (Boutsidis et al., 2011). In particular, given the constructed $C \in \mathbb{R}^{n\times\ell}$, we first orthonormalize the columns of C to construct a matrix $Q \in \mathbb{R}^{m\times\ell}$ with orthonormal columns, then compute the best rank-k approximation of $Q^\top A \in \mathbb{R}^{\ell\times n}$ denoted by $(Q^\top A)_k$, and finally construct the low-rank approximation as $Q(Q^\top A)_k$. It was shown that (Lemma 2.3 in (Boutsidis et al., 2011))

$$\|A - Q(Q^\top A)_k\|_2 \le \sqrt{2}\|A - \Pi^2_{C,k}(A)\|_2$$

where $\Pi^2_{C,k}(A)$ is the best approximation to A within the column space of C that has rank at most k. The running time of above procedure is $O(mn\ell + (m+n)\ell^2)$. Regarding its error bound, the above inequality together with the following theorem implies that its spectral error bound is only amplified by a factor of $\sqrt{2}$ compared to that of $P_C A$.

Theorem 3. *Let $A \in \mathbb{R}^{m\times n}$ have rank ρ and $C \in \mathbb{R}^{m\times \ell}$ contain the selected columns according to sampling scores in* $\mathbf{s}$. *With a probability $1-\delta-2k\exp(-\ell/[8kc(\mathbf{s})])$, we have*

$$\|A - \Pi^2_{C,k}(A)\|_2 \le \sigma_{k+1}(1+\epsilon(\mathbf{s}))$$

where $\epsilon(\mathbf{s})$ is given in Theorem 1.

Least Square Approximation. CSS has been used in least square approximation for developing faster and interpretable algorithms. In these applications, an over-constrained least square problem is considered, i.e., given $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ with $m \gg n$, to solve the following problem:

$$\mathbf{x}_{opt} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x}-\mathbf{b}\|_2^2 \tag{13}$$

The procedure for applying CSS to least square approximation is (i) to sample a set of $\ell > n$ rows from A and form a sampling-and-rescaling matrix denoted by $\Omega \in \mathbb{R}^{\ell\times m}$ [4]; (ii) to solve the following reduced least square problem:

$$\widehat{\mathbf{x}}_{opt} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\Omega A\mathbf{x}-\Omega\mathbf{b}\|_2^2 \tag{14}$$

It is worth pointing out that in this case the SLS $\mathbf{s}^* = (s_1^*,\ldots,s_m^*)$ are computed based on the the left singular vectors U of A by $s_i^* = \|U_{i*}\|_2^2$, where U_{i*} is the i-th row of U. One might be interested to see whether we can apply our analysis to derive a sampling dependent error bound for the approximation error $\|\mathbf{x}_{opt}-\widehat{\mathbf{x}}_{opt}\|_2$ similar to previous bounds of the form $\|\mathbf{x}_{opt}-\widehat{\mathbf{x}}_{opt}\|_2 \le \frac{\epsilon}{\sigma_{min}(A)}\|A\mathbf{x}_{top}-\mathbf{b}\|_2$. Unfortunately, naively combining our analysis with previous analysis is a worse case analysis, and consequentially yields a worse bound. The reason will become clear in our later discussions in Section 7. However, the statistical analysis in (Ma et al., 2014) does indicate that $\widehat{\mathbf{x}}_{opt}$ by using sqL-sampling could have smaller variance than that using L-sampling.

5. Numerical Experiments

Before delving into the detailed analysis, we present some experimental results. We consider synthetic data with the data matrix A generated from one of the three different classes of distributions introduced below, allowing the SLS vary from nearly uniform to very nonuniform.

- Nearly uniform SLS (GA). Columns of A are generated from a multivariate normal distribution $\mathcal{N}(\mathbf{1}_m, \Sigma)$, where $\Sigma_{ij} = 2 * 0.5^{|i-j|}$. This data is referred to as GA data.

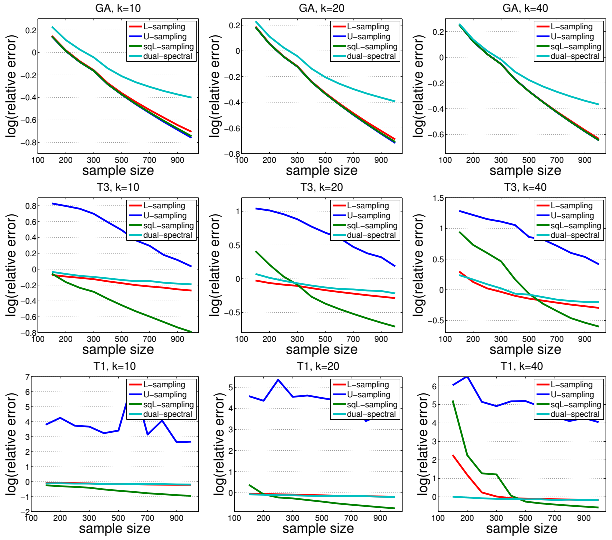

Figure 2. Comparison of the spectral error for different data, different samplings, different target rank and different sample size.

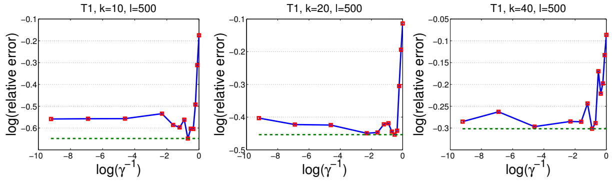

Figure 3. The spectral error for the sampling probabilities found by the constrained optimization approach with different values of $\gamma \ge 1$. The left most point corresponds to sqL-sampling and the right most point corresponds to L-sampling.

- Moderately nonuniform SLS (T_3). Columns of A are generated from a multivariate t-distribution with 3 degree of freedom and covariance matrix Σ as before. This data is referred to as T_3 data.

- Very nonuniform SLS (T_1). Columns of A are generated from a multivariate t-distribution with 1 degree of freedom and covariance matrix Σ as before. This data is referred to as T_1 data.

These distributions have been used in (Ma et al., 2014) to generate synthetic data for empirical evaluations.

We first compare the spectral norm reconstruction error of the three different samplings, namely L-sampling, U-sampling and the sqL-sampling, and the deterministic dual set spectral sparsification algorithm. We generate synthetic data with $n = m = 1000$ and repeat the experiments 1000 times. We note that the rank of the generated data matrix is 1000. The averaged results are shown in Figure 2. From these results we observe that (i) when the SLS are nearly uniform, the three sampling strategies perform similarly as expected; (ii) when the SLS become nonuniform, sqL-sampling performs always better than U-sampling and bet-

[4] We abuse the same notation Ω.

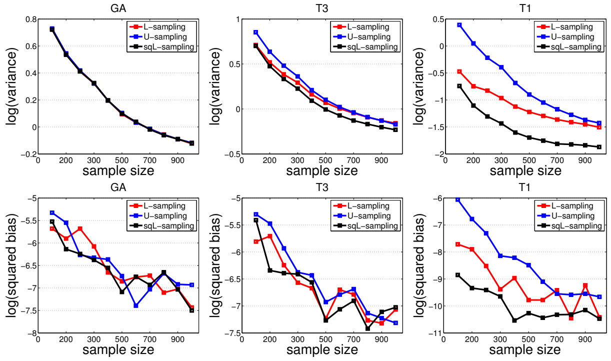

Figure 4. Comparison of variance and squared bias of the estimators for different data, different samplings and different sample size.

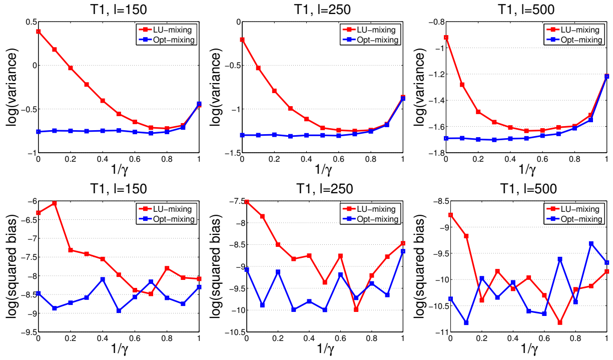

Figure 5. Comparison of variance and squared bias of the estimators for different mixing strategies. Opt refers to our optimization based approach and LU refers to a convex combination of L-sampling and U-sampling with γ^{-1} as the combination weight.

ter than the L-sampling when the target rank is small (e.g., $k = 10$) or the sample size ℓ is large; (iii) when the SLS are non-uniform, the spectral norm reconstruction error of sqL-sampling decreases faster than L-sampling w.r.t the sample size ℓ; (iv) randomized algorithms generally perform better than the deterministic dual set sparsification algorithm.

Second, we compare the sampling scores found the constrained optimization with L-sampling and sqL-sampling. We vary the value of γ from 1 (corresponding to L-sampling) to ∞ (corresponding to sqL-sampling). A result with sampling size $\ell = 500$ is shown in Figure 3. It demonstrates that intermediate samplings found by the proposed constrained optimization can perform better than both L-sampling and sqL-sampling.

Finally, we apply CSS to over-constrained least square regression. To this end, we generate a synthetic data matrix $A \in \mathbb{R}^{m\times n}$ with $m = 50$ and $n = 1000$ similarly to (Ma et al., 2014). The output is generated by $y = A^\top\beta + \epsilon$ where $\epsilon \sim (0, 9I_n)$ and $\beta = (\mathbf{1}_{10}, 0.1\mathbf{1}_{30}, \mathbf{1}_{10})^\top$. We compare the variance and bias of the obtained estimators over 1000 runs for different sampling distributions. The results shown in Figure 4 demonstrate the sqL-sampling gives smaller variance and better bias of the estimators than L-sampling and U-sampling. We also compare the proposed optimization approach with the simple mixing strategy (Ma et al., 2014) that uses a convex combination of the L-sampling and the U-sampling. The results are shown in Figure 5, which again support our approach.

More results including relative error versus varying size n of the target matrix and the Frobenius norm reconstruction error can be found in supplement.

6. Analysis

In this section, we present major analysis steps leading to Theorem 1 and Theorem 3 with detailed proofs included in supplement. The key to our analysis is the following lemma.

Lemma 3. *Let Y, Ω_1, Ω_2 be defined in (4). Assume that Ω_1 has full row rank. We have*

$$\|A - P_Y A\|_\xi^2 \le \|\Sigma_2\|_\xi^2 + \left\|\Sigma_2\Omega_2\Omega_1^\dagger\right\|_\xi^2$$

and

$$\left\|A - \Pi_{Y,k}^2(A)\right\|_\xi^2 \le \|\Sigma_2\|_\xi^2 + \left\|\Sigma_2\Omega_2\Omega_1^\dagger\right\|_\xi^2$$

where ξ could be 2 and F.

The first inequality was proved in (Halko et al., 2011) (Theorem 9.1) and the second inequality is credited to (Boutsidis et al., 2011) (Lemma 3.2) [5]. Previous work on the spectral norm analysis also start from a similar inequality as above. They bound the second term by using $\|\Sigma_2\Omega_2\Omega_1^\dagger\|_2 \le \|\Sigma_2\Omega_2\|_2\|\Omega_1^\dagger\|_2$ and then bound the two terms separately. However, we will first write $\left\|\Sigma_2\Omega_2\Omega_1^\dagger\right\|_2 = \|\Sigma_2\Omega_2\Omega_1^\top(\Omega_1\Omega_1^\top)^{-1}\|_2$ using the fact Ω_1 has full row rank, and then bound $\|(\Omega_1\Omega_1^\top)^{-1}\|_2$ and $\|\Omega_2\Omega_1^\top\|_2$ separately. To this end, we will apply the Matrix Chernoff bound as stated in Lemma 4 to bound $\|(\Omega_1\Omega_1^\top)^{-1}\|_2$ and apply the matrix Bernstein inequality as stated in Lemma 5 to bound $\|\Omega_2\Omega_1^\top\|_2$.

Lemma 4 (Matrix Chernoff (Tropp, 2012))**.** *Let $\mathcal{X}$ be a finite set of PSD matrices with dimension k, and suppose that $\max_{X\in\mathcal{X}} \lambda_{\max}(X) \le B$. Sample $\{X_1, \ldots, X_\ell\}$ independently from $\mathcal{X}$. Compute*

$$\mu_{\max} = \ell\lambda_{\max}(\mathrm{E}[X_1]), \quad \mu_{\min} = \ell\lambda_{\min}(\mathrm{E}[X_1])$$

Then

$$\Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{\ell} X_i\right) \ge (1+\delta)\mu_{\max}\right\} \le k\left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^{\frac{\mu_{\max}}{B}}$$

$$\Pr\left\{\lambda_{\min}\left(\sum_{i=1}^{\ell} X_i\right) \le (1-\delta)\mu_{\min}\right\} \le k\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu_{\min}}{B}}$$

Lemma 5 (Noncommutative Bernstein Inequality (Recht,

[5] In fact, the first inequality is implied by the second inequality.

2011)). *Let $Z_1, \ldots, Z_L$ be independent zero-mean random matrices of dimension $d_1 \times d_2$. Suppose $\tau_j^2 = \max\left\{\|\mathrm{E}[Z_j Z_j^\top]\|_2, \|\mathrm{E}[Z_j^\top Z_j\|_2\right\}$ and $\|Z_j\|_2 \leq M$ almost surely for all k. Then, for any $\epsilon > 0$,*

$$\Pr\left[\left\|\sum_{j=1}^{L} Z_j\right\|_2 > \epsilon\right] \leq (d_1+d_2)\exp\left[\frac{-\epsilon^2/2}{\sum_{j=1}^{L}\tau_j^2 + M\epsilon/3}\right]$$

Following immediately from Lemma 3, we have

$$\begin{aligned}\|A - P_Y A\|_2 &\leq \sigma_{k+1}\sqrt{1 + \|\Omega_2\Omega_1^\top(\Omega_1\Omega_1^\top)^{-1}\|_2^2} \\ &\leq \sigma_{k+1}\sqrt{1 + \|\Omega_2\Omega_1^\top\|_2^2\lambda_{\min}^{-2}(\Omega_1\Omega_1^\top)} \\ &\leq \sigma_{k+1}(1 + \|\Omega_2\Omega_1^\top\|_2\lambda_{\min}^{-1}(\Omega_1\Omega_1^\top)),\end{aligned}$$

where the last inequality uses the fact $\sqrt{a^2+b^2} \leq a + b$ for $a, b \geq 0$. Below we bound $\lambda_{\min}(\Omega_1\Omega_1^\top)$ from below and bound $\|\Omega_2\Omega_1^\top\|_2$ from above.

6.1. Bounding $\|(\Omega_1\Omega_1^\top)^{-1}\|_2$

We will utilize Lemma 4 to bound $\lambda_{\min}(\Omega_1\Omega_1^\top)$. Define $X_i = \mathbf{v}_i\mathbf{v}_i^\top/s_i$. It is easy to verify that

$$\Omega_1\Omega_1^\top = \sum_{j=1}^{\ell}\frac{1}{s_{i_j}}\mathbf{v}_{i_j}\mathbf{v}_{i_j}^\top = \sum_{j=1}^{\ell} X_{i_j}$$

and $\mathrm{E}[X_{i_j}] = \frac{1}{\sum_{i=1}^n s_i}\sum_{i=1}^n s_i X_i = \frac{1}{k}I_k$, where we use $\sum_{j=1}^n s_j = k$ and $V_1^\top V_1 = I_k$. Therefore we have $\lambda_{\min}(\mathrm{E}[X_{i_j}]) = \frac{1}{k}$. Then the theorem below will follow Lemma 4.

Theorem 4. *With a probability $1 - k\exp(-\delta^2\ell/[2kc(\mathbf{s})])$, we have $\lambda_{\min}(\Omega_1\Omega_1^\top) \geq (1-\delta)\frac{\ell}{k}$.*

Therefore, with a probability $1 - k\exp(-\delta^2\ell/[2kc(\mathbf{s})])$ we have $\|(\Omega_1\Omega_1^\top)^{-1}\|_2 \leq \frac{1}{1-\delta}\frac{k}{\ell}$.

6.2. Bounding $\|\Omega_2\Omega_1^\top\|_2$

We will utilize Lemma 5 to bound $\|\Omega_2\Omega_1^\top\|_2$. Define $Z_j = \mathbf{u}_{i_j}\mathbf{v}_{i_j}^\top/s_{i_j}$. Then

$$\Omega_2\Omega_1^\top = \sum_{j=1}^{\ell}\frac{1}{s_{i_j}}\mathbf{u}_{i_j}\mathbf{v}_{i_j}^\top = \sum_{j=1}^{l} Z_j$$

and $\mathrm{E}[Z_j] = 0$. In order to use the matrix Bernstein inequality, we will bound $\max_i \|Z_i\|_2 = \max_i \frac{\|\mathbf{u}_i\mathbf{v}_i^\top\|_2}{s_i} \leq q(\mathbf{s})$ and $\tau_j^2 \leq \frac{(\rho+1-k)c(\mathbf{s})}{k}$. Then we can prove the following theorem.

Theorem 5. *With a probability $1-\delta$, we have*

$$\|\Omega_2\Omega_1^\top\|_2 \leq \sqrt{2c(\mathbf{s})\frac{(\rho+1-k)\ell\log(\frac{\rho}{k})}{k}} + \frac{2q(\mathbf{s})\log(\frac{\rho}{k})}{3}.$$

We can complete the proof of Theorem 1 by combining the bounds for $\|\Omega_2\Omega_1^\top\|_2$ and $\lambda_{\min}^{-1}(\Omega_1\Omega_1^\top)$ and by setting $\delta = 1/2$ in Theorem 4 and using union bounds.

7. Discussions and Open Problems

From the analysis, it is clear that the matrix Bernstein inequality is the key to derive the sampling dependent bound for $\|\Omega_2\Omega_1^\top\|_2$. For bounding $\lambda_{\min}(\Omega_1\Omega_1^\top)$, similar analysis using matrix Chernoff bound has been exploited before for randomized matrix approximation (Gittens, 2011).

Since Lemma 3 also holds for the Frobenius norm, it might be interested to see whether we can derive a sampling dependent Frobenius norm error bound that depends on $c(\mathbf{s})$ and $q(\mathbf{s})$, which, however, still remains as an open problem for us. Nonetheless, in experiments (included in the supplement) we observe similar phenomena about the performance of L-sampling, U-sampling and sqL-sampling.

Finally, we briefly comment on the analysis for least square approximation using CSS. Previous results (Drineas et al., 2008; 2006b; 2011) were built on the structural conditions that are characterized by two inequalities

$$\begin{aligned}&\lambda_{\min}(\Omega UU^\top\Omega) \geq 1/\sqrt{2} \\ &\|U^\top\Omega^\top\Omega U^\perp U^{\perp\top}\mathbf{b}\|_2^2 \leq \frac{\epsilon}{2}\|U^\perp U^{\perp\top}\mathbf{b}\|_2^2\end{aligned}$$

The first condition can be guaranteed by Theorem 4 with a high probability. For the second condition, if we adopt a worse case analysis

$$\|U^\top\Omega^\top\Omega U^\perp U^{\perp\top}\mathbf{b}\|_2^2 \leq \|U^\top\Omega^\top\Omega U^\perp\|_2^2\|U^{\perp\top}\mathbf{b}\|_2^2$$

and bound the first term in R.H.S of the above inequality using Theorem 5, we would end up with a worse bound than existing ones that bound the left term as a whole. Therefore the naive combination can't yield a good sampling dependent error bound for the approximation error of least square regression.

8. Conclusions

In this paper, we have presented a sampling dependent spectral error bound for CSS. The error bound brings a new distribution with sampling probabilities proportional to the square root of the statistical leverage scores and exhibits more tradeoffs and insights than existing error bounds for CSS. We also develop a constrained optimization algorithm with an efficient bisection search to find better sampling probabilities for the spectral norm reconstruction. Numerical simulations demonstrate that the new sampling distributions lead to improved performance.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful and insightful comments. T. Yang was supported in part by NSF (IIS-1463988). R. Jin was partially supported by NSF IIS-1251031 and ONR N000141410631.

References

Boutsidis, Christos, Mahoney, Michael W., and Drineas, Petros. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 968–977, 2009.

Boutsidis, Christos, Drineas, Petros, and Magdon-Ismail, Malik. Near optimal column-based matrix reconstruction. In *The Annual Symposium on Foundations of Computer Science*, pp. 305–314, 2011.

Deshpande, Amit and Rademacher, Luis. Efficient volume sampling for row/column subset selection. *CoRR*, abs/1004.4057, 2010.

Drineas, Petros and Mahoney, Michael W. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2005, 2005.

Drineas, Petros, Mahoney, Michael W., and Muthukrishnan, S. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, volume 4110, pp. 316–326, 2006a.

Drineas, Petros, Mahoney, Michael W., and Muthukrishnan, S. Sampling algorithms for l2 regression and applications. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1127–1136, 2006b.

Drineas, Petros, Mahoney, Michael W., and Muthukrishnan, S. Relative-error cur matrix decompositions. *SIAM Journal Matrix Analysis Applications*, 30:844–881, 2008.

Drineas, Petros, Mahoney, Michael W., Muthukrishnan, S., and Sarlós, Tamàs. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, February 2011.

Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W., and Woodruff, David P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

Frieze, Alan, Kannan, Ravi, and Vempala, Santosh. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of ACM*, 51(6):1025–1041, 2004.

Gittens, Alex. The spectral norm errors of the naive nystrom extension. *CoRR*, abs/1110.5305, 2011.

Gittens, Alex and Mahoney, Michael W. Revisiting the nystrom method for improved large-scale machine learning. In *Proceedings of International Conference of Machine Learning*, volume 28, pp. 567–575, 2013.

Gu, Ming and Eisenstat, Stanley C. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

Guruswami, Venkatesan and Sinop, Ali Kemal. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1207–1214, 2012.

Halko, N., Martinsson, P. G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:217–288, 2011.

Ma, Ping, Mahoney, Michael W., and Yu, Bin. A statistical perspective on algorithmic leveraging. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pp. 91–99, 2014.

Mahoney, Michael W. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

Meyer, Carl D. (ed.). *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000. ISBN 0-89871-454-0.

Pan, C.-T. On the existence and computation of rank-revealing lu factorizations. *Linear Algebra and its Applications*, 316(13):199 – 222, 2000. Special Issue: Conference celebrating the 60th birthday of Robert J. Plemmons.

Pan, Ching-Tsuan and Tang, PingTakPeter. Bounds on singular values revealed by qr factorizations. *BIT Numerical Mathematics*, 39(4):740–756, 1999. ISSN 0006-3835.

Recht, Benjamin. A simpler approach to matrix completion. *Journal Machine Learning Research (JMLR)*, pp. 3413–3430, 2011.

Tropp, Joel A. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, August 2012. ISSN 1615-3375.

Wang, Shusen and Zhang, Zhihua. A scalable cur matrix decomposition algorithm: Lower time complexity and tighter bound. In *Advances in Neural Information Processing Systems 25*, pp. 656–664. 2012.

Wang, Shusen and Zhang, Zhihua. Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14(1):2729–2769, 2013.