Revisiting the Nyström method for improved large-scale machine learning

Alex Gittens GITTENS@CALTECH.EDU
Dept. of Applied and Computational Math, CalTech, Pasadena CA 91125 USA

Michael W. Mahoney MMAHONEY@CS.STANFORD.EDU
Dept. of Mathematics, Stanford University, Stanford, CA 9430 USA

Abstract

We reconsider randomized algorithms for the low-rank approximation of SPSD matrices such as Laplacian and kernel matrices that arise in data analysis and machine learning applications. Our main results consist of an empirical evaluation of the performance quality and running time of sampling and projection methods on a diverse suite of SPSD matrices. Our results highlight complementary aspects of sampling versus projection methods, and they point to differences between uniform and nonuniform sampling methods based on leverage scores. We complement our empirical results with a suite of worst-case theoretical bounds for both random sampling and random projection methods. These bounds are qualitatively superior to existing bounds—*e.g.*, improved additive-error bounds for spectral and Frobenius norm error and relative-error bounds for trace norm error.

1. Introduction

We reconsider randomized algorithms for the low-rank approximation of SPSD matrices such as Laplacian and kernel matrices that arise in data analysis and machine learning applications. Of particular interest are sampling-based versus projection-based methods as well as the use of uniform sampling versus nonuniform sampling based on the leverage score probabilities. Our main contributions are fourfold.

First, we provide an empirical evaluation of the complementary strengths and weaknesses of data-independent random projection methods and data-dependent random sampling methods when applied to SPSD matrices. We do so for a diverse class of SPSD matrices drawn from machine learning and more general data analysis applications, and we consider reconstruction error with respect to the spectral, Frobenius, as well as trace norms.

Second, we consider the running time of high-quality sampling and projection algorithms. By exploiting and extending recent work on "fast" random projections and related recent work on "fast" approximation of the statistical leverage scores, we illustrate that high-quality leverage-based random sampling and high-quality random projection algorithms have comparable running times.

Third, our main technical contribution is a set of deterministic structural results that hold for any "sketching matrix" applied to an SPSD matrix. (A precise statement of these results is given in Theorems 1, 2, and 3 in Section 4.1.) We call these "deterministic structural results" since there is no randomness involved in their statement or analysis and since they depend on structural properties of the input data matrix and the way the sketching matrix interacts with the input data.

Fourth, our main algorithmic contribution is to show that when the low-rank sketching matrix represents certain random projections or random sampling operations, then (by using our deterministic structural conditions) we obtain worst-case quality-of-approximation bounds that hold with high probability. (A precise statement of these results is given in Lemmas 1, 2, 3, and 4 in Section 4.2.) These bounds are qualitatively better than existing bounds (when nontrivial prior bounds even exist).

Our analysis is timely for at least two reasons. First, existing theory for the Nyström method is quite mod-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

est. For example, existing worst-case bounds such as those of (Drineas & Mahoney, 2005) are very weak, especially compared with existing bounds for least-squares regression and general low-rank matrix approximation problems (Drineas et al., 2008; 2010; Mahoney, 2011). Moreover, many other worst-case bounds make strong assumptions about the coherence properties of the input data (Kumar et al., 2012; Gittens, 2011). Second, there have been conflicting views about the usefulness of uniform sampling versus nonuniform sampling based on the empirical statistical leverage scores of the data in realistic data analysis and machine learning applications. For example, some work has concluded that the statistical leverage scores of realistic data matrices are fairly uniform, meaning that the coherence is small and thus uniform sampling is appropriate (Williams & Seeger, 2001; Kumar et al., 2012), while other work has demonstrated that leverage scores are often very nonuniform in ways that render uniform sampling inappropriate and that can be essential to highlight properties of downstream interest (Paschou et al., 2007; Mahoney & Drineas, 2009).

Remark. Space limitations prevent us from presenting more details from our empirical analysis (including results on additional data, results when the low-rank approximation is regularized to be better conditioned, detailed results on running times for different sketching methods, etc.) as well as additional theoretical analysis. These results, as well as additional discussion, are available in the technical report version of this paper (Gittens & Mahoney, 2013).

2. Preliminaries and Prior Work

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an arbitrary SPSD matrix with eigenvalue decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where we partition $\mathbf{U}$ and $\mathbf{\Sigma}$ as

$$\mathbf{U} = \begin{pmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{pmatrix} \text{ and } \mathbf{\Sigma} = \begin{pmatrix}\mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2\end{pmatrix}. \quad (1)$$

Here, $\mathbf{U}_1$ comprises k orthonormal columns spanning the top k-dimensional eigenspace of $\mathbf{A}$; likewise, $\mathbf{U}_2$ is an orthonormal basis for the bottom $n-k$ dimensional eigenspace of $\mathbf{A}$. The diagonal matrix $\mathbf{\Sigma}_1$ contains the largest k eigenvalues of $\mathbf{A}$; likewise, $\mathbf{\Sigma}_2$ is a diagonal matrix containing the smallest $n-k$ eigenvalues of $\mathbf{A}$. We assume $\mathbf{\Sigma}_1$ is full-rank. $\mathbf{A}_k = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T$ is the optimal rank-k approximation to $\mathbf{A}$ in any unitarily invariant norm.

The *statistical leverage scores of* $\mathbf{A}$ *relative to the best rank-k approximation to* $\mathbf{A}$ are the squared Euclidean norms of the rows of the $n \times k$ matrix $\mathbf{U}_1$:

$$\ell_j = \|(\mathbf{U}_1)_j\|^2. \quad (2)$$

We denote by $\mathbf{S}$ an arbitrary $n \times \ell$ *sketching matrix*. The matrices

$$\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S} \quad \text{and } \mathbf{\Omega}_2 = \mathbf{U}_2^T\mathbf{S} \quad (3)$$

capture the interaction of $\mathbf{S}$ with the top and bottom eigenspaces of $\mathbf{A}$, respectively. The orthogonal projection onto the range space of a matrix $\mathbf{M}$ is written $\mathbf{P}_\mathbf{M}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, let $\|\mathbf{x}\|_\xi$, for $\xi = 1, 2, \infty$, denote the 1-norm, the Euclidean norm, and the ∞-norm, respectively. Then, $\|\mathbf{A}\|_2 = \|\mathrm{Diag}(\mathbf{\Sigma})\|_\infty$ denotes the *spectral norm* of $\mathbf{A}$; $\|\mathbf{A}\|_\mathrm{F} = \|\mathrm{Diag}(\mathbf{\Sigma})\|_2$ denotes the *Frobenius norm* of $\mathbf{A}$; and $\|\mathbf{A}\|_\star = \|\mathrm{Diag}(\mathbf{\Sigma})\|_1$ denotes the *trace norm* (or nuclear norm) of $\mathbf{A}$.

The following model for the low-rank approximation of SPSD matrices subsumes sketches based on column-sampling (also known as Nyström extensions) as well as those based on mixtures of columns (also known as projection-based sketches, since the mixtures are often accomplished using Johnson–Lindenstrauss-type dimensionality reducing "projections").

- *SPSD Sketching Model.* Let $\mathbf{A}$ be an $n \times n$ positive semi-definite matrix, and let $\mathbf{S}$ be a matrix of size $n \times \ell$, where $\ell \ll n$. Take $\mathbf{C} = \mathbf{AS}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{AS}$. Then $\mathbf{CW}^\dagger\mathbf{C}^T$ is a low-rank approximation to $\mathbf{A}$ with rank at most ℓ.

The distribution of the (random) matrix $\mathbf{S}$ leads to different classes of low-rank approximations.

We point out that sketches formed using the so-called power method (Halko et al., 2011), for which $\mathbf{C} = \mathbf{A}^q\mathbf{S}_0$ and $\mathbf{W} = \mathbf{S}_0^T\mathbf{A}^{2q-1}\mathbf{S}_0$ for some integer $q \geq 2$ and sketching matrix $\mathbf{S}_0$, fit this model if one considers the sketching matrix to be $\mathbf{A}^{q-1}\mathbf{S}_0$. In (Gittens & Mahoney, 2013), we provide theoretical guarantees on the efficacy of the power method.

(Halko et al., 2011) considers SPSD sketches that can be written in the forms $\mathbf{P_{AS}AP_{AS}}$ and $\mathbf{A}(\mathbf{P_{AS}AP_{AS}})^\dagger\mathbf{A}$ and finds that the second scheme is empirically more effective, but it provides guarantees only for the performance of the first scheme. We note that both schemes fit into our SPSD Sketching Model; we provide error bounds for the second sketching scheme by establishing that it is an instance of the power method, with $q = 2$. See (Gittens & Mahoney, 2013) for details.

A large part of the recent body of work on randomized matrix algorithms has been summarized in the recent monograph of Mahoney (Mahoney, 2011) and the recent review of Halko, Martinsson, and Tropp (Halko et al., 2011). Much of the work in machine learning on the Nyström method has focused on new proposals for selecting columns (*e.g.*, (Zhang et al., 2008; Zhang &

Kwok, 2009; Liu et al., 2010; Arcolano & Wolfe, 2010)) and/or coupling the method with downstream applications. Ensemble Nyström methods, which mix several simpler Nyström extensions, and related schemes for improving the accuracy of Nyström extensions have also been investigated (Kumar et al., 2009; Li et al., 2010; Kumar et al., 2012).

On the theoretical side, much of the work has followed that of Drineas and Mahoney (Drineas & Mahoney, 2005), who provided the first rigorous bounds for the Nyström extension of a general SPSD matrix. Rather than summarize this work in detail, we simply refer to Table 1 (our results are from Lemma 4; we have similar improvements for Lemmas 1, 2 and 3) to provide an example of our improvement relative to related work.

3. Empirical Aspects of SPSD Low-rank Approximation

Here, we present our main empirical results, which consist of evaluating sampling and projection algorithms applied to a diverse set of SPSD matrices. We don't intend these results to be "comprehensive" but instead to be "illustrative" case-studies. That is, we illustrate the tradeoffs between these methods in different realistic applications.

3.1. SPSD Sketching Algorithms

We provide empirical results for four sketches, based on sampling columns uniformly at random, sampling columns using leverage scores, mixing columns using a subsampled randomized Fourier transform (SRFT), and taking Gaussian mixtures of columns.

In the case of Gaussian mixtures, $\mathbf{S}$ is a matrix of i.i.d. $\mathcal{N}(0,1)$ random variables. In the case of SRFT mixtures, $\mathbf{S} = \sqrt{\frac{n}{\ell}}\mathbf{DFR}$, where $\mathbf{D}$ is a diagonal matrix of Rademacher random variables (*i.e.*, random ± 1s with equal probability), $\mathbf{F}$ is the real Fourier transform matrix, and $\mathbf{R}$ restricts to ℓ columns.

Observe that $\{\ell_j/k\}_{j\in\{1,\ldots,n\}}$ is a probability distribution over the columns of $\mathbf{A}$. Sketches based on leverage scores take $\mathbf{S} = \mathbf{RD}$ where $\mathbf{R} \in \mathbb{R}^{n\times\ell}$ is a column selection matrix that samples columns of $\mathbf{A}$ from the given distribution and $\mathbf{D}$ is a diagonal rescaling matrix satisfying $\mathbf{D}_{jj} = \frac{1}{\sqrt{\ell p_i}}$ iff $\mathbf{R}_{ij} = 1$.

Exact computation of leverage scores is expensive, so we also consider two sketches that use approximate leverage scores: the "power" scheme iteratively approximates the leverage scores of $\mathbf{A}$ and uses the approximate leverage scores obtained once a specified convergence condition has been met; and the "frob lev" scheme uses the leverage scores of $\mathbf{A\Pi}$, where $\mathbf{\Pi}$ is a fast Johnson-Lindenstrauss transform (Drineas et al., 2012). See (Gittens & Mahoney, 2013) for the full details.

3.2. Data Sets

We consider four classes of matrices: normalized Laplacians of very sparse graphs drawn from "informatics graph" applications; dense matrices corresponding to Linear Kernels from machine learning applications; dense matrices constructed from a Gaussian Radial Basis Function Kernel (RBFK); and sparse RBFK matrices constructed using Gaussian radial basis functions, truncated to be nonzero only for nearest neighbors.

Recall that, given a graph with weighted adjacency matrix $\mathbf{W}$, the normalized graph Laplacian is $\mathbf{A} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal matrix of weighted degrees of the nodes of the graph, *i.e.*, $D_{ii} = \sum_{j\neq i} W_{ij}$. Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the Linear Kernel matrix $\mathbf{A}$ corresponding to those points is given by $A_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, and a Gaussian RBFK matrix $\mathbf{A}^\sigma$ is given by $A^\sigma_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma^2}\right)$, where σ is a nonnegative number. One can sparsify RBF Kernels while preserving their SPSD nature. See (Gittens & Mahoney, 2013) for the details of the method employed.

Table 2 illustrates the diverse range of properties exhibited by these four classes of data sets. Several observations are particularly relevant to our discussion below. First, the Laplacian Kernels drawn from informatics graph applications are extremely sparse, and tend to have very slow spectral decay. Second, both the Linear Kernels and the Dense RBF Kernels are much denser and are much more well-approximated by moderately to very low-rank matrices. In addition, both the Linear Kernels and the Dense RBF Kernels have more uniform leverage scores. Third, we consider two values of the σ parameter for the RBF Kernels, chosen (somewhat) arbitrarily. For AbaloneD, we see that decreasing σ from 1 to 0.15, *i.e.*, letting data points "see" fewer nearby points, has two important effects: first, it results in matrices that are much less well-approximated by low-rank matrices; and second, it results in matrices with much more heterogeneous leverage scores. Fourth, for the Sparse RBF Kernels, there are a range of sparsities, ranging from above the sparsity of the sparsest Linear Kernel, but all are much denser than the Laplacian Kernels. In addition, sparsifying a Dense RBF Kernel has the effects of making the matrix less well approximated by a low-rank ma-

Table 1. Comparison of our bounds on Nyström approximation errors with those of prior works. Here, $\mathbf{A}$ is an $n \times n$ SPSD matrix, opt_ξ is the smallest ξ-norm error possible when approximating $\mathbf{A}$ with a rank-k matrix, $r = \text{rank}(\mathbf{A})$, ℓ is the number of column samples sufficient for the stated bounds to hold, k is a target rank, and $\varepsilon \in (0, 1)$. With the exception of (Drineas & Mahoney, 2005), which samples columns with probabilities proportional to the square of the corresponding diagonal entries of $\mathbf{A}$, these bounds are for column sampling uniformly at random. All bounds hold with constant probability.

	ℓ	$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2$	$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_F$	$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_\star$
(Drineas & Mahoney, 2005)	$\Omega(k/\epsilon^4)$	$\text{opt}_2 + \epsilon\sum_{i=}^n A_{ii}^2$	$\text{opt}_F + \epsilon\sum_{i=1}^n A_{ii}^2$	–
(Belabbas & Wolfe, 2009)	$\Omega(1)$	–	–	$\text{O}\left(\frac{n-\ell}{n}\right)\|\mathbf{A}\|_\star$
(Talwalkar & Rostamizadeh, 2010)	$\Omega(\max_{i,j}\|(\mathbf{U}_1)_{i,j}\|^2 nr\ln r)$	0	0	0
(Kumar et al., 2012)	$\Omega(1)$	$\text{opt}_2 + \text{O}(\frac{2n}{\sqrt{\ell}})\|\mathbf{A}\|_2$	$\text{opt}_F + \text{O}(n(\frac{k}{\ell})^{1/4})\|\mathbf{A}\|_2$	–
this work, Lemma 4	$\Omega\left(\frac{\mu k\ln k}{(1-\epsilon)^2}\right)$	$\text{opt}_2(1 + \frac{n}{\epsilon\ell})$	$\text{opt}_F + \text{O}(\epsilon^{-1})\text{opt}_\star$	$\text{opt}_\star(1 + \text{O}(\epsilon^{-1}))$

Table 2. Summary statistics for the data sets that we used. Data sets are from (Leskovec et al., 2007; Klimt & Yang, 2004; Guyon et al., 2005; Gustafson et al., 2006; Nielsen et al., 2002; Corke, 1996; Asuncion & Newman, 2012).

Name	%nnz	$\left\lceil \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_2^2} \right\rceil$	k	$\frac{\lambda_{k+1}}{\lambda_k}$	$100\frac{\|\mathbf{A}-\mathbf{A}_k\|_F}{\|\mathbf{A}\|_F}$	$100\frac{\|\mathbf{A}-\mathbf{A}_k\|_\star}{\|\mathbf{A}\|_\star}$	kth-largest leverage score scaled by n/k
Laplacian Kernels							
HEP	0.06	3078	20	0.998	7.8	0.4	128.8
HEP	0.06	3078	60	0.998	13.2	1.1	41.9
GR	0.12	1679	20	0.999	10.5	0.74	71.6
GR	0.12	1679	60	1	17.9	2.16	25.3
Enron	0.22	2588	20	0.997	7.77	0.352	245.8
Enron	0.22	2588	60	0.999	12.0	0.94	49.6
Gnutella	0.09	2757	20	1	8.1	0.41	166.2
Gnutella	0.09	2757	60	0.999	13.7	1.20	49.4
Linear Kernels							
Dexter	83.8	176	8	0.963	14.5	.934	16.6
Protein	99.7	24	10	0.987	42.6	7.66	5.45
SNPs	100	3	5	0.928	85.5	37.6	2.64
Gisette	100	4	12	0.90	90.1	14.6	2.46
Dense RBF Kernels							
AbaloneD (dense, $\sigma = .15$)	100	41	20	0.992	42.1	3.21	18.11
AbaloneD (dense, $\sigma = 1$)	100	4	20	0.935	97.8	59	2.44
WineD (dense, $\sigma = 1$)	100	31	20	0.99	43.1	3.89	26.2
WineD (dense, $\sigma = 2.1$)	100	3	20	0.936	94.8	31.2	2.29
Sparse RBF Kernels							
AbaloneS (sparse, $\sigma = .15$)	82.9	400	20	0.989	15.4	1.06	48.4
AbaloneS (sparse, $\sigma = 1$)	48.1	5	20	0.982	90.6	21.8	3.57
WineS (sparse, $\sigma = 1$)	11.1	116	20	0.995	29.5	2.29	49.0
WineS (sparse, $\sigma = 2.1$)	88.0	39	20	0.992	41.6	3.53	24.1

trix and making the leverage scores more nonuniform.

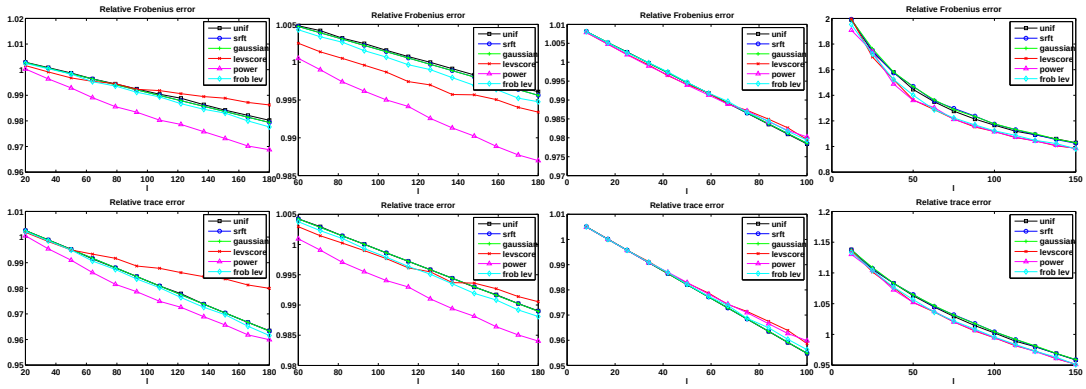
3.3. Reconstruction Accuracy of Sampling and Projection Algorithms

Here, we describe the performances of the SPSD sketches in terms of reconstruction accuracy for the data sets described in Section 3.2.

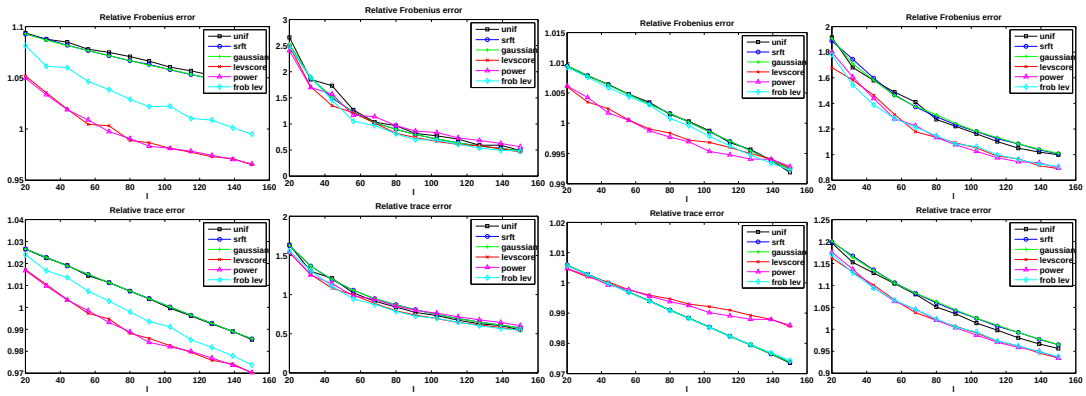
Abridged Empirical Evaluation. Figure 1 shows the Frobenius (top panel) and trace (bottom panel) norm errors of several Nyström schemes, as a function of the number of column samples ℓ, for datasets from Table 2. "unif" denotes uniform column sampling; "srft" denotes SRFT mixtures; "gaussian" denotes Gaussian mixtures; "levscore" denotes column sampling using the exact leverage scores; and "power" and "frob lev" are as described in Section 3.1. Figure 2 shows the times required to compute these sketches.

Summary of Comparison of Sampling and Projection Algorithms. Linear Kernels and to a lesser extent Dense RBF Kernels with larger σ parameter have relatively low-rank and relatively uniform leverage scores. In these circumstances uniform sampling does quite well. These data sets correspond most closely with those that have been studied previously in the machine learning literature; for these data sets our results are in agreement with those of prior work.

Sparsifying RBF Kernels and/or choosing a smaller σ parameter tends to make these kernels less well-approximated by low-rank matrices and to have more heterogeneous leverage scores. In general, these two

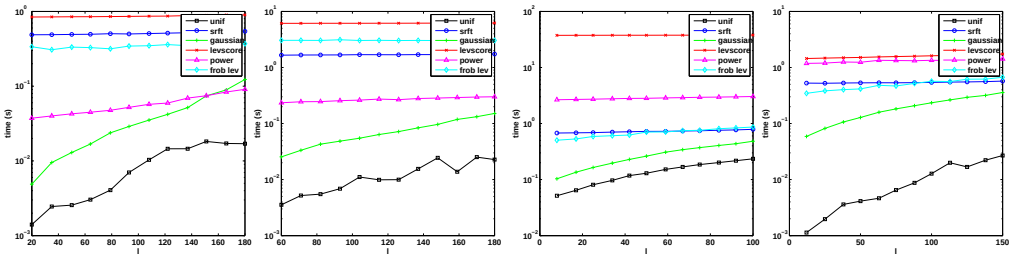


(a) GR, $k = 20$ (b) HEP, $k = 60$ (c) Dexter, $k = 8$ (d) Gisette, $k = 12$

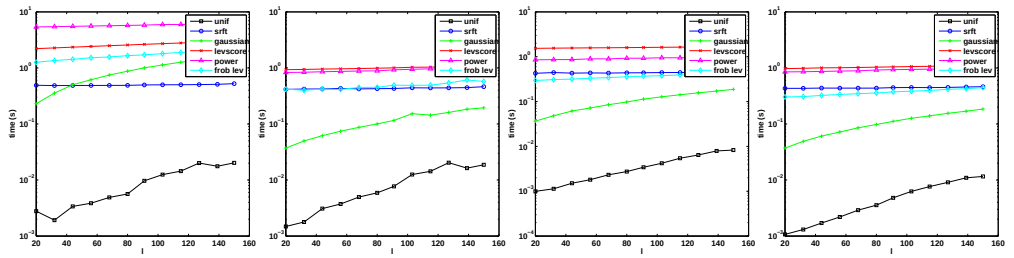


(e) AbaloneD, $\sigma = .15, k = 20$ (f) AbaloneD, $\sigma = 1, k = 20$ (g) AbaloneS, $\sigma = .15, k = 20$ (h) AbaloneS, $\sigma = 1, k = 20$

Figure 1. Frobenius and trace norm errors of several SPSD sketching schemes, as a function of number of column samples.



(a) GR, $k = 20$ (b) HEP, $k = 60$ (c) Dexter, $k = 8$ (d) Gisette, $k = 12$



(e) AbaloneD, $\sigma = .15, k = 20$ (f) AbaloneD, $\sigma = 1, k = 20$ (g) AbaloneS, $\sigma = .15, k = 20$ (h) AbaloneS, $\sigma = 1, k = 20$

Figure 2. The times required to compute several SPSD sketches, as a function of the number of column samples.

properties are not directly related—the spectrum is a property of eigenvalues, while the leverage scores are determined by the eigenvectors—but for the data we examined they are related, in that matrices with more slowly decaying spectra also often have more heterogeneous leverage scores.

For Dense RBF Kernels with smaller σ and Sparse RBF Kernels, leverage score sampling tends to outperform other methods. Interestingly, the Sparse RBF Kernels have many properties of very sparse Laplacian Kernels corresponding to relatively-unstructured informatics graphs, an observation which should be of interest for researchers who construct sparse graphs from data using, *e.g.*, "locally linear" methods to try to reconstruct hypothesized low-dimensional manifolds.

Reconstruction quality under exact leverage score sampling saturates, as a function of choosing more samples ℓ. As a consequence, the value of ℓ used determines whether leverage score sampling or other sketching methods result in lower errors.

Summary of Leverage Score Approximation Algorithms. The running time of computing the exact leverage scores is generally much worse than that of uniform sampling and both SRFT-based and Gaussian-based random projection methods. The running time of computing approximations to the leverage scores can, with appropriate choice of parameters, be much faster than the exact computation; and, especially for "'frob lev,"' it can be comparable to the time needed to execute the random projection used in the leverage score approximation algorithm (Drineas et al., 2012). For methods that involve $q > 1$ iterations to compute stronger approximations to the leverage scores, the running time can vary considerably depending on the stopping condition.

The leverage scores computed by the "frob lev" procedure are typically very different than the "exact" leverage scores, but they are leverage scores for a low-rank space that is near the best rank-k approximation to the matrix. This is often sufficient for good low-rank approximation, although the reconstruction accuracy can degrade in the rank-restricted cases (not presented here) as ℓ is increased. The approximate leverage scores computed from "power" approach those of the exact leverage scores, as q is increased; and they obtain reconstruction accuracy that is no worse, and in many cases is better, than those obtained using the exact leverage scores. This suggests that, by not fitting exactly to the empirical statistical leverage scores, we are observing a form of regularization.

4. Theoretical Aspects of SPSD Low-rank Approximation

In this section, we present our main theoretical results, which consist of a suite of bounds on the quality of low-rank approximation under several different sketching methods. These were motivated by our empirical observation that *all* of the sampling and projection methods we considered perform *much* better on the SPSD matrices we considered than previous worst-case bounds (*e.g.*, (Drineas & Mahoney, 2005; Kumar et al., 2012; Gittens, 2011)) would suggest.

Our results are based on the fact, established in (Gittens, 2011), that approximations which satisfy our SPSD Sketching Model satisfy

$$\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T} = \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}\mathbf{A}^{1/2}. \tag{4}$$

4.1. Deterministic Error Bounds

Here, we present theorems that bound the spectral, Frobenius, and trace norm approximation errors. Throughout, $\mathbf{A}$ is an $n \times n$ SPSD matrix with eigenvalue decomposition partitioned as in Equation (1), $\mathbf{S}$ is a sketching matrix of size $n \times \ell$, $\mathbf{C} = \mathbf{A}\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{A}\mathbf{S}$, and $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are defined in Equation (3).

Spectral Norm Bounds. We start with a bound on the spectral norm of the residual error.

Theorem 1. *If $\mathbf{\Omega}_1$ has full row-rank, then*

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_2 \leq \|\mathbf{\Sigma}_2\|_2 + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2.$$

Proof. It follows from Equation (4) that

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_2 = \left\|\mathbf{A}^{1/2} - \mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}\mathbf{A}^{1/2}\right\|_2^2. \tag{5}$$

Next, recall that $\mathbf{\Omega}_i = \mathbf{U}_i^T\mathbf{S}$, and that $\mathbf{\Omega}_1$ has full-row rank. It can be shown that

$$\|\mathbf{X} - \mathbf{P}_{\mathbf{X}\mathbf{S}}\mathbf{X}\|_2^2 \leq \|\mathbf{\Sigma}_2\|_2^2 + \left\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2, \tag{6}$$

for any matrix $\mathbf{X}$ (Boutsidis et al., 2009; Halko et al., 2011). □

Frobenius Norm Bounds. Next, we state the bound on the Frobenius norm of the residual error.

Theorem 2. *If $\mathbf{\Omega}_1$ has full row-rank, then*

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_{\mathrm{F}} \leq \|\mathbf{\Sigma}_2\|_{\mathrm{F}} + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2 \left(\sqrt{2\,\|\mathbf{\Sigma}_2\|_\star} + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}\right).$$

Proof. It follows from Equation (4) that

$$E := \left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} = \left\|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}})\mathbf{A}^{1/2}\right\|_{\mathrm{F}}.$$

To bound this, we first use the unitary invariance of the Frobenius norm and the fact that $\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}} = \mathbf{U}\mathbf{P}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{S}}\mathbf{U}^T$ to obtain

$$E = \left\|\mathbf{\Sigma}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{S}})\mathbf{\Sigma}^{1/2}\right\|_{\mathrm{F}}^2.$$

Then we take

$$\mathbf{Z} = \mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{S}\mathbf{\Omega}_1^{\dagger}\mathbf{\Sigma}_1^{-1/2} = \begin{pmatrix}\mathbf{I}\\ \mathbf{F}\end{pmatrix}, \tag{7}$$

where $\mathbf{I} \in \mathbb{R}^{k\times k}$ and $\mathbf{F} \in \mathbb{R}^{n-k\times k}$ is given by $\mathbf{F} = \mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\mathbf{\Sigma}_1^{-1/2}$. The latter equality holds because of our assumption that $\mathbf{\Omega}_1$ has full row-rank. Since the range of $\mathbf{Z}$ is contained in the range of $\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{S}$,

$$E \le \left\|\mathbf{\Sigma}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{\Sigma}^{1/2}\right\|_{\mathrm{F}}^2.$$

The fact that $\mathbf{Z}$ has full column-rank allows us to write

$$\begin{aligned}\mathbf{I} - \mathbf{P}_{\mathbf{Z}} &= \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\\ &= \begin{pmatrix}\mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & -(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\\ -\mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & \mathbf{I} - \mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\end{pmatrix}.\end{aligned} \tag{8}$$

This implies that

$$\begin{aligned}E &\le \left\|\mathbf{\Sigma}^{1/2}\begin{pmatrix}\mathbf{I} - (\mathbf{\Xi})^{-1} & -(\mathbf{\Xi})^{-1}\mathbf{F}^T\\ -\mathbf{F}(\mathbf{\Xi})^{-1} & \mathbf{I} - \mathbf{F}(\mathbf{\Xi})^{-1}\mathbf{F}^T\end{pmatrix}\mathbf{\Sigma}^{1/2}\right\|_{\mathrm{F}}^2\\ &= \left\|\mathbf{\Sigma}_1^{1/2}\left(\mathbf{I} - (\mathbf{\Xi})^{-1}\right)\mathbf{\Sigma}_1^{1/2}\right\|_{\mathrm{F}}^2\\ &\quad + 2\left\|\mathbf{\Sigma}_1^{1/2}(\mathbf{\Xi})^{-1}\mathbf{F}^T\mathbf{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2\\ &\quad + \left\|\mathbf{\Sigma}_2^{1/2}\left(\mathbf{I} - \mathbf{F}(\mathbf{\Xi})^{-1}\mathbf{F}^T\right)\mathbf{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2\\ &:= T_1 + T_2 + T_3.\end{aligned} \tag{9}$$

where $\mathbf{\Xi} = \mathbf{I} + \mathbf{F}^T\mathbf{F}$.

Next, we will provide bounds for T_1, T_2, and T_3. Using the fact that $\mathbf{0} \preceq \mathbf{I} - \mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \preceq \mathbf{I}$, we can bound T_3 with

$$T_3 \le \|\mathbf{\Sigma}_2\|_{\mathrm{F}}^2.$$

Likewise, the fact that $\mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} \preceq \mathbf{F}^T\mathbf{F}$ (readily verifiable with an SVD) implies that we can bound T_1 as

$$\begin{aligned}T_1 &\le \left\|\mathbf{\Sigma}_1^{1/2}\mathbf{F}^T\mathbf{F}\mathbf{\Sigma}_1^{1/2}\right\|_{\mathrm{F}}^2 \le \left\|\mathbf{\Sigma}_1^{1/2}\mathbf{F}^T\right\|_2^2\left\|\mathbf{\Sigma}_1^{1/2}\mathbf{F}^T\right\|_{\mathrm{F}}^2\\ &= \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^2\end{aligned}$$

To bound T_2, observe that

$$\begin{aligned}T_2 &\le 2\left\|\mathbf{\Sigma}_1^{1/2}\mathbf{\Xi}^{-1}\mathbf{F}^T\right\|_2^2\left\|\mathbf{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2\\ &= 2\left\|\mathbf{\Sigma}_1^{1/2}(\mathbf{I} + \mathbf{M})^{-1}\mathbf{M}(\mathbf{I} + \mathbf{M})^{-1}\mathbf{\Sigma}_1^{1/2}\right\|_2\|\mathbf{\Sigma}_2\|_{\star}\end{aligned}$$

where $\mathbf{M} = \mathbf{F}^T\mathbf{F}$. It is readily verifiable, using the SVD, that any SPSD matrix $\mathbf{M}$ satisfies the semidefinite inequality

$$(\mathbf{I} + \mathbf{M})^{-1}\mathbf{M}(\mathbf{I} + \mathbf{M})^{-1} \preceq \mathbf{M},$$

so we conclude that

$$T_2 \le 2\left\|\mathbf{F}\mathbf{\Sigma}_1^{1/2}\right\|_2^2\|\mathbf{\Sigma}_2\|_{\mathrm{F}}^2 = 2\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2\|\mathbf{\Sigma}_2\|_{\star}.$$

Combining our estimates for T_1, T_2, and T_3 with Equation (9) gives

$$\begin{aligned}E &\le \|\mathbf{\Sigma}_2\|_{\mathrm{F}}^2\\ &\quad + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2\left(2\|\mathbf{\Sigma}_2\|_{\star} + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^2\right)\end{aligned}$$

The claimed bound follows by applying the subadditivity of the square-root function. □

Trace Norm Bounds. Finally, we state the following bound on the trace norm of the residual error.

Theorem 3. *If $\mathbf{\Omega}_1$ has full row-rank, then*

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\star} \le \mathrm{Tr}\left(\mathbf{\Sigma}_2\right) + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^2,$$

Proof. Since $\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T = \mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}})\mathbf{A}^{1/2} \succeq \mathbf{0}$, its trace norm simplifies to its trace. Thus

$$\begin{aligned}\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\star} &= \mathrm{Tr}\left(\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right)\\ &= \mathrm{Tr}\left(\mathbf{\Sigma}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{\Sigma}^{1/2}\mathbf{S}})\mathbf{\Sigma}^{1/2}\right)\\ &\le \mathrm{Tr}\left(\mathbf{\Sigma}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{\Sigma}^{1/2}\right),\end{aligned}$$

where $\mathbf{Z}$ is defined in Equation (7). The expression for $\mathbf{P}_{\mathbf{Z}}$ supplied in Equation (8) implies that

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{\Sigma}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{\Sigma}^{1/2}\right) &\le \mathrm{Tr}\left(\mathbf{\Sigma}_1^{1/2}\mathbf{F}^T\mathbf{F}\mathbf{\Sigma}_1^{1/2}\right) + \mathrm{Tr}\left(\mathbf{\Sigma}_2\right)\\ &= \mathrm{Tr}\left(\mathbf{\Sigma}_2\right) + \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^2.\end{aligned}$$

The final equality follows from identifying $\mathbf{F}$ and the squared Frobenius norm. □

Remark. The assumption that $\mathbf{\Omega}_1$ has full row-rank is very non-trivial; it is false, for non-trivial parameter values, for common sketching methods such as uniform sampling. It is satisfied by our procedures in Section 4.2.

4.2. Stochastic Error Bounds for Low-rank SPSD Approximation

In this section, we apply the theorems from Section 4.1 to bound the reconstruction errors for several random sampling and random projection methods that conform to our SPSD Sketching Model. Throughout, $\mathbf{A}$ is an $n \times n$ SPSD matrix.

Lemma 1 (Leverage-based sketches). *Let $\mathbf{S}$ be a sketching matrix of size $n \times \ell$ corresponding to a leverage-based probability distribution derived from the top k-dimensional eigenspace of $\mathbf{A}$, satisfying*

$$p_j \geq \frac{\beta}{k} \|(\mathbf{U}_1)_j\|_2^2 \quad \text{and} \quad \sum_{j=1}^{n} p_j = 1,$$

for some $\beta \in (0,1]$. Fix a failure probability $\delta \in (0,1]$ and approximation factor $\varepsilon \in (0,1]$.

If $\ell \geq 3200(\beta\varepsilon^2)^{-1}k\log(4k/(\beta\delta))$, then

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 &\leq \|\mathbf{A}-\mathbf{A}_k\|_2 + \varepsilon^2 \|\mathbf{A}-\mathbf{A}_k\|_\star, \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} \\
&+ (\sqrt{2}\varepsilon+\varepsilon^2)\|\mathbf{A}-\mathbf{A}_k\|_\star, \text{ and} \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_\star &\leq (1+\varepsilon^2)\|\mathbf{A}-\mathbf{A}_k\|_\star
\end{aligned}$$

each hold, individually, with probability at least $1-4\delta-0.4$.

Lemma 2 (SRFT sketches). *Let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be an SRFT sketching matrix. Fix a failure probability $\delta \in (0, 1/9]$ and approximation factor $\varepsilon \in (0,1]$.*

If $k \geq \ln n$ and $\ell \geq 24\varepsilon^{-1}[\sqrt{k}+\sqrt{8\ln(8n/\delta)}]^2\ln(8k/\delta)$, then

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 &\leq \left(1+\frac{6}{1-\sqrt{\varepsilon}}\right)\|\mathbf{A}-\mathbf{A}_k\|_2 \\
&+ \frac{1}{(1-\sqrt{\varepsilon})\ln(8k/\delta)}\|\mathbf{A}-\mathbf{A}_k\|_\star, \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} \\
&+ (7\sqrt{\varepsilon}+22\varepsilon)\|\mathbf{A}-\mathbf{A}_k\|_\star, \text{ and} \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_\star &\leq (1+22\varepsilon)\|\mathbf{A}-\mathbf{A}_k\|_\star
\end{aligned}$$

each hold, individually, with probability at least $1-2\delta$.

Lemma 3 (Gaussian sketches). *Let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be a Gaussian sketching matrix. If $\ell = k+p$ where $p = k\varepsilon^{-2}$ for $\varepsilon \in (0,1]$ and $k > 4$, then*

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 &\leq \left(1+963\varepsilon^2\right)\|\mathbf{A}-\mathbf{A}_k\|_2 \\
&+ 219\frac{\varepsilon^2}{k}\|\mathbf{A}-\mathbf{A}_k\|_\star, \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} \\
&+ (11\varepsilon+544\varepsilon^2)\sqrt{\|\mathbf{A}-\mathbf{A}_k\|_2\|\mathbf{A}-\mathbf{A}_k\|_\star} \\
&+ 91\frac{\varepsilon}{\sqrt{k}}\|\mathbf{A}-\mathbf{A}_k\|_\star + 815\varepsilon^2\sqrt{\frac{\ln k}{k}}\|\mathbf{A}-\mathbf{A}_k\|_2, \text{ and} \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_\star &\leq (1+45\varepsilon^2)\|\mathbf{A}-\mathbf{A}_k\|_\star \\
&+ 874\varepsilon^2\frac{\ln k}{k}\|\mathbf{A}-\mathbf{A}_k\|_2
\end{aligned}$$

each hold, individually, with probability at least $1-2k^{-1}-4\mathrm{e}^{-k/\varepsilon^2}$.

Lemma 4 (Uniform column sampling). *Let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be a sketching matrix corresponding to sampling the columns of $\mathbf{A}$ uniformly at random (with or without replacement). Let*

$$\mu = \frac{n}{k}\cdot\max_{i\in\{1,\ldots,n\}}\|(\mathbf{U}_1)_i\|_2^2$$

denote the coherence *of the top k-dimensional eigenspace of $\mathbf{A}$ and fix a failure probability $\delta \in (0,1)$ and accuracy factor $\varepsilon \in (0,1)$. If $\ell \geq 2\varepsilon^{-2}\mu k\ln(k/\delta)$, then*

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 &\leq \left(1+\frac{n}{(1-\varepsilon)\ell}\right)\|\mathbf{A}-\mathbf{A}_k\|_2, \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} \\
&+ \frac{3}{\delta^2(1-\varepsilon)}\|\mathbf{A}-\mathbf{A}_k\|_\star, \text{ and} \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_\star &\leq \left(1+\frac{1}{\delta^2(1-\varepsilon)}\right)\|\mathbf{A}-\mathbf{A}_k\|_\star
\end{aligned}$$

each hold, individually, with probability at least $1-4\delta$.

Remark. The additive scale factors for the spectral and Frobenius norm bounds are much improved relative to the prior results of (Drineas & Mahoney, 2005). To our knowledge, our results supply the first relative-error trace norm approximation bounds.

As with previous bounds for uniform sampling, *e.g.*, (Kumar et al., 2012; Gittens, 2011), the results in Lemma 4 are much weaker than those for projecton-based and leverage score sampling-based SPSD sketches, since the sampling complexity depends on the coherence of the input matrix.

References

Arcolano, N. and Wolfe, P. J. Nyström approximation of Wishart matrices. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 3606–3609, 2010.

Asuncion, A. and Newman, D. J. UCI Machine Learning Repository, November 2012. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Belabbas, M.-A. and Wolfe, P. J. Spectral methods in machine learning and new strategies for very large datasets. *Proc. Natl. Acad. Sci. USA*, 106:369–374, 2009.

Boutsidis, C., Mahoney, M.W., and Drineas, P. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 968–977, 2009.

Corke, P. I. A Robotics Toolbox for MATLAB. *IEEE Robotics and Automation Magazine*, 3:24–32, 1996.

Drineas, P. and Mahoney, M.W. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Drineas, P., Mahoney, M.W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

Drineas, P., Mahoney, M.W., Muthukrishnan, S., and Sarlós, T. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.

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

Gittens, A. The spectral norm error of the naive Nystrom extension. Technical report, 2011. Preprint: arXiv:1110.5305 (2011).

Gittens, A. and Mahoney, M. W. Revisiting the Nyström Method for Improved Large-Scale Machine Learning. Technical report, 2013. Preprint: arXiv:1303.1849 (2013).

Gustafson, A. M., Snitkin, E. S., Parker, S. C. J., DeLisi, C., and Kasif, S. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics*, 7:265, 2006.

Guyon, I., Gunn, S. R., Ben-Hur, A., and Dror, G. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Klimt, B. and Yang, Y. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pp. 217–226, 2004.

Kumar, S., Mohri, M., and Talwalkar, A. Ensemble Nyström method. In *Annual Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference*, 2009.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

Leskovec, J., Kleinberg, J., and Faloutsos, C. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, 2007.

Li, M., Kwok, J.T., and Lu, B.-L. Making large-scale Nyström approximation possible. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 631–638, 2010.

Liu, S., Zhang, J., and Sun, K. Learning low-rank kernel matrices with column-based methods. *Communications in Statistics—Simulation and Computation*, 39(7):1485–1498, 2010.

Mahoney, M. W. *Randomized algorithms for matrices and data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.

Mahoney, M.W. and Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., Brown, P. O., Botstein, D., and van de Rijn, M. Molecular characterisation of soft tissue tumours: a gene expression study. *The Lancet*, 359: 1301–1307, 2002.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W., and Drineas, P. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, 3:1672–1686, 2007.

Talwalkar, A. and Rostamizadeh, A. Matrix coherence and the Nyström method. In *Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence*, 2010.

Williams, C.K.I. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Annual Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pp. 682–688, 2001.

Zhang, K. and Kwok, J. T. Density-weighted Nyström method for computing large kernel eigensystems. *Neural Computation*, 21(1):121–146, 2009.

Zhang, K., Tsang, I.W., and Kwok, J.T. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1232–1239, 2008.