Chapter 6

Theoretical and empirical aspects of SPSD sketches

6.1 Introduction

In this chapter we consider the accuracy of randomized low-rank approximations of symmetric positive-semidefinite matrices conforming to the following general model[1].

SPSD Sketching Model. Let $\mathbf{A}$ be an $n \times n$ SPSD matrix, and let $\mathbf{S}$ be matrix of size $n \times \ell$, where $\ell \ll n$. Form

$$\mathbf{C} = \mathbf{AS} \quad \text{and} \quad \mathbf{W} = \mathbf{S}^T\mathbf{AS}.$$

We call $\mathbf{CW}^\dagger\mathbf{C}^T$ the *SPSD sketch* of $\mathbf{A}$ associated with the *sketching matrix* $\mathbf{S}$. Note that this sketch is also SPSD, and has rank at most ℓ.

These sketches can be computed quickly, and as we see in Section 6.9, are accurate low-rank approximations for several classes of matrices that arise in machine learning and data analysis applications. This model subsumes both projection-based sketches (here, $\mathbf{S}$ mixes the columns and rows of $\mathbf{A}$), and sampling-based sketches (here, $\mathbf{S}$ selects columns and rows of $\mathbf{A}$).

[1]The content of this chapter is redacted from the technical report [Git11] and the technical report [GM13a] coauthored with Michael Mahoney. A preliminary version of some of these results appear in the conference paper [GM13b], also coauthored with Michael Mahoney.

The computation of an SPSD sketch requires only one pass over $\mathbf{A}$, because the matrix $\mathbf{C} = \mathbf{AS}$ uniquely determines the sketch. One should contrast this to the natural extension of the low-rank approximations considered in Chapter 5, namely $\mathbf{P}_{\mathbf{AS}}\mathbf{A}\mathbf{P}_{\mathbf{AS}}$, which requires two passes over $\mathbf{A}$: one to construct a basis for the range of $\mathbf{AS}$, and one to project $\mathbf{A}$ onto this basis.

In addition to one-pass sketches, low-rank approximations formed using the so-called power method [HMT11] fit the SPSD sketching model. For such sketches, $\mathbf{C} = \mathbf{A}^q\mathbf{S}_0$ and $\mathbf{W} = \mathbf{S}_0^T\mathbf{A}^{2q-1}\mathbf{S}_0$ for some integer $q \geq 2$ and sketching matrix $\mathbf{S}_0$. To see that these models fit the SPSD sketching model, simply consider the sketching matrix to be $\mathbf{S} = \mathbf{A}^{q-1}\mathbf{S}_0$. The approximation errors of these sketches decrease as p increases. The two-pass sketch is particularly of interest: we relate it to a low-rank approximation proposed in [HMT11]. As well, we show empirically that two-pass SPSD sketches are empirically significantly more accurate than the projection-based low-rank approximant $\mathbf{P}_{\mathbf{AS}}\mathbf{A}\mathbf{P}_{\mathbf{AS}}$.

When $\mathbf{S}$ comprises columns selected uniformly at random without replacement from the $n \times n$ identity matrix, the resulting SPSD sketch is called a Nyström extension. To form a Nyström extension, one needs only to be able to sample columns from $\mathbf{A}$. Further, the cost of forming the factored form of a Nyström extension is $\Omega(\ell^3 + n\ell)$, linear in the size of $\mathbf{A}$! For these two reasons, Nyström extensions are attractive in settings where it is costly or unreasonable to access all the elements of $\mathbf{A}$. However, as we see in this chapter, the use of Nyström extensions is only theoretically justified when the top k-dimensional eigenspace of $\mathbf{A}$ has low coherence. Recall from Chapter 4 that the coherence of this eigenspace is defined as

$$\mu = \max_{i=1,\ldots,n} \frac{n}{k} \left\| (\mathbf{P}_{\mathbf{U}_1})_{ii} \right\|_2^2,$$

where $\mathbf{U}_1$ is a basis for the eigenspace. This dependence on the coherence is not simply a

theoretical consideration: it is empirically observable. This motivates looking at the wider class of SPSD sketches, where, depending on the choice of $\mathbf{S}$, potentially *all* the columns of the matrix contribute to the approximation.

This chapter presents theoretical and empirical results for different choices of $\mathbf{S}$. Empirically, we find that the Nyström extension performs well in practice, but not as well as sketches that mix the columns of $\mathbf{A}$ before sampling or sketches that sample the columns according to a specific importance sampling distribution. Our theoretical bounds bear out this comparison, and are asymptotically superior to the bounds present in the literature for low-rank approximation schemes which fit into our SPSD sketching model. However, a large gap still remains between our bounds and the observed errors of SPSD sketches.

We develop a framework for the analysis of SPSD sketches that parallels the framework Lemma 4.8 provides for the analysis of projection-based low-rank approximations. Applied to Nyström extensions, our framework exposes a natural connection between Nyström extensions and the column subset selection problem that leads to an optimal worse-case bound for the spectral error of Nyström extensions. This is the first truly relative-error spectral norm bound available for Nyström extensions; the contemporaneous work [CD11] independently establishes this same bound in the broader context of CUR decompositions. More generally, we provide theoretical worst-case bounds for several sketches based on random column sampling and random projections. These bounds apply to both one-pass and multiple-pass sketches. In the case of multiple-pass sketches, we find that the errors of the sketches decrease proportionally to $\lambda_{k+1}(\mathbf{A})/\lambda_k(\mathbf{A})$ with each additional pass.

In general, the process of computing SPSD sketches is not numerically stable: if $\mathbf{W}$ is ill-conditioned, then instabilities may arise in solving the linear system $\mathbf{W}^\dagger\mathbf{C}^T$. This is of particular concern with Nyström extensions, since the particular submatrix of $\mathbf{A}$ selected may be quite

ill-conditioned. With other sketching schemes, the fact that $\mathbf{W}$ is formed as a mixture of the columns of $\mathbf{A}$ tends to provide implicit protection against $\mathbf{W}$ being poorly conditioned. The seminal paper on the use of Nyström extensions for low-rank approximation, [WS01], suggested regularizing Nyström extensions to avoid ill-conditioning. This algorithm can be used to regularize the computation of any SPSD sketch; we provide the first error bounds for these regularized sketches. Another regularization scheme is introduced in [CD11]. We compare the empirical performance of these two regularization schemes.

We supplement our theoretical results with a collection of empirical results intended to illustrate the performance of the considered SPSD sketches on matrices that arise in machine learning applications. We also provide empirical results for the rank-restricted sketches obtained by replacing $\mathbf{W}$ with $\mathbf{W}_k$ in the definition of an SPSD sketch. That is, we also consider sketches of the form $\mathbf{C}\mathbf{W}_k^{\dagger}\mathbf{C}^T$. These sketches do not fit into our SPSD sketching model, but are the natural way to ensure that the rank of the low-rank approximation does not exceed the target rank k. Further, since $\mathbf{W}_k$ has a smaller condition number than $\mathbf{W}$, the rank-restricted sketches can be computed more stably than the non-rank-restricted sketches.

6.1.1 Outline

In Section 6.2, we introduce the specific randomized SPSD sketches analyzed in this chapter and summarize the spectral, Frobenius, and trace-norm error bounds for the one-pass variants of these sketches. In Section 6.3, we compare our results with prior work on SPSD sketches, in particular Nyström extensions. In Section 6.4 we prove the deterministic error bounds that form the basis of our analyses. In Sections 6.5 and 6.6 we apply the deterministic bounds to the Nyström extension and several randomized mixture-based sketches. In Section 6.7, we recall two algorithms for computing regularized SPSD sketches; a novel error analysis is presented for

one of these algorithms. In Section 6.8 we provide experimental evidence of the efficacy of the two algorithms for computing regularized SPSD sketches. We provide a set of empirical results on the application of SPSD sketches to matrices drawn from data analysis and machine-learning applications in Section 6.9. Finally, we conclude in Section 6.10 with an empirical comparison of two-pass SPSD sketches with the low-rank approximation $\mathbf{P}_{\mathbf{AS}}\mathbf{A}\mathbf{P}_{\mathbf{AS}}$. In the same section, we show that a certain low-rank approximation introduced in [HMT11] is in fact a two-pass SPSD sketch.

6.2 Deterministic bounds on the errors of SPSD sketches

Recall the following partitioning of the eigenvalue decomposition of $\mathbf{A}$, which we use to state our results:

$$\mathbf{A} = \overset{\begin{matrix} k & n-k \end{matrix}}{\begin{bmatrix} & \\ \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}} \overset{\begin{matrix} k & n-k \end{matrix}}{\begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix}} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix}. \tag{6.2.1}$$

The matrix $[\mathbf{U}_1\ \mathbf{U}_2]$ is orthogonal, $\boldsymbol{\Sigma}_1$ contains the k largest eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively span a top k-dimensional eigenspace of $\mathbf{A}$ and the corresponding bottom $(n-k)$-dimensional eigenspace of $\mathbf{A}$. The interaction of the sketching matrix $\mathbf{S}$ with the eigenspaces spanned by $\mathbf{U}_1$ and $\mathbf{U}_2$ is captured by the matrices

$$\boldsymbol{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}, \quad \boldsymbol{\Omega}_2 = \mathbf{U}_2^T\mathbf{S}. \tag{6.2.2}$$

We now introduce the randomized sketching procedures considered in this chapter and summarize the bounds obtained on the spectral, Frobenius, and trace-norm approximation errors of each of these sketches. Our results bound the *additional error* of SPSD sketches. That is, they bound the amount by which the approximation errors of the SPSD sketches exceed the

approximation errors of $\mathbf{A}_k$, the optimal rank-k low-rank approximation.

Nyström extensions These sketches are formed by sampling columns from $\mathbf{A}$ uniformly at random, without replacement. The sketching matrix $\mathbf{S}$ comprises a set of columns sampled uniformly at random without replacement from the identity matrix.

Empirically, as we see in Section 6.9, Nyström extensions have surprisingly low error across a range of matrices with diverse properties. This is perhaps surprising, given that the column-sampling process does not take into consideration any properties of $\mathbf{A}$ other than its size.

Fix parameters ϵ and δ in $(0, 1)$. Theorem 6.9 finds that when $\ell \geq 2\mu\epsilon^{-2}k\log(k/\delta)$, the approximation errors of Nyström extensions satisfy

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

simultaneously, with probability at least $1 - 4\delta$.

Leverage-based sketches Similarly to Nyström extensions, these sketches are formed by sampling columns from $\mathbf{A}$. However, the columns are sampled, with replacement, according to a nonuniform distribution based on their *statistical leverage scores filtered through the top k-dimensional eigenspace of* $\mathbf{A}$. Recall that $\mathbf{U}_1 \in \mathbb{R}^{n\times k}$ is a basis for the top k-dimensional eigenspace of $\mathbf{A}$. The leverage score of the jth column of $\mathbf{A}$ is defined as the squared

Euclidean norm of the jth row of $\mathbf{U}_1$:

$$\ell_j = \|(\mathbf{U}_1)^{(j)}\|_2^2.$$

Since $\mathbf{U}_1$ has orthonormal columns, the leverage scores of the columns of $\mathbf{A}$ sum to k, so the quantities

$$p_j = \frac{1}{k}\|(\mathbf{U}_1)^{(j)}\|_2^2$$

define a nonuniform probability distribution over the columns of $\mathbf{A}$. This distribution is used in the construction of leverage-based sketches.

Previous work has established that the leverage scores reflect the nonuniformity properties of the top k-dimensional eigenspace of $\mathbf{A}$ [DM10]. The fact that the resulting sketch is expressed in terms of columns from the matrix rather than mixtures of columns, as in the case with truncated SVDs or mixture-based sketches, means that in data analysis applications, leverage-based sketches often lead to models with superior interpretability [PZB+07, DM09, Mah11, YMS+13].

The tradeoff for this interpretability is that it is expensive to compute the exact leverage score distribution. However, the leverage scores can be approximated in roughly the time it takes to apply a random projection to $\mathbf{A}$ [DMIMW12]. The error bounds we provide allow for sampling from approximate leverage score distributions.

The sketching matrix $\mathbf{S}$ associated with leverage-based sketches is factored into the product of a column selection matrix and a rescaling matrix, $\mathbf{S} = \mathbf{RD}$. Here $\mathbf{R} \in \mathbb{R}^{n\times\ell}$ samples columns of $\mathbf{A}$ according to the exact or approximate leverage score distribution, that is, $\mathbf{R}_{ij} = 1$ if and only if the ith column of $\mathbf{A}$ is the jth column selected; and $\mathbf{D} \in \mathbb{R}^{\ell\times\ell}$ is a

diagonal matrix satisfying $\mathbf{D}_{jj} = 1/\sqrt{\ell p_j}$.

In Section 6.9, we see that when ℓ is small, i.e. when $\ell \approx k$, leverage-based sketches consistently have the lowest error of all non-rank-restricted sketches considered. The errors of the restricted-rank leverage-based sketches are also usually lower, for *all* values of ℓ, than the errors of any other restricted-rank sketches considered. Both these facts match with the intuition that the leverage scores capture the nonuniformity properties of $\mathbf{A}$ that are relevant to forming an accurate rank-k approximation by sampling columns.

Fix parameters ϵ and δ in $(0,1)$. Theorem 6.12 guarantees that if $\ell \geq 3200\epsilon^{-2}k\log(4k/\delta)$, then leverage-based sketches satisfy

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

simultaneously, with probability at least $1 - 6\delta - 0.6$.

SRFT-based sketches These sketches are formed by mixing the columns of $\mathbf{A}$ using a subsampled fast transformation. Here $\mathbf{S} = \sqrt{n/\ell}\mathbf{DFR}$, where $\mathbf{D}$ is a diagonal matrix of Rademacher random variables, $\mathbf{F}$ is the normalized real Fourier transform matrix, and $\mathbf{R}$ restricts to ℓ columns. Of course, one could consider a slew of similar sketches where the Fourier transform is replaced with another unitary transform associated with a fast algorithm.

Fix parameters ϵ and δ in $(0,1)$. Theorem 6.16 implies that when

$$\ell \geq 24\epsilon^{-1}[\sqrt{k} + \sqrt{8\log(8n/\delta)}]^2 \log(8k/\delta),$$

the approximation errors of SRFT-based sketches satisfy

$$\begin{aligned}\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_2 &\le \left(1+\frac{1}{1-\sqrt{\epsilon}}\left(5+\frac{16\log(n/\delta)^2}{\ell}\right)\right)\|\mathbf{A}-\mathbf{A}_k\|_2 \\ &\quad + \frac{2\log(n/\delta)}{(1-\sqrt{\epsilon})\ell}\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right), \\ \|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_{\mathrm{F}} &\le \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + 29\sqrt{\epsilon}\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right), \text{ and} \\ \operatorname{Tr}\left(\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right) &\le (1+22\epsilon)\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right)\end{aligned}$$

simultaneously, with probability at least $1-2\delta$.

Gaussian sketches These sketches are formed by sampling with an $n\times\ell$ matrix of i.i.d. Gaussian entries. As we see in Section 6.9, for large ℓ, when the ranks of the sketches are not restricted, Gaussian sketches usually have the lowest error of all sketches considered in this chapter.

Fix an accuracy parameter $\epsilon\in(0,1]$ and choose $\ell\ge(1+\epsilon^2)k$. Theorem 6.17 implies the following:

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

simultaneously, with probability at least $1-2k^{-1}-4\mathrm{e}^{-k/\epsilon^2}$.

6.3 Comparison with prior work

Our bounds on the errors of the sketches just introduced are summarized in Table 6.1. They exhibit a common structure: for the spectral and Frobenius norms, we see that the additional error is on a larger scale than the optimal error, and the trace-norm bounds all guarantee relative error approximations. This follows from, as detailed in Section 6.4, the fact that the SPSD sketching procedure can be understood as forming column-sample/projection-based approximations to the *square root* of $\mathbf{A}$, then squaring this approximation to obtain the resulting approximation to $\mathbf{A}$. The squaring process results in *potentially* large additional errors in the case of the spectral and Frobenius norms. Whether or not the additional errors are large in practice depends upon the properties of the matrix and the form of stochasticity used in the sampling process. For instance, from our bounds it is clear that Gaussian-based SPSD sketches are expected to have a lower additional error in the spectral norm than any of the other sketches considered.

We also see, in the case of Nyström extensions, a necessary dependence on the coherence of the input matrix, since columns are sampled uniformly at random. However, we also see that the scales of the additional error of the Frobenius and trace-norm bounds are substantially improved over those in prior results. The large additional error in the spectral-norm error bound is necessary in the worst case (Section 6.8). The additional spectral-norm errors of the sketching methods which use more information about the matrix, namely the leverage-based, Fourier-based, and Gaussian-based sketches, are on a substantially smaller scale.

source	ℓ (column samples)	Spectral error	Frobenius error	Trace error
Prior works				
[DM05] column sampling	$\Omega(\epsilon^{-4}k)$	$\mathrm{opt}_2 + \epsilon\sum_{i=}^{n} A_{ii}^2$	$\mathrm{opt}_F + \epsilon\sum_{i=1}^{n} A_{ii}^2$	–
[BW09] Nyström	$\Omega(1)$	–	–	$(n-\ell)/n\,\mathrm{Tr}(\mathbf{A})$
[TR10] Nyström	$\Omega(\mu_r r \log r)$	0	0	0
[KMT12] Nyström	$\Omega(1)$	$\mathrm{opt}_2 + n/\sqrt{\ell}\,\|\mathbf{A}\|_2$	$\mathrm{opt}_F + n(k/\ell)^{1/4}\,\|\mathbf{A}\|_2$	–
This work				
Nyström, Thm. 6.9	$\Omega((1-\epsilon)^{-2}\mu_k k \log k)$	$\mathrm{opt}_2(1 + n/(\epsilon\ell))$	$\mathrm{opt}_F + \epsilon^{-1}\mathrm{opt}_{\mathrm{tr}}$	$\mathrm{opt}_{\mathrm{tr}}(1+\epsilon^{-1})$
Leverage-based, Thm. 6.12	$\Omega(\epsilon^{-2}k\log k)$	$\mathrm{opt}_2 + \epsilon^2\mathrm{opt}_{\mathrm{tr}}$	$\mathrm{opt}_F + \epsilon\mathrm{opt}_{\mathrm{tr}}$	$(1+\epsilon^2)\mathrm{opt}_{\mathrm{tr}}$
SRFT-based, Thm. 6.16	$\Omega(\epsilon^{-1}k\log n)$	$(1-\sqrt{\epsilon})^{-1}(1 + k^{-1}\log n)\mathrm{opt}_2 + \epsilon\mathrm{opt}_{\mathrm{tr}}/((1-\sqrt{\epsilon})k)$	$\mathrm{opt}_F + \sqrt{\epsilon}\mathrm{opt}_{\mathrm{tr}}$	$(1+\epsilon)\mathrm{opt}_{\mathrm{tr}}$
Gaussian-based, Thm. 6.17	$\Omega(\epsilon^{-1}k)$	$(1+\epsilon^2)\mathrm{opt}_2 + (\epsilon^2/k)\mathrm{opt}_{\mathrm{tr}}$	$\mathrm{opt}_F + \epsilon\mathrm{opt}_{\mathrm{tr}}$	$(1+\epsilon^2)\mathrm{opt}_{\mathrm{tr}}$

Table 6.1: ASYMPTOTIC COMPARISON OF OUR BOUNDS ON SPSD SKETCHES WITH PRIOR WORK. Here, ℓ is the number of column samples sufficient for the stated bounds to hold, μ_d indicates the coherence of the top d-dimensional eigenspace, opt_ξ with $\xi \in \{2, \mathrm{F}, \mathrm{tr}\}$ is the smallest ξ-norm error possible when approximating $\mathbf{A}$ with a rank-k matrix, $r = \mathrm{rank}(\mathbf{A})$, and k is the target rank. The sketch analyzed in [DM05] samples columns with probabilities proportional to their Euclidean norms. All bounds hold with constant probability.

6.4 Proof of the deterministic error bounds

In this section, we develop deterministic spectral, Frobenius, and trace norm error bounds for SPSD sketches. Along the way, we establish a connection between the accuracy of SPSD sketches and the performance of randomized *column subset selection*.

Our results are based on the observation that approximations which satisfy our SPSD sketching model can be written in terms of a projection onto a subspace of the range of the square root of the matrix being approximated.

Lemma 6.1. *Let* $\mathbf{A}$ *be an SPSD matrix and* $\mathbf{S}$ *be a conformal sketching matrix. Then when* $\mathbf{C} = \mathbf{AS}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{AS}$*, the corresponding SPSD sketch satisfies*

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

Proof. The orthoprojector onto the range of any matrix $\mathbf{X}$ satisfies $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^\dagger\mathbf{X}^T$. It follows that

$$\begin{aligned}
\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T &= \mathbf{AS}(\mathbf{S}^T\mathbf{AS})^\dagger\mathbf{S}^T\mathbf{A} \\
&= \mathbf{A}^{1/2}[\mathbf{A}^{1/2}\mathbf{S}(\mathbf{S}^T\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{S})^\dagger\mathbf{S}^T\mathbf{A}^{1/2}]\mathbf{A}^{1/2} \\
&= \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}\mathbf{A}^{1/2}.
\end{aligned}$$

□

6.4.1 Spectral-norm bounds

Our first theorem bounds the spectral-norm error of multiple-pass SPSD sketches.

Theorem 6.2. *Let* $\mathbf{A}$ *be an SPSD matrix of size* n*, and let* $\mathbf{S}$ *be an* $n \times \ell$ *matrix. Fix an integer* $p \geq 1$*. Partition* $\mathbf{A}$ *as in* (6.2.1) *and define* $\mathbf{\Omega}_1$ *and* $\mathbf{\Omega}_2$ *as in* (6.2.2)*.*

Assume $\mathbf{\Omega}_1$ *has full row-rank. Then when* $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$*, the corresponding SPSD sketch satisfies*

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 = \|\left(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}}\right)\mathbf{A}^{1/2}\|_2^2 \leq \left\|\mathbf{\Sigma}_2\right\|_2 + \|\mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^{2/(2p-1)}. \tag{6.4.1}$$

If $\mathbf{\Omega}_1$ *has full row-rank and, additionally,* $\operatorname{rank}(\mathbf{A}) < k$*, then in fact*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T.$$

Remark 6.3. We emphasize that the first relation in (6.4.1) is an equality. This equality holds when $\mathbf{A}^{1/2}$ is replaced with any generalized Cholesky factorization of $\mathbf{A}$: by appropriately modifying the proof of Theorem 6.2, it can be seen that if $\mathbf{\Pi}\mathbf{A}\mathbf{\Pi}^T = \mathbf{B}^T\mathbf{B}$ where $\mathbf{\Pi}$ is a permutation matrix, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{B}(\mathbf{B}^T\mathbf{B})^{p-1}\mathbf{\Pi}\mathbf{S}})\mathbf{B}\|_2^2$$

as well. We take $\mathbf{\Pi} = \mathbf{I}$ and $\mathbf{B} = \mathbf{A}^{1/2}$ in this chapter, but other factorizations may be of interest.

Remark 6.4. Given a matrix $\mathbf{M}$, the goal of *column subset selection* is to choose a small but informative subset $\mathbf{C}$ of the columns of $\mathbf{M}$. Informativity can be defined in many ways; in our context, $\mathbf{C}$ is informative if, after approximating $\mathbf{M}$ with the matrix obtained by projecting

$\mathbf{M}$ onto the span of $\mathbf{C}$, the residual $(\mathbf{I} - \mathbf{P}_{\mathbf{C}})\mathbf{M}$ is small in the spectral norm. In randomized column subset selection, the columns $\mathbf{C}$ are chosen randomly, either uniformly or according to some data-dependent distribution. Column subset selection has important applications in statistical data analysis and has been investigated by both the numerical linear algebra and the theoretical computer science communities. For an introduction to the column subset selection literature biased towards approaches involving randomization, we refer the interested reader to the surveys [Mah12, Mah11].

To see the connection of SPSD sketching to the column subset selection problem, model the column sampling operation as follows: let $\mathbf{S}$ be a random matrix with ℓ columns, each of which has exactly one nonzero element. Then right multiplication by $\mathbf{S}$ selects ℓ columns from $\mathbf{A}$. The distribution of $\mathbf{S}$ reflects the type of randomized column sampling being performed. In the case of the Nyström extension, $\mathbf{S}$ is distributed as the first ℓ columns of a matrix sampled uniformly at random from the set of all permutation matrices.

Let $p = 1$, then (6.4.1) states that

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

That is, the spectral-norm error of a Nyström extension is exactly the square of the spectral-norm error in approximating $\mathbf{A}^{1/2}$ with a projection onto its corresponding columns. Thus, the problem of constructing high-quality Nyström extensions of $\mathbf{A}$ is equivalent to the randomized column subset selection problem for $\mathbf{A}^{1/2}$.

To establish Theorem 6.2, we use the following bound on the error incurred by projecting a matrix onto a random subspace of its range (an immediate corollary of [HMT11, Theorems 9.1 and 9.2]). This result is similar to Lemma 4.8, but is better adapted to investigating multiple-pass

sketches, and it provides a guarantee of exact recovery when $\mathbf{M}$ is sufficiently low-rank.

Lemma 6.5. *Let $\mathbf{M}$ be an SPSD matrix of size n, and let $q \geq 0$ be an integer. Fix integers k and ℓ satisfying $1 \leq k \leq \ell \leq n$.*

Let $\mathbf{U}_1$ and $\mathbf{U}_2$ be matrices with orthonormal columns spanning, respectively, a top k-dimensional eigenspace of $\mathbf{M}$ and the corresponding bottom $(n-k)$-dimensional eigenspace of $\mathbf{M}$. Let $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ be the diagonal matrices of eigenvalues corresponding, respectively, to the top k-dimensional eigenspace of $\mathbf{M}$ and the bottom $(n-k)$-dimensional eigenspace of $\mathbf{M}$.

Given a matrix $\mathbf{S}$ of size $n \times \ell$, define $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ and $\mathbf{\Omega}_2 = \mathbf{U}_2^T\mathbf{S}$. Assume that $\mathbf{\Omega}_1$ has full row-rank. Then

$$\|(\mathbf{I} - \mathbf{P}_{\mathbf{M}^{2q+1}\mathbf{S}})\mathbf{M}\|_2^2 \leq \|\mathbf{\Sigma}_2\|_2^2 + \left\|\mathbf{\Sigma}_2^{2q+1}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^{2/(2q+1)}.$$

If $\mathbf{\Omega}_1$ has full row-rank and, additionally, $\mathbf{\Sigma}_1$ is singular, then

$$\mathbf{M} = \mathbf{P}_{\mathbf{M}^{2q+1}\mathbf{S}}\mathbf{M}.$$

Proof of Theorem 6.2. Define a sketching matrix $\mathbf{S}' = \mathbf{A}^{p-1}\mathbf{S}$, then

$$\mathbf{C} = \mathbf{A}\mathbf{S}' \quad \text{and } \mathbf{W} = (\mathbf{S}')^T\mathbf{A}\mathbf{S}'.$$

Apply Lemma 6.1 with the sketching matrix $\mathbf{S}'$ to see that

$$\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T = \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}(\mathbf{A}^{p-1}\mathbf{S})}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}}\mathbf{A}^{1/2}.$$

It follows that the spectral error of the Nyström extension satisfies

$$\begin{aligned}\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\|_2 &= \|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}})\mathbf{A}^{1/2}\|_2 = \|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}})^2\mathbf{A}^{1/2}\|_2 \\ &= \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}})\mathbf{A}^{1/2}\|_2^2 = \|(\mathbf{I} - \mathbf{P}_{(\mathbf{A}^{1/2})^{2p-1}\mathbf{S}})\mathbf{A}^{1/2}\|_2^2 \\ &= \|(\mathbf{I} - \mathbf{P}_{(\mathbf{A}^{1/2})^{2(p-1)+1}\mathbf{S}})\mathbf{A}^{1/2}\|_2^2.\end{aligned}$$

The second equality holds because orthogonal projections are idempotent. The third follows from the fact that $\|\mathbf{M}\mathbf{M}^T\|_2 = \|\mathbf{M}\|_2^2$ for any matrix $\mathbf{M}$. Partition $\mathbf{A}$ as in (6.2.1). Equation (6.4.1) and the following assertion hold by Lemma 6.5 with $\mathbf{M} = \mathbf{A}^{1/2}$ and $q = p - 1$. □

6.4.2 Frobenius-norm bounds

Next, we prove a bound on the Frobenius norm of the residual error of SPSD sketches. The proof parallels that for the spectral-norm bound, in that we use the connection between SPSD sketches and column-based approximations to $\mathbf{A}^{1/2}$, but the analysis is more involved. The starting point of our proof is the perturbation argument used in the proof of [HMT11, Theorem 9.1], which was in turn inspired by Stewart's work on the perturbation of projections [Ste77].

Theorem 6.6. *Let $\mathbf{A}$ be an SPSD matrix of size n, and let $\mathbf{S}$ be an $n \times \ell$ matrix. Fix an integer $p \geq 1$. Partition $\mathbf{A}$ as in (6.2.1) and let $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ be as defined in (6.2.2). Define*

$$\gamma = \frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}.$$

Assume $\mathbf{\Omega}_1$ has full row-rank. Then when $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$, the corresponding SPSD sketch satisfies

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

If $\mathbf{\Omega}_1$ *has full row-rank and, additionally,* $\operatorname{rank}(\mathbf{A}) < k$*, then in fact*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T.$$

Proof. First, we observe that if $\operatorname{rank}(\mathbf{A}) < k$ and $\mathbf{\Omega}_1$ has full row-rank, then by Theorem 6.2, $\mathbf{A} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$.

To establish the claimed inequality, apply Lemma 6.1 with the sketching matrix $\mathbf{A}^{p-1}\mathbf{S}$ to see that

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

From this it follows that

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

To bound the quantity on the right-hand side, we first recall the fact [HMT11, Proposition 8.4] that for an arbitrary matrix $\mathbf{M}$, when $\mathbf{U}$ is a unitary matrix, $\mathbf{P}_{\mathbf{UM}} = \mathbf{U}\mathbf{P}_{\mathbf{M}}\mathbf{U}^T$. Then we use the unitary invariance of the Frobenius norm to obtain

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

Then we take

$$\mathbf{Z} = \mathbf{\Sigma}^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\mathbf{\Sigma}_1^{-(p-1/2)} = \begin{pmatrix} \mathbf{I} \\ \mathbf{F} \end{pmatrix}, \tag{6.4.2}$$

where $\mathbf{I} \in \mathbb{R}^{k\times k}$ and $\mathbf{F} \in \mathbb{R}^{n-k\times k}$ is given by $\mathbf{F} = \mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\mathbf{\Sigma}_1^{-(p-1/2)}$. The last equality in 6.4.2

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 $\mathbf{\Sigma}^{p-1/2}\mathbf{U}^T\mathbf{S}$,

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

By construction, $\mathbf{Z}$ has full column rank, thus $\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1/2}$ is an orthonormal basis for the span of $\mathbf{Z}$, and

$$\begin{aligned}
\mathbf{I} - \mathbf{P}_\mathbf{Z} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T &= \mathbf{I} - \begin{pmatrix} \mathbf{I} \\ \mathbf{F} \end{pmatrix} (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{F}^T \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & -(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \\ -\mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & \mathbf{I} - \mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \end{pmatrix}. \qquad (6.4.3)
\end{aligned}$$

This implies that

$$\begin{aligned}
E^2 &\leq \left\| \mathbf{\Sigma}^{1/2} \begin{pmatrix} \mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & -(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \\ -\mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} & \mathbf{I} - \mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \end{pmatrix} \mathbf{\Sigma}^{1/2} \right\|_\mathrm{F}^2 \\
&= \left\| \mathbf{\Sigma}_1^{1/2}\left(\mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\right)\mathbf{\Sigma}_1^{1/2} \right\|_\mathrm{F}^2 + 2\left\| \mathbf{\Sigma}_1^{1/2}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{\Sigma}_2^{1/2} \right\|_\mathrm{F}^2 \qquad (6.4.4) \\
&\quad + \left\| \mathbf{\Sigma}_2^{1/2}\left(\mathbf{I} - \mathbf{F}(\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\right)\mathbf{\Sigma}_2^{1/2} \right\|_\mathrm{F}^2 \\
&= T_1 + T_2 + T_3.
\end{aligned}$$

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

$$T_3 \leq \left\| \mathbf{\Sigma}_2 \right\|_\mathrm{F}^2.$$

Likewise, the fact that $\mathbf{I} - (\mathbf{I} + \mathbf{F}^T\mathbf{F})^{-1} \preceq \mathbf{F}^T\mathbf{F}$ (easily seen with an SVD) implies that we can

bound T_1 with

$$\begin{aligned}
T_1 &\leq \left\|\boldsymbol{\Sigma}_1^{1/2}\mathbf{F}^T\mathbf{F}\boldsymbol{\Sigma}_1^{1/2}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{F}\boldsymbol{\Sigma}_1^{1/2}\right\|_2^2 \left\|\mathbf{F}\boldsymbol{\Sigma}_1^{1/2}\right\|_{\mathrm{F}}^2 \\
&= \left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Sigma}_1^{-(p-1)}\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Sigma}_1^{-(p-1)}\right\|_{\mathrm{F}}^2 \\
&\leq \left\|\boldsymbol{\Sigma}_2^{p-1}\right\|_2^4 \left\|\boldsymbol{\Sigma}_1^{-(p-1)}\right\|_2^4 \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2 \\
&= (\left\|\boldsymbol{\Sigma}_2\right\|_2 \left\|\boldsymbol{\Sigma}_1^{-1}\right\|_2)^{4(p-1)} \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2 \\
&= \left(\frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}\right)^{4(p-1)} \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2 \\
&= \gamma^{4(p-1)} \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2.
\end{aligned}$$

We proceed to bound T_2 by using the estimate

$$T_2 \leq 2\left\|\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\right\|_2^2 \left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2. \tag{6.4.5}$$

To develop the spectral norm term, observe that for any SPSD matrix $\mathbf{M}$ with eigenvalue decomposition $\mathbf{M} = \mathbf{V}\mathbf{D}\mathbf{V}^T$,

$$\begin{aligned}
(\mathbf{I}+\mathbf{M})^{-1}\mathbf{M}(\mathbf{I}+\mathbf{M})^{-1} &= (\mathbf{V}\mathbf{V}^T + \mathbf{V}\mathbf{D}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T + \mathbf{V}\mathbf{D}\mathbf{V}^T)^{-1} \\
&= \mathbf{V}(\mathbf{I}+\mathbf{D})^{-1}\mathbf{D}(\mathbf{I}+\mathbf{D})^{-1}\mathbf{V}^T \\
&\preceq \mathbf{V}\mathbf{D}\mathbf{V}^T = \mathbf{M}.
\end{aligned}$$

It follows that

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\right\|_2^2 &= \left\|\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{F}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\boldsymbol{\Sigma}_1^{1/2}\right\|_2 \\ &\leq \left\|\boldsymbol{\Sigma}_1^{1/2}\mathbf{F}^T\mathbf{F}\boldsymbol{\Sigma}_1^{1/2}\right\|_2 = \left\|\mathbf{F}\boldsymbol{\Sigma}_1^{1/2}\right\|_2^2 \\ &= \left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\boldsymbol{\Sigma}_1^{-(p-1)}\right\|_2^2 \\ &\leq \left\|\boldsymbol{\Sigma}_2^{p-1}\right\|_2^2\left\|\boldsymbol{\Sigma}_1^{-(p-1)}\right\|_2^2\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \\ &= \left(\frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}\right)^{2(p-1)}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2.\end{aligned}$$

Identifying γ and using this estimate in (6.4.5), we conclude that

$$T_2 \leq 2\gamma^{2(p-1)}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2\left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2.$$

Combining our estimates for T_1, T_2, and T_3 in (6.4.4) gives

$$\begin{aligned}E^2 &= \left\|\mathbf{A}^{1/2}\left(\mathbf{I}-\mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}}\right)\mathbf{A}^{1/2}\right\|_{\mathrm{F}}^2 \\ &\leq \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}}^2+\gamma^{2(p-1)}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2\cdot\left(2\left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2+\gamma^{2(p-1)}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2\right).\end{aligned}$$

Establish the claimed bound by applying the subadditivity of the square-root function:

$$E \leq \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}}+\gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2\cdot\left(\sqrt{2\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right)}+\gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}\right).$$

□

Remark. The quality of approximation guarantee provided by Theorem 6.6 depends on two quantities, $\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2$ and $\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}$, that depend in two slightly different ways on how the eigenstructure of $\mathbf{A}$ interacts with the sketching matrix. As we see in Sections 6.5 and 6.6,

the extent to which we can bound each of these for different sketching procedures is slightly different.

6.4.3 Trace-norm bounds

Finally, we prove a bound on the trace norm of the residual error of SPSD sketches. The proof method is analogous to that for the spectral and Frobenius norm bounds.

Theorem 6.7. *Let* $\mathbf{A}$ *be an SPSD matrix of size* n*, and let* $\mathbf{S}$ *be an* $n \times \ell$ *matrix. Fix an integer* $p \geq 1$*. Partition* $\mathbf{A}$ *as in* (6.2.1)*, and let* $\mathbf{\Omega}_1$ *and* $\mathbf{\Omega}_2$ *be as defined in* (6.2.2)*. Define*

$$\gamma = \frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}.$$

Assume $\mathbf{\Omega}_1$ *has full row-rank. Then when* $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$*, the corresponding SPSD sketch satisfies*

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

If $\mathbf{\Omega}_1$ *has full row-rank and, additionally,* $\operatorname{rank}(\mathbf{A}) < k$*, then in fact*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T.$$

Proof. First, we observe that if $\operatorname{rank}(\mathbf{A}) < k$ and $\mathbf{\Omega}_1$ has full row-rank, then by Theorem 6.2, $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$.

Now observe that

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

where $\mathbf{Z}$ is defined in (6.4.2). The expression for $\mathbf{I}-\mathbf{P}_{\mathbf{Z}}$ given in (6.4.3) implies that

$$\operatorname{Tr}\left(\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\boldsymbol{\Sigma}^{1/2}\right) = \operatorname{Tr}\left(\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}-(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1})\boldsymbol{\Sigma}_1^{1/2}\right) + \operatorname{Tr}\left(\boldsymbol{\Sigma}_2^{1/2}(\mathbf{I}-\mathbf{F}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T)\boldsymbol{\Sigma}_2^{1/2}\right).$$

Recall the estimate $\mathbf{I}-(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1} \preceq \mathbf{F}^T\mathbf{F}$ and the basic estimate $\mathbf{I}-\mathbf{F}(\mathbf{I}+\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \preceq \mathbf{I}$. Together these imply that

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

The first equality follows from substituting the definition of $\mathbf{F}$ and identifying the squared Frobenius norm. The last equality follows from identifying γ. We have established the claimed bound.

□

6.5 Error bounds for Nyström extensions

In this section, we use the structural results from Section 6.4 to bound the approximation errors of Nyström extensions. To obtain our results, we use Theorems 6.2, 6.6, and 6.7 in conjunction with the estimate of $\|\mathbf{\Omega}_1^\dagger\|_2^2$ provided by the following lemma.

Lemma 6.8. *Let* $\mathbf{U}$ *be an* $n \times k$ *matrix with orthonormal columns. Take* μ *to be the coherence of* $\mathbf{U}$. *Select* $\epsilon \in (0,1)$ *and a nonzero failure probability* δ. *Let* $\mathbf{S}$ *be a random matrix distributed as the first* ℓ *columns of a uniformly random permutation matrix of size* n, *where*

$$\ell \geq \frac{2\mu}{(1-\epsilon)^2} k \log \frac{k}{\delta}.$$

Then with probability exceeding $1-\delta$, *the matrix* $\mathbf{U}^T\mathbf{S}$ *has full row rank and satisfies*

$$\left\|(\mathbf{U}^T\mathbf{S})^\dagger\right\|_2^2 \leq \frac{n}{\epsilon l}.$$

Proof of Lemma 6.8. Note that $\mathbf{U}^T\mathbf{S}$ has full row rank if $\lambda_k(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}) > 0$. Furthermore,

$$\left\|(\mathbf{U}^T\mathbf{S})^\dagger\right\|_2^2 = \lambda_k^{-1}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}).$$

Thus to obtain both conclusions of the lemma, it is sufficient to verify that

$$\lambda_k(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}) \geq \frac{\epsilon\ell}{n}$$

when ℓ is as stated.

We apply the lower Chernoff bound given as (4.1.1) to bound the probability that this inequality is not satisfied. Let $\mathbf{u}_i$ denote the ith column of $\mathbf{U}^T$ and $\mathscr{X} = \{\mathbf{u}_i \mathbf{u}_i^T\}_{i=1,\ldots,n}$. Then

$$\lambda_k(\mathbf{U}^T \mathbf{S}\mathbf{S}^T \mathbf{U}) = \lambda_k \left(\sum_{i=1}^{\ell} \mathbf{X}_i \right),$$

where the $\mathbf{X}_i$ are chosen uniformly at random, without replacement, from $\mathscr{X}$. Clearly

$$B = \max_i \|\mathbf{u}_i\|^2 = \frac{k}{n}\mu \quad \text{and} \quad \mu_{\min} = \ell \cdot \lambda_k(\mathbb{E}\mathbf{X}_1) = \frac{\ell}{n}\lambda_k(\mathbf{U}^T\mathbf{U}) = \frac{\ell}{n}.$$

The Chernoff bound yields

$$\mathbb{P}\left\{ \lambda_k \left(\mathbf{U}^T \mathbf{S}\mathbf{S}^T \mathbf{U} \right) \leq \epsilon \frac{\ell}{n} \right\} \leq k \cdot \mathrm{e}^{-(1-\epsilon)^2 \ell/(2k\mu)}.$$

We require enough samples that

$$\lambda_k(\mathbf{U}^T \mathbf{S}\mathbf{S}^T \mathbf{U}) \geq \epsilon \frac{\ell}{n}$$

with probability greater than $1 - \delta$, so we set

$$k \cdot \mathrm{e}^{-(1-\epsilon)^2 \ell/(2k\mu)} \leq \delta$$

and solve for ℓ, finding

$$\ell \geq \frac{2\mu}{(1-\epsilon)^2} k \log \frac{k}{\delta}.$$

Thus, for values of ℓ satisfying this inequality, we achieve the stated spectral error bound and ensure that $\mathbf{U}^T \mathbf{S}$ has full row rank. □

Theorem 6.9 establishes that the error incurred by the simple Nyström extension process is small when an appropriate number of columns is sampled. Specifically, if the number of columns sampled is proportional to the coherence of the top eigenspace of the matrix, and grows with the desired target rank as $k \log k$, then the bounds provided in the theorem hold.

Note that the theorem provides two spectral norm error bounds. The first is a relative-error bound which compares the spectral norm error with the smallest error achievable when approximating $\mathbf{A}$ with a rank-k matrix. It does not use any information about the spectrum of $\mathbf{A}$ other than the value of the $(k+1)$st eigenvalue. The second bound uses information about the entire tail of the spectrum of $\mathbf{A}$. If the spectrum of $\mathbf{A}$ decays fast, the second bound is much tighter than the first. If the spectrum of $\mathbf{A}$ is flat, then the first bound is tighter.

Theorem 6.9. *Let $\mathbf{A}$ be an SPSD matrix of size n, and $p \geq 1$ be an integer. Given an integer $k \leq n$, partition $\mathbf{A}$ as in (6.2.1). Let μ denote the coherence of $\mathbf{U}_1$. Fix a failure probability $\delta \in (0,1)$ and accuracy factor $\epsilon \in (0,1)$. If $\mathbf{S}$ comprises*

$$\ell \geq 2\mu\epsilon^{-2} k \log(k/\delta)$$

columns of the identity matrix, sampled uniformly at random without replacement, and $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$, then the corresponding SPSD sketch satisfies

$$\begin{aligned}
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 &\leq \left(1 + \left(\frac{n}{(1-\epsilon)\ell}\right)^{1/(2p-1)}\right)\left\|\mathbf{A} - \mathbf{A}_k\right\|_2, \\
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \left(\frac{1}{\delta^2(1-\epsilon)}\operatorname{Tr}\left((\mathbf{A} - \mathbf{A}_k)^{2p-1}\right)\right)^{1/(2p-1)}, \\
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} + \left(\frac{\gamma^{p-1}}{\delta}\sqrt{\frac{2}{1-\epsilon}} + \frac{\gamma^{2p-2}}{\delta^2(1-\epsilon)}\right)\operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k\right), \text{ and} \\
\operatorname{Tr}\left(\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right) &\leq \left(1 + \frac{\gamma^{2p-2}}{\delta^2(1-\epsilon)}\right)\operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k\right),
\end{aligned}$$

simultaneously, with probability at least $1 - 4\delta$.

If, additionally, $k \geq \text{rank}(\mathbf{A})$, *then*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}$$

with probability exceeding $1 - \delta$.

Remark 6.10. Theorem 6.9, like the main result of [TR10], promises exact recovery with probability at least $1 - \delta$ when $\mathbf{A}$ is exactly rank k and has small coherence, with a sample of $O(k \log(k/\delta))$ columns. Unlike the result in [TR10], Theorem 6.9 is applicable in the case that $\mathbf{A}$ is full-rank but has a sufficiently fastly decaying spectrum.

Remark 6.11. Let $p = 1$. The first spectral-norm error bound in Theorem 6.9 simplifies to

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_2 \leq \left(1 + \frac{n}{(1-\epsilon)\ell}\right) \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 .$$

The multiplicative factor in this relative error bound is optimal in terms of its dependence on n and ℓ. This fact follows from the connection between Nyström extensions and the column subset selection problem.

Indeed, [BDMI11] establishes the following lower bound for the column subset selection problem: for any $\alpha > 0$ there exist matrices $\mathbf{M}_\alpha$ such that for any $k \geq 1$ and any $\ell \geq 1$, the error of approximating $\mathbf{M}_\alpha$ with $\mathbf{P}_\mathbf{D}\mathbf{M}_\alpha$, where $\mathbf{D}$ may be *any* subset of ℓ columns of $\mathbf{M}_\alpha$, satisfies

$$\left\|\mathbf{M}_\alpha - \mathbf{P}_\mathbf{D}\mathbf{M}_\alpha\right\|_2 \geq \sqrt{\frac{n+\alpha^2}{\ell+\alpha^2}} \cdot \left\|\mathbf{M}_\alpha - (\mathbf{M}_\alpha)_k\right\|_2 ,$$

where $(\mathbf{M}_\alpha)_k$ is the rank-k matrix that best approximates $\mathbf{M}_\alpha$ in the spectral norm. We get a lower bound on the error of the Nyström extension by taking $\mathbf{A}_\alpha = \mathbf{M}_\alpha^T\mathbf{M}_\alpha$: it follows from the remark following Theorem 6.2 that for any $k \geq 1$ and $\ell \geq 1$, any Nyström extension formed

using $\mathbf{C} = \mathbf{A}_\alpha \mathbf{S}$ consisting of ℓ columns sampled from $\mathbf{A}_\alpha$ satisfies

$$\left\| \mathbf{A}_\alpha - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T \right\|_2 \geq \frac{n+\alpha^2}{\ell+\alpha^2} \cdot \lambda_{k+1}(\mathbf{A}_\alpha).$$

Proof of Theorem 6.9. The sketching matrix $\mathbf{S}$ is formed by taking the first ℓ columns of a uniformly sampled random permutation matrix. Recall that $\mathbf{\Omega}_1 = \mathbf{U}_1^T \mathbf{S}$ and $\mathbf{\Omega}_2 = \mathbf{U}_2^T \mathbf{S}$. By Lemma 6.8, $\mathbf{\Omega}_1$ has full row-rank with probability at least $1-\delta$, so the bounds in Theorems 6.2, 6.6, and 6.7 are applicable. In particular, if $k > \operatorname{rank}(\mathbf{A})$ then $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T$.

First we use Theorems 6.2 and 6.7 to develop the first spectral-norm error bound and the trace-norm error bound. To apply these theorems, we need estimates of the quantities

$$\left\| \mathbf{\Sigma}_2^{p-1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger \right\|_2^2 \quad \text{and} \quad \left\| \mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger \right\|_{\mathrm{F}}.$$

Applying Lemma 6.8, we see that $\|\mathbf{\Omega}_1^\dagger\|_2^2 \leq n/((1-\epsilon)\ell)$ with probability exceeding $1-\delta$. Observe that $\left\|\mathbf{\Omega}_2\right\|_2 \leq \left\|\mathbf{U}_2\right\|_2 \|\mathbf{S}\|_2 \leq 1$, so

$$\left\| \mathbf{\Sigma}_2^{p-1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger \right\|_2^2 \leq \left\| \mathbf{\Sigma}_2^{p-1/2} \right\|_2^2 \left\| \mathbf{\Omega}_1^\dagger \right\|_2^2 \leq \frac{n}{(1-\epsilon)\ell} \left\| \mathbf{\Sigma}_2 \right\|_2^{2p-1}. \tag{6.5.1}$$

Likewise,

$$\left\| \mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger \right\|_{\mathrm{F}} \leq \left\| \mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \right\|_{\mathrm{F}} \left\| \mathbf{\Omega}^\dagger \right\|_2 \leq \sqrt{\frac{n}{(1-\epsilon)\ell}} \left\| \mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \right\|_{\mathrm{F}}. \tag{6.5.2}$$

These estimates hold simultaneously with probability at least $1-\delta$.

To further develop (6.5.2), observe that since $\mathbf{S}$ selects ℓ columns uniformly at random,

$$\mathbb{E} \left\| \mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \right\|_{\mathrm{F}}^2 = \mathbb{E} \left\| \mathbf{\Sigma}_2^{1/2} \mathbf{U}_2^T \mathbf{S} \right\|_{\mathrm{F}}^2 = \sum\nolimits_{i=1}^{\ell} \mathbb{E} \|\mathbf{x}_i\|^2,$$

where the summands $\mathbf{x}_i$ are distributed uniformly at random over the columns of $\mathbf{\Sigma}_2^{1/2}\mathbf{U}_2^T$. The summands all have the same expectation:

$$\mathbb{E}\|\mathbf{x}_i\|^2 = \frac{1}{n}\sum_{j=1}^{n}\|(\mathbf{\Sigma}_2\mathbf{U}_2^T)_{(j)}\|^2 = \frac{1}{n}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{U}_2\right\|_{\mathrm{F}}^2 = \frac{1}{n}\left\|\mathbf{\Sigma}_2^{1/2}\right\|_{\mathrm{F}}^2 = \operatorname{Tr}\left(\mathbf{\Sigma}_2\right).$$

Consequently,

$$\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_{\mathrm{F}}^2 = \frac{\ell}{n}\operatorname{Tr}\left(\mathbf{\Sigma}_2\right),$$

so by Jensen's inequality

$$\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_{\mathrm{F}} \leq \left(\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_{\mathrm{F}}^2\right)^{1/2} = \sqrt{\frac{\ell}{n}\operatorname{Tr}\left(\mathbf{\Sigma}_2\right)}.$$

Now applying Markov's inequality to (6.5.2), we see that

$$\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \frac{1}{\delta}\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \frac{1}{\delta}\sqrt{\frac{1}{(1-\epsilon)}\operatorname{Tr}\left(\mathbf{\Sigma}_2\right)}$$

with probability at least $1-2\delta$. Clearly one can similarly show that

$$\left\|\mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \frac{1}{\delta}\sqrt{\frac{1}{(1-\epsilon)}\operatorname{Tr}\left(\mathbf{\Sigma}_2^{2p-1}\right)} \tag{6.5.3}$$

with probability at least $1-2\delta$.

Thus far, we have established that $\mathbf{\Omega}_1$ has full row-rank and the estimates

$$\left\|\mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2^2 \leq \frac{n}{(1-\epsilon)\ell}\left\|\mathbf{\Sigma}_2\right\|_2^{2p-1} \text{ and} \tag{6.5.4}$$

$$\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \frac{1}{\delta}\sqrt{\frac{1}{(1-\epsilon)}\operatorname{Tr}\left(\mathbf{\Sigma}_2\right)} \tag{6.5.5}$$

hold simultaneously with probability at least $1 - 2\delta$.

These estimates used in Theorems 6.2 and 6.7 yield the trace-norm error bound and the first spectral-norm error bound.

To develop the second spectral-norm error bound, observe that by (6.5.3), we also have the estimate

$$\left\| \boldsymbol{\Sigma}_2^{p-1/2} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger \right\|_2^2 \leq \left\| \boldsymbol{\Sigma}_2^{p-1/2} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger \right\|_{\mathrm{F}}^2 \leq \frac{1}{\delta^2(1-\epsilon)} \operatorname{Tr}\left(\boldsymbol{\Sigma}_2^{2p-1} \right)$$

which holds with probability at least $1 - 2\delta$. This estimate used in Theorem 6.2 yields the second spectral-norm bound.

To develop the Frobenius-norm bound, observe that Theorem 6.6 can be weakened to the assertion that, when $\boldsymbol{\Omega}_1$ has full row-rank, the estimate

$$\left\| \mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T \right\|_{\mathrm{F}} \leq \left\| \boldsymbol{\Sigma}_2 \right\|_{\mathrm{F}} + \gamma^{p-1} \left\| \boldsymbol{\Sigma}_2^{1/2} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger \right\|_{\mathrm{F}} \sqrt{2 \operatorname{Tr}\left(\boldsymbol{\Sigma}_2 \right)} + \gamma^{2p-2} \left\| \boldsymbol{\Sigma}_2^{1/2} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger \right\|_{\mathrm{F}}^2 \tag{6.5.6}$$

holds. Recall that with probability at least $1 - 2\delta$, the estimate (6.5.5) holds and $\boldsymbol{\Omega}_1$ has full row-rank. Insert the estimate (6.5.5) into (6.5.6) to establish the claimed Frobenius-norm error bound.

□

6.6 Error bounds for random mixture-based SPSD sketches

In this section, we apply Theorems 6.2, 6.6, and 6.7 to bound the reconstruction errors for several random mixture-based sketches that conform to our SPSD sketching model. In particular, we consider the following schemes, corresponding to different choices of the sketching matrix:

- sketches formed by sampling columns according to an importance sampling distribution

that depends on the statistical leverage scores (in Section 6.6.1);

- sketches formed from mixtures of the columns formed using subsampled randomized Fourier transformations (in Section 6.6.2); and

- sketches formed from Gaussian mixtures of the columns (in Section 6.6.3).

6.6.1 Sampling with leverage-based importance sampling probabilities

We first consider a Nyström-like scheme that approximates $\mathbf{A}$ using column sampling. Specifically, we consider sketches where the columns of $\mathbf{A}$ are sampled with replacement according to a nonuniform probability distribution determined by the (exact or approximate) statistical leverage scores of $\mathbf{A}$ relative to the best rank-k approximation to $\mathbf{A}$. Previous work has highlighted the fact that the leverage scores reflect the nonuniformity properties of the top k-dimensional eigenspace of $\mathbf{A}$ [PZB+07, DM09, Mah11, YMS+13]. Thus it is reasonable to expect that sketches formed in this manner should better approximate $\mathbf{A}$ than Nyström extensions.

Fix β in $(0, 1]$. A probability distribution on the columns of $\mathbf{A}$ is considered to be β-close to the leverage score distribution if it satisfies

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

Our bounds apply to probability distributions that are β-close to the leverage score distribution.

The sketching matrices associated with a fixed β-close leverage-based probability distribution are factored into the product of two random matrices, $\mathbf{S} = \mathbf{RD}$. Here $\mathbf{R} \in \mathbb{R}^{n \times \ell}$ is a column selection matrix that samples columns of $\mathbf{A}$ according to the given distribution. That is, $\mathbf{R}_{ij} = 1$ if and only if the ith column of $\mathbf{A}$ is the jth column selected. The matrix $\mathbf{D}$ is a diagonal rescaling matrix with entries satisfying $\mathbf{D}_{jj} = 1/\sqrt{\ell p_i}$ if and only if $\mathbf{R}_{ij} = 1$. We have the following bounds

on the error of approximations formed using these sketching matrices.

Theorem 6.12. *Let* $\mathbf{A}$ *be an SPSD matrix of size* n*, and let* $p \geq 1$ *be an integer. Given an integer* $k \leq n$*, partition* $\mathbf{A}$ *as in* (6.2.1)*. 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*

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

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

$$\gamma = \frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}.$$

If $\ell \geq 3200(\beta\epsilon^2)^{-1} k \log(4k/(\beta\delta))$*, then, when* $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$*, the corresponding low-rank SPSD approximation satisfies*

$$\left\| \mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T \right\|_2 \leq \left\| \mathbf{A} - \mathbf{A}_k \right\|_2 + \left(\epsilon^2 \operatorname{Tr}\left((\mathbf{A} - \mathbf{A}_k)^{2p-1} \right) \right)^{1/(2p-1)}, \tag{6.6.1}$$

$$\left\| \mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T \right\|_{\mathrm{F}} \leq \left\| \mathbf{A} - \mathbf{A}_k \right\|_{\mathrm{F}} + \left(\sqrt{2}\epsilon\gamma^{p-1} + \epsilon^2\gamma^{2(p-1)} \right) \operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k \right), \text{ and} \tag{6.6.2}$$

$$\operatorname{Tr}\left(\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T \right) \leq (1 + \gamma^{2(p-1)}\epsilon^2) \operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k \right), \tag{6.6.3}$$

simultaneously, with probability at least $1 - 6\delta - 0.6$*.*

Proof. Recall that $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ and $\mathbf{\Omega}_2 = \mathbf{U}_2^T\mathbf{S}$. To apply our deterministic error bounds, we need estimates for the quantities

$$\left\| \mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger \right\|_{\mathrm{F}}^{2/(2p-1)}, \quad \left\| \mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger \right\|_2, \quad \text{and} \quad \left\| \mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger \right\|_{\mathrm{F}}.$$

In [MTJ12, proof of Proposition 22] it is shown that if ℓ satisfies the given bound and the

samples are drawn from an approximate subspace probability distribution, then for any SPSD diagonal matrix $\mathbf{D}$,

$$\|\mathbf{D}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}} \leq \epsilon \|\mathbf{D}\|_{\mathrm{F}}$$

with probability at least $1 - 2\delta - 0.2$; further, $\boldsymbol{\Omega}_1$ has full row-rank when this estimate holds. Thus, the estimates

$$\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \epsilon \left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}} = \epsilon\sqrt{\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)} = \epsilon\sqrt{\mathrm{Tr}\left(\mathbf{A} - \mathbf{A}_k\right)},$$

and

$$\begin{aligned}
\left(\left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_2\right)^{2/(2p-1)} &\leq \left(\left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}\right)^{2/(2p-1)} \\
&\leq \left(\epsilon^2\left\|\boldsymbol{\Sigma}_2^{p-1/2}\right\|_{\mathrm{F}}^2\right)^{1/(2p-1)} \\
&= \left(\epsilon^2\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_2^{2p-1}\right)\right)^{1/(2p-1)} \\
&= \left(\epsilon^2\,\mathrm{Tr}\left((\mathbf{A} - \mathbf{A}_k)^{2p-1}\right)\right)^{1/(2p-1)}
\end{aligned}$$

each hold, individually, with probability at least $1 - 2\delta - 0.2$. Taking $p = 1$ in this last estimate, we see that

$$\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_2 \leq \epsilon\sqrt{\mathrm{Tr}\left(\mathbf{A} - \mathbf{A}_k\right)}$$

also holds with probability at least $1 - 2\delta - 0.2$. Thus these three estimates hold and $\boldsymbol{\Omega}_1$ has full row-rank, simultaneously, with probability at least $1 - 6\delta - 0.6$. These three estimates used in Theorems 6.2, 6.6, and 6.7 yield the bounds given in the statement of the theorem. □

Remark 6.13. The additive scale factors for the spectral and Frobenius norm bounds are much improved relative to the prior results of [DM05]. At root, this is because the leverage score

importance sampling probabilities expose the structure of the data, e.g., which columns to select so that $\mathbf{\Omega}_1$ has full row rank, in a more refined way than the sampling probabilities used in [DM05].

Remark 6.14. These improvements come at additional computational expense, but leverage-based sampling probabilities of the form used by Theorem 6.12 can sometimes be computed faster than the time needed to compute the basis $\mathbf{U}_1$. In [DMIMW12], for example, it is shown that the leverage scores of $\mathbf{A}$ can be approximated; the cost of this computation is determined by the time required to perform a random projection-based low-rank approximation to the matrix.

Remark 6.15. Not surprisingly, constant factors such as 3200 (as well as other similarly large factors below) and a failure probability bounded away from zero are artifacts of the analysis; the empirical behavior of this sampling method is much better. This has been observed previously [DMM08, DM09] and is seen in the experimental results presented in this chapter.

6.6.2 Random projections with subsampled randomized Fourier transforms

We now consider sketches in which the columns of $\mathbf{A}$ are randomly mixed using a subsampled randomized Fourier transform before sampling. That is, $\mathbf{S} = \sqrt{n/\ell}\mathbf{DFR}$, where $\mathbf{D}$ is a diagonal matrix of Rademacher random variables, $\mathbf{F}$ is the normalized Fourier transform matrix, and $\mathbf{R}$ restricts to ℓ columns. We prove the following bounds on the errors of approximations formed using such sketching matrices.

Theorem 6.16. *Let $\mathbf{A}$ be an SPSD matrix of size n, and let $p \geq 1$ be an integer. Given an integer k satisfying $4 \leq k \leq n$, partition $\mathbf{A}$ as in (6.2.1). Let $\mathbf{S} = \sqrt{n/\ell}\mathbf{DFR}$ be a sketching matrix of size $n \times \ell$, where $\mathbf{D}$ is a diagonal matrix of Rademacher random variables, $\mathbf{F}$ is a normalized Fourier matrix of size $n \times n$, and $\mathbf{R}$ restricts to ℓ columns. Fix a failure probability $\delta \in (0,1)$ and*

approximation factor $\epsilon \in (0,1)$*, and define*

$$\gamma = \frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}.$$

If $\ell \geq 24\epsilon^{-1}[\sqrt{k} + \sqrt{8\log(8n/\delta)}]^2 \log(8k/\delta)$*, then, when* $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$*, the corresponding low-rank SPSD approximation satisfies*

$$\begin{aligned}
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 &\leq \left[1 + \left(\frac{1}{1-\sqrt{\epsilon}} \cdot \left(5 + \frac{16\log(n/\delta)^2}{\ell}\right)\right)^{1/(2p-1)}\right] \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 \\
&\quad + \left(\frac{2\log(n/\delta)}{(1-\sqrt{\epsilon})\ell}\right)^{1/(2p-1)} \operatorname{Tr}\left((\mathbf{A} - \mathbf{A}_k)^{2p-1}\right)^{1/(2p-1)}, \\
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} + \left(5\gamma^{p-1}\sqrt{\epsilon} + 11\gamma^{2p-2}\epsilon\right) \operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k\right), \text{ and} \\
\operatorname{Tr}\left(\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right) &\leq (1 + 11\epsilon\gamma^{2p-2}) \operatorname{Tr}\left(\mathbf{A} - \mathbf{A}_k\right)
\end{aligned} \tag{6.6.4}$$

simultaneously, with probability at least $1 - 2\delta$*.*

Proof. Recall that $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ and $\mathbf{\Omega}_2 = \mathbf{U}_2^T\mathbf{S}$.

In Chapter 5 (cf. the proof of Theorem 5.13), it is shown that for this choice of $\mathbf{S}$ and number of samples ℓ,

$$\begin{aligned}
\left\|\mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\right\|_2^2 &\leq \frac{1}{1-\sqrt{\epsilon}} \cdot \left(5\left\|\mathbf{\Sigma}_2^{p-1/2}\right\|_2^2 \right. \\
&\qquad \left. + \frac{\log(n/\delta)}{\ell}\left(\left\|\mathbf{\Sigma}_2^{p-1/2}\right\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\left\|\mathbf{\Sigma}_2^{p-1/2}\right\|_2\right)^2\right) \\
&= \frac{1}{1-\sqrt{\epsilon}} \cdot \left(5\left\|\mathbf{\Sigma}_2\right\|_2^{2p-1} \right. \\
&\qquad \left. + \frac{\log(n/\delta)}{\ell}\left(\operatorname{Tr}\left(\mathbf{\Sigma}_2^{2p-1}\right)^{1/2} + \sqrt{8\log(n/\delta)}\left\|\mathbf{\Sigma}_2\right\|_2^{p-1/2}\right)^2\right) \\
&\leq \frac{1}{1-\sqrt{\epsilon}} \cdot \left(\left(5 + \frac{16\log(n/\delta)^2}{\ell}\right)\left\|\mathbf{\Sigma}_2\right\|_2^{2p-1} + \frac{2\log(n/\delta)}{\ell}\operatorname{Tr}\left(\mathbf{\Sigma}_2^{2p-1}\right)\right)
\end{aligned}$$

and

$$\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}} \leq \sqrt{11\epsilon}\left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}} = \sqrt{11\epsilon\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)}$$

each hold, individually, with probability at least $1-\delta$. When either estimate holds, $\boldsymbol{\Omega}_1$ has full row-rank. These estimates used in Theorems 6.2 and 6.7 yield the stated bounds for the spectral and trace-norm errors.

The Frobenius-norm bound follows from the same estimates and a simplification of the bound stated in Theorem 6.6:

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_2\left(\sqrt{2\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}\right)\\
&\leq \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}\sqrt{2\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \gamma^{2(p-1)}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^2\\
&\leq \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \left(\gamma^{p-1}\sqrt{22\epsilon} + 11\gamma^{2p-2}\epsilon\right)\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right).
\end{aligned}$$

The stated failure probability comes from the fact that the two estimates used hold simultaneously with probability at least $1-2\delta$. □

Remark. Suppressing the dependence on δ and ϵ, the spectral norm bound ensures that when $p=1$, $k=\Omega(\log n)$ and $\ell=\Omega(k\log k)$, then

$$\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 = \mathrm{O}\left(\frac{\log n}{\log k}\left\|\mathbf{A}-\mathbf{A}_k\right\|_2 + \frac{1}{\log k}\mathrm{Tr}\left(\mathbf{A}-\mathbf{A}_k\right)\right).$$

This should be compared to the guarantee established in Theorem 6.17 below for Gaussian-based SPSD sketches constructed using just $\ell=\mathrm{O}(k)$:

$$\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 = \mathrm{O}\left(\left\|\mathbf{A}-\mathbf{A}_k\right\|_2 + \frac{1}{k}\mathrm{Tr}\left(\mathbf{A}-\mathbf{A}_k\right)\right).$$

Theorem 6.16 guarantees that errors on this order can be achieved by SRFT sketches if one increases the number of samples by a logarithm factor in the dimension: specifically, such a bound is achieved when $k = \Omega(\log n)$ and $\ell = \Omega(k \log k \log n)$. The difference between the number of samples necessary for Fourier-based sketches and Gaussian-based sketches is reflective of the difference in our understanding of the relevant random projections: the geometry of any k-dimensional subspace is preserved under projection onto the span of $\ell = \mathrm{O}(k)$ Gaussian random vectors [HMT11], but the sharpest analysis available suggests that to preserve the geometry of such a subspace under projection onto the span of ℓ SRFT vectors, ℓ must satisfy $\ell = \Omega(\max\{k, \log n\} \log k)$ [Tro11b]. We note, however, that in practice the Fourier-based and Gaussian-based SPSD sketches have similar reconstruction errors.

6.6.3 Random projections with i.i.d. Gaussian random matrices

The final class of SPSD sketches we consider are mixture-based sketches in which the columns of $\mathbf{A}$ are randomly mixed using Gaussian random variables before sampling. That is, the entries of the sketching matrix $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ are i.i.d. standard Gaussian random variables. We consider the case where the number of column samples is comparable to and only slightly larger than the desired rank, i.e., $\ell = \mathrm{O}(k)$.

Theorem 6.17. *Let $\mathbf{A}$ be an SPSD matrix of size n, and let $p \geq 1$ be an integer. Given an integer k satisfying $4 < k \leq n$, partition $\mathbf{A}$ as in (6.2.1). Let $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ be a sketching matrix of i.i.d standard Gaussians. Define*

$$\gamma = \frac{\lambda_{k+1}(\mathbf{A})}{\lambda_k(\mathbf{A})}.$$

Assume $\ell = (1 + \epsilon^{-2})k$, where $\epsilon \in (0, 1]$. Then, when $\mathbf{C} = \mathbf{A}^p\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{A}^{2p-1}\mathbf{S}$, the

corresponding low-rank SPSD approximation satisfies

$$\begin{aligned}
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_2 &\leq \left(1+\left(89\epsilon^2+874\epsilon^2\frac{\log k}{k}\right)^{1/(2p-1)}\right)\left\|\mathbf{A}-\mathbf{A}_k\right\|_2 \\
&\quad+\left(219\frac{\epsilon^2}{k}\right)^{1/(2p-1)}\cdot\operatorname{Tr}\left((\mathbf{A}-\mathbf{A}_k)^{2p-1}\right)^{1/(2p-1)}, \\
\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_{\mathrm{F}} &\leq \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}+\left[\gamma^{p-1}\epsilon\left(5+6\sqrt{\frac{\log k}{k}}\right)\right. \\
&\quad\left.+\gamma^{2p-2}\epsilon^2\left(45+190\sqrt{\frac{\log k}{k}}+309\frac{\sqrt{\log k}}{k}\right)\right]\sqrt{\left\|\mathbf{A}-\mathbf{A}_k\right\|_2\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right)} \\
&\quad+\left(21\gamma^{p-1}\frac{\epsilon}{\sqrt{k}}+70\gamma^{2p-2}\frac{\epsilon^2}{\sqrt{k}}\right)\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right) \\
&\quad+\gamma^{2p-2}\epsilon^2\left(197\sqrt{\frac{\log k}{k}}+618\frac{\log k}{k}\right)\left\|\mathbf{A}-\mathbf{A}_k\right\|_2, \\
\operatorname{Tr}\left(\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right) &\leq (1+45\gamma^{2p-2}\epsilon^2)\operatorname{Tr}\left(\mathbf{A}-\mathbf{A}_k\right)+874\gamma^{2p-2}\epsilon^2\frac{\log k}{k}\left\|\mathbf{A}-\mathbf{A}_k\right\|_2
\end{aligned}$$

simultaneously, with probability at least $1-2k^{-1}-4\mathrm{e}^{-k/\epsilon^2}$.

Proof. As before, this result is established by bounding the quantities involved in the deterministic error bounds of Theorems 6.2, 6.6, and 6.7. Recall that $\mathbf{\Omega}_1=\mathbf{U}_1^T\mathbf{S}$ and $\mathbf{\Omega}_2=\mathbf{U}_2^T\mathbf{S}$. We need to develop bounds on the quantities

$$\left\|\mathbf{\Sigma}_2^{p-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}^{2/(2p-1)}, \quad \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2, \quad \text{and} \quad \left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}.$$

The following deviation bounds, established in [HMT11, Section 10], are useful in that regard: if $\mathbf{D}$ is a diagonal matrix, $\ell=k+s$ with $s>4$ and $u,t\geq 1$, then

$$\begin{aligned}
\mathbb{P}\left\{\left\|\mathbf{D}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_2>\|\mathbf{D}\|_2\left(\sqrt{\frac{3k}{s+1}}\cdot t+\frac{\mathrm{e}\sqrt{\ell}}{s+1}\cdot tu\right)+\|\mathbf{D}\|_{\mathrm{F}}\frac{\mathrm{e}\sqrt{\ell}}{s+1}\cdot t\right\}&\leq 2t^{-s}+\mathrm{e}^{-u^2/2}, \text{ and} \\
\mathbb{P}\left\{\left\|\mathbf{D}\mathbf{\Omega}_2\mathbf{\Omega}_1^{\dagger}\right\|_{\mathrm{F}}>\|\mathbf{D}\|_{\mathrm{F}}\sqrt{\frac{3k}{s+1}}\cdot t+\|\mathbf{D}\|_2\frac{\mathrm{e}\sqrt{\ell}}{s+1}\cdot tu\right\}&\leq 2t^{-s}+\mathrm{e}^{-u^2/2}. \quad (6.6.5)
\end{aligned}$$

Define $s = k\epsilon^{-2}$, so that $\ell = k + s$. Estimate the terms in (6.6.5) with

$$\sqrt{\frac{3k}{s+1}} \le \sqrt{\frac{3k}{s}} = \sqrt{3}\epsilon \quad \text{and}$$
$$\frac{\sqrt{\ell}}{s+1} \le \frac{\epsilon^2\sqrt{k(1+1/\epsilon^2)}}{k} \le \epsilon\sqrt{\frac{2}{k}}$$

and take $t = \mathrm{e}$ and $u = \sqrt{2\log k}$ in (6.6.5) to obtain that

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_2^{p-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2^2 &\le \left[\left(\sqrt{3}\mathrm{e} + 2\mathrm{e}^2\sqrt{\frac{\log k}{k}}\right)\epsilon\cdot\left\|\boldsymbol{\Sigma}_2^{p-1/2}\right\|_2 + \frac{2\mathrm{e}^2\epsilon}{\sqrt{k}}\cdot\left\|\boldsymbol{\Sigma}_2^{p-1/2}\right\|_{\mathrm{F}}\right]^2 \\ &\le 2\left(\sqrt{3}\mathrm{e} + 2\mathrm{e}^2\sqrt{\frac{\log k}{k}}\right)^2\epsilon^2\cdot\left\|\boldsymbol{\Sigma}_2\right\|_2^{2p-1} + \frac{4\mathrm{e}^4\epsilon^2}{k}\cdot\mathrm{Tr}\left(\boldsymbol{\Sigma}_2^{2p-1}\right) \\ &\le \left(12\mathrm{e}^2 + 16\mathrm{e}^4\frac{\log k}{k}\right)\epsilon^2\cdot\left\|\boldsymbol{\Sigma}_2\right\|_2^{2p-1} + \frac{4\mathrm{e}^4\epsilon^2}{k}\cdot\mathrm{Tr}\left(\boldsymbol{\Sigma}_2^{2p-1}\right)\end{aligned}$$

with probability at least $1 - k^{-1} - 2\mathrm{e}^{-k/\epsilon^2}$ and

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}} &\le \sqrt{3}\epsilon\mathrm{e}\cdot\left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_{\mathrm{F}} + \mathrm{e}^2\epsilon\sqrt{\frac{8\log k}{k}}\cdot\left\|\boldsymbol{\Sigma}_2^{1/2}\right\|_2 \\ &= \sqrt{3}\epsilon\mathrm{e}\cdot\sqrt{\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \mathrm{e}^2\epsilon\sqrt{\frac{8\log k}{k}\left\|\boldsymbol{\Sigma}_2\right\|_2}\end{aligned}$$

with the same probability. Likewise,

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}^2 &\le \left(\sqrt{3}\epsilon\mathrm{e}\cdot\sqrt{\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \mathrm{e}^2\epsilon\sqrt{\frac{8\log k}{k}\left\|\boldsymbol{\Sigma}_2\right\|_2}\right)^2 \\ &\le 6\epsilon^2\mathrm{e}^2\cdot\mathrm{Tr}\left(\boldsymbol{\Sigma}_2\right) + 16\mathrm{e}^4\epsilon^2\frac{\log k}{k}\cdot\left\|\boldsymbol{\Sigma}_2\right\|_2\end{aligned}$$

with the same probability.

These estimates used in Theorems 6.2 and 6.7 yield the stated spectral and trace-norm

bounds. To obtain the corresponding Frobenius-norm bound, define the quantities

$$F_1 = \left(12e^2 + 16e^4\frac{\log k}{k}\right)\epsilon^2, \qquad F_3 = 3e^2\epsilon^2,$$
$$F_2 = 4e^4\frac{\epsilon^2}{k}, \qquad F_4 = 8e^4\frac{\log k}{k}\epsilon^2$$

for notational convenience. By Theorem 6.6 and our estimates for the quantities $\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2$ and $\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}$,

$$\begin{aligned}
\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2 \cdot \left(\sqrt{2\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \gamma^{p-1}\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}\right) \\
&\le \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \gamma^{p-1}(F_1\left\|\boldsymbol{\Sigma}_2\right\|_2 + F_2\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right))^{1/2}\times \\
&\qquad \left(\sqrt{2\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \gamma^{p-1}\sqrt{F_3\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right)} + \gamma^{p-1}\sqrt{F_4\left\|\boldsymbol{\Sigma}_2\right\|_2}\right) \\
&\le \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}} + \left(\gamma^{p-1}\sqrt{2F_1} + \gamma^{2p-2}(\sqrt{F_1F_3} + \sqrt{F_2F_4})\right)\cdot\sqrt{\left\|\boldsymbol{\Sigma}_2\right\|_2\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right)} \\
&\quad + \left(\gamma^{p-1}\sqrt{2F_2} + \gamma^{2p-2}\sqrt{F_2F_3}\right)\cdot\operatorname{Tr}\left(\boldsymbol{\Sigma}_2\right) \\
&\quad + \gamma^{2p-2}\sqrt{F_1F_4}\left\|\boldsymbol{\Sigma}_2\right\|_2.
\end{aligned} \tag{6.6.6}$$

The following estimates hold for the coefficients in this inequality:

$$\sqrt{2F_1} \le \left(5 + 6\sqrt{\frac{\log k}{k}}\right)\epsilon, \qquad \sqrt{F_1F_3} \le \left(45 + 140\sqrt{\frac{\log k}{k}}\right)\epsilon^2,$$
$$\sqrt{F_2F_4} \le 309\frac{\sqrt{\log k}}{k}\epsilon^2, \qquad \sqrt{2F_2} \le 21\frac{\epsilon}{\sqrt{k}},$$
$$\sqrt{F_2F_3} \le 70\frac{\epsilon^2}{\sqrt{k}}, \qquad \sqrt{F_1F_4} \le \left(197\sqrt{\frac{\log k}{k}} + 618\frac{\log k}{k}\right)\epsilon^2.$$

The Frobenius norm bound follows from using these estimates in (6.6.6) and grouping terms

appropriately:

$$\left\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_{\mathrm{F}} \leq \left\|\mathbf{\Sigma}_2\right\|_{\mathrm{F}} + \left[\gamma^{p-1}\epsilon\left(5+6\sqrt{\frac{\log k}{k}}\right) + \gamma^{2p-2}\epsilon^2\left(45+190\sqrt{\frac{\log k}{k}}+309\frac{\sqrt{\log k}}{k}\right)\right]\sqrt{\left\|\mathbf{\Sigma}_2\right\|_2 \operatorname{Tr}\left(\mathbf{\Sigma}_2\right)} + \left(21\gamma^{p-1}\frac{\epsilon}{\sqrt{k}}+70\gamma^{2p-2}\frac{\epsilon^2}{\sqrt{k}}\right)\operatorname{Tr}\left(\mathbf{\Sigma}_2\right) + \gamma^{2p-2}\epsilon^2\left(197\sqrt{\frac{\log k}{k}}+618\frac{\log k}{k}\right)\left\|\mathbf{\Sigma}_2\right\|_2.$$

□

Remark 6.18. The way we have parameterized these bounds for Gaussian-based projections makes explicit the dependence on various parameters, but obscures the structural simplicity of these bounds. In particular, since $\|\cdot\|_2 \leq \|\cdot\|_{\mathrm{F}} \leq \operatorname{Tr}(\cdot)$, note that the Frobenius norm bounds are upper bounded by a term that depends on the Frobenius norm of $\mathbf{A}-\mathbf{A}_k$ and a term that depends on the trace norm of $\mathbf{A}-\mathbf{A}_k$; and that, similarly, the trace norm bounds are upper bounded by a multiplicative factor times the optimal rank-k approximation error. This factor can be set to $1+\nu$ with ν arbitrarily small.

6.7 Stable algorithms for computing regularized SPSD sketches

The error bounds we have provided for SPSD sketches assume that the calculations are carried out in exact arithmetic. In reality, since the formation of SPSD sketches requires the computation of a linear system, i.e. the computation of $\mathbf{W}^{\dagger}\mathbf{C}^{T}$, a direct application of the SPSD sketching procedure may not produce a result that is close to a valid SPSD approximation. Specifically, if $\mathbf{W}$ is ill-conditioned, the product $\mathbf{W}^{\dagger}\mathbf{C}^{T}$ may not be computed accurately.

Algorithm 6.1: SPSD sketch, regularized via additive perturbation [WS01]

Input: an $n \times n$ SPSD matrix $\mathbf{A}$; a regularization parameter $\rho > 0$; and an $n \times \ell$ sketching matrix $\mathbf{S}$.

Output: an $n \times \ell$ matrix $\mathbf{C}_\rho$; an $\ell \times \ell$ SPSD matrix $\mathbf{W}_\rho$; and the sketch $\tilde{\mathbf{A}} = \mathbf{C}_\rho \mathbf{W}_\rho^\dagger \mathbf{C}_\rho^T$.

1: Let $\mathbf{A}_\rho = \mathbf{A} + \rho \mathbf{I}$.
2: Form the matrix $\mathbf{C}_\rho = \mathbf{A}_\rho \mathbf{S}$.
3: Form the matrix $\mathbf{W}_\rho = \mathbf{S}^T \mathbf{C}_\rho$.
4: Compute $\mathbf{Y} = \mathbf{W}_\rho^\dagger \mathbf{C}_\rho^T$.
5: Form the matrix $\tilde{\mathbf{A}} = \mathbf{C}_\rho \mathbf{Y}$.
6: Return $\mathbf{C}_\rho$, $\mathbf{W}_\rho$, and $\tilde{\mathbf{A}}$.

In the seminal paper [WS01], the authors propose Algorithm 6.1, an algorithm for computing regularized SPSD sketches. Algorithm 6.1 returns the SPSD sketch of the matrix $\mathbf{A}_\rho = \mathbf{A} + \rho \mathbf{I}$, where $\rho > 0$ is a regularization parameter. Note that $\mathbf{A}_\rho$ and $\mathbf{A}$ have the same eigenvectors, so the eigenvectors of the sketch returned by Algorithm 6.1 approximate those of $\mathbf{A}$. This suggests that the sketch returned by Algorithm 6.1 may be relevant in applications where the goal is to approximate the eigenvectors of $\mathbf{A}$ with those of an SPSD sketch [FBCM04, FNL+09, BF12]. In this section, we provide the first theoretical analysis of Algorithm 6.2.

The paper [CD11] investigates the performance of the column and row sampling-based CUR decomposition, an approximate matrix decomposition for rectangular matrices that is analogous to the Nyström extension. The results apply immediately to the Nyström extension, because Nyström extensions are simply CUR decompositions of SPSD matrices, where the columns and rows sampled are constrained to be the same. In particular, [CD11] introduces an algorithm for computing CUR decompositions stably; in the context of SPSD sketches, this algorithm becomes Algorithm 6.2. Algorithm 6.2 replaces $\mathbf{W}$ with a regularized matrix $\mathbf{W}_\rho$ in which all eigenvalues of $\mathbf{W}$ smaller than the regularization parameter ρ are set to zero. We compare Algorithms 6.1 and 6.2 empirically in Section 6.8.

We begin our analysis of Algorithm 6.1 by observing that the product $\mathbf{W}_\rho^\dagger \mathbf{C}_\rho^T$ can be computed

Algorithm 6.2: SPSD sketch, regularized via a truncated eigendecomposition [CD11, Algorithm 1]

Input: an $n \times n$ SPSD matrix $\mathbf{A}$; a regularization parameter $\rho > 0$; and an $n \times \ell$ sketching matrix $\mathbf{S}$.

Output: an $n \times \ell$ matrix $\mathbf{C}$; an $\ell \times \ell$ SPSD matrix $\mathbf{W}_\rho$; and the SPSD sketch $\tilde{\mathbf{A}} = \mathbf{C}\mathbf{W}_\rho^\dagger \mathbf{C}^T$.

1: Form the matrix $\mathbf{C} = \mathbf{A}\mathbf{S}$.
2: Form the matrix $\mathbf{W} = \mathbf{S}^T\mathbf{C}$.
3: Compute the SVD of $\mathbf{W}$. Set all components with eigenvalues smaller than ρ to zero to obtain $\mathbf{W}_\rho$.
4: Compute $\mathbf{Y} = \mathbf{W}_\rho^\dagger \mathbf{C}^T$.
5: Form the matrix $\tilde{\mathbf{A}} = \mathbf{C}\mathbf{Y}$.
6: Return $\mathbf{C}$, $\mathbf{W}_\rho$, and $\tilde{\mathbf{A}}$.

stably if the two-norm condition number

$$\kappa_2(\mathbf{W}_\rho) = \|\mathbf{W}_\rho\|_2 \|\mathbf{W}_\rho^\dagger\|_2 = \frac{\lambda_1(\mathbf{W}_\rho)}{\lambda_{\min}(\mathbf{W}_\rho)}$$

is small [GV96]. Here, $\lambda_{\min}(\mathbf{W}_\rho)$ denotes the smallest nonzero eigenvalue of $\mathbf{W}_\rho$. It follows that Algorithm 6.1 will stably compute a regularized SPSD sketch when $\kappa_2(\mathbf{W}_\rho)$ is small. The following lemma relates $\kappa_2(\mathbf{W}_\rho)$ to the regularization parameter ρ and the sketching matrix $\mathbf{S}$.

Lemma 6.19. *Let $\mathbf{A}$ be an SPSD matrix of size n, and let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be a sketching matrix. Fix a regularization parameter $\rho > 0$. The two-norm condition number of the matrix $\mathbf{W}_\rho$ returned by Algorithm 6.1 satisfies*

$$\kappa_2(\mathbf{W}_\rho) \leq \left(\frac{\lambda_1(\mathbf{A})}{\rho} + 1\right) \kappa_2(\mathbf{S})^2.$$

Proof. Observe that, because $\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{I}$ is full rank, $\operatorname{rank}(\mathbf{A}_\rho^{1/2}\mathbf{S}) = \operatorname{rank}(\mathbf{S})$. It follows that $\operatorname{rank}(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S}) = \operatorname{rank}(\mathbf{S}^T\mathbf{S})$. Let r denote this common rank. Then, because $\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{I} \succeq \rho\mathbf{I}$,

we have

$$\lambda_{\min}(\mathbf{W}_\rho) = \lambda_{\min}(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S}) = \lambda_r(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S}) \geq \lambda_r(\rho\mathbf{S}^T\mathbf{S})$$
$$= \lambda_{\min}(\rho\mathbf{S}^T\mathbf{S}) = \rho\|\mathbf{S}^\dagger\|_2^{-2}.$$

Similarly, from the observation that $\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{I} \preceq (\lambda_1(\mathbf{A}) + \rho)\mathbf{I}$, we argue

$$\lambda_1(\mathbf{W}_\rho) = \lambda_1(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S}) \leq (\lambda_1(\mathbf{A}) + \rho)\|\mathbf{S}\|_2^2.$$

We reach the claimed bound on $\kappa_2(\mathbf{W}_\rho)$ by combining these estimates of $\lambda_{\min}(\mathbf{W}_\rho)$ and $\lambda_1(\mathbf{W}_\rho)$:

$$\kappa_2(\mathbf{W}) = \frac{\lambda_1(\mathbf{W}_\rho)}{\lambda_{\min}(\mathbf{W}_\rho)} \leq \rho^{-1}(\lambda_1(\mathbf{A}) + \rho)\|\mathbf{S}\|_2^2\|\mathbf{S}^\dagger\|_2^2 = \left(\frac{\lambda_1(\mathbf{A})}{\rho} + 1\right)\kappa_2(\mathbf{S})^2.$$

□

Remark 6.20. Let $\sigma_1(\cdot)$ and $\sigma_{\min}(\cdot)$ denote, respectively, the largest singular value and the smallest nonzero singular value of their arguments. Then

$$\kappa_2(\mathbf{W}_\rho) = \frac{\lambda_1(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S})}{\lambda_{\min}(\mathbf{S}^T\mathbf{A}_\rho\mathbf{S})} = \frac{\sigma_1(\mathbf{S}^T\mathbf{A}_\rho^{1/2})^2}{\sigma_{\min}(\mathbf{S}^T\mathbf{A}_\rho^{1/2})^2},$$

so good bounds on $\sigma_1(\mathbf{S}^T\mathbf{A}_\rho^{1/2})$ and $\sigma_{\min}(\mathbf{S}^T\mathbf{A}_\rho^{1/2})$ would lead to a sharper estimate of $\kappa_2(\mathbf{W}_\rho)$ than the one stated in Lemma 6.19. Such bounds could be developed for the sketches considered in this chapter.

Lemma 6.19 shows that the condition number of $\mathbf{W}_\rho$ is small when the sketching matrix has small condition number, and when the regularization parameter is large. However, it is clear

that taking ρ too large will result in a sketch which does a poor job of approximating $\mathbf{A}$. The following lemma quantifies this observation.

Lemma 6.21. *Let $\mathbf{A}$ be an SPSD matrix of size n, and let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be a sketching matrix. Fix a regularization parameter $\rho > 0$. Let $\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{I}$, $\mathbf{C}_\rho = \mathbf{A}_\rho\mathbf{S}$, and $\mathbf{W}_\rho = \mathbf{S}^T\mathbf{A}_\rho\mathbf{S}$. Then the SPSD sketch $\tilde{\mathbf{A}}$ returned by Algorithm 6.1 satisfies*

$$\begin{aligned}
\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2 &\leq \left\|\mathbf{A}_\rho - \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T\right\|_2 + \rho, \\
\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_{\mathrm{F}} &\leq \left\|\mathbf{A}_\rho - \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T\right\|_{\mathrm{F}} + \sqrt{n}\rho, \text{ and} \\
\mathrm{Tr}\left(\mathbf{A} - \tilde{\mathbf{A}}\right) &\leq \mathrm{Tr}\left(\mathbf{A}_\rho - \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T\right) + n\rho.
\end{aligned}$$

This lemma relates the errors of the regularized sketch of $\mathbf{A}$ to the error of an SPSD sketch of $\mathbf{A}+\rho\mathbf{I}$. Therefore, given a particular sketching model, this lemma can be used in conjunction with the results of Sections 6.5 and 6.6 to predict the errors of the regularized sketch. Concurrently, Lemma 6.19 can be used to quantify the stability of the sketch.

Proof. Recall that $\tilde{\mathbf{A}} = \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T$. By the triangle inequality,

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_\xi \leq \|\mathbf{A} - \mathbf{A}_\rho\|_\xi + \|\mathbf{A}_\rho - \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T\|_\xi = \|\mathbf{A}_\rho - \mathbf{C}_\rho\mathbf{W}_\rho^\dagger\mathbf{C}_\rho^T\|_\xi + \rho\|\mathbf{I}\|_\xi.$$

Calculate $\|\mathbf{I}\|_\xi$ to reach the stated error bounds. □

6.8 Computational investigations of the spectral-norm bound for Nyström extensions

In this section we demonstrate the tightness of the relative-error spectral-norm bound provided for the Nyström extension in Theorem 6.9 and compare Algorithms 6.1 and 6.2 for the computation of regularized Nyström extensions.

6.8.1 Optimality

In the first experiment, we use a matrix introduced in [BDMI11] to demonstrate the worst-case optimality of the dependence on n and ℓ in the relative-error spectral-norm bound provided in Theorem 6.9. Let $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ be defined by

$$\mathbf{A} = \mathbf{M}^T\mathbf{M} \quad \text{where} \quad \mathbf{M} = [\mathbf{e}_2 + \mathbf{e}_1, \quad \mathbf{e}_3 + \mathbf{e}_1, \quad \cdots, \quad \mathbf{e}_{1001} + \mathbf{e}_1]; \tag{6.8.1}$$

here $\mathbf{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^{1001}$. By construction $\lambda_{\min}(\mathbf{A}) = 1$, so Nyström extensions of $\mathbf{A}$ are always stably computable.

Figure 6.1 plots the ratio of the spectral-norm error of the Nyström extensions to the optimal rank-10 approximation error, as a function of the number of column samples ℓ. The ratio n/ℓ is provided for comparison. To capture the worst-case behavior of the Nyström extension, each point in the plot is the worst ratio observed in 60 trials. It is clear that the n/ℓ term present in the error bound is necessary.

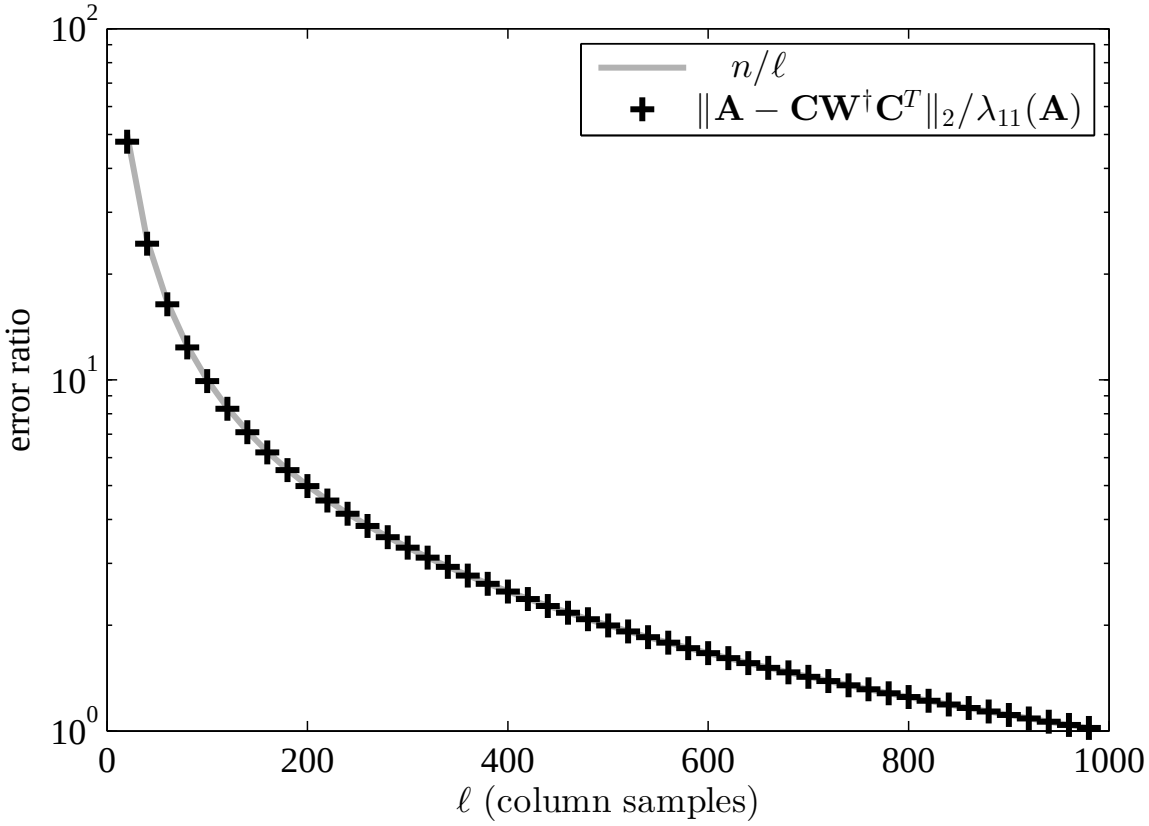


Figure 6.1: EMPIRICAL DEMONSTRATION OF THE OPTIMALITY OF THEOREM 6.9. The empirical spectral-norm error of Nyström extensions of $\mathbf{A}$, the matrix defined in (6.8.1), relative to the spectral-norm error of the optimal rank-10 approximation of $\mathbf{A}$. Each point is the worst relative error observed in 60 trials. The ratio n/ℓ is plotted; this is the dependence on n and ℓ of the bound given in Theorem 6.9.

6.8.2 Dependence on coherence

In the following experiments, we use 500×500 matrices $\mathbf{A}$ with eigendecompositions of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} \tag{6.8.2}$$

where $\mathbf{U}_1$ is a 500×10 matrix with orthonormal columns and specified coherence and the matrix $\mathbf{U}_2$ is chosen so that $\begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}$ is an orthogonal matrix. The 20 largest eigenvalues of $\mathbf{A}$ range logarithmically from 10 to 10^{-3}, and the remaining eigenvalues are identically 10^{-15}. Routines

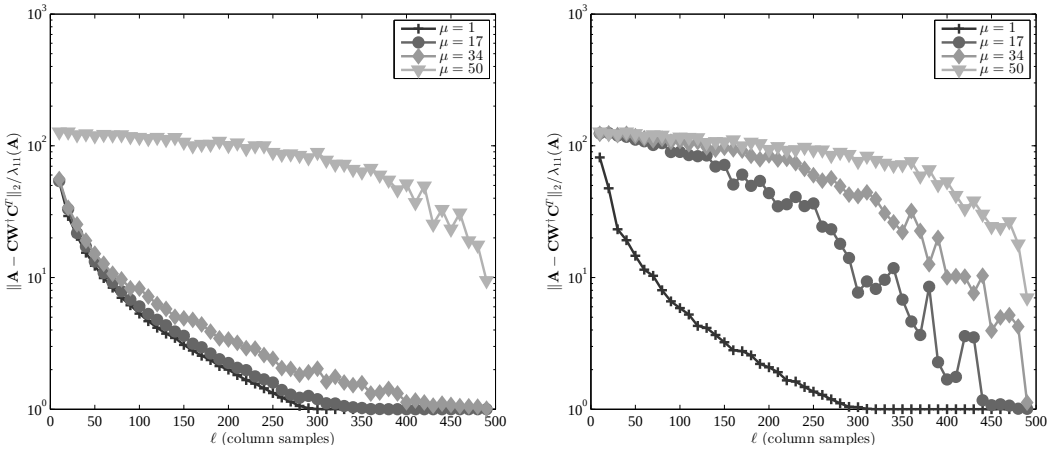
from the *kappaSQ* Matlab package introduced in [IW12] are used to generate $\mathbf{U}_1$ with specified coherences. For each value of coherence, we consider two types of matrices $\mathbf{U}_1$ achieving this coherence: dense $\mathbf{U}_1$, in which many rows of $\mathbf{U}_1$ are nonzero, and sparse $\mathbf{U}_1$, in which many rows of $\mathbf{U}_1$ are zero. Dense $\mathbf{U}_1$ are generated using the `mtxGenMethod1` routine, and sparse $\mathbf{U}_1$ are generated using the `mtxGenMethod3` routine.

We compare the accuracies of regularized Nyström extensions constructed using Algorithm 6.1, to those of regularized Nyström extensions constructed using Algorithm 6.2. In Figure 6.2 we plot the ratio of the approximation errors of the two regularized Nyström extensions to the approximation error of the optimal rank-10 approximant, as the coherence and sparsity of $\mathbf{U}_1$ vary. The regularization parameter ρ is assigned the value $\lambda_{11}(\mathbf{A})$.

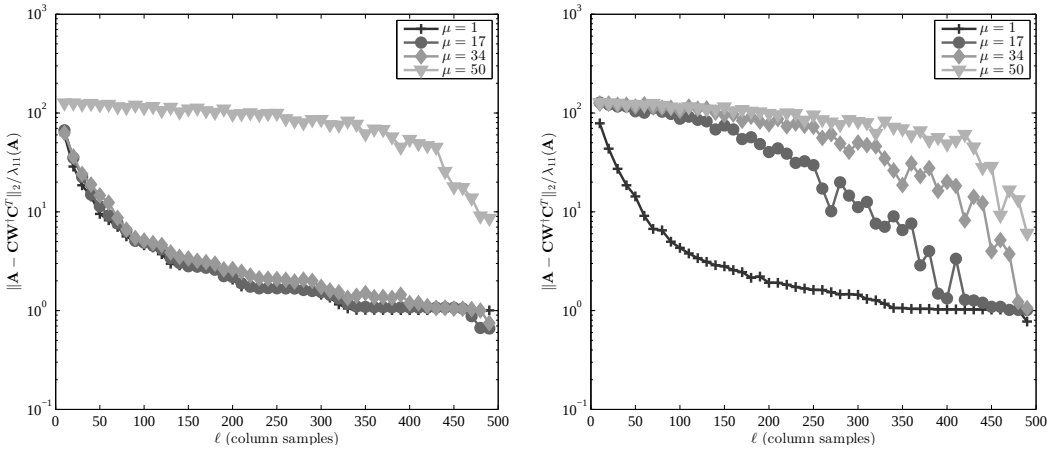
Both algorithms perform as suggested by Theorem 6.9: as the coherence of the top k-dimensional eigenspace increases, the number of samples needed to obtain a small relative error increases. Additionally, Figure 6.2 shows that the structure of the eigenvectors is as important as the coherence of the eigenspace: when the eigenvectors are dense, the number of samples needed to obtain a small relative error is much less sensitive to the coherence than when the eigenvectors are sparse. That is, for a fixed coherence and number of column samples, the Nyström extensions give lower errors when the eigenvectors are dense than they do when the eigenvectors are sparse.

Dependence on the regularization parameter

Both algorithms require the choice of a regularization parameter ρ. In Figure 6.3, we observe the effect of the regularization parameter ρ on the errors of the Nyström extensions. Here the



(a) Relative spectral-norm errors of Algorithm 6.1



(b) Relative spectral-norm errors of Algorithm 6.2

Figure 6.2: SPECTRAL-NORM ERRORS OF REGULARIZED NYSTRÖM EXTENSIONS AS COHERENCE VARIES. The relative spectral-norm errors of Nyström extensions of $\mathbf{A}$, the matrix defined in (6.8.2), generated using Algorithms 6.1 and 6.2, as a function of the coherence of the dominant 10-dimensional eigenspace. The errors are measured relative to the error of the optimal rank-10 approximation, and averaged over 60 runs for each value of ℓ. The eigenvectors spanning the dominant eigenspace of the matrices used in the experiments on the left-hand side are dense, and the corresponding eigenvectors of the matrices used in the experiments on the right-hand side are sparse. The coherences range from the minimum possible, 1, to the maximum of 50.

matrix $\mathbf{A}$ is again a 500×500 matrix with eigendecomposition

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix}, \tag{6.8.3}$$

where $\mathbf{U}_1$ is a 500×20 matrix with orthonormal columns and $\mathbf{U}_2$ is chosen so that $[\mathbf{U}_1 \quad \mathbf{U}_2]$ is an orthogonal matrix. The 40 dominant eigenvalues of $\mathbf{A}$ range logarithmically from 1 to 10^{-10} and all remaining eigenvalues are identically 10^{-10}. The `mtxGenMethod1` routine is used to construct $\mathbf{U}_1$ with coherence 1.

Figure 6.3 shows the ratios of the spectral-norm errors of the Nyström extension and the regularized extensions computed by Algorithms 6.1 and 6.2 to the optimal rank-20 approximation error. The number of columns used to form the extensions is fixed at $\ell = 200$, and the regularization parameter is varied from the minimum possible value of 1 to the maximum possible value of 50. We see that both regularization algorithms exhibit the same behavior: for large values of ρ, they have higher error than the Nyström extension; as ρ decreases, their errors become orders of magnitude smaller than that of the Nyström extension, and as ρ continues to decrease, their errors once again approach that of the Nyström extension. This behavior highlights the importance of choosing an appropriate regularization parameter: if ρ is too small then there is no benefit gained from the regularization, and if it is too large then the regularization has a deliterious effect. We also observe that Algorithm 6.2 can be orders of magnitude more accurate than Algorithm 6.1.

6.9 Empirical aspects of SPSD low-rank approximation

In this section, we examine the empirical performance of the SPSD sketches for which theoretical bounds were provided in Sections 6.5 and 6.6, on a diverse set of SPSD matrices.

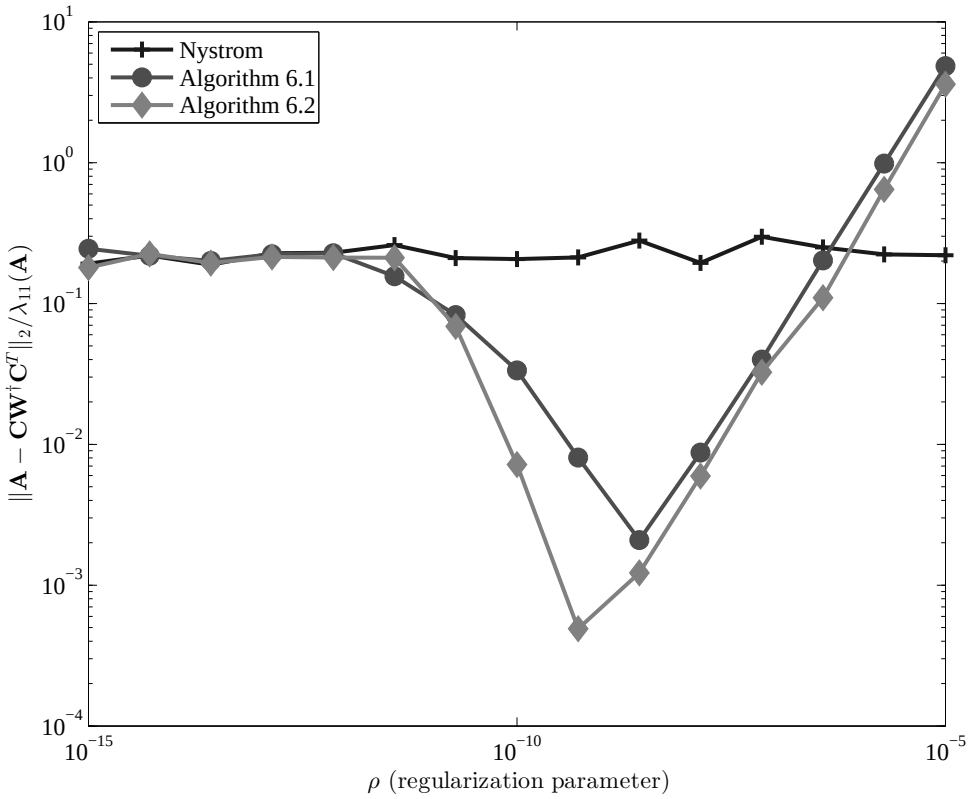


Figure 6.3: SPECTRAL-NORM ERROR OF REGULARIZED NYSTRÖM EXTENSIONS AS REGULARIZATION PARAMETER VARIES. For the matrix $\mathbf{A}$ defined in (6.8.3), the spectral-norm errors of the Nyström extension and the extensions generated using Algorithms 6.1 and 6.2, as a function of the regularization parameter ρ. The errors are averaged over 60 runs for each value of ρ and plotted relative to the optimal spectral-norm rank-10 approximation error.

6.9.1 Test matrices

Table 6.2 provides summary statistics for the test matrices used in our computational experiments. In order to illustrate the strengths and weaknesses of encountered in machine learning and data analysis applications, we drawn our test matrices from the following 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

Name	Description	n	d	%nnz
Laplacian kernels				
HEP	arXiv High Energy Physics collaboration graph	9877	NA	0.06
GR	arXiv General Relativity collaboration graph	5242	NA	0.12
Enron	subgraph of the Enron email graph	10000	NA	0.22
Gnutella	Gnutella peer to peer network on Aug. 6, 2002	8717	NA	0.09
Linear kernels				
Dexter	bag of words	2000	20000	83.8
Protein	derived feature matrix for S. cerevisiae	6621	357	99.7
SNPs	DNA microarray data from cancer patients	5520	43	100
Gisette	images of handwritten digits	6000	5000	100
Dense RBF kernels				
AbaloneD	physical measurements of abalones	4177	8	100
WineD	chemical measurements of wine	4898	12	100
Sparse RBF kernels				
AbaloneS	physical measurements of abalones	4177	8	82.9/48.1
WineS	chemical measurements of wine	4898	12	11.1/88.0

Table 6.2: INFORMATION ON THE SPSD MATRICES USED IN OUR EMPIRICAL EVALUATIONS. The matrices used in our empirical evaluation ([LKF07], [KY04], [GGBHD05], [GSP+06], [NWL+02], [Cor96], [BL13]). Here, n is the number of data points, and d is the number of features in the input space before kernelization. For Laplacian "kernels," n is the number of nodes in the graph (and thus there is no d since the graph is "given" rather than "constructed"). The %nnz for the Sparse RBF kernels depends on the σ parameter; see Table 6.3.

- sparse RBFK matrices constructed using Gaussian radial basis functions, truncated to be nonzero only for nearest neighbors.

We briefly review the construction of normalized graph Laplacians, linear kernel matrices, RBFK matrices, and sparse RBFK matrices.

Given a graph with weighted adjacency matrix $\mathbf{W}$, its 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., $D_{ii} = \sum_{j \neq i} W_{ij}$.

The remaining classes of matrices are constructed using 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

$$A_{ij} = \langle \mathbf{x}_i, \mathbf{x}_k \rangle.$$

A Gaussian RBFK matrix $\mathbf{A}^\sigma$ corresponding to these same points is given by

$$A^\sigma_{ij} = \exp\left(\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2}{\sigma^2} \right),$$

where σ, a nonnegative number, determines the scale of the kernel. Informally, σ defines the "size scale" over which pairs of points $\mathbf{x}_i$ and $\mathbf{x}_j$ "see" each other. Typically σ is determined by a global cross-validation criterion, as $\mathbf{A}^\sigma$ is generated for some specific machine learning task. Thus, one may have no *a priori* knowledge of the behavior of the spectrum or leverage scores of $\mathbf{A}^\sigma$ as σ is varied. Accordingly, we consider Gaussian RBFK matrices with different values of σ. Finally, given the same data points, one can construct sparse Gaussian RBFK matrices using the formula

$$A^{(\sigma,\nu,C)}_{ij} = \left[\left(1 - \frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2}{C} \right)^\nu \right]^+ \cdot \exp\left(\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2}{\sigma^2} \right),$$

where $[x]^+ = \max\{0, x\}$. When ν is larger than $(d+1)/2$, this matrix is positive semidefinite; and as the cutoff point C decreases this matrix becomes more sparse [Gen02]. For simplicity, in our experiments we fix $\nu = \lceil (d+1)/2 \rceil$ and $C = 3\sigma$ and we vary σ. As with the effect of varying σ, the effect of varying the sparsity parameter C is not obvious *a priori*. The parameter C is typically chosen according to a global criterion to ensure good performance at a specific machine learning task, without consideration for its effect on the spectrum or leverage scores of $A^{(\sigma,\nu,C)}_{ij}$.

Name	%nnz	$\left\lceil \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_2^2} \right\rceil$	k	$\frac{\lambda_{k+1}}{\lambda_k}$	$100\frac{\|\mathbf{A}-\mathbf{A}_k\|_F}{\|\mathbf{A}\|_F}$	kth-largest leverage score
HEP	0.06	3078	20	0.998	7.8	0.261
HEP	0.06	3078	60	0.998	13.2	0.278
GR	0.12	1679	20	0.999	10.5	0.286
GR	0.12	1679	60	1	17.9	0.289
Enron	0.22	2588	20	0.997	7.77	0.492
Enron	0.22	2588	60	0.999	12.0	0.298
Gnutella	0.09	2757	20	1	8.1	0.381
Gnutella	0.09	2757	60	0.999	13.7	0.340
Dexter	83.8	176	8	0.963	14.5	0.067
Protein	99.7	24	10	0.987	42.6	0.008
SNPs	100	3	5	0.928	85.5	0.002
Gisette	100	4	12	0.90	90.1	0.005
AbaloneD (dense, $\sigma = .15$)	100	41	20	0.992	42.1	0.087
AbaloneD (dense, $\sigma = 1$)	100	4	20	0.935	97.8	0.012
WineD (dense, $\sigma = 1$)	100	31	20	0.99	43.1	0.107
WineD (dense, $\sigma = 2.1$)	100	3	20	0.936	94.8	0.009
AbaloneS (sparse, $\sigma = .15$)	82.9	400	20	0.989	15.4	0.232
AbaloneS (sparse, $\sigma = 1$)	48.1	5	20	0.982	90.6	0.017
WineS (sparse, $\sigma = 1$)	11.1	116	20	0.995	29.5	0.200
WineS (sparse, $\sigma = 2.1$)	88.0	39	20	0.992	41.6	0.098

Table 6.3: STATISTICS OF OUR TEST MATRICES. Summary statistics for the matrices from Table 6.2 used in our computational experiments.

To illustrate the diverse range of properties exhibited by these four classes of matrices, consider Table 6.3. Several observations are particularly relevant to our discussion below.

- All of the Laplacian kernels drawn from informatics graph applications are extremely sparse in terms of number of nonzeros, and tend to have very slow spectral decay, as illustrated both by the quantity $\lceil \|\mathbf{A}\|_F^2 / \|\mathbf{A}\|_2^2 \rceil$ (this is the *stable rank*, a numerically stable (under)estimate of the rank of $\mathbf{A}$ also utilized in Chaper 5) as well as by the relatively small fraction of the Frobenius norm that is captured by the best rank-k approximation to $\mathbf{A}$. For the Laplacian kernels we considered two values of the rank parameter k that were chosen (somewhat) arbitrarily; many of the results we report continue to hold qualitatively if k is chosen to be (say) an order of magnitude larger.

- Both the linear kernels and the dense RBF kernels are much denser and are much more

well-approximated by moderate to very low-rank matrices. In addition, both the linear kernels and the dense RBF kernels have statistical leverage scores that are much more uniform—there are several ways to illustrate this, none of them perfect, and here, we illustrate this by considering the kth largest leverage score. For the linear kernels and the dense RBF kernels, this quantity is one to two orders of magnitude smaller than for the Laplacian kernels.

- For the dense RBF kernels, we consider two values of the σ parameter, again chosen (somewhat) arbitrarily. For both AbaloneD and WineD, 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 that have much more heterogeneous leverage scores. For example, for AbaloneD, the fraction of the Frobenius norm that is captured decreases from 97.8 to 42.1 and the kth largest leverage score increases from 0.012 to 0.087.

- For the sparse RBF kernels, there are a range of sparsities, ranging from above the sparsity of the sparsest linear kernel, but all are denser than the Laplacian kernels. Changing the σ parameter has the same effect (although it is even more pronounced) for sparse RBF kernels as it has for dense RBF kernels. In addition, "sparsifying" a dense RBF kernel also has the effect of making the matrix less well approximated by a low-rank matrix and of making the leverage scores more nonuniform. For example, for AbaloneD with $\sigma = 1$ (respectively, $\sigma = 0.15$), the fraction of the Frobenius norm that is captured decreases from 97.8 (respectively, 42.1) to 90.6 (respectively, 15.4), and the kth largest leverage score increases from 0.012 (respectively, 0.087) to 0.017 (respectively, 0.232).

As we see below, when we consider the RBF kernels as the width parameter and sparsity are varied, we observe a range of intermediate cases between the extremes of linear kernels and Laplacian kernels.

6.9.2 A comparison of empirical errors with the theoretical error bounds

Table 6.4 illustrates the gap between the theoretical results currently available in the literature, the bounds derived in this chapter, and what is observed in practice: it depicts the ratio between the error bounds summarized in Table 6.1 and the average errors observed over 10 trials of SPSD sketching. The error bound from [TR10] is not considered in the table, as it does not apply at the number of samples ℓ used in the experiments.

Several trends can be identified; among them, we note that the bounds provided in this chapter for Gaussian-based sketches come quite close to capturing the errors seen in practice, and the Frobenius and trace-norm error guarantees of the leverage-based and Fourier-based sketches tend to more closely reflect the empirical behavior than the error guarantees provided in prior work for Nyström sketches. Overall, the trace-norm error bounds are quite accurate. On the other hand, prior bounds are sometimes more informative in the case of the spectral norm (with the notable exception of the Gaussian sketches). Several important points can be gleaned from these observations.

First, the accuracy of the Gaussian error bounds suggests that the main theoretical contribution of this work, the deterministic structural results given as Theorems 6.2, 6.6, and 6.7, captures the underlying behavior of the SPSD sketching process. This supports our belief that our deterministic framework provides a foundation for truly informative error bounds. Second, it is clear that the analysis of the stochastic elements of the SPSD sketching process is much sharper in the Gaussian case than in the leverage-score, Fourier, and Nyström cases. We expect

that, at least in the case of leverage and Fourier-based sketches, the stochastic analysis can and will be sharpened to produce error guarantees almost as informative as the ones we have provided for Gaussian-based sketches.

source, sketch	pred./obs. spec. error	pred./obs. Frob. error	pred./obs. trace error
Enron, $k = 60$			
[BW09], Nyström	–	–	2.0
[KMT12], Nyström	331.2	77.7	–
Thm 6.12, leverage-based	12888	21	1.2
Thm 6.16, Fourier-based	201.0	42.7	1.6
Thm 6.17, Gaussian-based	10.1	5.6	1.2
Thm 6.9, Nyström	9.4	385.2	5.4
Protein, $k = 10$			
[BW09], Nyström	–	–	3.6
[KMT12], Nyström	33.4	20.5	–
Thm 6.12, leverage-based	42.5	6.9	2.0
Thm 6.16, Fourier-based	297.5	21.7	3.1
Thm 6.17, Gaussian-based	3.8	3.3	1.8
Thm 6.9, Nyström	86.3	91.3	8
AbaloneD, $\sigma = .15, k = 20$			
[BW09], Nyström	–	–	2.0
[KMT12], Nyström	62.9	46.7	–
Thm 6.12, leverage-based	235.3	14.6	1.3
Thm 6.16, Fourier-based	139.4	36.9	1.7
Thm 6.17, Gaussian-based	5.2	4.7	1.1
Thm 6.9, Nyström	12.9	228.3	5.1
WineS, $\sigma = 1, k = 20$			
[BW09], Nyström	–	–	2.1
[KMT12], Nyström	72.8	44.2	–
Thm 6.12, leverage-based	244.9	13.4	1.2
Thm 6.16, Fourier-based	186.7	36.8	1.7
Thm 6.17, Gaussian-based	6.6	4.7	1.2
Thm 6.9, Nyström	13.7	222.6	5.1

Table 6.4: COMPARISON OF EMPIRICAL ERRORS OF SPSD SKETCHES WITH PREDICTED ERRORS. We compare the empirically observed approximation errors to the guarantees provided in this and other works, for several matrices. Each approximation was formed using $\ell = 6k \log k$ samples. To evaluate the error guarantees, $\delta = 1/2$ was taken and all constants present in the statements of the bounds were replaced with ones. The observed errors were taken to be the average errors over 10 runs of the approximation algorithms. The matrices, described in Table 6.2, are representative of several classes of matrices prevalent in machine learning applications.

6.9.3 Reconstruction accuracy of sampling and projection-based sketches

Here, we describe the performances of the various SPSD sketches in terms of reconstruction accuracy on the matrices described in Section 6.9.1. Recall that the sketches considered are Nyström extensions, leverage-based sketches, and sketches formed using Gaussian and SRFT mixtures of columns.

We describe general observations we have made about each class of matrices in turn, and then we summarize our observations. We present results for both the rank-restricted and non-rank-restricted sketches. That is, we plot the errors

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_{\xi} / \|\mathbf{A} - \mathbf{A}_{k}\|_{\xi} \tag{6.9.1}$$

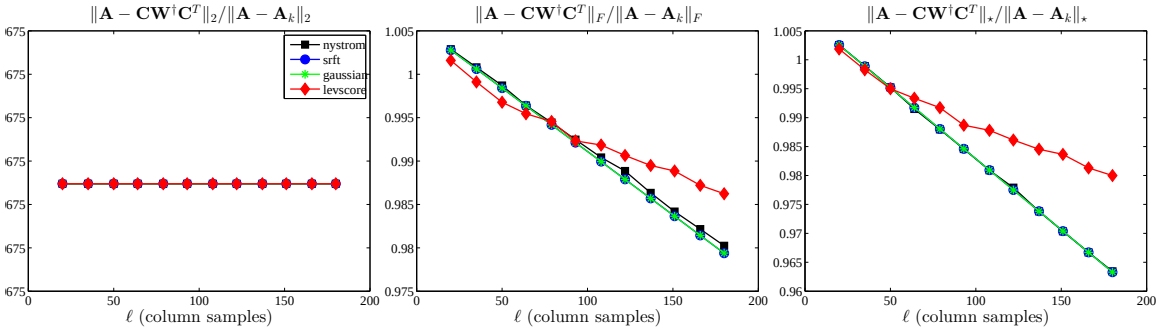
for the non-rank-restricted sketches, and the errors

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}_{k}^{\dagger}\mathbf{C}^{T}\|_{\xi} / \|\mathbf{A} - \mathbf{A}_{k}\|_{\xi} \tag{6.9.2}$$

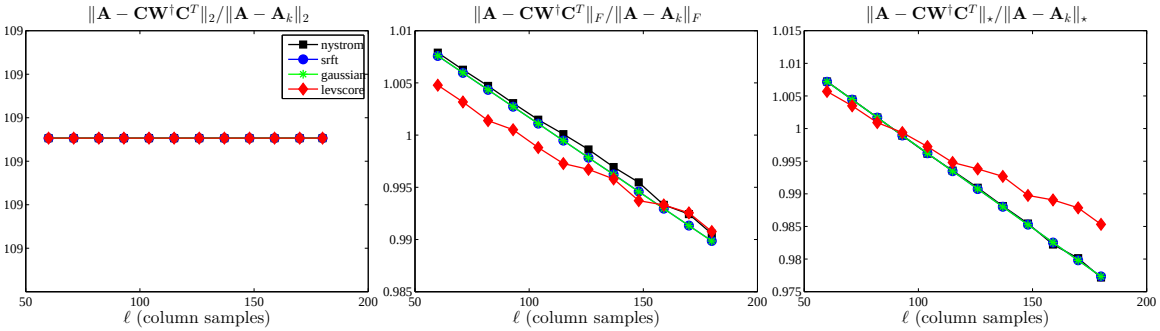
for the rank-restricted sketches.

6.9.3.1 Graph Laplacians

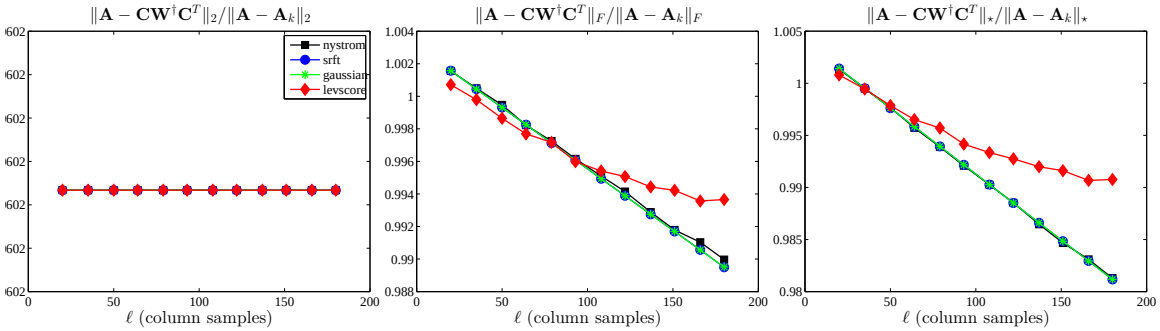
Figures 6.4–6.7 show the reconstruction error results for sampling and mixture methods applied to several normalized graph Laplacians. Figures 6.4 and 6.6 show GR and HEP, each for two values of the rank parameter. The remaining two show Enron and Gnutella, again each for two values of the rank parameter. The first two figures show the ratios of the spectral, Frobenius, and trace-norm approximation errors of non-rank-restricted sketches to the optimal rank-k approximation errors, as a function of the number of column samples ℓ. The remaining two



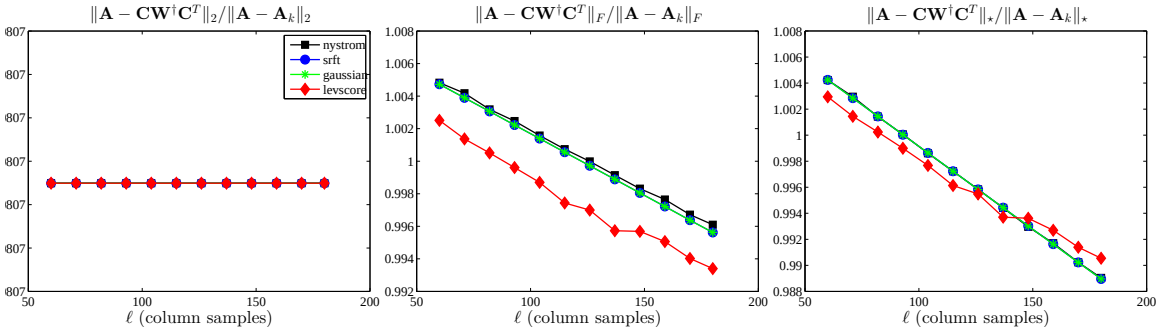
(a) GR, $k = 20$



(b) GR, $k = 60$

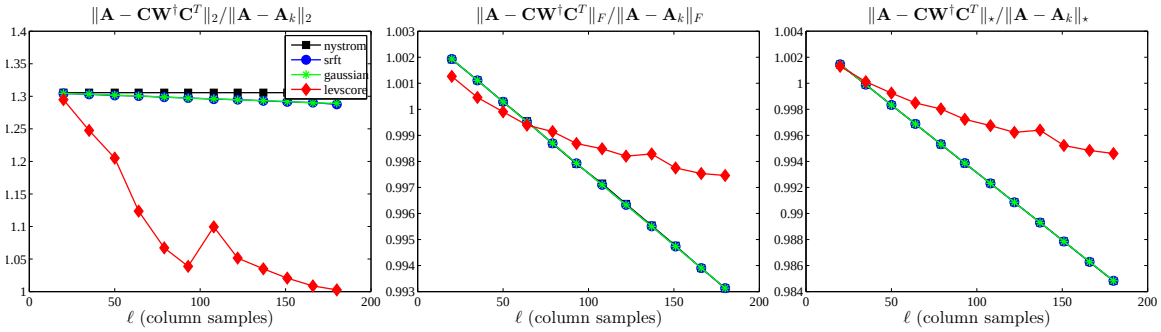


(c) HEP, $k = 20$

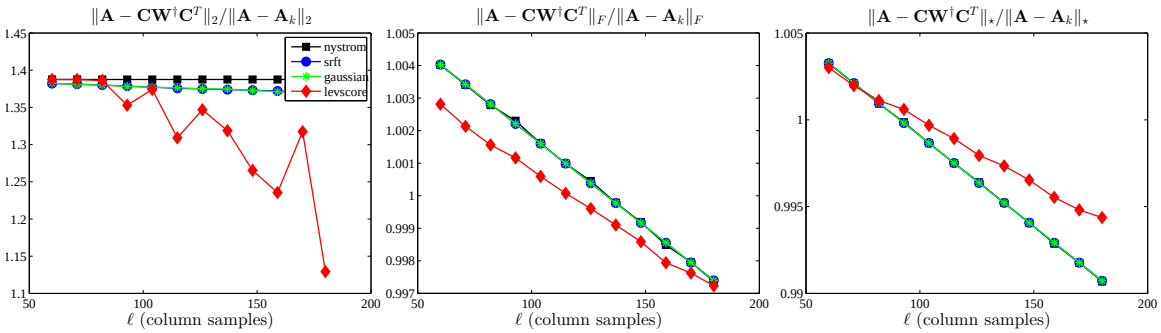


(d) HEP, $k = 60$

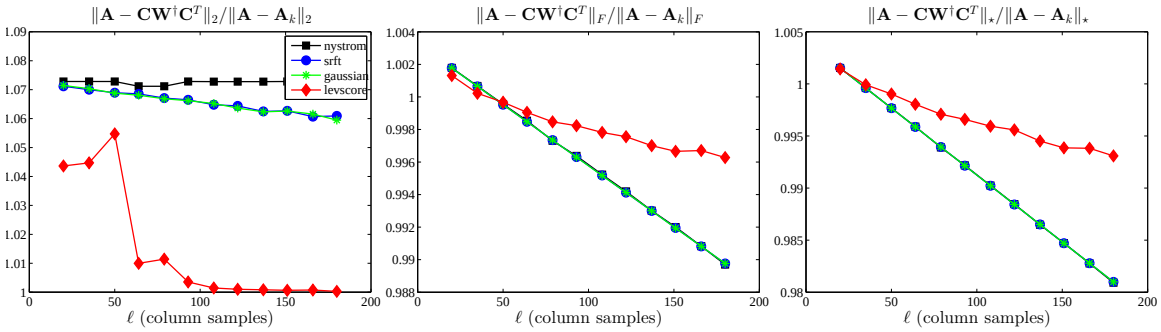
Figure 6.4: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE GR AND HEP LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the GR and HEP Laplacian matrices, with two choices of the rank parameter k.



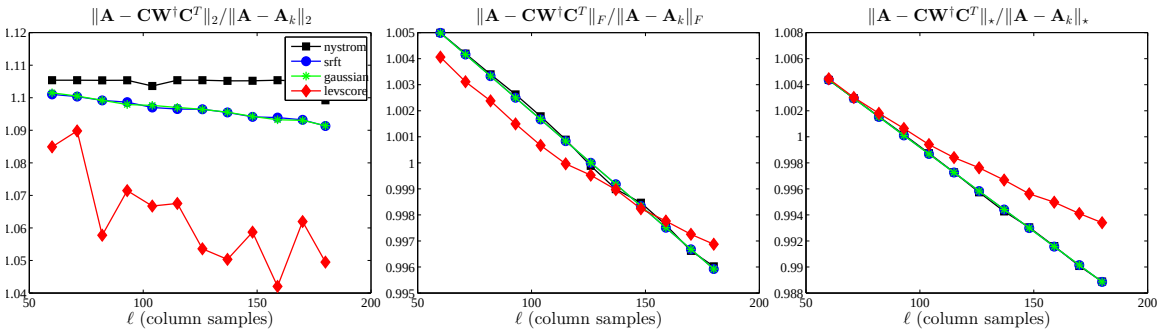
(a) Enron, $k = 20$



(b) Enron, $k = 60$

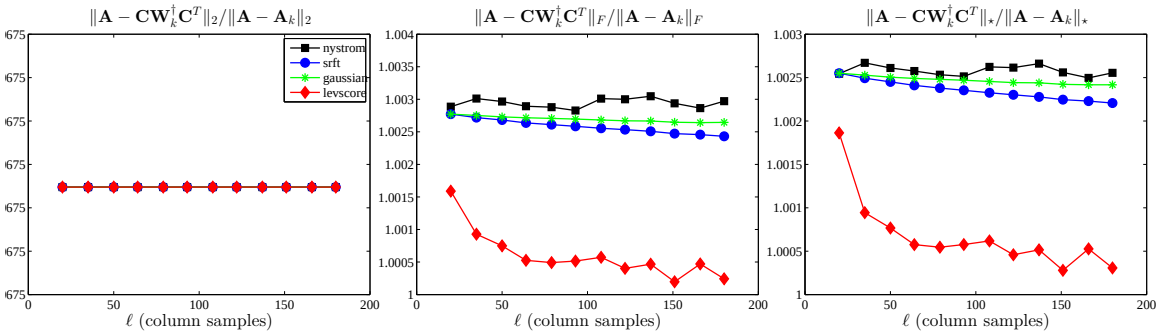


(c) Gnutella, $k = 20$

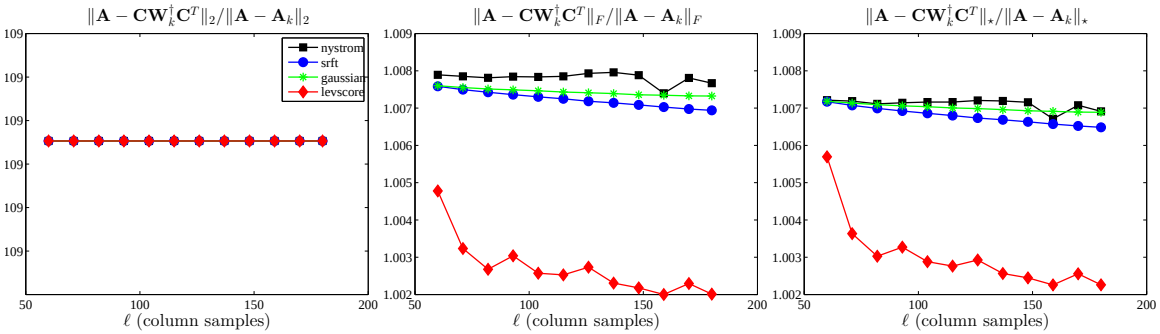


(d) Gnutella, $k = 60$

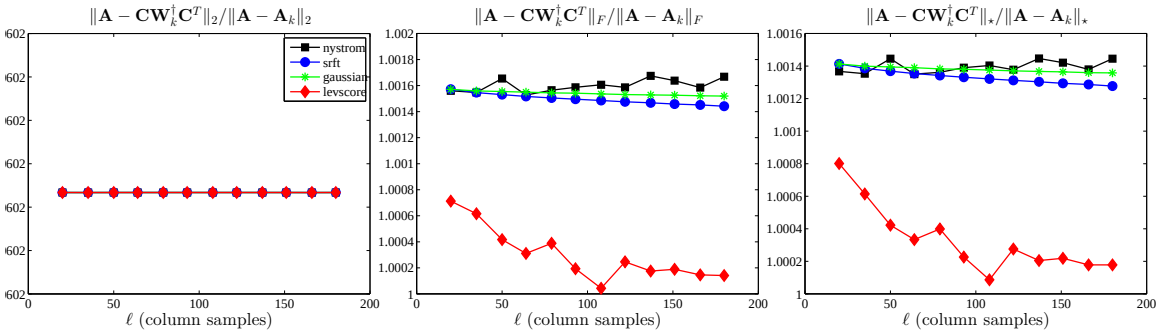
Figure 6.5: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE ENRON AND GNUTELLA LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the Enron and Gnutella Laplacian matrices, with two choices of the rank parameter k.



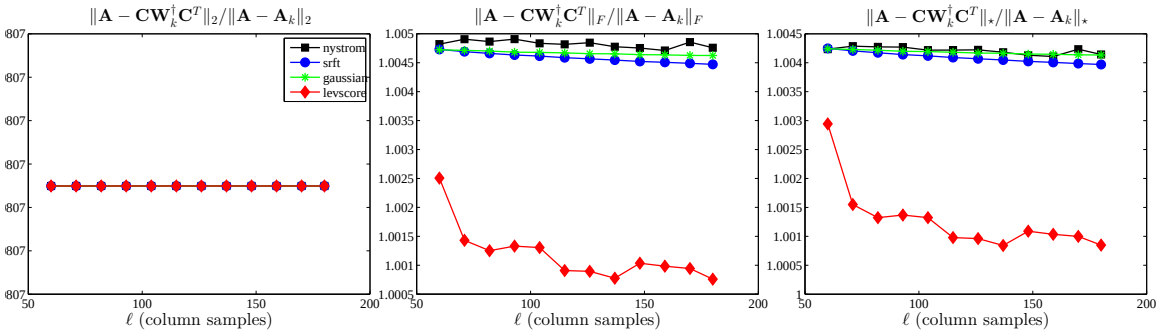
(a) GR, $k = 20$



(b) GR, $k = 60$

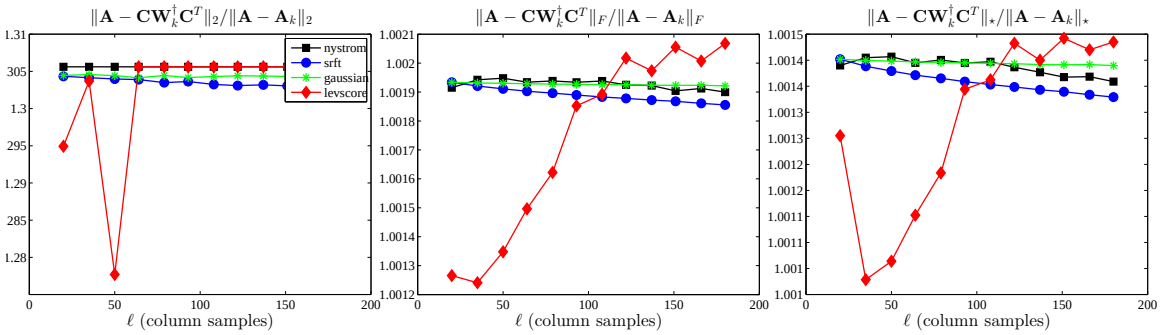


(c) HEP, $k = 20$

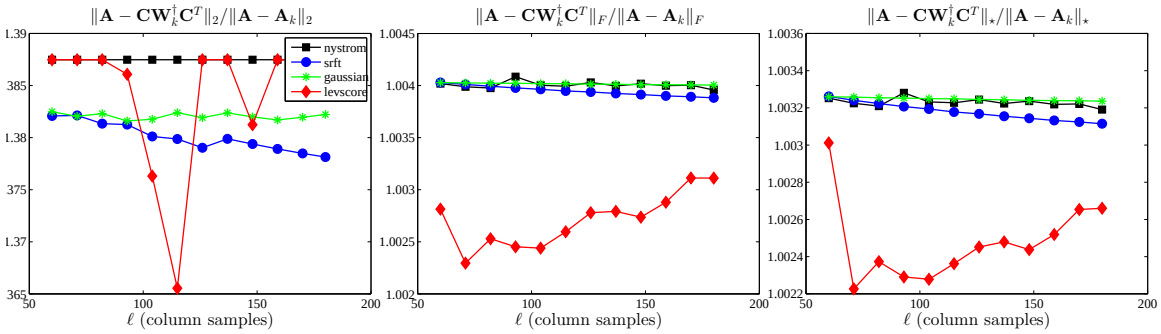


(d) HEP, $k = 60$

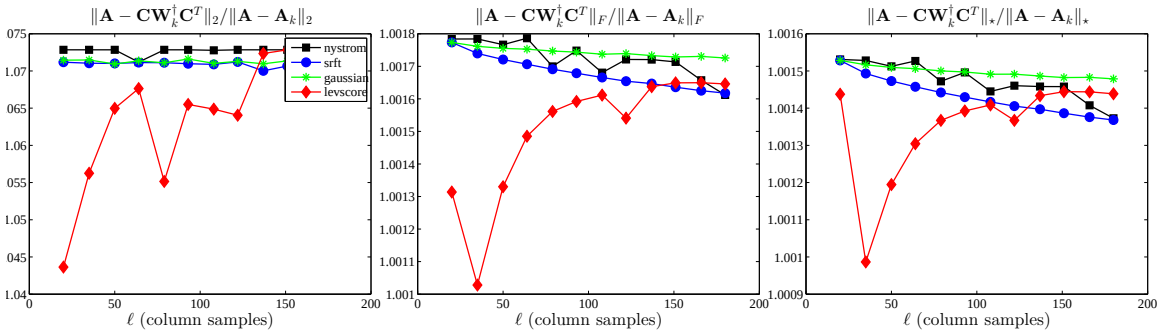
Figure 6.6: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE GR AND HEP LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the GR and HEP Laplacian matrices, with two choices of the rank parameter k.



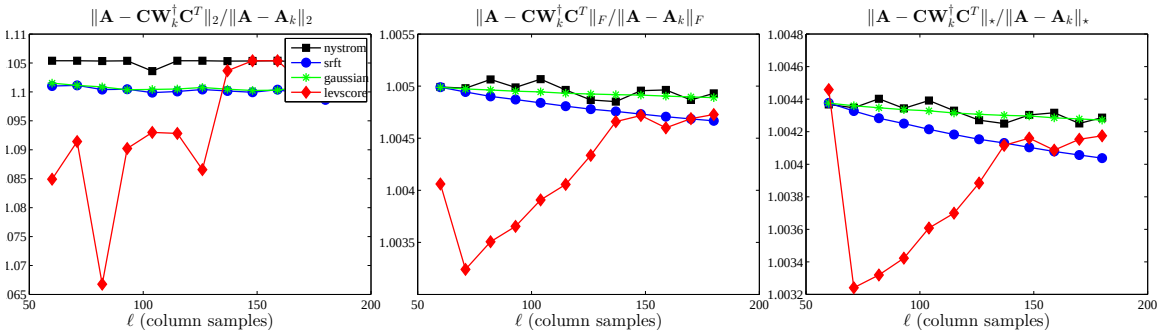
(a) Enron, $k = 20$



(b) Enron, $k = 60$



(c) Gnutella, $k = 20$



(d) Gnutella, $k = 60$

Figure 6.7: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE ENRON AND GNUTELLA LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the Enron and Gnutella Laplacian matrices, with two choices of the rank parameter k.

similarly show the errors of the rank-restricted sketches.

These and subsequent figures contain a lot of information, some of which is peculiar to the given matrices and some of which is more general. In light of subsequent discussion, several observations are worth making about the results presented in these figures.

- All of the SPSD sketches provide quite accurate approximations even with only k column samples (or in the case of the Gaussian and SRFT mixtures, with only k linear combinations of vectors). Upon examination, this is partly due to the extreme sparsity and extremely slow spectral decay of these matrices which means, as shown in Table 6.2, that only a small fraction of the (spectral or Frobenius or trace) mass is captured by the optimal rank 20 or 60 approximation. Thus although an SPSD sketch constructed from 20 or 60 vectors also only captures a small portion of the mass of the matrix, the relative error is small.

- The scale of the vertical axes is different between different figures and subfigures. This is to highlight properties within a given plot, but it can hide several things. In particular, note that the scale for the spectral norm is generally larger than for the Frobenius norm, which is generally larger than for the trace norm, consistent with the size of those norms; and that the scale is larger for higher-rank approximations, e.g. compare GR $k = 20$ with GR $k = 60$, also consistent with the larger amount of mass captured by higher-rank approximations.

- Both the non-rank-restricted and rank-restricted results are the same for $\ell = k$. For $\ell > k$, the non-rank-restricted errors tend to decrease (or at least not increase, as for GR and HEP the spectral norm error is flat as a function of ℓ), which is intuitive. While the rank-restricted errors also tend to decrease for $\ell > k$, the decrease is much less (since the rank-restricted plots are bounded below by unity) and the behavior is much more

complicated as a function of increasing ℓ.

- The horizontal axes range from k to $9k$ for the $k = 20$ plots and to $3k$ for the $k = 60$ plots. As a practical matter, choosing ℓ between k and (say) $2k$ or $3k$ is probably of greatest interest. In this regime, there is an interesting tradeoff for the non-rank-restricted plots: for moderately large values of ℓ in this regime, the error for leverage-based sampling is moderately better than for uniform sampling or random mixtures, while if one chooses ℓ to be much larger then the improvements from leverage-based sampling saturate and the uniform sampling and random mixture methods are better. This is most obvious in the Frobenius-norm plots, although it is also seen in the trace norm plots, and it suggests that some combination of leverage-based sampling and uniform sampling might be best.

- For the rank-restricted plots, in some cases, e.g., with GR and HEP, the errors for leverage-based sampling are much better than for the other methods and quickly improve with increasing ℓ until they saturate; while in other cases, e.g., with Enron and Gnutella, the errors for leverage-based sampling improve quickly and then degrade with increasing ℓ. Upon examination, the former phenomenon is similar to what was observed in the non-rank-restricted case and is due to the strong "bias" provided by the leverage score importance sampling distribution to the top part of the spectrum, allowing the sampling process to focus very quickly on the low-rank part of the input matrix. (In some cases, this is due to the fact that the heterogeneity of the leverage score importance sampling distribution means that one is likely to choose the same high leverage columns multiple times, rather than increasing the accuracy of the extension by adding new columns whose leverage scores are lower.) The latter phenomenon of degrading error quality as ℓ is increased is more complex and seems to be due to some sort of "overfitting" caused by

this strong bias and by choosing many more than k columns.

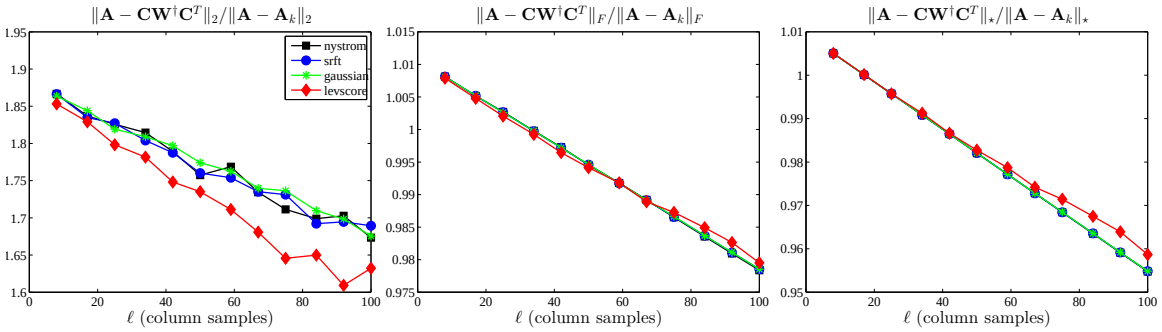
- The behavior of the approximations with respect to the spectral norm is quite different from the behavior in the Frobenius and trace norms. In the latter, as the number of samples ℓ increases, the errors tend to decrease, although in an erratic manner for some of the rank-restricted plots; while for the former, the errors tend to be much flatter as a function of increasing ℓ for at least the Gaussian, SRFT, and uniformly column sampled (i.e., Nyström) sketches.

All in all, there seems to be quite complicated behavior for low-rank sketches for these Laplacian matrices. Several of these observations can also be made for subsequent figures; but in some other cases the (very sparse and not very low rank) structural properties of the data are primarily responsible.

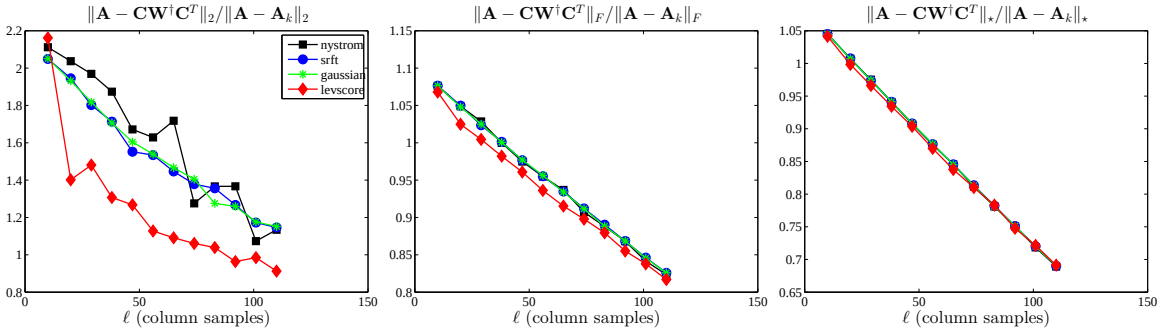
6.9.3.2 Linear kernels

Figures 6.8 and 6.9 show the reconstruction error results for sampling and mixture methods applied to several linear kernels. The matrices (Dexter, Protein, SNPs, and Gisette) are all quite low-rank and have fairly uniform leverage scores. Several observations are worth making about the results presented in these figures.

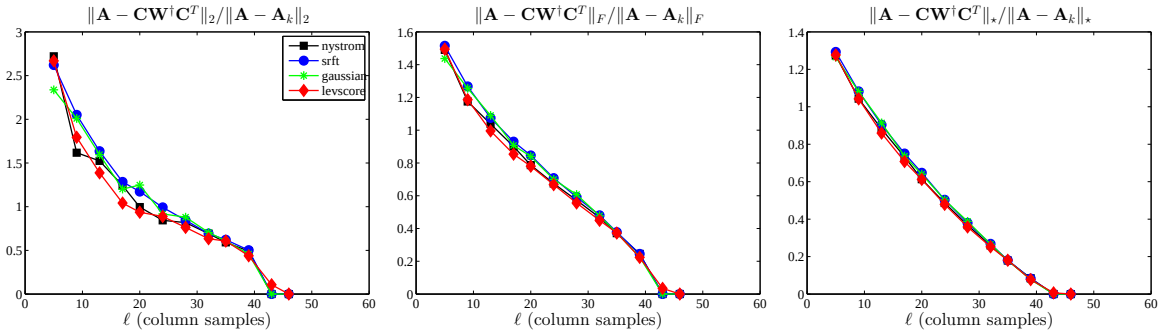
- All of the methods perform quite similarly for the non-rank-restricted case: all have errors that decrease smoothly with increasing ℓ, and in this case there is little advantage to using methods other than uniform sampling (since they perform similarly and are more expensive). Also, since the ranks are so low and the leverage scores are so uniform, the leverage score extension is no longer significantly distinguished by its tendency to saturate quickly.



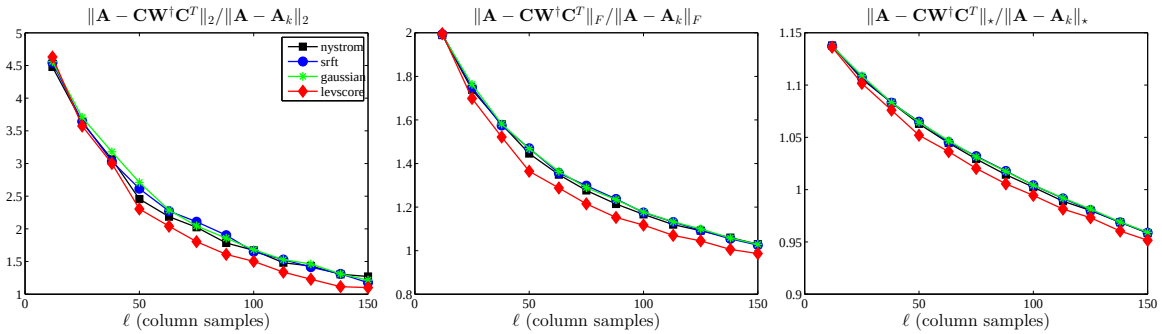
(a) Dexter, $k = 8$



(b) Protein, $k = 10$

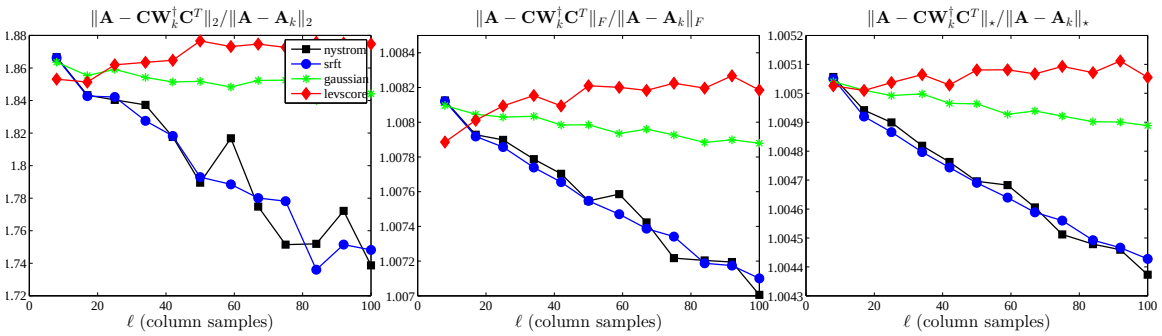


(c) SNPs, $k = 5$

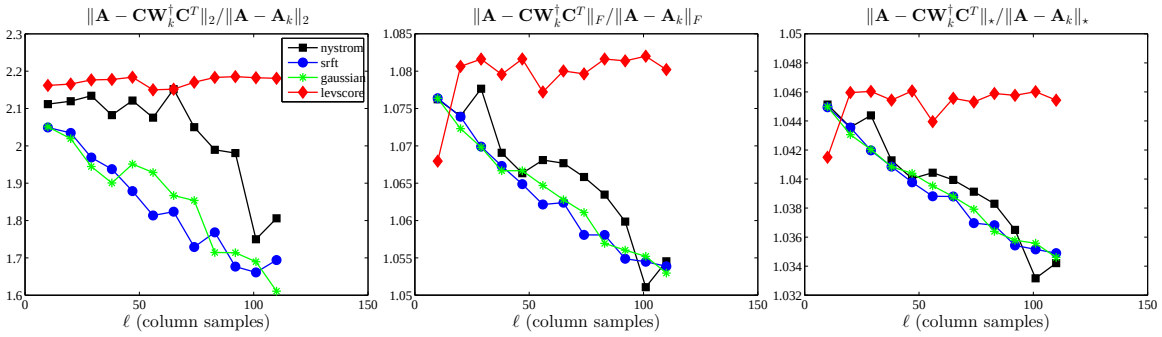


(d) Gisette, $k = 12$

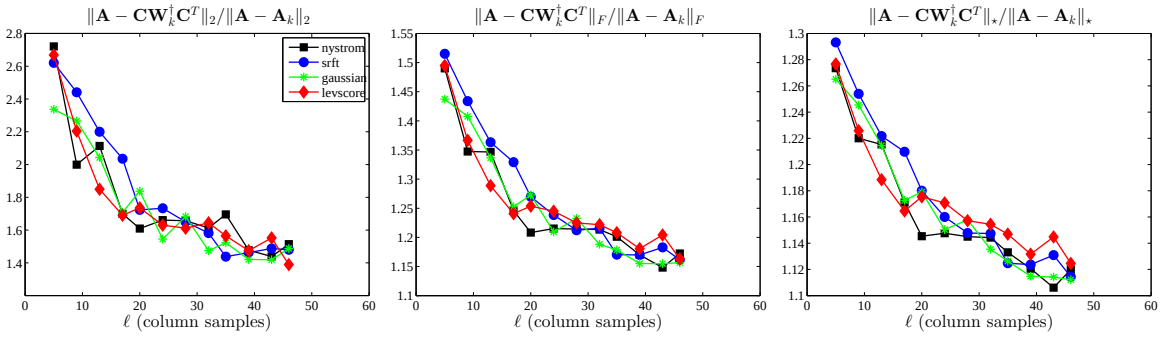
Figure 6.8: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE LINEAR KERNEL MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the linear kernel matrices.



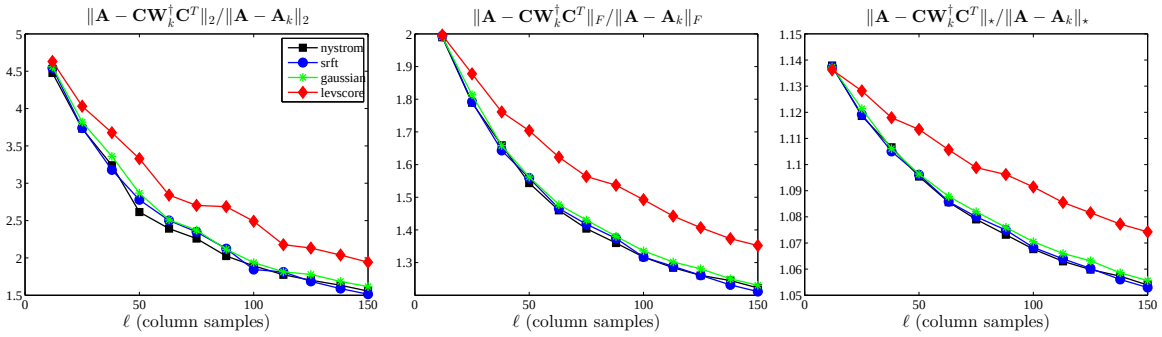
(a) Dexter, $k = 8$



(b) Protein, $k = 10$



(c) SNPs, $k = 5$



(d) Gisette, $k = 12$

Figure 6.9: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE LINEAR KERNEL MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the linear kernel matrices.

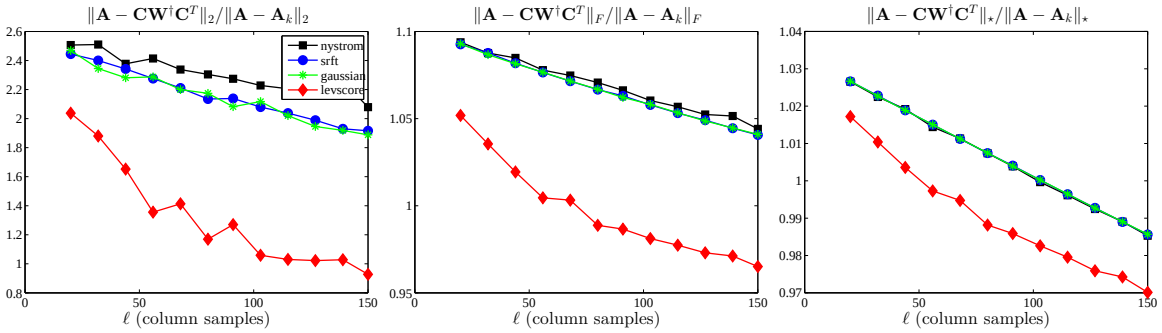
- The scale of the vertical axes is much larger than for the Laplacian matrices, mostly since the matrices are much better approximated by low-rank matrices, although the scale decreases as one goes from spectral to Frobenius to trace reconstruction error, as before.

- For SNPs and Gisette, the rank-restricted reconstruction results are very similar for all four methods, with a smooth decrease in error as ℓ is increased, although interestingly using leverage scores is slightly worse for Gisette. For Dexter and Protein, the situation is more complicated: using the SRFT always leads to smooth decrease as ℓ is increased, and uniform sampling generally behaves the same way also; Gaussian mixtures behave this way for Protein, but for Dexter Gaussian mixtures are noticably worse than SRFT and uniform sampling; and, except for very small values of ℓ, leverage-based sampling is worse still and gets noticably worse as ℓ is increased. Even this poor behavior of leverage score sampling on the linear kernels is notably worse than for the rank-restricted Laplacians, where there was a range of moderately small ℓ where leverage score sampling was much superior to other methods.

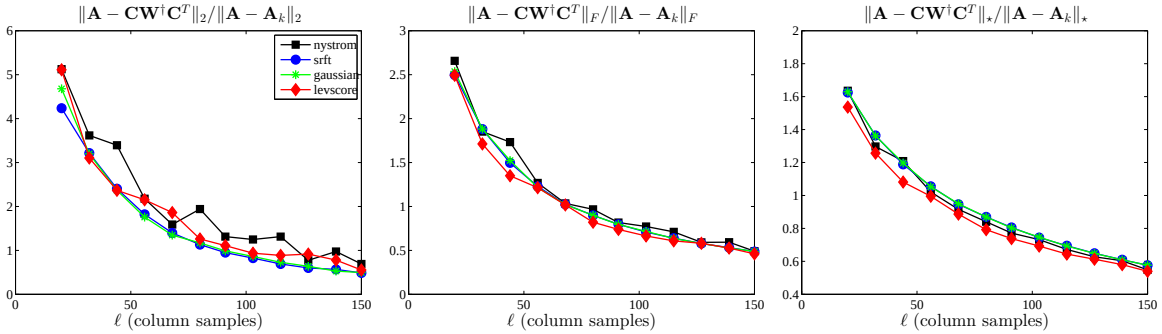
These linear kernels (and also to some extent the dense RBF kernels below that have larger σ parameter) are examples of relatively "nice" machine learning matrices that are similar to matrices where uniform sampling has been shown to perform well previously [TKR08, KMT09a, KMT09b, KMT12]; and for these matrices our empirical results agree with these prior works.

6.9.3.3 Dense and sparse RBF kernels

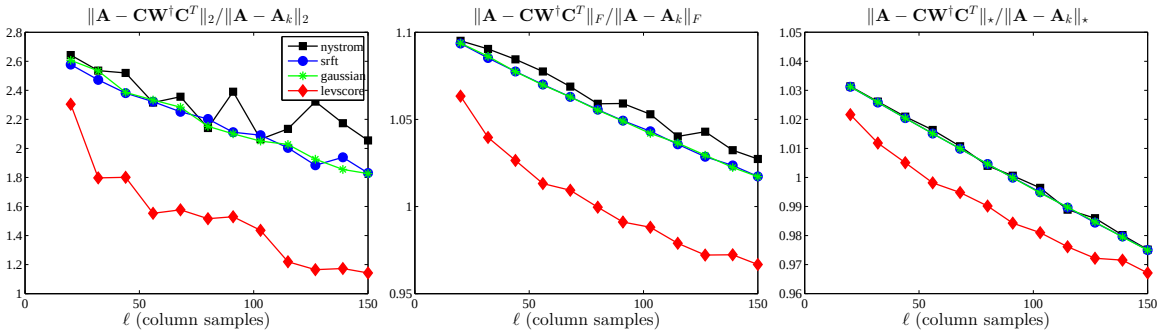
Figures 6.10–6.13 present the reconstruction error results for sampling and mixture methods applied to several dense RBF and sparse RBF kernels. Several observations are worth making about the results presented in these figures.



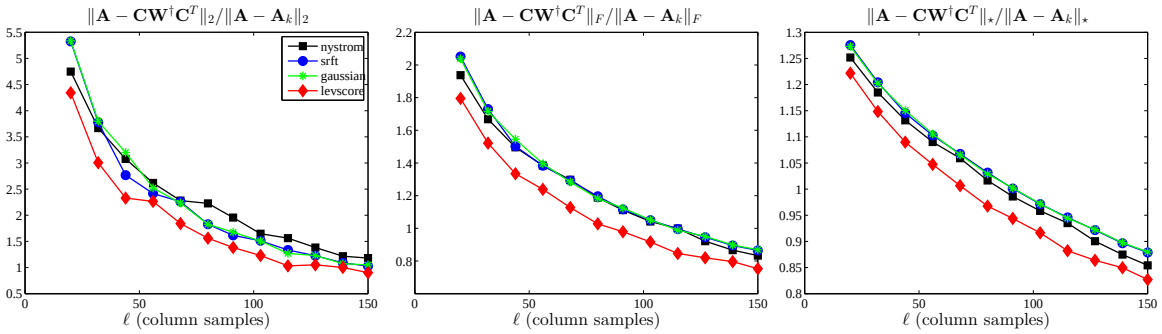
(a) AbaloneD, $\sigma = .15, k = 20$



(b) AbaloneD, $\sigma = 1, k = 20$

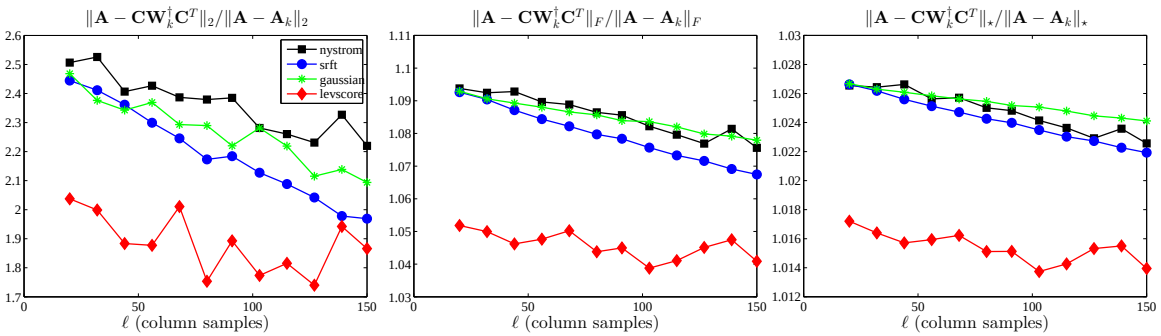


(c) WineD, $\sigma = 1, k = 20$

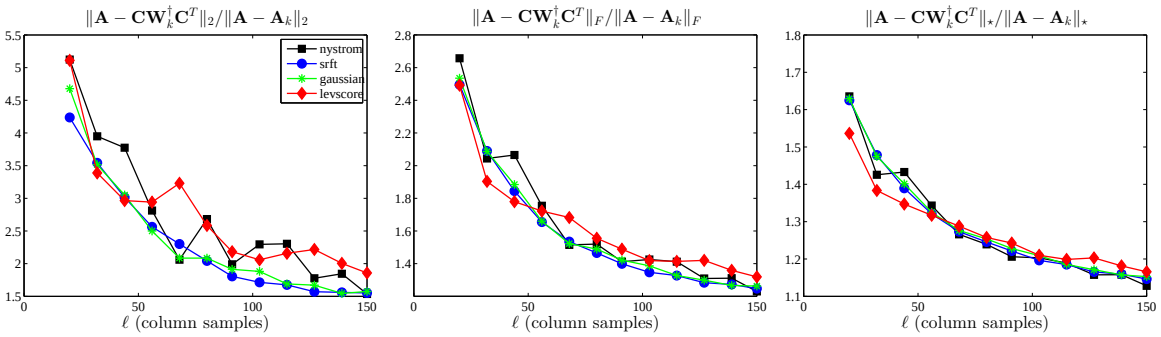


(d) WineD, $\sigma = 2.1, k = 20$

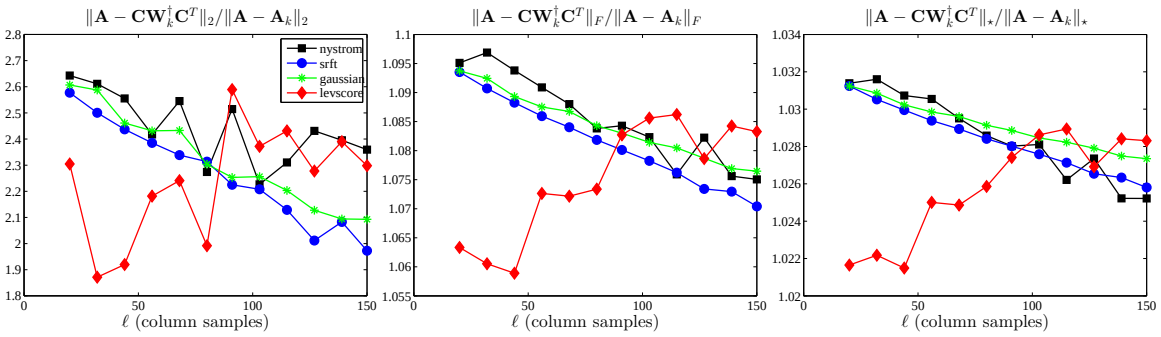
Figure 6.10: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE DENSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the dense RBFK matrices.



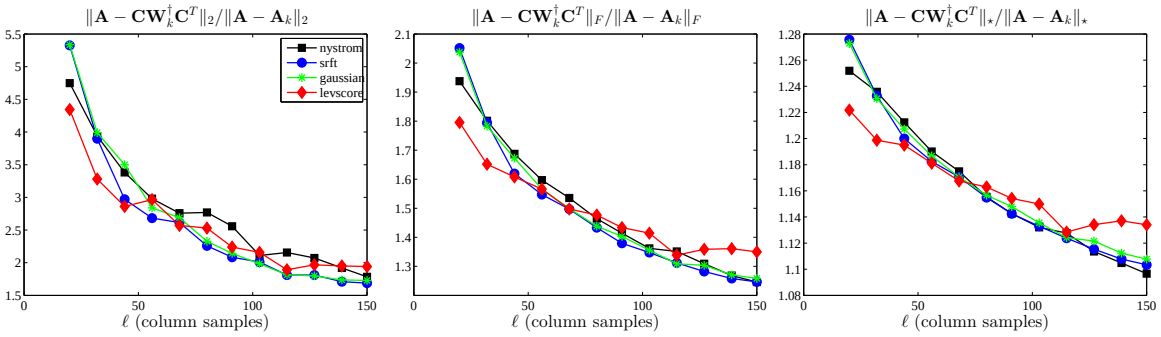
(a) AbaloneD, $\sigma = .15, k = 20$



(b) AbaloneD, $\sigma = 1, k = 20$

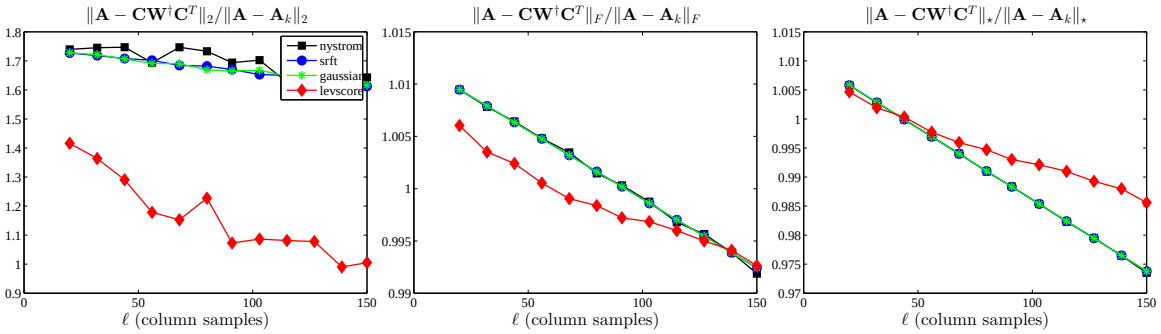


(c) WineD, $\sigma = 1, k = 20$

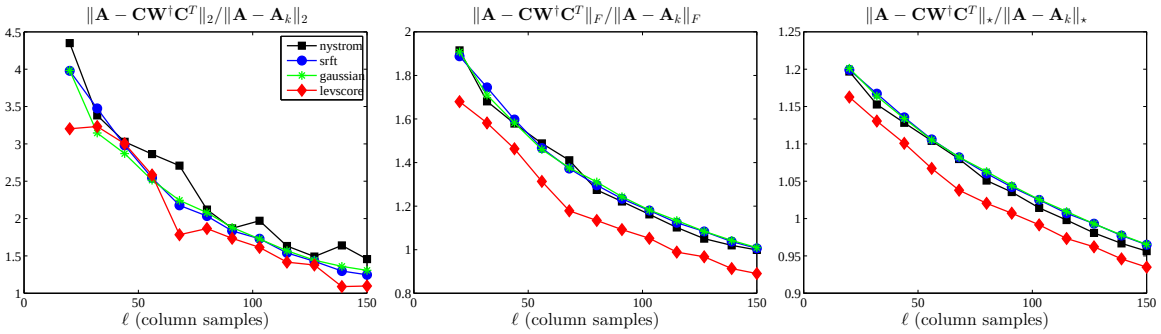


(d) WineD, $\sigma = 2.1, k = 20$

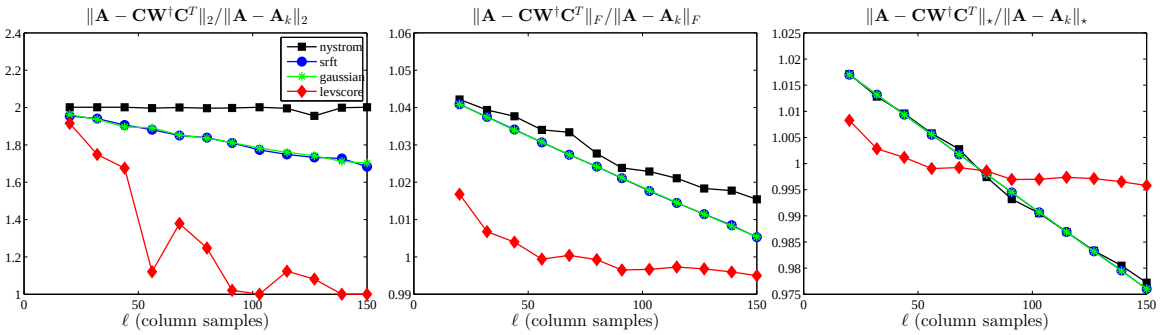
Figure 6.11: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE DENSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the dense RBFK matrices.



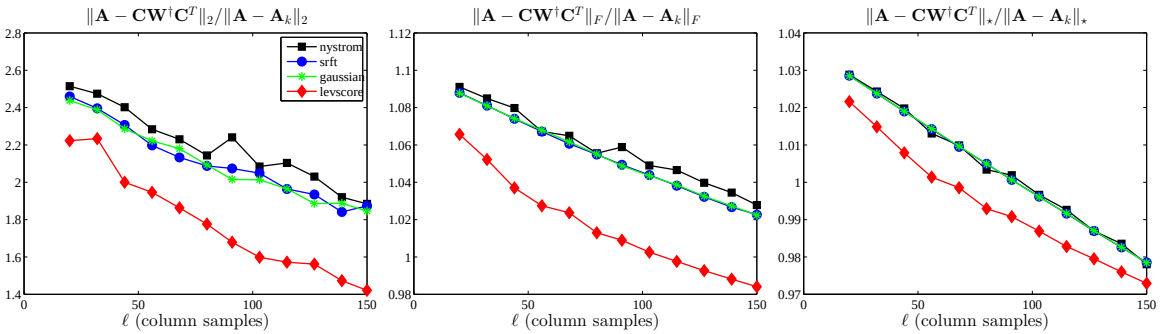
(a) AbaloneS, $\sigma = .15, k = 20$



(b) AbaloneS, $\sigma = 1, k = 20$

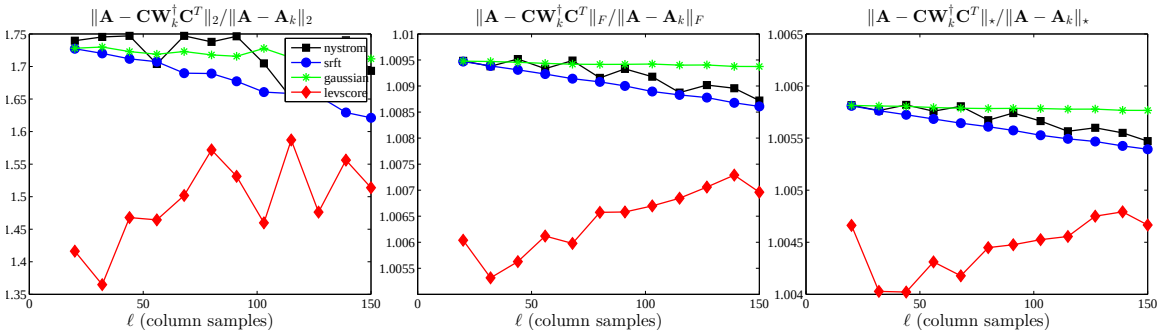


(c) WineS, $\sigma = 1, k = 20$

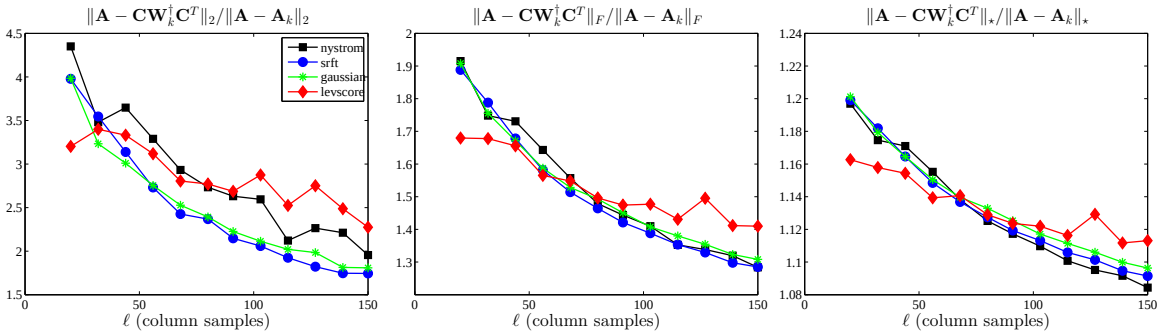


(d) WineS, $\sigma = 2.1, k = 20$

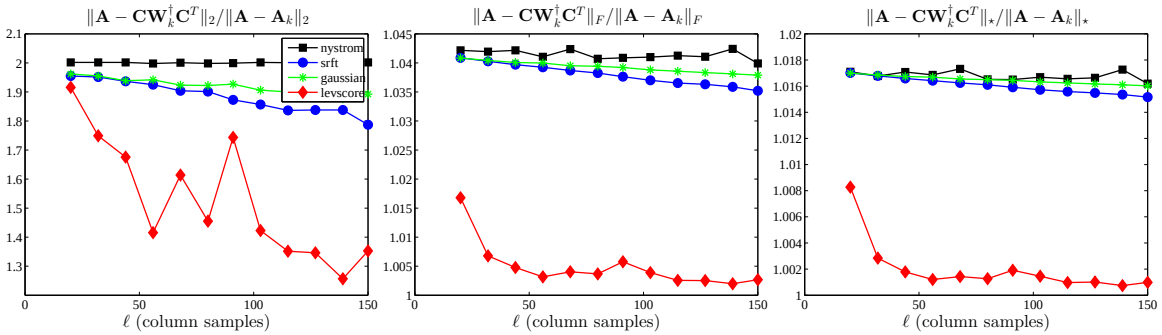
Figure 6.12: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE SPARSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the sparse RBFK matrices.



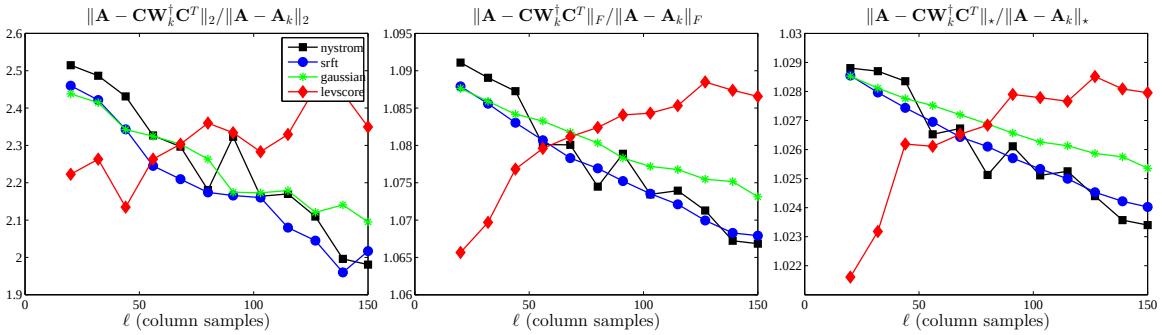
(a) AbaloneS, $\sigma = .15, k = 20$



(b) AbaloneS, $\sigma = 1, k = 20$



(c) WineS, $\sigma = 1, k = 20$



(d) WineS, $\sigma = 2.1, k = 20$

Figure 6.13: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE SPARSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the sparse RBFK matrices.

- For the non-rank-restricted results, all of the methods have errors that decrease with increasing ℓ. In particular, for larger values of σ and for denser matrices, the decrease is somewhat more regular, and the four methods tend to perform similarly. For larger values of σ and sparser matrices, leverage score sampling is somewhat better. This parallels what we observed with the linear kernels, except that here the leverage score sampling is somewhat better for all values of ℓ.

- For the non-rank-restricted results for the smaller values of σ, leverage score sampling tends to be much better than uniform sampling and mixture-based methods. For the sparse matrices, however, this effect saturates. We again observe (especially when σ is smaller in AbaloneS and WineS) the tradeoff we observed previously with the Laplacian matrices: leverage score sampling is better when ℓ is moderately larger than k, while uniform sampling and random mixtures are better when ℓ is much larger than k.

- For the rank-restricted results, we see that when σ is large, all of the results tend to perform similarly. (The exception to this is WineS, for which leverage score sampling starts out much better than other methods and then gets worse as ℓ is increased.) On the other hand, when σ is small, the results are more complex. Leverage score sampling is typically much better than other methods, although the results are quite choppy as a function of ℓ, and in some cases the effect diminishes as ℓ is increased.

Recall from Table 6.3 that for smaller values of σ and for sparser kernels, the SPSD matrices are less well-approximated by low-rank matrices, and they have more heterogeneous leverage scores. Thus, they are more similar to the Laplacian matrices than the linear kernel matrices; and this suggests (as we have observed) that leverage score sampling should perform better than uniform column sampling and mixture-based schemes in these two cases. In particular,

nowhere do we see that leverage score sampling performs much worse than other methods, as we saw with the rank-restricted linear kernel results.

6.9.3.4 Summary of comparison of sampling and mixture-based SPSD Sketches

Several summary observations can be made about sampling versus mixture-based SPSD sketches for the matrices we have considered.

- Linear kernels and to a lesser extent dense RBF kernels with larger σ parameter have relatively low rank and relatively uniform leverage scores, and in these cases uniform sampling does quite well. These matrices correspond most closely with those that have been studied previously in the machine learning literature, and for these matrices our results are in agreement with that prior work.

- Sparsifying RBF kernels and/or choosing a smaller σ parameter tends to make these kernels worse approximated by low-rank matrices and to have more heterogeneous leverage scores. In general, these two properties need not be directly related: the spectrum is a property of eigenvalues, while the leverage scores are determined by the eigenvectors. However, in the matrices 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 do much better than other methods. Interestingly, the sparse RBF kernels have many properties of very sparse Laplacian kernels corresponding to relatively unstructured informatics graphs.

- Reconstruction quality under leverage score sampling saturates, as a function of choosing more samples ℓ; this is seen both for non-rank-restricted and rank-restricted situations.

As a consequence, there is often a transition between leverage score sampling or other methods being better as ℓ increases.

- Although they are potentially ill-conditioned, non-rank-restricted approximations behave better in terms of reconstruction quality. Rank-constrained approximations tend to have much more complicated behavior as a function of increasing the number of samples ℓ, including choppier and non-monotonic behavior. This is particularly severe for leverage score sampling, but it occurs with other methods. Other forms of regularization might be appropriate.

In general, *all* of the sampling and mixture-based sketches we considered perform *much* better on the SPSD matrices we considered than both the previous worst-case bounds (e.g., [DM05, KMT12]) and the bounds derived in this chapter would suggest. Even the worst results correspond to single-digit approximation factors in relative scale.

6.10 A comparison with projection-based low-rank approximations

Finally, we consider the performance of two projection-based SPSD sketches proposed in [HMT11]. Recall from Chapter 5 that these low-rank approximations are constructed by forming an approximate basis $\mathbf{Q}$ for the top k-dimensional eigenspace of $\mathbf{A}$ and then restricting $\mathbf{A}$ to that eigenspace.

Given a sampling matrix $\mathbf{S}$, form the matrix $\mathbf{Y} = \mathbf{AS}$ and take the QR decomposition of $\mathbf{Y}$ to obtain $\mathbf{Q}$, a matrix with orthonormal columns. The first projection-based approximant mentioned in [HMT11], which we eponymously refer to as the *pinched* approximant, is simply $\mathbf{A}$

pinched to the space spanned by $\mathbf{Q}$:

$$\mathbf{P}_{\mathbf{AS}}\mathbf{A}\mathbf{P}_{\mathbf{AS}} = \mathbf{Q}(\mathbf{Q}^T\mathbf{A}\mathbf{Q}^T)\mathbf{Q}.$$

Note that this approximant requires two passes over $\mathbf{A}$. The second approximant, which we refer to as the *prolonged* approximant, is

$$\mathbf{A}\mathbf{Q}(\mathbf{Q}^T\mathbf{A}\mathbf{Q})^\dagger\mathbf{Q}^T\mathbf{A}.$$

The computation of a prolonged approximant also requires two passes over $\mathbf{A}$.

It is clear that the prolonged approximant can be constructed using our SPSD sketching model by taking $\mathbf{Q}$ as the sketching matrix. In fact, a stronger statement can be made. Recall, from Lemma 6.1, that for any sketching matrix $\mathbf{X}$, when $\mathbf{C} = \mathbf{A}\mathbf{X}$ and $\mathbf{W} = \mathbf{X}^T\mathbf{A}\mathbf{X}$,

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

By considering the two choices $\mathbf{X} = \mathbf{A}\mathbf{S}$ and $\mathbf{X} = \mathbf{Q}$, we see that in fact the prolonged approximant is exactly the two-pass SPSD sketch:

$$\begin{aligned}\mathbf{A}\mathbf{Q}(\mathbf{Q}^T\mathbf{A}\mathbf{Q})^\dagger\mathbf{A}\mathbf{Q} &= \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{Q}}\mathbf{A}^{1/2} \\ &= \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}(\mathbf{A}\mathbf{S})}\mathbf{A}^{1/2} \\ &= \mathbf{A}^2\mathbf{S}(\mathbf{S}^T\mathbf{A}^3\mathbf{S})^\dagger\mathbf{S}^T\mathbf{A}^2.\end{aligned}$$

It follows that the bounds we provide in Section 6.4 on the performance of multi-pass sketches pertain also to prolonged approximants. In particular, the additional errors of these approximants

are expected to be at least a factor of $\lambda_{k+1}(\mathbf{A})/\lambda_k(\mathbf{A})$ smaller than the additional errors of one-pass sketches.

In Figure 6.14, we compare the empirical performances of several of the SPSD sketches considered earlier with pinched and prolonged approximants constructed using the same matrix $\mathbf{S}$. Specifically, we plot the errors of pinched and prolonged approximants for choices of sketching matrices corresponding to uniform column sampling, gaussian column mixtures, and SRFT-based column mixtures, along with the errors of one-pass SPSD sketches constructed using the same choices of $\mathbf{S}$. In the interest of brevity, we provide results only for a subset of the matrices listed in Table 6.2 and consider only the nonfixed-rank variants of the sketches.

Some trends are clear from Figure 6.14.

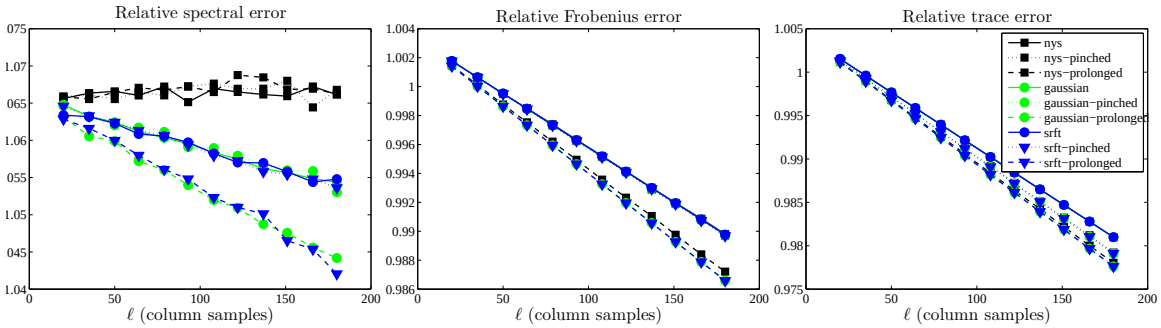
- In the spectral norm, the prolonged approximants are considerably more accurate than the pinched approximants and one-pass sketches for all the matrices considered. Without exception, the prolonged Gaussian and SRFT column-mixture approximants are the most accurate in the spectral norm, of all the schemes considered. Only in the case of the Dexter linear kernel is the prolonged uniformly column-sampled approximant nearly as accurate in the spectral norm as the prolonged Gaussian and SRFT approximants. To a lesser extent, the prolonged approximants are also more accurate in the Frobenius and trace norms than the other schemes considered. The increased Frobenius and trace norm accuracy is particularly notable for the two RBF kernel matrices; again, the prolonged Gaussian and SRFT approximants are considerably more accurate than the prolonged uniformly column-sampled approximants.

- After the prolonged approximants, the pinched Gaussian and SRFT column-mixture approximants have the smallest spectral, Frobenius, and trace-norm errors. Again however,

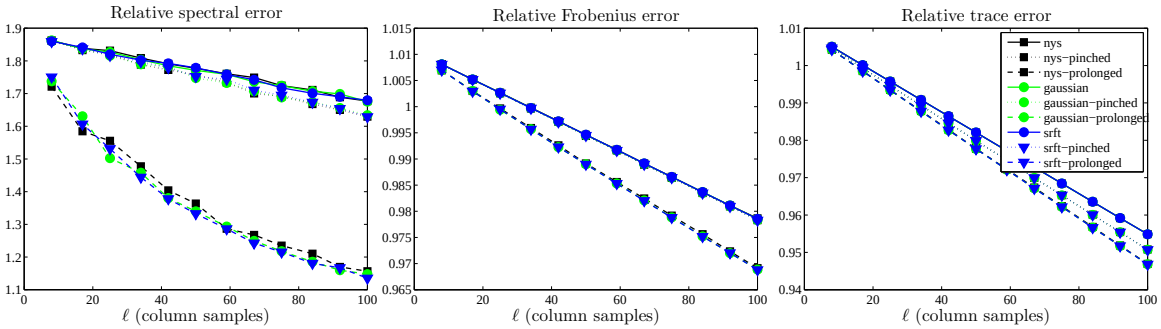
we see that the pinched uniformly column-sampled approximants are considerably less accurate than the pinched Gaussian and SRFT column-mixture approximants. Particularly in the spectral and Frobenius norms, the pinched uniformly column-sampled approximants are not any more accurate than the uniformly column-sampled sketches.

It is evident that the benefits of pinched and prolonged approximants are most dramatic when the spectral norm is the error metric, and that Nyström extensions do not benefit as much from multiple passes as do other sketching schemes.

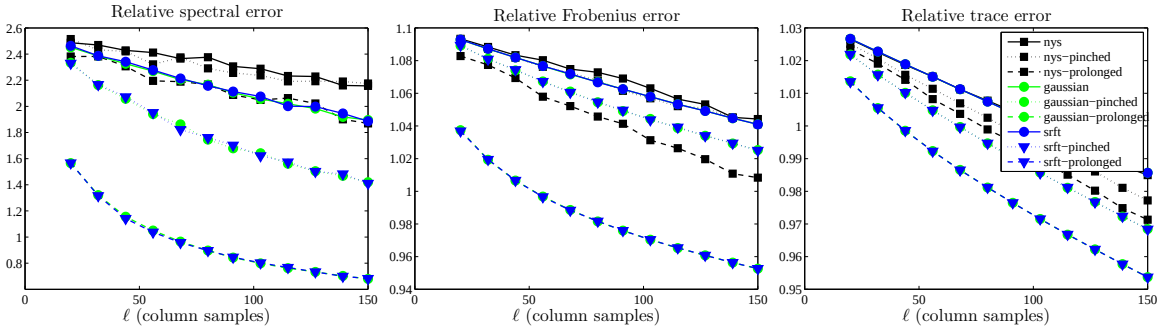
It is also evident that the pinched approximants often yield a much slighter increase in accuracy over the one-pass sketches than do the prolonged approximants. Recall that the prolonged approximants are simply two-pass sketches. Our investigations point to the conclusion that two-pass sketches are significantly more accurate than the projection-based low-rank approximations that also require two passes over $\mathbf{A}$. Of course, one should temper this comparison with the knowledge that projection-based low-rank approximations, unlike SPSD sketches, are stably computable.



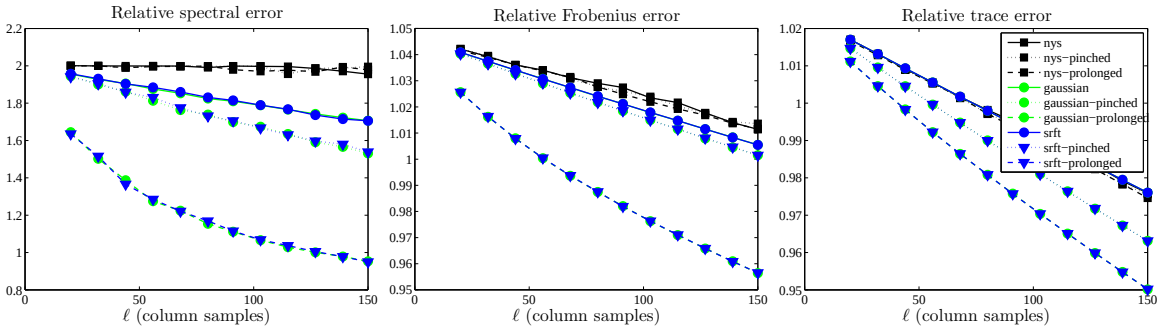
(a) Gnutella, $k = 20$



(b) Dexter, $k = 8$



(c) AbaloneD, $\sigma = .15, k = 20$



(d) WineS, $\sigma = 1, k = 20$

Figure 6.14: COMPARISON OF PROJECTION-BASED LOW-RANK APPROXIMATIONS WITH ONE-PASS SPSD SKETCHES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, including the pinched and prolonged low-rank approximants, as a function of the number of columns samples ℓ, for several matrices from Table 6.2.