Topics in Randomized Numerical Linear Algebra

Thesis by
Alex Gittens

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2013
(Defended May 31, 2013)

Acknowledgements

I have been privileged to befriend and work with exceptional individuals.

I owe a debt of thank to Drs. Emmanuel Papadakis and Gordon Johnson of the University of Houston for helping me get started on my academic path. I consider myself lucky to have had the guidance of my PhD advisor, Joel Tropp, and remain humbled by his intuitive grasp of the ideas and tools I have wrestled with throughout my graduate career. In addition to Joel, I am grateful to the other collaborators I was privileged to work with: Christos Boutsidis, Michael Mahoney, and Richard Chen. It was a pleasure working with them.

I benefited greatly from interactions with my fellow students; in particular, I thank Catherine Beni, Chia-Chieh Chu, Derek Leong, Jinghao Huang, and Yamuna Phal for being great office mates and friends. My dear friend and amateur herpetologist Yekaterina Pavlova I thank for being her very unique self. Stephen Becker, Michael McCoy, Peter Stobbe, Patrick Sanan, Zhiyi Li, George Chen, and Yaniv Plan may have provided answers to the odd math question, but I appreciate them more for the hours of wide-ranging discussions on everything under the sun.

Finally, I thank my family for their support: my mother Mura; my father Erskine; my sister Lauren; and my grandfather Chesterfield, who was a formative influence on my young mind.

Abstract

This thesis studies three classes of randomized numerical linear algebra algorithms, namely: (i) randomized matrix sparsification algorithms, (ii) low-rank approximation algorithms that use randomized unitary transformations, and (iii) low-rank approximation algorithms for positive-semidefinite (PSD) matrices.

Randomized matrix sparsification algorithms set randomly chosen entries of the input matrix to zero. When the approximant is substituted for the original matrix in computations, its sparsity allows one to employ faster sparsity-exploiting algorithms. This thesis contributes bounds on the approximation error of nonuniform randomized sparsification schemes, measured in the spectral norm and two NP-hard norms that are of interest in computational graph theory and subset selection applications.

Low-rank approximations based on randomized unitary transformations have several desirable properties: they have low communication costs, are amenable to parallel implementation, and exploit the existence of fast transform algorithms. This thesis investigates the tradeoff between the accuracy and cost of generating such approximations. State-of-the-art spectral and Frobenius-norm error bounds are provided.

The last class of algorithms considered are SPSD "sketching" algorithms. Such sketches can be computed faster than approximations based on projecting onto mixtures of the columns of the matrix. The performance of several such sketching schemes is empirically evaluated using a suite of canonical matrices drawn from machine learning and data analysis applications, and a framework is developed for establishing theoretical error bounds.

In addition to studying these algorithms, this thesis extends the Matrix Laplace Transform framework to derive Chernoff and Bernstein inequalities that apply to *all* the eigenvalues of certain classes of random matrices. These inequalities are used to investigate the behavior of the singular values of a matrix under random sampling, and to derive convergence rates for each individual eigenvalue of a sample covariance matrix.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and contributions

Massive datasets are common: among other places, they arise in data-analysis and machine learning applications. These datasets are often represented as matrices, so the fundamental tools of linear algebra are indispensable in their analysis. For instance, modeling and data analysis methods based on low-rank approximation have become popular because they capture the low-dimensional structure implicit in massive high-dimensional modern datasets. Low-rank approximations are also used for their noise-elimination and regularization properties [Han90]. Among many applications, we mention PCA [HTF08], multidimensional scaling [CC00], collaborative filtering [SAJ10], manifold learning [HLMS04], and latent semantic indexing [DDF$^+$90].

The truncated singular value decomposition (SVD) and the rank-revealing QR decomposition are classical decompositions used to construct low-rank approximants. However, the construction of both of these decompositions costs $O(n^\omega)$ operations for an $n \times n$ matrix [CH92] (where ω is the exponent for matrix multiplication). For small k and large n, Krylov space methods can potentially provide truncated SVDs in much less time. In practice, the number of operations required varies considerably depending upon the specifics of the method and the spectral properties of the matrix, but since one must perform at least k dense matrix–vector multiplies (assuming the matrix is unstructured), computing the rank-k truncated SVD using a Krylov method requires at least $\Omega(kn^2)$ operations. Further, iterative schemes like Krylov methods require multiple passes over the matrix, which may incur high communication costs if the matrix is stored in a distributed fashion, or if the data has to percolate through a hierarchical memory architecture [CW09].

Much interest has been expressed in finding $o(kn^2)$ low-rank approximation schemes that offer approximation guarantees comparable with those of the truncated SVD. *Randomized numerical linear algebra* (RNLA) refers to a field of research that arose in the early 2000s at the intersection of several research communities, including the theoretical computer science and numerical linear algebra communities, in response to the desire for fast, efficient algorithms for manipulating large matrices. RNLA algorithms for matrix approximation focus on reducing the number of arithmetic operations and the communications costs of algorithms by judiciously exploiting randomness. Typically, these algorithms take one of two approaches. The sampling approach advocates using information obtained by randomly sampling the columns, rows, or entries of the matrix to form an approximation to the matrix. The random projection approach randomly mixes the entries of the matrix before employing the sampling approach. The analysis of both classes of algorithms requires the use of tools from the nonasymptotic theory of random matrices.

This thesis contributes to both approaches to forming randomized matrix approximants, and

it extends the toolset available to researchers working in the field of RNLA.

- Chapter 2 builds upon the matrix Laplace transform originated by Ahlswede and Winter to provide eigenvalue analogs of classical exponential tail bounds for *all* eigenvalues of a sum of random Hermitian matrices. Such sums arise often in the analysis of RNLA algorithms.
- Chapter 3 develops bounds on the norms of random matrices with independent mean-zero entries, and it applies these bounds to investigate the performance of randomized entry-wise sparsification algorithms.
- Chapter 5 provides guarantees on the quality of low-rank approximations generated using a class of random projections that exploit fast unitary transformations.
- Chapter 6 concludes by providing a framework for the analysis of a diverse class of low-rank approximations to positive-semidefinite matrices, as well as empirical evidence of the efficacy of these approximations over a wide range of matrices. The class of approximations considered includes both sampling-based approximations as well as projection-based approximations.

In the remainder of this introductory chapter, we survey the sampling and projection-based approaches to randomized matrix approximation and the tools currently available to researchers for the interrogation of the properties of random matrices. We conclude with an overview of the contributions of this thesis.

1.1 The sampling approach to matrix approximation

Sparse approximants are of interest because they be used in lieu of the original matrix to reduce the cost of calculations. Randomized sparsified approximations to matrices have found applications in approximate eigenvector computations [AM01, AHK06, AM07] and semidefinite optimization algorithms [AHK05, d'A11].

The first randomized element-wise matrix sparsification algorithms are due to Achlioptas and McSherry [AM01, AM07], who considered schemes in which a matrix is replaced with a randomized approximant that has far fewer nonzero entries. Their motivation for considering randomized sparsification was the desire to use the fast algorithms available for computing the SVDs of large sparse matrices to approximate the SVDs of large dense matrices. In the same work, they presented a scheme that randomly quantizes the entries of the matrix to $\pm \max_{ij} |A_{ij}|$. Such quantization schemes are of interest because they reduce the cost of storing and working with the matrix. Note that this quantization scheme requires two passes over the matrix: one to compute b, then another to quantize. The bounds given in [AM07] for both schemes guarantee that the spectral norm error of the approximations to a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ remain on the order of $\sqrt{\max\{m,n\}}\max_{ij}|A_{ij}|$ with high probability. If each entry in the matrix is replaced by zero with probability $1-p$, the expected number of nonzeros in the approximant is shown to be at most $p\|\mathbf{A}\|_{\mathrm{F}}^2/\max_{ij}|A_{ij}|^2 + 4096m\log^4(n)$. These bounds are quite weak: the algorithms perform much better on average.

Arora et al. presented an alternative quantization and sparsification scheme in [AHK06] that has the advantage of requiring only one pass over the input matrix. The schemes of both Arora et. al and Achlioptas and McSherry involve entrywise calculations on the matrix being approximated, and have the property that the entries in the random approximant are independent of each other. Succeeding works on entry-wise matrix sparsification include [NDT10, DZ11, AKL13];

the algorithms given in these works also produce approximants with independent entries. The sharpest available bound on randomized element-wise sparsification is satisfied by the algorithm given in [DZ11]: given an accuracy parameter $\epsilon > 0$, this algorithm produces an approximant that satisfies $\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \leq \epsilon$ with high probability and has at most $28\epsilon^2 n \log(\sqrt{2n})\|\mathbf{A}\|_{\mathrm{F}}^2$ nonzero entries; the approximant can be calculated in one pass. The paper [NDT10] goes beyond matrix sparsification, addressing randomized element-wise tensor sparsification.

The natural next step after entry-wise sampling is the sampling of entire columns and rows. An influential portion of the first wave of RNLA algorithms employed such a sampling approach, in the form of Monte Carlo approximation algorithms. In [FKV98], Frieze, Kannan, and Vempala introduce the first algorithm of this type for calculating approximate SVDs of large matrices. They propose judiciously sampling a submatrix from $\mathbf{A}$ and using the SVD of this submatrix to find an approximation of the top singular spaces of $\mathbf{A}$. The projection of $\mathbf{A}$ onto this subspace is then used as the low-rank approximation. This algorithm of course requires two passes over the matrix. The original idea in [FKV98] was refined in a series of papers providing increasingly strong guarantees on the quality of the approximation [DK01, DK03, FKV04, DKM06a, DKM06b].

Rudelson and Vershynin take a different approach to the analysis of the Monte Carlo methodology for low-rank approximation in [RV07]. They consider $\mathbf{A}$ as a linear operator between finite-dimensional Banach spaces and apply techniques of probability in Banach spaces: decoupling, symmetrization, Slepian's lemma for Rademacher random variables, and a law of large numbers for operator-valued random variables. They show that, if $\mathbf{A}$ has numerical rank close to k, then it is possible to obtain an accurate rank-k approximation to $\mathbf{A}$ by sampling $\mathrm{O}(k \log k)$ rows of $\mathbf{A}$. Specifically, if one projects $\mathbf{A}$ onto the span of $\ell = \mathrm{O}(\epsilon^{-4} k \log k)$ of its rows, then the approximant satisfies $\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2 + \epsilon\|\mathbf{A}\|_2$ with high probability. Here $\mathbf{A}_k$ denotes the optimal rank-k approximation to $\mathbf{A}$, obtainable as the rank-k truncated SVD of $\mathbf{A}$.

Other researchers forwent the SVD entirely, considering instead alternative column and row-based matrix decompositions. In one popular class of approximations, the matrix is approximated with a product $\mathbf{CUR}$, where $\mathbf{C}$ and $\mathbf{R}$ are respectively small subsets of the columns and rows of the matrix and $\mathbf{U}$, the *coupling matrix*, is computed from $\mathbf{C}$ and $\mathbf{R}$ [DKM06c]. Accordingly, these schemes are known as CUR decompositions. Nyström extensions, introduced by Williams and Seeger in [WS01], are a similar class of low-rank approximations to positive-semidefinite matrices. They can be thought of as CUR decompositions constructed with the additional constraint that $\mathbf{C} = \mathbf{R}^T$, to preserve the positive-semidefiniteness of the approximant. Both CUR and Nyström decompositions can be constructed in one pass over the matrix.

The paper [DMM08] introduced a "subspace sampling" method of sampling the columns and rows to form $\mathbf{C}$ and $\mathbf{R}$ and showed that approximations formed with $\mathrm{O}(k \log k)$ columns and rows in this manner achieve Frobenius norm errors close to the optimal rank-k approximation error: $\|\mathbf{A} - \mathbf{CUR}\|_{\mathrm{F}} \leq (1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$. The *leverage scores* of the columns of $\mathbf{A}$ are used to generate the probability distribution used for column sampling: given $\mathbf{P}$, a projection onto the dominant k-dimensional right singular space of $\mathbf{A}$, the leverage score of the jth column of $\mathbf{A}$ is proportional to $(\mathbf{P})_{ii}$. The intuition is that the magnitude of the leverage score of a particular column reflects its influence in determining the dominant k-dimensional singular spaces of $\mathbf{A}$ [DM10].

In [MRT06, MRT11], Tygert et al. introduced randomized Interpolative Decompositions (ID) as an alternative low-rank factorization to the SVD. In IDs, the columns of $\mathbf{A}$ are represented as linear combinations of some small subset of the columns of $\mathbf{A}$. The algorithm of [MRT06] is accelerated in [WLRT08]. With high probability, it constructs matrices $\mathbf{B}$ and $\mathbf{\Pi}$ such that $\mathbf{B}$ consists of k columns sampled from $\mathbf{A}$, some subset of the columns of $\mathbf{\Pi}$ make up the $k \times k$ identity matrix, and $\|\mathbf{A} - \mathbf{B\Pi}\|_2 = \mathrm{O}(\sqrt{kmn}\|\mathbf{A} - \mathbf{A}_k\|_2)$.

The works of Har-Peled [HP06], and Deshpande et al. [DRVW06] use more intricate approaches based on column sampling to produce low-rank approximations with relative-error Frobenius norm guarantees. These algorithms require, respectively, $O(k^2 \log k)$ and $O(k)$ column samples.

Boutsidis et al. develop a general framework for analyzing the error of matrix approximation schemes based on column sampling in [BDMI11], where they establish optimal bounds on the errors of approximants produced by projecting a matrix onto the span of some subset of its columns. In particular, they show that there are matrices that cannot be efficiently approximated in the spectral norm using the sampling paradigm; specifically, given positive integers $k \leq \ell \leq n$, they demonstrate the existence of a matrix $\mathbf{A}$ such that

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \geq \left(1 + \sqrt{\frac{n^2 + \alpha}{\ell^2 + \alpha}}\right) \|\mathbf{A} - \mathbf{A}_k\|_2$$

when $\tilde{\mathbf{A}}$ is *any* approximation obtained by projecting $\mathbf{A}$ onto the span of ℓ of its columns. Because this bound holds regardless of how the columns are selected, it is clear that, at least in the spectral norm, the sampling paradigm is not sufficient to obtain near optimal approximation errors. Stronger spectral norm guarantees can be obtained using the random projection approach to matrix approximation.

1.2 The random-projection approach to matrix approximation

A wide range of results in RNLA have been inspired by the work of Johnson and Lindenstrauss in geometric functional analysis, who showed that embeddings into random low-dimensional spaces can preserve the geometry of point sets. The celebrated Johnson–Lindenstrauss lemma states that, given n points in a high-dimensional space, a random projection into a space of dimension $\Omega(\log n)$ preserves the distance between the points. Such geometry-preserving, dimension-reducing maps are known as Johnson–Lindenstrauss transforms (JLT).

The work of Papadimitriou et al. in [PRTV00] on the algorithmic application of randomness to facilitate information retrieval popularized the use of JLTs in RNLA. Unlike sample-based methods like the CUR decomposition that project the matrix onto the span of a subset of its *columns* (and/or rows), random projection methods produce approximations to the matrix by projecting it onto some subspace of its entire *range*. The intuition behind these methods is similar to that behind the power method, or orthogonal iteration: one can approximately capture the top left singular space of a matrix by applying it to a sufficiently large number of random vectors. One then obtains a low-rank approximation of the matrix by projecting it onto this approximate singular space. Projection-based matrix approximation algorithms require at least two passes over the matrix: one to form an approximate basis for the top left singular space of the matrix, and one to project the matrix onto that basis.

In the influential paper [Sar06], Sarlós developed fast approximation algorithms for SVDs, least squares, and matrix multiplication under the randomized projection paradigm. His algorithms take advantage of Ailon and Chazelle's work, which establish that certain structured randomized transformations can be used to quickly compute dimension reductions [AC06]. At around the same time, Martinsson, Rohklin, and Tygert introduced a randomized projection-based algorithm for the calculation of approximate SVDs [MRT06, MRT11]. In this algorithm, to obtain an approximate rank-k SVD of $\mathbf{A}$, one applies $k + p$ gaussian vectors to $\mathbf{A}$ then projects $\mathbf{A}$ onto the resulting subspace. Here, p is a small integer known as the *oversampling parameter*.

The approximation returned by the algorithm can be written as $\tilde{\mathbf{A}} = \mathbf{P}_{\mathbf{AS}}\mathbf{A}$, where $\mathbf{S}$ is a Gaussian matrix and the notation $\mathbf{P}_{\mathbf{M}}$ denotes the projection onto the range of the matrix $\mathbf{M}$. The spectral norm error of the approximant is guaranteed to be at most $\sqrt{\max m, n}\|\mathbf{A} - \mathbf{A}_k\|_2$ with high probability, and if $\mathbf{A}$ is unstructured and dense, the algorithm costs $\mathrm{O}(mnk)$ time. Despite the fact that its runtime is asymptotically the same as those of classical Krylov iteration schemes (e.g. the Lanczos method), this algorithm is of interest because it requires only two passes over the matrix. Moreover, the algorithm performs well in the presence of degenerate singular values, a situation which often causes Lanczos methods to stagnate [MRT11]. Finally, this algorithm is more readily parallelizable than iterative schemes.

In [WLRT08], inspired by Sarlós's work in [Sar06], Woolfe et al. observed that the runtime of the algorithm of [MRT06, MRT11] could be reduced to $\mathrm{O}(mn\log(k) + k^4(m+n))$ by substituting a structured random matrix for the Gaussian matrix used in the original algorithm. Specifically, they show that if the "sampling matrix" $\mathbf{S}$ consists of $\mathrm{O}(k^2)$ uniformly randomly selected columns of the product of the discrete Fourier transform matrix and a diagonal matrix of random signs, then the error guarantees of the algorithm remain unchanged while the worst-case runtime decreases. Nguyen et al. consider the same approximation in [NDT09], $\tilde{\mathbf{A}} = \mathbf{P}_{\mathbf{AS}}\mathbf{A}$, and obtain improved results: if $\mathbf{S}$ has $\mathrm{O}(k\log k)$ columns constructed as in the algorithm of [WLRT08], then with constant probability $\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \leq \sqrt{m/(k\log k)}\|\mathbf{A} - \mathbf{A}_k\|_2$.

The paper [BDMI11] and the survey article [HMT11] constituted a significant step forward in the analysis of random projection-based matrix approximation algorithms, because they provided a framework for the analysis of the Frobenius and spectral norm errors of approximants of the form $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$ using arbitrary sampling matrices $\mathbf{S}$. In [HMT11], this framework is used to provide guarantees on the errors of approximants of the form $\tilde{\mathbf{A}} = \mathbf{P}_{\mathbf{AS}}\mathbf{A}$ for $\mathbf{S}$ Gaussian and for $\mathbf{S}$ consisting of uniformly randomly selected columns of the product of the Walsh–Hadamard transform matrix and a diagonal matrix of random signs.

1.3 Nonasymptotic random matrix theory

The behavior of RNLA algorithms can often be analyzed in terms of the behavior of a sum of random matrices. As an example, consider the entry-wise sparsification schemes described earlier in the chapter: there, the approximants can be considered to be a sum of random matrices, where each term in the sum contributes one entry to the approximant. In each of the works cited, the design of the sparsification algorithm was crucially influenced by the particular tool used to analyze its performance. Achlioptas and McSherry used a concentration inequality due to Talagrand [AM01, AM07], Arora et al. used scalar Chernoff bounds [AHK06], Drineas et al. used the non-commutative Khintchine inequalities [NDT10], and Drineas and Zouzias used matrix Bernstein inequalities [DZ11]. As the tools available to researchers increased in their generality, the sparsification algorithms became more sophisticated, and the analysis of their errors became sharper.

The study of the spectra of random matrices is naturally divided into two subfields: the nonasymptotic theory, which gives probability bounds that hold for finite-dimensional matrices but may not be sharp, and the asymptotic theory, which precisely describes the behavior of certain families of matrices as their dimensions go to infinity. Unfortunately, the strength of the asymptotic techniques lies in the determination of convergence and the development of asymptotically sharp bounds, rather than the development of tail bounds which hold at a fixed dimension. Accordingly, the nonasymptotic theory is of most relevance in RNLA applications.

The sharpest and most comprehensive results available in the nonasymptotic theory concern

the behavior of Gaussian matrices. The amenability of the Gaussian distribution makes it possible to obtain results such as Szarek's nonasymptotic analog of the Wigner semicircle theorem for Gaussian matrices [Sza90] and Chen and Dongarra's bounds on the condition number of Gaussian matrices [CD05]. The properties of less well-behaved random matrices can sometimes be related back to those of Gaussian matrices using probabilistic tools, such as symmetrization; see, e.g., the derivation of Latała's bound on the norms of zero-mean random matrices [Lat05].

More generally, bounds on extremal eigenvalues can be obtained from knowledge of the moments of the entries. For example, the smallest singular value of a square matrix with i.i.d. zero-mean subgaussian entries is $O(n^{-1/2})$ with high probability [RV08]. Concentration of measure results, such as Talagrand's concentration inequality for product spaces [Tal95], have also contributed greatly to the nonasymptotic theory. We mention in particular the work of Achlioptas and McSherry on randomized sparsification of matrices [AM01, AM07], that of Meckes on the norms of random matrices [Mec04], and that of Alon, Krivelevich and Vu [AKV02] on the concentration of the largest eigenvalues of random symmetric matrices, all of which are applications of Talagrand's inequality. In cases where geometric information on the distribution of the random matrices is available, the tools of empirical process theory—such as generic chaining, also due to Talagrand [Tal05]—can be used to convert this geometric information into information on the spectra. One natural example of such a case consists of matrices whose rows are independently drawn from a log-concave distribution [MP06, ALPTJ11].

One of the most general tools in the nonasymptotic theory toolbox is the Noncommutative Khintchine Inequality (NCKI), which bounds the moments of the norm of a sum of randomly signed matrices [LPP91]. Despite its power and generality, the NCKI is unwieldy. To use it, one must reduce the problem to a suitable form by applying symmetrization and decoupling arguments and exploiting the equivalence between moments and tail bounds. It is often more convenient to apply the NCKI in the guise of a lemma, due to Rudelson [Rud99], that provides an analog of the law of large numbers for sums of rank-one matrices. This result has found many applications, including column-subset selection [RV07] and the fast approximate solution of least-squares problems [DMMS11]. The NCKI and its corollaries do not always yield sharp results because parasitic logarithmic factors arise in many settings.

Classical exponential tail bounds for sums of independent random variables can be developed using the machinery of moment-generating functions (mgfs), by exploiting the fact that the mgf of a sum of independent random variables is the product of the mgfs of the summands. Ahlswede and Winter [AW02] extended this technique to produce tail bounds for the eigenvalues of sums of independent Hermitian random variables. Because matrices are non-commutative, the matrix mgf of a sum of independent random matrices does not factor nicely as in the scalar case. The influential work of Ahlswede and Winter, as well as the immediately following works developing exponential matrix probability inequalities, relied upon trace inequalities to circument the difficulty of noncommutativity [CM08, Rec11, Oli09, Oli10, Gro11]. Tropp showed that these matrix probability inequalities can be sharpened considerably by working with cumulant generating functions instead of mgfs [Tro12, Tro11c, Tro11a].

Chatterjee established that in the scalar case, powerful concentration inequalities could be recovered from arguments based on the method of exchangeable pairs [Cha07]. Mackey and collaborators extended the method of exchangeable pairs to matrix-valued functions [MJC+12]. The resulting bounds are sufficiently sharp to recover the NCKI, and can even be used to interrogate the behavior of matrix-valued functions of dependent random variables. Most recently, Paulin et al. have further extended the matrix method of exchangeable pairs to apply to an even larger class of matrix-valued functions [PMT13].

Despite the diversity of the tools mentioned here, all share a common limitation: they provide bounds only on the extremal eigenvalues of the relevant classes of random matrices.

1.4 Contributions

We conclude with a summary of the main contributions of this thesis.

1.4.1 Nonasymptotic random matrix theory

The matrix Laplace transform technique pioneered by Ahlswede and Winter, which applies to sums of independent random matrices [AW02, Tro12], is one of the most generally applicable techniques in the arsenal of nonasymptotic random matrix theory.

However, the matrix Laplace transform technique yields bounds on only the extremal eigenvalues of Hermitian random matrices. Chapter 2 describes an extension of the matrix Laplace transform technique, based upon the variational characterization of the eigenvalues of Hermitian matrices, for bounding *all* eigenvalues of sums of independent random Hermitian matrices. This is the first general purpose tool for bounding interior eigenvalues of such a wide class of random matrices.

The minimax Laplace transform introduced in Chapter 2 relates the behavior of the k-th eigenvalue of a random self-adjoint matrix to the behavior of its compressions to subspaces:

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

where the minimization is taken over an appropriate set of matrices $\mathbf{V}$ with orthonormal columns. We show that when one has sufficiently strong semidefinite bounds on the matrix cumulant generating functions $\log \mathbb{E}\mathrm{e}^{\theta \mathbf{V}^*\mathbf{X}_i\mathbf{V}}$ of the compressions of the summands $\mathbf{X}_i$, the minimax Laplace transform technique yields exponential probability bounds for all the eigenvalues of $\mathbf{Y} = \sum_i \mathbf{X}_i$.

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

1.4.2 Matrix sparsification

Chapter 3 analyzes the approximation errors of randomized schemes that approximate a fixed $m \times n$ matrix $\mathbf{A}$ with a random matrix $\mathbf{X}$ having the properties that the entries of $\mathbf{X}$ are independent and average to the corresponding entries of $\mathbf{A}$. This investigation was initiated by the observation that several algorithms for random matrix quantization and sparsification are based on approximations that have these properties [AM01, AHK06, AM07]. A generic framework for the analysis of such approximation schemes is established, and this essentially recapitulates the known guarantees for the referenced algorithms.

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

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|_2 \le \mathrm{C}\left[\max_j\left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2} + \max_k\left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2}\left(\sum_{jk}\mathbb{E}(X_{jk}-a_{jk})^4\right)^{1/4}\right], \quad (1.4.1)$$

where C is a universal constant. This expectation bound is obtained by leveraging work done by Latała on the spectral norm of random matrices with zero mean entries [Lat05]. When the entries of $\mathbf{A}$ are bounded (so that the variances of the entries of $\mathbf{X}$ are small), an argument based on a bounded difference inequality shows that the approximation error does not exceed this expectation by much.

Inequality (1.4.1) identifies properties desirable in randomized approximation schemes: namely, that they minimize the maximum column and row norms of the variances of the entries, as well as the fourth moments of all entries. Thus our results supply guidance in the design of future approximation schemes. The results also yield comparable analyses of the quantization and sparsification schemes introduced in [AM01, AM07] and recover error bounds for the quantization/sparsification scheme proposed by Arora, Hazan, and Kale in [AHK06] that are comparable to those supplied in [AHK06]. However, for the more recent sparsification schemes presented in [NDT10, DZ11, AKL13], our results do not provide sparsification guarantees as strong as those offered in the originating papers.

Chapter 3 also analyzes the performance of randomized matrix approximation schemes as measured using non-unitary invariant norms. The literature on randomized matrix approximation has, with few exceptions, focused on the behavior of the spectral and Frobenius norms. However, depending on the application, other norms are of more interest; for instance, the $p \to q$ norms naturally arise when one considers $\mathbf{A}$ as a map from $\ell_p(\mathbb{R}^n)$ to $\ell_q(\mathbb{R}^m)$. Consider, in particular, the $\infty \to 1$ and $\infty \to 2$ norms, both of which are NP-hard to compute. The $\infty \to 1$ norm has applications in graph theory and combinatorics. The $\infty \to 2$ norm has applications in numerical linear algebra. In particular, it is a useful tool in the column subset selection problem: that of, given a matrix $\mathbf{A}$ with unit norm columns, choosing a large subset of the columns of $\mathbf{A}$ so that the resulting submatrix has a norm smaller than some fixed constant (larger than one).

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

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

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

$$\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} \le 2\mathbb{E}(\|\mathbf{Z}\|_{\mathrm{col}} + \left\|\mathbf{Z}^T\right\|_{\mathrm{col}}) \quad \text{and}$$
$$\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} \le 2\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} + 2\min_{\mathbf{D}}\mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty},$$

where the minimization is taken over the set of positive diagonal matrices satisfying $\mathrm{Tr}(\mathbf{D}^2)=1$. The norm $\|\mathbf{Z}\|_{2\to\infty}$ is the largest of the Euclidean norms of the rows of the matrix, $\|\mathbf{Z}\|_{\mathrm{F}}$ is the

Frobenius norm, and the column norm $\|\mathbf{Z}\|_{\text{col}}$ is the sum of the Euclidean norms of the columns of the matrix. As in the case of the spectral norm, a bounded differences inequality guarantees that if the entries of $\mathbf{A}$ are bounded, then the errors $\|\mathbf{A}-\mathbf{X}\|_{\infty\to\xi}$ for $\xi \in \{1,2\}$ concentrate about these expectations. Thus we have bounds on norms which are NP-hard to compute, in terms of much simpler quantities. Both these bounds are optimal in the sense that each term in the bound can be shown to be necessary. In the case of the $\infty \to 1$ norm, a matching lower bound establishes the sharpness of the bound.

1.4.3 Low-rank approximation using fast unitary transformations

Chapter 5 offers a new analysis of the subsampled randomized Hadamard transform (SRHT) approach to low-rank approximation. This is a specific instance of a class of low-rank approximation algorithms based on fast unitary transformations, and the analysis provided applies, *mutatis mutandis*, to other low-rank approximation algorithms which use fast unitary transformations.

Let $\ell > k$ be a positive integer and let $\mathbf{S} \in \mathbb{R}^{n\times\ell}$ be a matrix whose columns are random vectors, then projection methods approximate $\mathbf{A}$ with $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$, which has rank at most ℓ. Here, the notation $\mathbf{P}_{\mathbf{M}}$ denotes the projection onto the range of $\mathbf{M}$. One can reduce the cost of the algorithm by using random matrices $\mathbf{S}$ whose structure allows for fast multiplication. Specifically, one can reduce the cost of forming the product $\mathbf{AS}$ from $\mathrm{O}(mn\ell)$ to $\mathrm{O}(mn\log\ell)$. One choice of a structured random matrix is the transpose of the subsampled randomized Hadamard transform (SRHT),

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

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

The previous tightest bound on the spectral-norm error of SRHT-based low-rank approximations is given in [HMT11], where it is shown that

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

with probability at least $1-\mathrm{O}(1/k)$ when ℓ is at least on the order of $k\log k$. In some situations, this bound is close to optimal. But when $\mathbf{A}$ is rank-deficient or has fast spectral decay, this result does not reflect the correct behavior. In Chapter 5 we establish that

$$\left\|\mathbf{A}-\mathbf{P}_{\mathbf{AS}}\mathbf{A}\right\|_2 \le \mathrm{O}\left(\sqrt{\frac{\log(n)\log(\operatorname{rank}(\mathbf{A}))}{\ell}}\right)\left\|\mathbf{A}-\mathbf{A}_k\right\|_2 + \mathrm{O}\left(\sqrt{\frac{\log(\operatorname{rank}(\mathbf{A}))}{\ell}}\right)\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}$$

with constant failure probability. The factor in front of the optimal error has been reduced at the cost of the introduction of a Frobenius term. This Frobenius term is small when $\mathbf{A}$ has fast spectral decay. We also find Frobenius-norm error bounds.

1.4.4 Randomized SPSD sketches

Chapter 6 considers the problem of forming a low-rank approximation to a symmetric positive-semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ using "SPSD sketches." Let $\mathbf{S}$ be a matrix of size $n \times \ell$, where $\ell \ll n$. Then the SPSD sketch of $\mathbf{A}$ corresponding to $\mathbf{S}$ is $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$, where

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

Sketches formed according to this model have rank at most ℓ and are also symmetric positive-semidefinite. The simplest such SPSD sketches are formed by taking $\mathbf{S}$ to contain random columns sampled uniformly without replacement from the appropriate identity matrix. These sketches, known as Nyström extensions, are popular in applications where it is expensive or undesirable to have full access to $\mathbf{A}$: Nyström extensions require only knowledge of ℓ columns of $\mathbf{A}$.

The accuracy of SPSD sketches can be increased using the so-called power method, wherein one takes the sketching matrix to be $\mathbf{S} = \mathbf{A}^p\mathbf{S}_0$ for some integer $p \geq 2$ and $\mathbf{S}_0$ is a sketching matrix. The corresponding SPSD sketch is $\mathbf{A}^p\mathbf{S}_0(\mathbf{S}_0^T\mathbf{A}^{2p-1}\mathbf{S}_0)^{\dagger}\mathbf{S}_0^T\mathbf{A}^p$.

Chapter 6 establishes a framework for the analysis of SPSD sketches, and supplies spectral, Frobenius, and trace-norm error bounds for SPSD sketches corresponding to random $\mathbf{S}$ sampled from several distributions. The error bounds obtained are asymptotically smaller than the other bounds available in the literature for SPSD sketching schemes. Our bounds apply to sketches constructed using the power method, and we see that the errors of these sketches decrease like $(\lambda_{k+1}(\mathbf{A})/\lambda_k(\mathbf{A}))^p$.

In particular, our framework supplies an optimal spectral-norm error bound for Nyström extensions. Because they are based on uniform column sampling, Nyström extensions perform best when the information in the top k-dimensional eigenspace is distributed evenly throughout the columns of $\mathbf{A}$. One way to quantify this idea uses the concept of *coherence*, taken from the matrix completion literature [CR09]. Let $\mathscr{S}$ be a k-dimensional subspace of $\mathbb{R}^n$. The coherence of $\mathscr{S}$ is

$$\mu(\mathscr{S}) = \frac{n}{k}\max_i(\mathbf{P}_{\mathscr{S}})_{ii}.$$

The coherence of the dominant k-dimensional eigenspace of $\mathbf{A}$ is a measure of how much comparative influence the individual columns of $\mathbf{A}$ have on this subspace: if μ is small, then all columns have essentially the same influence; if μ is large, then it is possible that there is a single column in $\mathbf{A}$ which alone determines one of the top k eigenvectors of $\mathbf{A}$.

Talwalkar and Rostamizadeh were the first to use coherence in the analysis of Nyström extensions. Let $\mathbf{A}$ be exactly rank-k and μ denote the coherence of its top k-dimensional eigenspace. In [TR10], they show that if one samples on the order of $\mu k \log(k/\delta)$ columns to form a Nyström extension, then with probability at least $1-\delta$ the Nyström extension is *exactly* $\mathbf{A}$. The framework provided in Chapter 6 allows us to expand this result to apply to matrices with arbitrary rank. Specifically, we show that when $\ell = \mathrm{O}(\mu k \log k)$, then

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

with constant probability. This bound is shown to be optimal in the worst case.

Low-rank approximations computed using the SPSD sketching model are *not* guaranteed to be numerically stable: if $\mathbf{W}$ is ill-conditioned, then instabilities may arise in forming the product $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$. A regularization scheme proposed in [WS01] suggests avoiding numerical ill-conditioning issues by using an SPSD sketch constructed from the matrix $\mathbf{A} + \rho\mathbf{I}$, where

$\rho > 0$ is a regularization parameter. In Chapter 6, we provide the first error analysis of this regularization scheme, and compare it empirically to another regularization scheme introduced in [CD11].

Finally, in addition to theoretical results, Chapter 6 provides a detailed suite of empirical results on the performance of SPSD sketching schemes applied to matrices culled from data analysis and machine learning applications.

Chapter 2

Bounds for all eigenvalues of sums of Hermitian random matrices

2.1 Introduction

The classical tools of nonasymptotic random matrix theory can sometimes give quite sharp estimates of the extreme eigenvalues of a Hermitian random matrix, but they are not readily adapted to the study of the interior eigenvalues. This is because, while the extremal eigenvalues are the maxima and minima of a random process, more delicate and challenging minimax problems must be solved to obtain the interior eigenvalues.

This chapter introduces a simple method, based upon the variational characterization of eigenvalues, that parlays bounds on the extreme eigenvalues of sums of random Hermitian matrices into bounds that apply to all the eigenvalues[1]. This technique extends the matrix Laplace transform method detailed in [Tro12]. We combine these ideas to extend several of the inequalities in [Tro12] to address the fluctuations of interior eigenvalues. Specifically, we provide eigenvalue analogs of the classical multiplicative Chernoff bounds and Bennett and Bernstein inequalities.

In this technique, the delicacy of the minimax problems which implicitly define the eigenvalues of Hermitian matrices is encapsulated in terms that reflect the fluctuations of the summands in the appropriate eigenspaces. In particular, we see that the fluctuations of the kth eigenvalue of the sum above and below the kth eigenvalue of the expected sum are controlled by two different quantities. This satisfies intuition: for instance, given samples from a nondegenerate stationary random process with finite covariance matrix, one expects that the smallest eigenvalue of the sample covariance matrix is more likely to be an underestimate of the smallest eigenvalue of the covariance matrix than it is to be an overestimate.

We provide two illustrative applications of our eigenvalue tail bounds: Theorem 2.14 quantifies the behavior of the singular values of matrices obtained by sampling columns from a short, fat matrix; and Theorem 2.15 quantifies the convergence of the eigenvalues of Wishart matrices.

[1] The content of this chapter is adapted from the technical report [GT09] co-authored with Joel Tropp.

2.2 Notation

We define $\mathbb{M}_{\text{sa}}^n$ to be the set of Hermitian matrices with dimension n. We often compare Hermitian matrices using the semidefinite ordering. In this ordering, $\mathbf{A}$ is greater than or equal to $\mathbf{B}$, written $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \preceq \mathbf{A}$, when $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

The eigenvalues of a matrix $\mathbf{A}$ in $\mathbb{M}_{\text{sa}}^n$ are arranged in weakly decreasing order: $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$. Likewise, the singular values of a rectangular matrix $\mathbf{A}$ with rank ρ are ordered $\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_\rho(\mathbf{A}) = \sigma_{\min}(\mathbf{A})$. The spectral norm of a matrix $\mathbf{B}$ is written $\|\mathbf{B}\|_2$.

2.3 The Courant–Fisher Theorem

In this chapter, we work over the complex field $\mathbb{C}$. One of our central tools is the variational characterization of the eigenvalues of a Hermitian matrix given by the Courant–Fischer Theorem. For integers d and n satisfying $1 \leq d \leq n$, the complex Stiefel manifold

$$\mathbb{V}_d^n = \{\mathbf{V} \in \mathbb{C}^{n \times d} : \mathbf{V}^*\mathbf{V} = \mathbf{I}\}$$

is the collection of orthonormal bases for the d-dimensional subspaces of $\mathbb{C}^n$, or, equivalently, the collection of all isometric embeddings of $\mathbb{C}^d$ into $\mathbb{C}^n$. Let $\mathbf{A}$ be a Hermitian matrix with dimension n, and let $\mathbf{V} \in \mathbb{V}_d^n$ be an orthonormal basis for a subspace of $\mathbb{C}^n$. Then the matrix $\mathbf{V}^*\mathbf{A}\mathbf{V}$ can be interpreted as the compression of $\mathbf{A}$ to the space spanned by $\mathbf{V}$.

Proposition 2.1 (Courant–Fischer ([HJ85, Theorem 4.2.11])). *Let $\mathbf{A}$ be a Hermitian matrix with dimension n. Then*

$$\lambda_k(\mathbf{A}) = \min_{\mathbf{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\mathbf{V}^*\mathbf{A}\mathbf{V}\right) \quad \text{and} \tag{2.3.1}$$

$$\lambda_k(\mathbf{A}) = \max_{\mathbf{V} \in \mathbb{V}_k^n} \lambda_{\min}\left(\mathbf{V}^*\mathbf{A}\mathbf{V}\right). \tag{2.3.2}$$

A matrix $\mathbf{V}_- \in \mathbb{V}_k^n$ achieves equality in (2.3.2) if and only if its columns span a top k-dimensional invariant subspace of $\mathbf{A}$. Likewise, a matrix $\mathbf{V}_+ \in \mathbb{V}_{n-k+1}^n$ achieves equality in (2.3.1) if and only if its columns span a bottom $(n-k+1)$-dimensional invariant subspace of $\mathbf{A}$.

The $\pm$ subscripts in Proposition 2.1 are chosen to reflect the fact that $\lambda_k(\mathbf{A})$ is the *minimum* eigenvalue of $\mathbf{V}_-^*\mathbf{A}\mathbf{V}_-$ and the *maximum* eigenvalue of $\mathbf{V}_+^*\mathbf{A}\mathbf{V}_+$. As a consequence of Proposition 2.1, when $\mathbf{A}$ is Hermitian, $\lambda_k(-\mathbf{A}) = -\lambda_{n-k+1}(\mathbf{A})$. This fact allows us to use the same techniques we develop for bounding the eigenvalues from above to bound them from below.

2.4 Tail bounds for interior eigenvalues

In this section we develop a generic bound on the tail probabilities of eigenvalues of sums of independent, random, Hermitian matrices. We establish this bound by supplementing the matrix Laplace transform methodology of [Tro12] with Proposition 2.1 and a result, due to Lieb and Seiringer [LS05], on the concavity of a certain trace function on the cone of positive-definite matrices.

First we observe that the Courant–Fischer Theorem allows us to relate the behavior of the kth eigenvalue of a matrix to the behavior of the largest eigenvalue of an appropriate compression of the matrix.

Theorem 2.2. *Let* $\mathbf{Y}$ *be a random Hermitian matrix with dimension* n*, and let* $k \le n$ *be an integer. Then, for all* $t \in \mathbb{R}$*,*

$$\mathbb{P}\{\lambda_k(\mathbf{Y}) \ge t\} \le \inf_{\theta>0} \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \left\{ \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \mathbf{V}^*\mathbf{Y}\mathbf{V}} \right\}. \tag{2.4.1}$$

Proof. Let θ be a fixed positive number. Then

$$\begin{aligned}\mathbb{P}\{\lambda_k(\mathbf{Y}) \ge t\} = \mathbb{P}\{\lambda_k(\theta\mathbf{Y}) \ge \theta t\} &= \mathbb{P}\left\{\mathrm{e}^{\lambda_k(\theta\mathbf{Y})} \ge \mathrm{e}^{\theta t}\right\} \\ &\le \mathrm{e}^{-\theta t} \cdot \mathbb{E}\mathrm{e}^{\lambda_k(\theta\mathbf{Y})} = \mathrm{e}^{-\theta t} \cdot \mathbb{E}\exp\left\{ \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\theta \mathbf{V}^*\mathbf{Y}\mathbf{V}\right) \right\}.\end{aligned}$$

The first identity follows from the positive homogeneity of eigenvalue maps and the second from the monotonicity of the scalar exponential function. The final two relations are Markov's inequality and (2.3.1).

To continue, we need to bound the expectation. Use monotonicity to interchange the order of the exponential and the minimum; then apply the spectral mapping theorem to see that

$$\begin{aligned}\mathbb{E}\exp\left\{ \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\theta \mathbf{V}^*\mathbf{Y}\mathbf{V}\right) \right\} &= \mathbb{E} \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\exp(\theta \mathbf{V}^*\mathbf{Y}\mathbf{V})\right) \\ &\le \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \mathbb{E}\lambda_{\max}\left(\exp(\theta \mathbf{V}^*\mathbf{Y}\mathbf{V})\right) \\ &\le \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \mathbb{E}\,\mathrm{tr}\exp(\theta \mathbf{V}^*\mathbf{Y}\mathbf{V}).\end{aligned}$$

The first inequality is Jensen's. The second inequality follows because the exponential of a Hermitian matrix is positive definite, so its largest eigenvalue is smaller than its trace.

Combine these observations and take the infimum over all positive θ to complete the argument. □

In most cases it is prohibitively difficult to compute the quantity $\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \mathbf{V}^*\mathbf{Y}\mathbf{V}}$ exactly. The main contribution of [Tro12] is a bound on this quantity, when $\mathbf{V} = \mathbf{I}$, in terms of the cumulant generating functions of the summands. The main tool in the proof is a classical result due to Lieb [Lie73, Thm. 6] that establishes the concavity of the function

$$\mathbf{A} \longmapsto \mathrm{tr}\exp\left(\mathbf{H} + \log(\mathbf{A})\right) \tag{2.4.2}$$

on the positive-definite cone, where $\mathbf{H}$ is Hermitian.

We are interested in the case where $\mathbf{V} \ne \mathbf{I}$ and the matrix $\mathbf{Y}$ in Theorem 2.2 can be expressed as a sum of independent random matrices. In this case, we use the following result to develop the right-hand side of the Laplace transform bound (2.4.1).

Theorem 2.3. *Consider a finite sequence* $\{\mathbf{X}_j\}$ *of independent, random, Hermitian matrices with dimension* n *and a sequence* $\{\mathbf{A}_j\}$ *of fixed Hermitian matrices with dimension* n *that satisfy the relations*

$$\mathbb{E}\mathrm{e}^{\mathbf{X}_j} \preceq \mathrm{e}^{\mathbf{A}_j}. \tag{2.4.3}$$

Let $\mathbf{V} \in \mathbb{V}_k^n$ *be an isometric embedding of* $\mathbb{C}^k$ *into* $\mathbb{C}^n$ *for some* $k \le n$*. Then*

$$\mathbb{E}\,\mathrm{tr}\exp\left\{ \sum\nolimits_j \mathbf{V}^*\mathbf{X}_j\mathbf{V} \right\} \le \mathrm{tr}\exp\left\{ \sum\nolimits_j \mathbf{V}^*\mathbf{A}_j\mathbf{V} \right\}. \tag{2.4.4}$$

In particular,

$$\mathbb{E}\operatorname{tr}\exp\left\{\sum\nolimits_j \mathbf{X}_j\right\} \le \operatorname{tr}\exp\left\{\sum\nolimits_j \mathbf{A}_j\right\}. \tag{2.4.5}$$

Theorem 2.3 is an extension of Lemma 3.4 of [Tro12], which establishes the special case (2.4.5). The proof depends upon a result due to Lieb and Seiringer [LS05, Thm. 3] that extends Lieb's earlier result (2.4.2) by showing that the functional remains concave when the $\log(\mathbf{A})$ term is compressed.

Proposition 2.4 (Lieb–Seiringer 2005)**.** *Let $\mathbf{H}$ be a Hermitian matrix with dimension k. Let $\mathbf{V} \in \mathbb{V}_k^n$ be an isometric embedding of $\mathbb{C}^k$ into $\mathbb{C}^n$ for some $k \le n$. Then the function*

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

is concave on the cone of positive-definite matrices in $\mathbb{M}_{\mathrm{sa}}^n$.

Proof of Theorem 2.3. First, note that (2.4.3) and the operator monotonicity of the matrix logarithm yield the following inequality for each k:

$$\log \mathbb{E}\mathrm{e}^{\mathbf{X}_k} \preceq \mathbf{A}_k. \tag{2.4.6}$$

Let $\mathbb{E}_k$ denote expectation conditioned on the first k summands, $\mathbf{X}_1$ through $\mathbf{X}_k$. Then

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\exp\left(\sum_{j\le\ell}\mathbf{V}^*\mathbf{X}_j\mathbf{V}\right) &= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-1}\operatorname{tr}\exp\left(\sum_{j\le\ell-1}\mathbf{V}^*\mathbf{X}_j\mathbf{V} + \mathbf{V}^*\left(\log \mathrm{e}^{\mathbf{X}_\ell}\right)\mathbf{V}\right)\\
&\le \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left(\sum_{j\le\ell-1}\mathbf{V}^*\mathbf{X}_j\mathbf{V} + \mathbf{V}^*\left(\log \mathbb{E}\mathrm{e}^{\mathbf{X}_\ell}\right)\mathbf{V}\right)\\
&\le \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left(\sum_{j\le\ell-1}\mathbf{V}^*\mathbf{X}_j\mathbf{V} + \mathbf{V}^*\left(\log \mathrm{e}^{\mathbf{A}_\ell}\right)\mathbf{V}\right)\\
&= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left(\sum_{j\le\ell-1}\mathbf{V}^*\mathbf{X}_j\mathbf{V} + \mathbf{V}^*\mathbf{A}_\ell\mathbf{V}\right).
\end{aligned}$$

The first inequality follows from Proposition 2.4 and Jensen's inequality, and the second depends on (2.4.6) and the monotonicity of the trace exponential. Iterate this argument to complete the proof. □

Our main result follows from combining Theorem 2.2 and Theorem 2.3.

Theorem 2.5 (Minimax Laplace Transform)**.** *Consider a finite sequence $\{\mathbf{X}_j\}$ of independent, random, Hermitian matrices with dimension n, and let $k \le n$ be an integer.*

(i) Let $\{\mathbf{A}_j\}$ be a sequence of Hermitian matrices that satisfy the semidefinite relations

$$\mathbb{E}\mathrm{e}^{\theta\mathbf{X}_j} \preceq \mathrm{e}^{g(\theta)\mathbf{A}_j}$$

where $g : (0,\infty) \to [0,\infty)$. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \ge t\right\} \le \inf_{\theta>0}\min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n}\left[\mathrm{e}^{-\theta t}\cdot\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \mathbf{V}^*\mathbf{A}_j\mathbf{V}\right\}\right].$$

(ii) Let $\{\mathbf{A}_j : \mathbb{V}_{n-k+1}^n \to \mathbb{M}_{\mathrm{sa}}^n\}$ *be a sequence of functions that satisfy the semidefinite relations*

$$\mathbb{E}\mathrm{e}^{\theta \mathbf{V}^* \mathbf{X}_j \mathbf{V}} \preceq \mathrm{e}^{g(\theta)\mathbf{A}_j(\mathbf{V})}$$

for all $\mathbf{V} \in \mathbb{V}_{n-k+1}^n$, *where* $g : (0,\infty) \to [0,\infty)$. *Then, for all* $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq t\right\} \leq \inf_{\theta>0} \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \left[\mathrm{e}^{-\theta t} \cdot \operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \mathbf{A}_j(\mathbf{V})\right\}\right].$$

The first bound in Theorem 2.5 requires less detailed information on how compression affects the summands but correspondingly does not give as sharp results as the second. For most cases we consider, we use the second inequality because it is straightforward to obtain semidefinite bounds for the compressed summands. The exception occurs in the proof of the subexponential Bernstein inequality (Theorem 2.12 in Section 2.6); here we use the first bound, because in this case there are no nontrivial semidefinite bounds for the compressed summands.

In the following two sections, we use the minimax Laplace transform method to derive Chernoff and Bernstein inequalities for the interior eigenvalues of a sum of independent random matrices. Tail bounds for the eigenvalues of matrix Rademacher and Gaussian series, eigenvalue Hoeffding, and matrix martingale eigenvalue tail bounds can all be derived in a similar manner; see [Tro12] for the details of the arguments leading to such tail bounds for the maximum eigenvalue.

2.5 Chernoff bounds

Classical Chernoff bounds establish that the tails of a sum of independent nonnegative random variables decay subexponentially. [Tro12] develops Chernoff bounds for the maximum and minimum eigenvalues of a sum of independent positive semidefinite matrices. We extend this analysis to study the interior eigenvalues.

Intuitively, the eigenvalue tail bounds should depend on how concentrated the summands are; e.g., the maximum eigenvalue of a sum of operators whose ranges are aligned is likely to vary more than that of a sum of operators whose ranges are orthogonal. To measure how much a finite sequence of random summands $\{\mathbf{X}_j\}$ concentrates in a given subspace, we define a function $\Psi : \bigcup_{1\leq k\leq n} \mathbb{V}_k^n \to \mathbb{R}$ that satisfies

$$\max\nolimits_j \lambda_{\max}\left(\mathbf{V}^* \mathbf{X}_j \mathbf{V}\right) \leq \Psi(\mathbf{V}) \qquad \text{almost surely for each } \mathbf{V} \in \bigcup_{1\leq k\leq n} \mathbb{V}_k^n. \tag{2.5.1}$$

The sequence $\{\mathbf{X}_j\}$ associated with Ψ will always be clear from context. We have the following result.

Theorem 2.6 (Eigenvalue Chernoff Bounds)**.** *Consider a finite sequence* $\{\mathbf{X}_j\}$ *of independent, random, positive-semidefinite matrices with dimension* n. *Given an integer* $k \leq n$, *define*

$$\mu_k = \lambda_k\left(\sum\nolimits_j \mathbb{E}\mathbf{X}_j\right),$$

and let $\mathbf{V}_+ \in \mathbb{V}_{n-k+1}^n$ *and* $\mathbf{V}_- \in \mathbb{V}_k^n$ *be isometric embeddings that satisfy*

$$\mu_k = \lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right) = \lambda_{\min}\left(\sum\nolimits_j \mathbf{V}_-^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\right).$$

Then

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq (1+\delta)\mu_k\right\} \leq (n-k+1)\cdot\left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_k/\Psi(\mathbf{V}_+)} \qquad \text{for } \delta > 0, \text{ and}$$

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \leq (1-\delta)\mu_k\right\} \leq k\cdot\left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_k/\Psi(\mathbf{V}_-)} \qquad \text{for } \delta \in [0,1),$$

where Ψ *is a function that satisfies* (2.5.1).

Theorem 2.6 tells us how the tails of the kth eigenvalue are controlled by the variation of the random summands in the top and bottom invariant subspaces of $\sum_j \mathbb{E}\mathbf{X}_j$. Up to the dimensional factors k and $n-k+1$, the eigenvalues exhibit binomial-type tails. When $k=1$ (respectively, $k=n$) Theorem 2.6 controls the probability that the largest eigenvalue of the sum is small (respectively, the probability that the smallest eigenvalue of the sum is large), thereby complementing the one-sided Chernoff bounds of [Tro12].

Remark 2.7. The results in Theorem 2.6 have the following standard simplifications:

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq t\mu_k\right\} \leq (n-k+1)\cdot\left[\frac{\mathrm{e}}{t}\right]^{t\mu_k/\Psi(\mathbf{V}_+)} \qquad \text{for } t \geq \mathrm{e}, \text{ and}$$

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \leq t\mu_k\right\} \leq k\cdot\mathrm{e}^{-(1-t)^2\mu_k/(2\Psi(\mathbf{V}_-))} \qquad \text{for } t \in [0,1].$$

Remark 2.8. If it is difficult to estimate $\Psi(\mathbf{V}_+)$ or $\Psi(\mathbf{V}_-)$ and the summands are uniformly bounded, one can resort to the weaker estimates

$$\Psi(\mathbf{V}_+) \leq \max_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n} \max\nolimits_j \left\|\mathbf{V}^*\mathbf{X}_j\mathbf{V}\right\| = \max\nolimits_j \left\|\mathbf{X}_j\right\| \text{ and}$$

$$\Psi(\mathbf{V}_-) \leq \max_{\mathbf{V}\in\mathbb{V}_k^n} \max\nolimits_j \left\|\mathbf{V}^*\mathbf{X}_j\mathbf{V}\right\| = \max\nolimits_j \left\|\mathbf{X}_j\right\|.$$

Theorem 2.6 follows from Theorem 2.5 using an appropriate bound on the matrix moment-generating functions. The following lemma is due to Ahlswede and Winter [AW02]; see also [Tro12, Lem. 5.8].

Lemma 2.9. *Suppose that* $\mathbf{X}$ *is a random positive-semidefinite matrix that satisfies* $\lambda_{\max}(\mathbf{X}) \leq 1$. *Then*

$$\mathbb{E}\mathrm{e}^{\theta\mathbf{X}} \preceq \exp\left((\mathrm{e}^{\theta}-1)(\mathbb{E}\mathbf{X})\right) \qquad \text{for } \theta \in \mathbb{R}.$$

Proof of Theorem 2.6, upper bound. We consider the case where $\Psi(\mathbf{V}_+) = 1$; the general case follows by homogeneity. Define

$$\mathbf{A}_j(\mathbf{V}_+) = \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+ \quad \text{and} \quad g(\theta) = \mathrm{e}^{\theta} - 1.$$

Theorem 2.5(ii) and Lemma 2.9 imply that

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq (1+\delta)\mu_k\right\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta(1+\delta)\mu_k}\cdot\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right\}.$$

Bound the trace by the maximum eigenvalue, taking into account the reduced dimension of the summands:

$$\begin{aligned}\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right\} &\le (n-k+1)\cdot\lambda_{\max}\left(\exp\left\{g(\theta)\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right\}\right)\\ &= (n-k+1)\cdot\exp\left\{g(\theta)\cdot\lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right)\right\}.\end{aligned}$$

The equality follows from the spectral mapping theorem. Identify the quantity μ_k; then combine the last two inequalities to obtain

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \ge (1+\delta)\mu_k\right\} \le (n-k+1)\cdot\inf_{\theta>0} \mathrm{e}^{[g(\theta)-\theta(1+\delta)]\mu_k}.$$

The right-hand side is minimized when $\theta = \log(1+\delta)$, which gives the desired upper tail bound. □

Proof of Theorem 2.6, lower bound. As before, we consider the case where $\Psi(\mathbf{V}_-) = 1$. Clearly,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \le (1-\delta)\mu_k\right\} = \mathbb{P}\left\{\lambda_{n-k+1}\left(\sum\nolimits_j -\mathbf{X}_j\right) \ge -(1-\delta)\mu_k\right\}. \tag{2.5.2}$$

Apply Lemma 2.9 to see that, for $\theta > 0$,

$$\mathbb{E}\mathrm{e}^{\theta(-\mathbf{V}_-^*\mathbf{X}_j\mathbf{V}_-)} = \mathbb{E}\mathrm{e}^{(-\theta)\mathbf{V}_-^*\mathbf{X}_j\mathbf{V}_-} \preceq \exp\bigl(g(\theta)\cdot\mathbf{V}_-^*(-\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\bigr),$$

where $g(\theta) = 1-\mathrm{e}^{-\theta}$. Theorem 2.5(ii) thus implies that the latter probability in (2.5.2) is bounded by

$$\inf_{\theta>0} \mathrm{e}^{\theta(1-\delta)\mu_k}\cdot\operatorname{tr}\exp\left(g(\theta)\sum\nolimits_j \mathbf{V}_-^*(-\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\right).$$

Using reasoning analogous to that in the proof of the upper bound, we justify the first of the following inequalities:

$$\begin{aligned}\operatorname{tr}\exp\left(g(\theta)\sum\nolimits_j \mathbf{V}_-^*(-\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\right) &\le k\cdot\exp\left\{\lambda_{\max}\left(g(\theta)\sum\nolimits_j \mathbf{V}_-^*(-\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\right)\right\}\\ &= k\cdot\exp\left\{-g(\theta)\cdot\lambda_{\min}\left(\sum\nolimits_j \mathbf{V}_-^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_-\right)\right\}\\ &= k\cdot\exp\left\{-g(\theta)\mu_k\right\}.\end{aligned}$$

The remaining equalities follow from the fact that $-g(\theta) < 0$ and the definition of μ_k.

This argument establishes the bound

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \le (1-\delta)\mu_k\right\} \le k\cdot\inf_{\theta>0} \mathrm{e}^{[\theta(1-\delta)-g(\theta)]\mu_k}.$$

The right-hand side is minimized when $\theta = -\log(1-\delta)$, which gives the desired lower tail bound. □

2.6 Bennett and Bernstein inequalities

The classical Bennett and Bernstein inequalities use the variance or knowledge of the moments of the summands to control the probability that a sum of independent random variables deviates from its mean. In [Tro12], matrix Bennett and Bernstein inequalities are developed for the extreme eigenvalues of Hermitian random matrix sums. We establish that the interior eigenvalues satisfy analogous inequalities.

As in the derivation of the Chernoff inequalities of Section 2.5, we need a measure of how concentrated the random summands are in a given subspace. Recall that the function $\Psi : \bigcup_{1\le k\le n} \mathbb{V}_k^n \to \mathbb{R}$ satisfies

$$\max_j \lambda_{\max}\left(\mathbf{V}^*\mathbf{X}_j\mathbf{V}\right) \le \Psi(\mathbf{V}) \qquad \text{almost surely for each } \mathbf{V} \in \bigcup_{1\le k\le n} \mathbb{V}_k^n. \tag{2.6.1}$$

The sequence $\{\mathbf{X}_j\}$ associated with Ψ will always be clear from context.

Theorem 2.10 (Eigenvalue Bennett Inequality)**.** *Consider a finite sequence $\{\mathbf{X}_j\}$ of independent, random, Hermitian matrices with dimension n, all of which have zero mean. Given an integer $k \le n$, define*

$$\sigma_k^2 = \lambda_k\left(\sum\nolimits_j \mathbb{E}(\mathbf{X}_j^2)\right).$$

Choose $\mathbf{V}_+ \in \mathbb{V}_{n-k+1}^n$ to satisfy

$$\sigma_k^2 = \lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*\mathbb{E}(\mathbf{X}_j^2)\mathbf{V}_+\right).$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \ge t\right\} \le (n-k+1)\cdot\exp\left\{-\frac{\sigma_k^2}{\Psi(\mathbf{V}_+)^2}\cdot h\left(\frac{\Psi(\mathbf{V}_+)t}{\sigma_k^2}\right)\right\} \tag{i}$$

$$\le (n-k+1)\cdot\exp\left\{\frac{-t^2/2}{\sigma_k^2 + \Psi(\mathbf{V}_+)t/3}\right\} \tag{ii}$$

$$\le \begin{cases} (n-k+1)\cdot\exp\left\{-\frac{3}{8}t^2/\sigma_k^2\right\} & \text{for } t \le \sigma_k^2/\Psi(\mathbf{V}_+) \\ (n-k+1)\cdot\exp\left\{-\frac{3}{8}t/\Psi(\mathbf{V}_+)\right\} & \text{for } t \ge \sigma_k^2/\Psi(\mathbf{V}_+), \end{cases} \tag{iii}$$

where the function $h(u) = (1+u)\log(1+u) - u$ for $u \ge 0$. The function Ψ satisfies (2.6.1) above.

Results (i) and (ii) are, respectively, matrix analogs of the classical Bennett and Bernstein inequalities. As in the scalar case, the Bennett inequality reflects a Poisson-type decay in the tails of the eigenvalues. The Bernstein inequality states that small deviations from the eigenvalues of the expected matrix are roughly normally distributed while larger deviations are subexponential. The split Bernstein inequalities (iii) make explicit the division between these two regimes.

As stated, Theorem 2.10 controls the probability that the eigenvalues of a sum are large. Using the identity

$$\lambda_k\left(-\sum_j \mathbf{X}_j\right) = -\lambda_{n-k+1}\left(\sum_j \mathbf{X}_j\right),$$

Theorem 2.10 can also be applied to control the probability that eigenvalues of a sum are small.

To prove Theorem 2.10, we use the following lemma (Lemma 6.7 in [Tro12]) to control the moment-generating function of a random matrix with bounded maximum eigenvalue.

Lemma 2.11. *Let* $\mathbf{X}$ *be a random Hermitian matrix satisfying* $\mathbb{E}\mathbf{X} = \mathbf{0}$ *and* $\lambda_{\max}(\mathbf{X}) \le 1$ *almost surely. Then*

$$\mathbb{E}e^{\theta \mathbf{X}} \preceq \exp((e^{\theta} - \theta - 1) \cdot \mathbb{E}(\mathbf{X}^2)) \quad \text{for } \theta > 0.$$

Proof of Theorem 2.10. Using homogeneity, we assume without loss that $\Psi(\mathbf{V}_+) = 1$. This implies that $\lambda_{\max}\left(\mathbf{X}_j\right) \le 1$ almost surely for all the summands. By Lemma 2.11,

$$\mathbb{E}e^{\theta \mathbf{X}_j} \preceq \exp\big(g(\theta) \cdot \mathbb{E}(\mathbf{X}_j^2)\big),$$

with $g(\theta) = e^{\theta} - \theta - 1$.

Theorem 2.5(i) then implies

$$\begin{aligned}
\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \ge t\right\} &\le \inf_{\theta>0} e^{-\theta t} \cdot \operatorname{tr}\exp\left(g(\theta)\sum\nolimits_j \mathbf{V}_+^* \mathbb{E}(\mathbf{X}_j^2)\mathbf{V}_+\right) \\
&\le (n-k+1) \cdot \inf_{\theta>0} e^{-\theta t} \cdot \lambda_{\max}\left(\exp\left\{g(\theta)\sum\nolimits_j \mathbf{V}_+^* \mathbb{E}(\mathbf{X}_j^2)\mathbf{V}_+\right\}\right) \\
&= (n-k+1) \cdot \inf_{\theta>0} e^{-\theta t} \cdot \exp\left\{g(\theta) \cdot \lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^* \mathbb{E}(\mathbf{X}_j^2)\mathbf{V}_+\right)\right\}.
\end{aligned}$$

The maximum eigenvalue in this expression equals σ_k^2, thus

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \ge t\right\} \le (n-k+1) \cdot \inf_{\theta>0} e^{g(\theta)\sigma_k^2 - \theta t}.$$

The Bennett inequality (i) follows by substituting $\theta = \log(1 + t/\sigma_k^2)$ into the right-hand side and simplifying.

The Bernstein inequality (ii) is a consequence of (i) and the fact that

$$h(u) \ge \frac{u^2/2}{1 + u/3} \quad \text{for } u \ge 0,$$

which can be established by comparing derivatives.

The subgaussian and subexponential portions of the split Bernstein inequalities (iii) are verified through algebraic comparisons on the relevant intervals. □

Occasionally, as in the application in Section 2.8 to the problem of covariance matrix estimation, one desires a Bernstein-type tail bound that applies to summands that do not have bounded maximum eigenvalues. In this case, if the moments of the summands satisfy sufficiently strong growth restrictions, one can extend classical scalar arguments to obtain results such as the following Bernstein bound for subexponential matrices.

Theorem 2.12 (Eigenvalue Bernstein Inequality for Subexponential Matrices)**.** *Consider a finite sequence* $\{\mathbf{X}_j\}$ *of independent, random, Hermitian matrices with dimension* n*, all of which satisfy the subexponential moment growth condition*

$$\mathbb{E}(\mathbf{X}_j^m) \preceq \frac{m!}{2} B^{m-2} \mathbf{\Sigma}_j^2 \quad \text{for } m = 2, 3, 4, \ldots,$$

where B *is a positive constant and* $\mathbf{\Sigma}_j^2$ *are positive-semidefinite matrices. Given an integer* $k \le n$*, set*

$$\mu_k = \lambda_k\left(\sum\nolimits_j \mathbb{E}\mathbf{X}_j\right).$$

Choose $\mathbf{V}_+ \in \mathbb{V}_{n-k+1}^n$ *that satisfies*

$$\mu_k = \lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right),$$

and define

$$\sigma_k^2 = \lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*\boldsymbol{\Sigma}_j^2\mathbf{V}_+\right).$$

Then, for any $t \geq 0$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq \mu_k + t\right\} \leq (n-k+1)\cdot\exp\left\{-\frac{t^2/2}{\sigma_k^2 + Bt}\right\} \tag{i}$$

$$\leq \begin{cases} (n-k+1)\cdot\exp\left\{-\frac{1}{4}t^2/\sigma_k^2\right\} & \text{for } t \leq \sigma_k^2/B \\ (n-k+1)\cdot\exp\left\{-\frac{1}{4}t/B\right\} & \text{for } t \geq \sigma_k^2/B. \end{cases} \tag{ii}$$

This result is an extension of [Tro12, Theorem 6.2], which, in turn, generalizes a classical scalar argument [DG98].

As with the other matrix inequalities, Theorem 2.12 follows from an application of Theorem 2.5 and appropriate semidefinite bounds on the moment-generating functions of the summands. Thus, the key to the proof lies in exploiting the moment growth conditions of the summands to majorize their moment-generating functions. The following lemma, a trivial extension of Lemma 6.8 in [Tro12], provides what we need.

Lemma 2.13. *Let* $\mathbf{X}$ *be a random Hermitian matrix satisfying the subexponential moment growth conditions*

$$\mathbb{E}(\mathbf{X}^m) \preceq \frac{m!}{2}\boldsymbol{\Sigma}^2 \quad \text{for } m = 2, 3, 4, \ldots.$$

Then, for any θ *in* $[0, 1)$,

$$\mathbb{E}\exp(\theta\mathbf{X}) \preceq \exp\left(\theta\mathbb{E}\mathbf{X} + \frac{\theta^2}{2(1-\theta)}\boldsymbol{\Sigma}^2\right).$$

Proof of Theorem 2.12. We note that $\mathbf{X}_j$ satisfies the growth condition

$$\mathbb{E}(\mathbf{X}_j^m) \preceq \frac{m!}{2}B^{m-2}\boldsymbol{\Sigma}_j^2 \quad \text{for } m \geq 2$$

if and only if the scaled matrix $\mathbf{X}_j/B$ satisfies

$$\mathbb{E}\left(\frac{\mathbf{X}_j}{B}\right)^m \preceq \frac{m!}{2}\cdot\frac{\boldsymbol{\Sigma}_j^2}{B^2} \quad \text{for } m \geq 2.$$

Thus, by rescaling, it suffices to consider the case $B = 1$.

By Lemma 2.13, the moment-generating functions of the summands satisfy

$$\mathbb{E}\exp(\theta\mathbf{X}_j) \preceq \exp\left(\theta\mathbb{E}\mathbf{X}_j + g(\theta)\boldsymbol{\Sigma}_j^2\right),$$

where $g(\theta) = \theta^2/(2-2\theta)$. Now we apply Theorem 2.5(i):

$$\begin{aligned}
\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \geq \mu_k + t\right\} &\leq \inf_{\theta\in[0,1)} \mathrm{e}^{-\theta(\mu_k+t)} \cdot \operatorname{tr}\exp\left(\theta\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+ + g(\theta)\sum\nolimits_j \mathbf{V}_+^*\boldsymbol{\Sigma}_j^2\mathbf{V}_+\right) \\
&\leq \inf_{\theta\in[0,1)} (n-k+1)\cdot\exp\Big\{-\theta(\mu_k+t) + \theta\cdot\lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*(\mathbb{E}\mathbf{X}_j)\mathbf{V}_+\right) \\
&\qquad + g(\theta)\cdot\lambda_{\max}\left(\sum\nolimits_j \mathbf{V}_+^*\boldsymbol{\Sigma}_j^2\mathbf{V}_+\right)\Big\} \\
&= \inf_{\theta\in[0,1)} (n-k+1)\cdot\exp\left(-\theta t + g(\theta)\sigma_k^2\right).
\end{aligned}$$

To achieve the final simplification, we identified μ_k and σ_k^2. Now, select $\theta = t/(t+\sigma_k^2)$. Then simplication gives the Bernstein inequality (i).

Algebraic comparisons on the relevant intervals yield the split Bernstein inequalities (ii). □

2.7 An application to column subsampling

As an application of our Chernoff bounds, we examine how sampling columns from a matrix with orthonormal rows affects the spectrum. This question has applications in numerical linear algebra and compressed sensing. The special cases of the maximum and minimum eigenvalues have been studied in the literature [Tro08, RV07]. The limiting spectral distributions of matrices formed by sampling columns from similarly structured matrices have also been studied: the results of [GH08] apply to matrices formed by sampling columns from any fixed orthogonal matrix, and [Far10] studies matrices formed by sampling columns and rows from the discrete Fourier transform matrix.

Let $\mathbf{U}$ be an $n \times r$ matrix with orthonormal rows. We model the sampling operation using a random diagonal matrix $\mathbf{D}$ whose entries are independent Bern(p) random variables. Then the random matrix

$$\widehat{\mathbf{U}} = \mathbf{U}\mathbf{D} \tag{2.7.1}$$

can be interpreted as a random column submatrix of $\mathbf{U}$ with an average of pr nonzero columns. Our goal is to study the behavior of the spectrum of $\widehat{\mathbf{U}}$.

Recall that the decay of the Chernoff tail bounds is influenced by the variation of the random summands when compressed to invariant subspaces of the expected sum, as measured by $\Psi(\mathbf{V})$. In this application, the choice of invariant subspace is arbitrary, so we choose that which gives the smallest variations and hence the fastest decay. This gives rise to a coherence-like quantity associated with the matrix $\mathbf{U}$: Recall that the jth column of $\mathbf{U}$ is written $\mathbf{u}_j$. Consider the following coherence-like quantity associated with $\mathbf{U}$:

$$\tau_k = \min_{\mathbf{V}\in\mathbb{V}_k^n} \max\nolimits_j \left\|\mathbf{V}^*\mathbf{u}_j\right\|^2 \quad \text{for } k = 1,\dots,n. \tag{2.7.2}$$

There does not seem to be a simple expression for τ_k. However, by choosing $\mathbf{V}^*$ to be the restriction to an appropriate k-dimensional coordinate subspace, we see that τ_k always satisfies

$$\tau_k \leq \min_{|I|\leq k} \max\nolimits_j \sum_{i\in I} u_{ij}^2.$$

The following theorem shows that the behavior of $\sigma_k(\widehat{\mathbf{U}})$, the kth singular value of $\widehat{\mathbf{U}}$, can be explained in terms of τ_k.

Theorem 2.14 (Column Subsampling of Matrices with Orthonormal Rows)**.** *Let* $\mathbf{U}$ *be an* $n \times r$ *matrix with orthonormal rows, and let* p *be a sampling probability. Define the sampled matrix* $\widehat{\mathbf{U}}$ *according to* (2.7.1)*, and the numbers* $\{\tau_k\}$ *according to* (2.7.2)*. Then, for each* $k = 1, \dots, n$*,*

$$\mathbb{P}\left\{\sigma_k(\widehat{\mathbf{U}}) \geq \sqrt{(1+\delta)p}\right\} \leq (n-k+1) \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{p/\tau_{n-k+1}} \qquad \text{for } \delta > 0, \text{ and}$$

$$\mathbb{P}\left\{\sigma_k(\widehat{\mathbf{U}}) \leq \sqrt{(1-\delta)p}\right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{p/\tau_k} \qquad \text{for } \delta \in [0,1).$$

Proof. Observe, using (2.7.1), that

$$\sigma_k(\widehat{\mathbf{U}})^2 = \lambda_k(\mathbf{U}\mathbf{D}^2\mathbf{U}^*) = \lambda_k\left(\sum_j d_j \mathbf{u}_j \mathbf{u}_j^*\right),$$

where $\mathbf{u}_j$ is the jth column of $\mathbf{U}$ and $d_j \sim \mathrm{Bern}(p)$. Compute

$$\mu_k = \lambda_k\left(\sum\nolimits_j \mathbb{E} d_j \mathbf{u}_j \mathbf{u}_j^*\right) = p \cdot \lambda_k(\mathbf{U}\mathbf{U}^*) = p \cdot \lambda_k(\mathbf{I}) = p.$$

It follows that, for *any* $\mathbf{V} \in \mathbb{V}^n_{n-k+1}$,

$$\lambda_{\max}\left(\sum\nolimits_j \mathbf{V}^*(\mathbb{E} d_j \mathbf{u}_j \mathbf{u}_j^*)\mathbf{V}\right) = p \cdot \lambda_{\max}\left(\mathbf{V}^*\mathbf{V}\right) = p = \mu_k,$$

so the choice of $\mathbf{V}_+ \in \mathbb{V}^n_{n-k+1}$ is arbitrary. Similarly, the choice of $\mathbf{V}_- \in \mathbb{V}^n_k$ is arbitrary. We select $\mathbf{V}_+$ to be an isometric embedding that achieves τ_{n-k+1} and $\mathbf{V}_-$ to be an isometric embedding that achieves τ_k. Accordingly,

$$\begin{aligned}
\Psi(\mathbf{V}_+) &= \max\nolimits_j \|\mathbf{V}_+^* \mathbf{u}_j \mathbf{u}_j^* \mathbf{V}_+\| = \max\nolimits_j \|\mathbf{V}_+^* \mathbf{u}_j\|^2 = \tau_{n-k+1}, \quad \text{and} \\
\Psi(\mathbf{V}_-) &= \max\nolimits_j \|\mathbf{V}_-^* \mathbf{u}_j \mathbf{u}_j^* \mathbf{V}_-\| = \max\nolimits_j \|\mathbf{V}_-^* \mathbf{u}_j\|^2 = \tau_k.
\end{aligned}$$

Theorem 2.6 delivers the upper bound

$$\begin{aligned}
\mathbb{P}\left\{\sigma_k(\widehat{\mathbf{U}}) \geq \sqrt{(1+\delta)p}\right\} &= \mathbb{P}\left\{\lambda_k\left(\sum_j d_j \mathbf{u}_j \mathbf{u}_j^*\right) \geq (1+\delta)p\right\} \\
&\leq (n-k+1) \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{p/\tau_{n-k+1}}
\end{aligned}$$

for $\delta > 0$, and the lower bound

$$\mathbb{P}\left\{\sigma_k(\widehat{\mathbf{U}}) \leq \sqrt{(1-\delta)p}\right\} = \mathbb{P}\left\{\lambda_k\left(\sum_j d_j \mathbf{u}_j \mathbf{u}_j^*\right) \leq (1-\delta)p\right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{p/\tau_k}$$

for $\delta \in [0,1)$. □

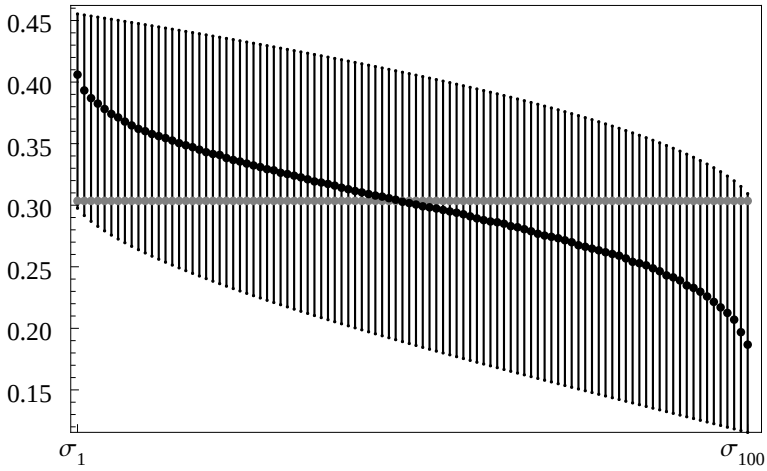


Figure 2.1: SPECTRUM OF A RANDOM SUBMATRIX OF A UNITARY DFT MATRIX. The matrix $\mathbf{U}$ is a $10^2 \times 10^4$ submatrix of the unitary DFT matrix with dimension 10^4, and the sampling probability $p = 10^{-4} \log(10^4)$. The kth vertical bar, calculated using Theorem 2.14, describes an interval containing the median value of the kth singular value of the sampled matrix $\widehat{\mathbf{U}}$. The black circles denote the empirical medians of the singular values of $\widehat{\mathbf{U}}$, calculated from 500 trials. The gray circles represent the singular values of $\mathbb{E}\widehat{\mathbf{U}}$.

To illustrate the discriminatory power of these bounds, let $\mathbf{U}$ be an $n \times n^2$ matrix consisting of n rows of the $n^2 \times n^2$ Fourier matrix and choose $p = (\log n)/n$ so that, on average, sampling reduces the aspect ratio from n to $\log n$. For $n = 100$, we determine upper and lower bounds for the median value of $\sigma_k(\widehat{\mathbf{U}})$ by numerically finding the value of δ where the probability bounds in Theorem 2.14 equal one-half. Figure 2.1 plots the empirical median value along with the computed interval. We see that these ranges reflect the behavior of the singular values more faithfully than the simple estimates $\sigma_k(\mathbb{E}\widehat{\mathbf{U}}) = p$.

2.8 Covariance estimation

We conclude with an extended example that illustrates how this circle of ideas allows one to answer interesting statistical questions. Specifically, we investigate the convergence of the individual eigenvalues of sample covariance matrices. Our results establish conditions under which the eigenvalues can be recovered to relative precision, and furthermore reflect the difference in the probabilities of the kth eigenvalue of the sample covariance matrix over- or underestimating that of the covariance matrix.

Covariance estimation is a basic and ubiquitious problem that arises in signal processing, graphical modeling, machine learning, and genomics, among other areas. Let $\{\eta_j\}_{j=1}^n \subset \mathbb{R}^p$ be i.i.d. samples drawn from some distribution with zero mean and covariance matrix $\mathbf{C}$. Define the sample covariance matrix

$$\widehat{\mathbf{C}}_n = \frac{1}{n} \sum_{j=1}^{n} \eta_j \eta_j^*.$$

An important challenge is to determine how many samples are needed to ensure that the empirical covariance estimator has a fixed relative accuracy in the spectral norm. That is, given

a fixed ε, how large must n be so that

$$\left\|\widehat{\mathbf{C}}_n - \mathbf{C}\right\|_2 \leq \varepsilon \|\mathbf{C}\|_2? \tag{2.8.1}$$

This estimation problem has been studied extensively. It is now known that for distributions with a finite second moment, $\Omega(p \log p)$ samples suffice [Rud99], and for log-concave distributions, $\Omega(p)$ samples suffice [ALPTJ11]. More broadly, Vershynin [Ver11b] conjectures that, for distributions with finite fourth moment, $\Omega(p)$ samples suffice; he establishes this result to within iterated log factors. In [SV], Srivastava and Vershynin establish that $\Omega(p)$ samples suffice for distributions which have finite $2+\varepsilon$ moments, for some $\varepsilon > 0$, and satisfy an additional regularity condition.

Inequality (2.8.1) ensures that the difference between the kth eigenvalues of $\widehat{\mathbf{C}}_n$ and $\mathbf{C}$ is small, but it requires $\mathrm{O}(p)$ samples to obtain estimates of even a few of the eigenvalues. Specifically, letting $\kappa_\ell = \lambda_1(\mathbf{C})/\lambda_\ell(\mathbf{C})$, we see that $\mathrm{O}(\varepsilon^{-2}\kappa_\ell^2 p)$ samples are required to obtain relative error estimates of the largest ℓ eigenvalues of $\mathbf{C}$ using the results of [ALPTJ11, Ver11b, SV]. However, it is reasonable to expect that when the spectrum of $\mathbf{C}$ exhibits decay and $\ell \ll p$, far fewer than $\mathrm{O}(p)$ samples should suffice to ensure relative error recovery of the largest ℓ eigenvalues.

In fact, Vershynin shows this is the case when the random vector is subgaussian: in [Ver11a], he defines the effective rank of $\mathbf{C}$ to be $r = \big(\sum_{i=1}^p \lambda_i(\mathbf{C})\big)/\lambda_1(\mathbf{C})$ and uses r to provide bounds of the form (2.8.1). It follow from his arguments that, with high probability, the largest ℓ eigenvalues of $\mathbf{C}$ are estimated to relative precision when $n = \mathrm{O}(\varepsilon^{-2} r \kappa_\ell^2 \log p)$ samples are taken. Clearly this result is most of interest when the effective rank is small: e.g. when r is $\mathrm{O}(1)$, we see that $\mathrm{O}(\varepsilon^{-2}\kappa_\ell^2 \log p)$ samples suffice to give relative error accuracy in the largest ℓ eigenvalues of $\mathbf{C}$. Note, however, that this result does not supply the rates of convergence of the *individual* eigenvalues, and it requires the effective rank to be small. To the best of the author's knowledge, there are no nonasymptotic estimates of the relative errors of individual eigenvalues that do not require the assumption that $\mathbf{C}$ has low effective rank.

In this section, we derive a relative approximation bound for each eigenvalue of $\mathbf{C}$. For simplicity, we assume the samples are drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{C})$ distribution where $\mathbf{C}$ is full-rank, but we expect that the arguments can be extended to cover other subgaussian distributions.

Theorem 2.15. *Assume that $\mathbf{C} \in \mathbb{M}_{\mathrm{sa}}^p$ is positive definite. Let $\{\eta_j\}_{j=1}^n \subset \mathbb{R}^p$ be i.i.d. samples drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{C})$ distribution. Define*

$$\widehat{\mathbf{C}}_n = \frac{1}{n} \sum\nolimits_{j=1}^n \eta_j \eta_j^*.$$

Write λ_k for the kth eigenvalue of $\mathbf{C}$, and write $\hat{\lambda}_k$ for the kth eigenvalue of $\widehat{\mathbf{C}}_n$. Then for $k = 1, \ldots, p$,

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq (p-k+1) \cdot \exp\left(\frac{-nt^2}{32\lambda_k \sum_{i=k}^p \lambda_i}\right) \quad \text{for } t \leq 4n\lambda_k,$$

and

$$\mathbb{P}\left\{\hat{\lambda}_k \leq \lambda_k - t\right\} \leq k \cdot \exp\left(\frac{-3nt^2}{8\lambda_1\big(\lambda_1 + \sum_{i=1}^k \lambda_i\big)}\right) \quad \text{for } t \leq n\big(\lambda_1 + \sum_{i=1}^k \lambda_i\big).$$

The following corollary provides an answer to our question about relative error estimates.

Corollary 2.16. *Let λ_k and $\hat{\lambda}_k$ be as in Theorem 2.15. Then*

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq (p-k+1)\cdot \exp\left(\frac{-cn\varepsilon^2}{\sum_{i=k}^{p}\frac{\lambda_i}{\lambda_k}}\right) \quad \textit{for } \varepsilon \leq 4n,$$

and

$$\mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\} \leq k\cdot \exp\left(\frac{-cn\varepsilon^2}{\frac{\lambda_1}{\lambda_k}\left(\sum_{i=1}^{k}\frac{\lambda_i}{\lambda_k}\right)}\right) \quad \textit{for } \varepsilon \in (0,1],$$

where the constant c is at least $1/32$.

The first bound in Corollary 2.16 tells us how many samples are needed to ensure that $\hat{\lambda}_k$ does not overestimate λ_k. Likewise, the second bound tells us how many samples ensure that $\hat{\lambda}_k$ does not underestimate λ_k.

Corollary 2.16 suggests that the relationship of $\hat{\lambda}_k$ to λ_k is determined by the spectrum of $\mathbf{C}$ in the following manner. When the eigenvalues below λ_k are small compared with λ_k, the quantity

$$\sum_{i=k}^{p}\lambda_i/\lambda_k$$

is small (viz., it is no larger than $p-k+1$), and so $\hat{\lambda}_k$ is not likely to overestimate λ_k. Similarly, when the eigenvalues above λ_k are comparable with λ_k, the quantity

$$\frac{\lambda_1}{\lambda_k}\left(\sum_{i=1}^{k}\lambda_i/\lambda_k\right)$$

is small (viz., it is no larger than $k\cdot\kappa_k^2$), and so $\hat{\lambda}_k$ is not likely to underestimate λ_k.

Remark 2.17. The results in Theorem 2.15 and Corollary 2.16 also apply when $\mathbf{C}$ is rank-deficient: simply replace each occurence of the dimension p in the bounds with $\operatorname{rank}(\mathbf{C})$.

Indeed, assume that $\mathbf{C}$ is rank-deficient and take its truncated eigenvalue decomposition to be $\mathbf{C}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$. If $\eta_j \sim \mathcal{N}(\mathbf{0},\mathbf{C})$, then η_j lies in the span of $\mathbf{C}$. It follows that $\hat{\lambda}_k=\lambda_k=0$ for all $k>\operatorname{rank}(\mathbf{C})$. When $k\leq\operatorname{rank}(\mathbf{C})$, we observe that

$$\lambda_k(\mathbf{C})=\lambda_k(\mathbf{\Lambda}) \quad \text{and} \quad \lambda_k\left(\sum_j \eta_j\eta_j^*\right)=\lambda_k\left(\sum_j \xi_j\xi_j^*\right),$$

where $\xi_j=\mathbf{U}^*\eta_j$ is distributed $\mathcal{N}(\mathbf{0},\mathbf{\Lambda})$. Thus,

$$\left|\lambda_k\left(\sum_j \eta_j\eta_j^*\right)-\lambda_k(\mathbf{C})\right| = \left|\lambda_k\left(\sum_j \xi_j\xi_j^*\right)-\lambda_k(\mathbf{\Lambda})\right|.$$

Consequently, the problem of estimating the eigenvalues of $\mathbf{C}$ to relative error using the samples $\{\eta_j\}$ is equivalent to that of estimating the eigenvalues of the full-rank covariance matrix $\mathbf{\Lambda}$ to relative error using the samples $\{\xi_j\}$.

It is reasonable to expect that one should be able to use Corollary 2.16 to recover Vershynin's result in [Ver11a] for Wishart matrices: that $\Omega(\varepsilon^{-2} r \kappa_\ell^2 \log p)$ samples suffice to estimate the eigenvalues of the covariance matrix of a Gaussian random variable to within a relative precision of $1 \pm \varepsilon$. Indeed, this result follows from Corollary 2.16 and a simple union bound argument.

Corollary 2.18. *Assume* $\mathbf{C}$ *is positive semidefinite. Let* $\{\eta_j\}_{j=1}^n \subset \mathbb{R}^p$ *be i.i.d. samples drawn from a* $\mathcal{N}(\mathbf{0}, \mathbf{C})$ *distribution. If* $n = \Omega(\varepsilon^{-2} r \kappa_\ell^2 \log p)$*, then with high probability*

$$|\lambda_k(\widehat{\mathbf{C}}_n) - \lambda_k(\mathbf{C})| \le \varepsilon \lambda_k(\mathbf{C}) \quad \text{for } k = 1, \dots, \ell.$$

Proof. From Corollary 2.16, we see that

$$\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}_n) \le (1-\varepsilon)\lambda_k\right\} \le p^{-\beta} \quad \text{when } n \ge 32\varepsilon^{-2}\left(\frac{\lambda_1}{\lambda_k}\sum\nolimits_{i \le k}\frac{\lambda_i}{\lambda_k}\right)(\log k + \beta \log p).$$

Recall that $\kappa_k = \lambda_1(\mathbf{C})/\lambda_k(\mathbf{C})$ and $r = \left(\sum_{i=1}^p \lambda_i(\mathbf{C})\right)/\lambda_1(\mathbf{C})$, so

$$\left(\frac{\lambda_1}{\lambda_k}\sum\nolimits_{i \le k}\frac{\lambda_i}{\lambda_k}\right) \le \kappa_k^2 r.$$

Clearly, taking $n = \Omega(\varepsilon^{-2} r \kappa_\ell^2 \log p)$ samples ensures that, with high probability, each of the top ℓ eigenvalues of the sample covariance matrix satisfies $\lambda_k(\widehat{\mathbf{C}}_n) > (1-\varepsilon)\lambda_k$.

Likewise,

$$\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}_n) \ge (1-\varepsilon)\lambda_k\right\} \le p^{-\beta} \quad \text{when } n \ge 32\varepsilon^{-2}\left(\sum\nolimits_{i \ge k}\frac{\lambda_i}{\lambda_k}\right)(\log(p-k+1) + \beta \log p)$$

and

$$\sum\nolimits_{i \ge k}\frac{\lambda_i}{\lambda_k} = \frac{\lambda_1}{\lambda_k}\frac{\left(\sum_{i \ge k}\lambda_i\right)}{\lambda_1} \le \kappa_k \frac{\left(\sum_{i=1}^p \lambda_i\right)}{\lambda_1} = \kappa_k r,$$

so we see that taking $n = \Omega(\varepsilon^{-2} r \kappa_\ell \log p)$ samples ensures that, with high probability, each of the top ℓ eigenvalues of the sample covariance matrix satisfies $\lambda_k(\widehat{\mathbf{C}}_n) < (1+\varepsilon)\lambda_k$.

Combining these two results, we conclude that $n = \Omega(\varepsilon^{-2} r \kappa_\ell^2 \log p)$ ensures that the top ℓ eigenvalues of $\mathbf{C}$ are estimated to within relative precision $1 \pm \varepsilon$ with probability at least $1 - 2\ell p^{-\beta}$. □

2.8.1 Proof of Theorem 2.15

We now prove Theorem 2.15. This result requires a number of supporting lemmas; we defer their proofs until after a discussion of extensions to Theorem 2.15.

We study the error $|\lambda_k(\widehat{\mathbf{C}}_n) - \lambda_k(\mathbf{C})|$. To apply the methods developed in this chapter, we pass to a question about the eigenvalues of a difference of two matrices. The first lemma accomplishes this goal by compressing both the population covariance matrix and the sample covariance matrix to a fixed invariant subspace of the population covariance matrix.

Lemma 2.19. *Let* $\mathbf{X}$ *be a random Hermitian matrix with dimension* p*, and let* $\mathbf{A}$ *be a fixed Hermitian matrix with dimension* p*. Choose* $\mathbf{W}_+ \in \mathbb{V}_{p-k+1}^p$ *and* $\mathbf{W}_- \in \mathbb{V}_k^p$ *for which*

$$\lambda_k(\mathbf{A}) = \lambda_{\max}\left(\mathbf{W}_+^* \mathbf{A} \mathbf{W}_+\right) = \lambda_{\min}\left(\mathbf{W}_-^* \mathbf{A} \mathbf{W}_-\right).$$

Then, for all $t > 0$,

$$\mathbb{P}\left\{\lambda_k(\mathbf{X}) \geq \lambda_k(\mathbf{A}) + t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_+^*\mathbf{X}\mathbf{W}_+\right) \geq \lambda_k(\mathbf{A}) + t\right\} \tag{2.8.2}$$

and

$$\mathbb{P}\left\{\lambda_k(\mathbf{X}) \leq \lambda_k(\mathbf{A}) - t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_-^*(\mathbf{A}-\mathbf{X})\mathbf{W}_-\right) \geq t\right\}. \tag{2.8.3}$$

We apply this result with $\mathbf{A} = \mathbf{C}$ and $\mathbf{X} = \widehat{\mathbf{C}}_n$. The first estimate (2.8.2) and the second estimate (2.8.3) are handled using different arguments. The second estimate is easier because the maximum eigenvalue of the matrix $\mathbf{C} - \widehat{\mathbf{C}}_n$ is bounded. Indeed,

$$\lambda_{\max}\left(\mathbf{W}_+^*(\mathbf{C}-\widehat{\mathbf{C}}_n)\mathbf{W}_+\right) \leq \lambda_{\max}\left(\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+\right).$$

Thus, we may use Theorem 2.10 to complete the second estimate. The next lemma gives the matrix variances that we need to apply this theorem.

Lemma 2.20. *Let* $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$. *Then*

$$\mathbb{E}(\xi\xi^* - \mathbf{G})^2 = \mathbf{G}^2 + \operatorname{tr}(\mathbf{G}) \cdot \mathbf{G}.$$

The first inequality (2.8.2) is harder because $\widehat{\mathbf{C}}_n$ is unbounded. In this case, we may apply Theorem 2.12. To use this theorem, we need the following moment growth estimate for rank-one Wishart matrices.

Lemma 2.21. *Let* $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$. *Then for any integer* $m \geq 2$,

$$\mathbb{E}\left(\xi\xi^*\right)^m \preceq 2^m m!(\operatorname{tr}\mathbf{G})^{m-1} \cdot \mathbf{G}.$$

With these preliminaries addressed, we prove Theorem 2.15.

Proof of lower estimate in Theorem 2.15. First we consider the probability that $\hat{\lambda}_k$ underestimates λ_k. Let $\mathbf{W}_- \in \mathbb{V}_k^p$ satisfy

$$\lambda_k(\mathbf{C}) = \lambda_{\min}\left(\mathbf{W}_-^*\mathbf{C}\mathbf{W}_-\right).$$

Then Lemma 2.19 implies

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}_n) \leq \lambda_k(\mathbf{C}) - t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_-^*(\mathbf{C}-\widehat{\mathbf{C}}_n)\mathbf{W}_-\right) \geq t\right\} \\ &= \mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \mathbf{W}_-^*(\mathbf{C}-\eta_j\eta_j^*)\mathbf{W}_-\right) \geq nt\right\}.\end{aligned}$$

The factor n comes from the normalization of the sample covariance matrix. Each term in the sum is zero mean and bounded above by $\mathbf{W}_-^*\mathbf{C}\mathbf{W}_-$ in the semidefinite order, so Theorem 2.10 applies. As we desire a bound on the maximum eigenvalue of the sum, we take $\mathbf{V}_+ = \mathbf{I}$ when we invoke Theorem 2.10. Then

$$\sigma_1^2 = \lambda_{\max}\left(\sum\nolimits_j \mathbb{E}\left[\mathbf{W}_-^*(\mathbf{C}-\eta_j\eta_j^*)\mathbf{W}_-\right]^2\right) = n\lambda_{\max}\left(\mathbb{E}\left[\mathbf{W}_-^*(\mathbf{C}-\eta_1\eta_1^*)\mathbf{W}_-\right]^2\right).$$

The covariance matrix of η_1 is $\mathbf{C}$, so that of $\mathbf{W}_-^*\eta_1$ is $\mathbf{W}_-^*\mathbf{C}\mathbf{W}_-$. It follows from Lemma 2.20 that

$$\mathbb{E}\left[\mathbf{W}_-^*(\mathbf{C}-\eta_1\eta_1^*)\mathbf{W}_-\right]^2 = (\mathbf{W}_-^*\mathbf{C}\mathbf{W}_-)^2 + \operatorname{tr}(\mathbf{W}_-^*\mathbf{C}\mathbf{W}_-) \cdot \mathbf{W}_-^*\mathbf{C}\mathbf{W}_-.$$

Observe that $\mathbf{W}_-^* \mathbf{C} \mathbf{W}_-$ is the restriction of $\mathbf{C}$ to its top k-dimensional invariant subspace, so

$$\sigma_1^2 = n\lambda_{\max}\left(\mathbb{E}\left[\mathbf{W}_-^*(\mathbf{C} - \eta_1\eta_1^*)\mathbf{W}_-\right]^2\right) = n\lambda_1(\mathbf{C})\left(\lambda_1(\mathbf{C}) + \sum_{i=1}^k \lambda_i(\mathbf{C})\right)$$

and we can take $\Psi(\mathbf{V}_+) = \lambda_{\max}(\mathbf{C})$.

The subgaussian branch of the split Bernstein inequality of Theorem 2.10 shows that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \mathbf{W}_-^*(\mathbf{C} - \eta_j\eta_j^*)\mathbf{W}_-\right) \geq nt\right\} \leq k \cdot \exp\left(\frac{-3nt^2}{8\lambda_1(\mathbf{C})\left(\lambda_1(\mathbf{C}) + \sum_{i=1}^k \lambda_i(\mathbf{C})\right)}\right)$$

when $t \leq n\left(\lambda_1(\mathbf{C}) + \sum_{i=1}^k \lambda_i(\mathbf{C})\right)$. This inequality provides the desired bound on the probability that $\lambda_k(\widehat{\mathbf{C}}_n)$ underestimates $\lambda_k(\mathbf{C})$. □

Proof of upper estimate in Theorem 2.15. Now we consider the probability that $\hat{\lambda}_k$ overestimates λ_k. Let $\mathbf{W}_+ \in \mathbb{V}_{p-k+1}^p$ satisfy

$$\lambda_k(\mathbf{C}) = \lambda_{\max}\left(\mathbf{W}_+^* \mathbf{C} \mathbf{W}_+\right).$$

Then Lemma 2.19 implies

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}_n) \geq \lambda_k(\mathbf{C}) + t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_+^* \widehat{\mathbf{C}}_n \mathbf{W}_+\right) \geq \lambda_k(\mathbf{C}) + t\right\} \\ &= \mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \mathbf{W}_+^*(\eta_j\eta_j^*)\mathbf{W}_+\right) \geq n\lambda_k(\mathbf{C}) + nt\right\}. \qquad (2.8.4)\end{aligned}$$

The factor n comes from the normalization of the sample covariance matrix.

The covariance matrix of η_j is $\mathbf{C}$, so that of $\mathbf{W}_+^*\eta_j$ is $\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+$. Apply Lemma 2.21 to verify that $\mathbf{W}_+^*\eta_j$ satisfies the subexponential moment growth bound required by Theorem 2.12 with

$$B = 2\operatorname{tr}(\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+) \quad \text{and} \quad \mathbf{\Sigma}_j^2 = 8\operatorname{tr}(\mathbf{W}_+^*\mathbf{C}\mathbf{W}) \cdot \mathbf{W}_+^*\mathbf{C}\mathbf{W}_+.$$

In fact, $\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+$ is the compression of $\mathbf{C}$ to the invariant subspace corresponding with its bottom $p - k + 1$ eigenvalues, so

$$B = 2\sum\nolimits_{i=k}^p \lambda_i(\mathbf{C}) \quad \text{and} \quad \lambda_{\max}\left(\mathbf{\Sigma}_j^2\right) = 8\lambda_k(\mathbf{C})\sum\nolimits_{i=k}^p \lambda_i(\mathbf{C}).$$

We are concerned with the maximum eigenvalue of the sum in (2.8.4), so we take $\mathbf{V}_+ = \mathbf{I}$ in the statement of Theorem 2.12 to find that

$$\begin{aligned}\sigma_1^2 = \lambda_{\max}\left(\sum\nolimits_j \mathbf{\Sigma}_j^2\right) = n\lambda_{\max}\left(\mathbf{\Sigma}_1^2\right) = 8n\lambda_k(\mathbf{C})\sum\nolimits_{i=k}^p \lambda_i(\mathbf{C}) \quad \text{and} \\ \mu_1 = \lambda_{\max}\left(\sum_j \mathbf{W}_+^*\mathbb{E}(\eta_j\eta_j^*)\mathbf{W}_+\right) = n\lambda_{\max}\left(\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+\right) = n\lambda_k(\mathbf{C}).\end{aligned}$$

It follows from the subgaussian branch of the split Bernstein inequality of Theorem 2.12 that

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{W}_+^*(\eta_j\eta_j^*)\mathbf{W}_+\right) \geq n\lambda_k(\mathbf{C}) + nt\right\} \leq (p - k + 1) \cdot \exp\left(\frac{-nt^2}{32\lambda_k(\mathbf{C})\sum_{i=k}^p \lambda_i(\mathbf{C})}\right)$$

when $t \leq 4n\lambda_k(\mathbf{C})$. This provides the desired bound on the probability that $\lambda_k(\widehat{\mathbf{C}}_n)$ overestimates $\lambda_k(\mathbf{C})$. □

2.8.2 Extensions of Theorem 2.15

Results analogous to Theorem 2.15 can be established for other distributions. If the distribution is bounded, the possibility that $\hat{\lambda}_k$ deviates above or below λ_k can be controlled using the Bernstein inequality of Theorem 2.10. If the distribution is unbounded but has matrix moments that satisfy a sufficiently nice growth condition, the probability that $\hat{\lambda}_k$ deviates below λ_k can be controlled with the Bernstein inequality of Theorem 2.10 and the probability that it deviates above λ_k can be bounded using a Bernstein inequality analogous to that in Theorem 2.12.

We established Theorem 2.15 using this technique to demonstrate the simplicity of the Laplace transform machinery. However, the results of [ALPTJ11] on the convergence of empirical covariance matrices of isotropic log-concave random vectors lead to tighter bounds on the probability that $\hat{\lambda}_k$ overestimates λ_k. There does not seem to be an analogous reduction for handling the probability that $\hat{\lambda}_k$ is an underestimate.

To see the relevance of the results in [ALPTJ11], first observe the following consequence of the subadditivity of the maximum eigenvalue mapping:

$$\begin{aligned}\lambda_{\max}\left(\mathbf{W}_+^*(\mathbf{X}-\mathbf{A})\mathbf{W}_+\right) &\geq \lambda_{\max}\left(\mathbf{W}_+^*\mathbf{X}\mathbf{W}_+\right) - \lambda_{\max}\left(\mathbf{W}_+^*\mathbf{A}\mathbf{W}_+\right) \\ &= \lambda_{\max}\left(\mathbf{W}_+^*\mathbf{X}\mathbf{W}_+\right) - \lambda_k(\mathbf{A}).\end{aligned}$$

In conjunction with (2.8.2), this gives us the following control on the probability that $\lambda_k(\mathbf{X})$ overestimates $\lambda_k(\mathbf{A})$:

$$\mathbb{P}\left\{\lambda_k(\mathbf{X}) \geq \lambda_k(\mathbf{A}) + t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_+^*(\mathbf{X}-\mathbf{A})\mathbf{W}_+\right) \geq t\right\}.$$

In our application, $\mathbf{X}$ is the empirical covariance matrix and $\mathbf{A}$ is the actual covariance matrix. The spectral norm dominates the maximum eigenvalue, so

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}_n) \geq \lambda_k(\mathbf{C}) + t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_+^*(\widehat{\mathbf{C}}_n - \mathbf{C})\mathbf{W}_+\right) \geq t\right\} \\ &\leq \mathbb{P}\left\{\|\mathbf{W}_+^*(\widehat{\mathbf{C}}_n - \mathbf{C})\mathbf{W}_+\| \geq t\right\} = \mathbb{P}\left\{\|\mathbf{W}_+^*\widehat{\mathbf{C}}_n\mathbf{W}_+ - \mathbf{S}^2\| \geq t\right\},\end{aligned}$$

where $\mathbf{S}$ is the square root of $\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+$. Now factor out $\mathbf{S}^2$ and identify $\lambda_k(\mathbf{C}) = \|\mathbf{S}^2\|$ to obtain

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\mathbf{C}}) \geq \lambda_k(\mathbf{C}) + t\right\} &\leq \mathbb{P}\left\{\|\mathbf{S}^{-1}\mathbf{W}_+^*\widehat{\mathbf{C}}_n\mathbf{W}_+\mathbf{S}^{-1} - \mathbf{I}\|\|\mathbf{S}^2\| \geq t\right\} \\ &= \mathbb{P}\left\{\|\mathbf{S}^{-1}\mathbf{W}_+^*\widehat{\mathbf{C}}_n\mathbf{W}_+\mathbf{S}^{-1} - \mathbf{I}\| \geq t/\lambda_k(\mathbf{C})\right\}.\end{aligned}$$

Note that if η is drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{C})$ distribution, then the covariance matrix of the transformed sample $\mathbf{S}^{-1}\mathbf{W}_+^*\eta$ is the identity:

$$\mathbb{E}\left(\mathbf{S}^{-1}\mathbf{W}_+^*\eta\eta^*\mathbf{W}_+\mathbf{S}^{-1}\right) = \mathbf{S}^{-1}\mathbf{W}_+^*\mathbf{C}\mathbf{W}_+\mathbf{S}^{-1} = \mathbf{I}.$$

Thus $\mathbf{S}^{-1}\mathbf{W}_+^*\widehat{\mathbf{C}}_n\mathbf{W}_+\mathbf{S}^{-1}$ is the empirical covariance matrix of a standard Gaussian vector in $\mathbb{R}^{p-k+1}$. By Theorem 1 of [ALPTJ11], it follows that $\hat{\lambda}_k$ is unlikely to overestimate λ_k in relative error when the number n of samples is $\Omega(p-k+1)$.

Similarly, for more general distributions, the bounds on the probability of $\hat{\lambda}_k$ exceeding λ_k can be tightened beyond those suggested in Theorem 2.15 by using the results in [ALPTJ11] or [Ver11b].

Finally, we note that the techniques developed in the proof of Theorem 2.15 can be used to investigate the spectrum of the error matrices $\widehat{\mathbf{C}}_n - \mathbf{C}$.

2.8.3 Proofs of the supporting lemmas

We now establish the lemmas used in the proof of Theorem 2.15.

Proof of Lemma 2.19. The probability that $\lambda_k(\mathbf{X})$ overestimates $\lambda_k(\mathbf{A})$ is controlled with the sequence of inequalities

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\mathbf{X}) \geq \lambda_k(\mathbf{A}) + t\right\} &= \mathbb{P}\left\{\inf_{\mathbf{W}\in\mathbb{V}_{p-k+1}^p} \lambda_{\max}\left(\mathbf{W}^*\mathbf{X}\mathbf{W}\right) \geq \lambda_k(\mathbf{A}) + t\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_+^*\mathbf{X}\mathbf{W}_+\right) \geq \lambda_k(\mathbf{A}) + t\right\}.\end{aligned}$$

We use a related approach to study the probability that $\lambda_k(\mathbf{X})$ underestimates $\lambda_k(\mathbf{A})$. Our choice of $\mathbf{W}_-$ implies that

$$\lambda_{p-k+1}(-\mathbf{A}) = -\lambda_k(\mathbf{A}) = -\lambda_{\min}\left(\mathbf{W}_-^*\mathbf{A}\mathbf{W}_-\right) = \lambda_{\max}\left(\mathbf{W}_-^*(-\mathbf{A})\mathbf{W}_-\right).$$

It follows that

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\mathbf{X}) \leq \lambda_k(\mathbf{A}) - t\right\} &= \mathbb{P}\left\{\lambda_{p-k+1}(-\mathbf{X}) \geq \lambda_{p-k+1}(-\mathbf{A}) + t\right\} \\ &= \mathbb{P}\left\{\inf_{\mathbf{W}\in\mathbb{V}_k^p} \lambda_{\max}\left(\mathbf{W}^*(-\mathbf{X})\mathbf{W}\right) \geq \lambda_{\max}\left(\mathbf{W}_-^*(-\mathbf{A})\mathbf{W}_-\right) + t\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_-^*(-\mathbf{X})\mathbf{W}_-\right) - \lambda_{\max}\left(\mathbf{W}_-^*(-\mathbf{A})\mathbf{W}_-\right) \geq t\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\max}\left(\mathbf{W}_-^*(\mathbf{A}-\mathbf{X})\mathbf{W}_-\right) \geq t\right\}.\end{aligned}$$

The final inequality follows from the subadditivity of the maximum eigenvalue mapping. □

Proof of Lemma 2.20. We begin by taking $\mathbf{S}$ to be the positive-semidefinite square root of $\mathbf{G}$. Let $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ be the eigenvalue decomposition of $\mathbf{S}$, and let γ be a $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ random variable. Recalling that $\mathbf{G}$ is the covariance matrix of ξ, we see that ξ and $\mathbf{U}\mathbf{\Lambda}\gamma$ are identically distributed. Thus,

$$\begin{aligned}\mathbb{E}(\xi\xi^* - \mathbf{G})^2 &= \mathbb{E}(\mathbf{U}\mathbf{\Lambda}\gamma\gamma^*\mathbf{\Lambda}\mathbf{U}^* - \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^*)^2 \\ &= \mathbf{U}\mathbf{\Lambda}\mathbb{E}(\gamma\gamma^*\mathbf{\Lambda}^2\gamma\gamma^*)\mathbf{\Lambda}\mathbf{U}^* - \mathbf{G}^2. \qquad (2.8.5)\end{aligned}$$

Consider the (i,j) entry of the matrix being averaged:

$$\mathbb{E}(\gamma\gamma^*\mathbf{\Lambda}^2\gamma\gamma^*)_{ij} = \sum_k \mathbb{E}(\gamma_i\gamma_j\gamma_k^2)\lambda_k^2.$$

The (i,j) entry of this matrix is zero because the entries of γ are independent and symmetric. Furthermore, the (i,i) entry satisfies

$$\mathbb{E}(\gamma\gamma^*\mathbf{\Lambda}^2\gamma\gamma^*)_{ii} = \mathbb{E}(\gamma_i^4)\lambda_i^2 + \sum_{k\neq i}\mathbb{E}(\gamma_k^2)\lambda_k^2 = 2\lambda_i^2 + \operatorname{tr}(\mathbf{\Lambda}^2).$$

We have shown

$$\mathbb{E}(\gamma\gamma^*\mathbf{\Lambda}^2\gamma\gamma^*) = 2\mathbf{\Lambda}^2 + \operatorname{tr}(\mathbf{G})\cdot\mathbf{I}.$$

This equality and (2.8.5) imply the desired result. □

Proof of Lemma 2.21. Factor the covariance matrix of ξ as $\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ where $\mathbf{U}$ is orthogonal and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ is the matrix of eigenvalues of $\mathbf{G}$. Let γ be a $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ random variable. Then ξ and $\mathbf{U}\mathbf{\Lambda}^{1/2}\gamma$ are identically distributed, so

$$\begin{aligned}\mathbb{E}(\xi\xi^*)^m &= \mathbb{E}\left[(\xi^*\xi)^{m-1}\xi\xi^*\right] = \mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\mathbf{U}\mathbf{\Lambda}^{1/2}\gamma\gamma^*\mathbf{\Lambda}^{1/2}\mathbf{U}^*\right] \\ &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\gamma\gamma^*\right]\mathbf{\Lambda}^{1/2}\mathbf{U}^*. \end{aligned} \tag{2.8.6}$$

Consider the (i, j) entry of the bracketed matrix in (2.8.6):

$$\mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\gamma_i\gamma_j\right] = \mathbb{E}\left[\left(\sum\nolimits_{\ell=1}^{p}\lambda_\ell\gamma_\ell^2\right)^{m-1}\gamma_i\gamma_j\right]. \tag{2.8.7}$$

From this expression, and the independence and symmetry of the Gaussian variables $\{\gamma_i\}$, we see that this matrix is diagonal.

To bound the diagonal entries, use a multinomial expansion to further develop the sum in (2.8.7) for the (i, i) entry:

$$\mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\gamma_i^2\right] = \sum_{\ell_1+\cdots+\ell_p=m-1}\binom{m-1}{\ell_1,\dots,\ell_p}\lambda_1^{\ell_1}\cdots\lambda_p^{\ell_p}\mathbb{E}\left[\gamma_1^{2\ell_1}\cdots\gamma_p^{2\ell_p}\gamma_i^2\right].$$

Now we use the generalized AM–GM inequality to replace the expectation of the product of Gaussians with the $2m$th moment of a single standard Gaussian g. Denote the L_r norm of a random variable X by

$$\|X\|_{L_r} = (\mathbb{E}|X|^r)^{1/r}.$$

Since $\ell_1, \dots, \ell_p$ are nonnegative integers summing to $m-1$, the generalized AM-GM inequality justifies the first of the following inequalities:

$$\begin{aligned}\mathbb{E}\gamma_1^{2\ell_1}\cdots\gamma_p^{2\ell_p}\gamma_i^2 &\le \mathbb{E}\left(\frac{\ell_1|\gamma_1| + \cdots + \ell_p|\gamma_p| + |\gamma_i|}{m}\right)^{2m} = \left\|\frac{1}{m}\left(|\gamma_i| + \sum_{j=1}^{p}\ell_j|\gamma_j|\right)\right\|_{L_{2m}}^{2m} \\ &\le \left(\frac{1}{m}\left(\left\|\gamma_i\right\|_{L_{2m}} + \sum_{j=1}^{p}\ell_j\left\|\gamma_j\right\|_{L_{2m}}\right)\right)^{2m} \\ &= \left(\frac{1+\ell_1+\ldots+\ell_p}{m}\right)^{2m}\left\|g\right\|_{L_{2m}}^{2m} = \mathbb{E}(g^{2m}).\end{aligned}$$

The second inequality is the triangle inequality for L_r norms. Now we reverse the multinomial expansion to see that the diagonal terms satisfy the inequality

$$\begin{aligned}\mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\gamma_i^2\right] &\le \sum_{\ell_1+\cdots+\ell_p=m-1}\binom{m-1}{\ell_1,\dots,\ell_p}\lambda_1^{\ell_1}\cdots\lambda_p^{\ell_p}\mathbb{E}(g^{2m}) \\ &= (\lambda_1+\ldots+\lambda_p)^{m-1}\mathbb{E}(g^{2m}) = \operatorname{tr}(\mathbf{G})^{m-1}\mathbb{E}(g^{2m}).\end{aligned} \tag{2.8.8}$$

Estimate $\mathbb{E}(g^{2m})$ using the fact that $\Gamma(x)$ is increasing for $x \ge 1$:

$$\mathbb{E}\left(g^{2m}\right) = \frac{2^m}{\sqrt{\pi}}\Gamma(m+1/2) < \frac{2^m}{\sqrt{\pi}}\Gamma(m+1) = \frac{2^m}{\sqrt{\pi}}m! \quad \text{for } m \ge 1.$$

Combine this result with (2.8.8) to see that

$$\mathbb{E}\left[(\gamma^*\mathbf{\Lambda}\gamma)^{m-1}\gamma\gamma^*\right] \preceq \frac{2^m}{\sqrt{\pi}} m!\,\mathrm{tr}(\mathbf{G})^{m-1}\cdot\mathbf{I}.$$

Complete the proof by using this estimate in (2.8.6). □

Chapter 3

Randomized sparsification in NP-hard norms

Massive matrices are ubiquitous in modern data processing. Classical dense matrix algorithms are poorly suited to such problems because their running times scale superlinearly with the size of the matrix. When the dataset is sparse, one prefers to use sparse matrix algorithms, whose running times depend more on the sparsity of the matrix than on the size of the matrix. Of course, in many applications the matrix is *not* sparse. Accordingly, one may wonder whether it is possible to approximate a computation on a large dense matrix with a related computation on a sparse approximant to the matrix.

Let $\|\cdot\|$ be a norm on matrices. Here is one way to frame this challenge mathematically: Given a matrix $\mathbf{A}$, how can one efficiently generate a sparse matrix $\mathbf{X}$ for which the approximation error $\|\mathbf{A}-\mathbf{X}\|$ is small?

The literature has concentrated on the behavior of the approximation error in the spectral and Frobenius norms; however, these norms are not always the most natural choice. Sometimes it is more appropriate to consider the matrix as an operator from a finite-dimensional ℓ_p space to a finite-dimensional ℓ_q space, and investigate the behavior of the approximation error in the associated $p \to q$ operator norm. As an example, the problem of graph sparsification is naturally posed as a question of preserving the so-called *cut norm* of a matrix associated with the graph. The strong equivalency of the cut norm and the $\infty \to 1$ norm suggests that, for graph-theoretic applications, it may be fruitful to consider the behavior of the $\infty \to 1$ norm under sparsification. In other applications, e.g., the column subset selection algorithm in [Tro09], the $\infty \to 2$ norm is the norm of interest.

This chapter investigates the errors incurred by approximating a fixed real matrix with a random matrix[1]. Our results apply to any scheme in which the entries of the approximating matrix are independent and average to the corresponding entries of the fixed matrix. Our main contribution is a bound on the expected $\infty \to p$ norm error, which we specialize to the case of the $\infty \to 1$ and $\infty \to 2$ norms. We also use a result of Latała [Lat05] to bound the expected spectral approximation error, and we establish the subgaussianity of the spectral approximation error.

Our methods are similar to those of Rudelson and Vershynin in [RV07] in that we treat $\mathbf{A}$ as a linear operator between finite-dimensional Banach spaces and use some of the same tools of probability in Banach spaces. Whereas Rudelson and Vershynin consider the behavior of the norms of random submatrices of $\mathbf{A}$, we consider the behavior of the norms of matrices formed by randomly sparsifying (or quantizing) the entries of $\mathbf{A}$. This yields error bounds applicable to

[1] The content of this chapter is adapted from the technical report [GT11] co-authored with Joel Tropp.

schemes that sparsify or quantize matrices entrywise. Since some graph algorithms depend more on the number of edges in the graph than the number of vertices, such schemes may be useful in developing algorithms for handling large graphs. In particular, the algorithm of [BSS09] is not suitable for sparsifying graphs with a large number of vertices. Part of our motivation for investigating the $\infty \to 1$ approximation error is the belief that the equivalence of the cut norm with the $\infty \to 1$ norm means that matrix sparsification in the $\infty \to 1$ norm might be useful for efficiently constructing optimal sparsifiers for such graphs.

3.1 Notation

We establish the notation particular to this chapter.

All quantities are real. For $1 \leq p \leq \infty$, the ℓ_p^n norm of $\mathbf{x} \in \mathbb{R}^n$ is written as $\|\mathbf{x}\|_p$. Each space ℓ_p^n has an associated dual space $\ell_{p'}^n$, where p', the *conjugate exponent* to p, is determined by the relation $p^{-1} + (p')^{-1} = 1$. The dual space of ℓ_1^n (respectively, ℓ_∞^n) is ℓ_∞^n (respectively, ℓ_1^n).

The kth column of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{(k)}$, and the (j,k)th element is denoted by a_{jk}. We treat $\mathbf{A}$ as an operator from ℓ_p^n to ℓ_q^m, and the $p \to q$ operator norm of $\mathbf{A}$ is written as $\|\mathbf{A}\|_{p\to q}$. The spectral norm, i.e. the $2 \to 2$ operator norm, is written $\|\mathbf{A}\|_2$. Recall that given an operator $\mathbf{A} : \ell_p^n \to \ell_q^m$, the associated adjoint operator ($\mathbf{A}^T$, in the case of a matrix) maps from $\ell_{q'}^m$ to $\ell_{p'}^n$. Further, the $p \to q$ and $q' \to p'$ norms are dual in the sense that

$$\|\mathbf{A}\|_{p\to q} = \left\|\mathbf{A}^T\right\|_{q'\to p'}.$$

This chapter is concerned primarily with the spectral norm and the $\infty \to 1$ and $\infty \to 2$ norms. The $\infty \to 1$ and $\infty \to 2$ norms are not unitarily invariant, so do not have simple interpretations in terms of singular values; in fact, they are NP-hard to compute for general matrices [Roh00]. We remark that $\|\mathbf{A}\|_{\infty\to 1} = \|\mathbf{Ax}\|_1$ and $\|\mathbf{A}\|_{\infty\to 2} = \left\|\mathbf{Ay}\right\|_2$ for certain vectors $\mathbf{x}$ and $\mathbf{y}$ whose components take values ± 1. An additional operator norm, the $2 \to \infty$ norm, is of interest: it is the largest ℓ_2 norm achieved by a row of $\mathbf{A}$. In the sequel we also encounter the column norm

$$\|\mathbf{A}\|_{\text{col}} = \sum\nolimits_k \|\mathbf{A}_{(k)}\|_2.$$

The variance of X is written $\text{Var}X = \mathbb{E}(X - \mathbb{E}X)^2$. The expectation taken with respect to one variable X, with all others fixed, is written $\mathbb{E}_X$. The expression $X \sim Y$ indicates the random variables X and Y are identically distributed. Given a random variable X, the symbol X' denotes a random variable independent of X such that $X' \sim X$. The indicator variable of the event $X > Y$ is written $\mathbb{1}_{X>Y}$. The Bernoulli distribution with expectation p is written $\text{Bern}(p)$ and the binomial distribution of n independent trials each with success probability p is written $\text{Bin}(n,p)$. We write $X \sim \text{Bern}(p)$ to indicate X is Bernoulli with mean p.

3.1.0.1 Graph sparsification

Graphs are often represented and fruitfully manipulated in terms of matrices, so the problems of graph sparsification and matrix sparsification are strongly related. We now introduce the relevant notation before surveying the literature.

Let $G = (V, E, \omega)$ be a weighted simple undirected graph with n vertices, m edges, and adjacency matrix $\mathbf{A}$ given by

$$a_{jk} = \begin{cases} \omega_{jk} & (j,k) \in E \\ 0 & \text{otherwise} \end{cases}.$$

Orient the edges of G in an arbitrary manner. Then define the corresponding $2m \times n$ *oriented incidence matrix* $\mathbf{B}$ in the following manner: $b_{2i-1,j} = b_{2i,k} = \omega_i$ and $b_{2i-1,k} = b_{2i,j} = -\omega_i$ if edge i is oriented from vertex j to vertex k, and all other entries of B are identically zero.

A *cut* is a partition of the vertices of G into two blocks: $V = S \cup \overline{S}$. The *cost* of a cut is the sum of the weights of all edges in E which have one vertex in S and one vertex in $\overline{S}$. Several problems relating to cuts are of considerable practical interest. In particular, the MAXCUT problem, to determine the cut of maximum cost in a graph, is common in computer science applications. The cuts of maximum cost are exactly those that correspond to the *cut-norm* of the oriented incidence matrix $\mathbf{B}$, which is defined as

$$\|\mathbf{B}\|_{\mathrm{C}} = \max_{I \subset \{1,\ldots,2m\}, J \subset \{1,\ldots,n\}} \left| \sum_{i \in I, j \in J} b_{ij} \right|.$$

Finding the cut-norm of a general matrix is NP-hard, but in [AN04], the authors offer a randomized polynomial-time algorithm which finds a submatrix $\tilde{\mathbf{B}}$ of $\mathbf{B}$ for which $|\sum_{jk} \tilde{b}_{jk}| \geq 0.56 \|\mathbf{B}\|_{\mathrm{C}}$. This algorithm thereby gives a feasible means of approximating the MAXCUT value for arbitrary graphs. A crucial point in the derivation of the algorithm is the fact that for general matrices the $\infty \to 1$ operator norm is strongly equivalent with the cut-norm:

$$\|\mathbf{A}\|_{\mathrm{C}} \leq \|\mathbf{A}\|_{\infty \to 1} \leq 4 \|\mathbf{A}\|_{\mathrm{C}};$$

in fact, in the particular case of oriented incidence matrices, $\|\mathbf{B}\|_{\mathrm{C}} = \|\mathbf{B}\|_{\infty \to 1}$.

In his thesis [Kar95] and the sequence of papers [Kar94a, Kar94b, Kar96], Karger introduces the idea of random sampling to increase the efficiency of calculations with graphs, with a focus on cuts. In [Kar96], he shows that by picking each edge of the graph with a probability inversely proportional to the density of edges in a neighborhood of that edge, one can construct a *sparsifier*, i.e., a graph with the same vertex set and significantly fewer edges that preserves the value of each cut to within a factor of $(1 \pm \epsilon)$.

In [SS08], Spielman and Srivastava improve upon this sampling scheme, instead keeping an edge with probability proportional to its *effective resistance*—a measure of how likely it is to appear in a random spanning tree of the graph. They provide an algorithm which produces a sparsifier with $\mathrm{O}\left((n \log n)/\epsilon^2\right)$ edges, where n is the number of vertices in the graph. They obtain this result by reducing the problem to the behavior of projection matrices $\mathbf{\Pi}_G$ and $\mathbf{\Pi}_{G'}$ associated with the original graph and the sparsifier, and then appealing to a spectral-norm concentration result.

The $\log n$ factor in [SS08] seems to be an unavoidable consequence of using spectral-norm concentration. In [BSS09], Batson et al. prove that the $\log n$ factor is not intrinsic: they establish that every graph has a sparsifier that has $\Omega(n)$ edges. The existence proof is constructive and provides a deterministic algorithm for constructing such sparsifiers in $\mathrm{O}(n^3 m)$ time, where m is the number of edges in the original graph.

3.2 Preliminaries

Bounded differences inequalities are useful tools for establishing measure concentration for functions of independent random variables that are insensitive to changes in a single argument. In this chapter, we use a bounded differences inequality to show that the norms of the random matrices that we encounter exhibit measure concentration.

Before stating the inequality of interest to us, we establish some notation. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a measurable function of n random variables. Let $X_1, \ldots, X_n$ be independent random variables,

and write $W = g(X_1, \dots, X_n)$. Let W_i denote the random variable obtained by replacing the ith argument of g with an independent copy: $W_i = g(X_1, \dots, X_i', \dots, X_n)$.

The following bounded differences inequality states that if g is insensitive to changes of a single argument, then W does not deviate much from its mean.

Lemma 3.1 ([BLM03, Corollary 3])**.** *Let W and $\{W_i\}$ be random variables defined as above. Assume that there exists a positive number C such that, almost surely,*

$$\sum_{i=1}^{n} (W - W_i)^2 \mathbb{1}_{W > W_i} \leq C.$$

Then, for all $t > 0$,

$$\mathbb{P}\{W > \mathbb{E}W + t\} \leq \mathrm{e}^{-t^2/(4C)}.$$

Rademacher random variables take the values ± 1 with equal probability. Rademacher vectors are vectors of i.i.d. Rademacher random variables. We make use of Rademacher variables in this chapter to simplify our analyses through the technique of Rademacher symmetrization. Essentially, given a random variable Z, Rademacher symmetrization allows us to estimate the behavior of Z in terms of that of the random variable $Z_{\text{sym}} = \varepsilon(Z - Z')$, where Z' is an i.i.d. copy of Z and ε is a Rademacher random variable that is independent of the pair (Z, Z'). The variable Z_{sym} is often easier to manipulate than Z, since it is guaranteed to be symmetric (i.e., Z_{sym} and $-Z_{\text{sym}}$ are identically distributed); in particular, $\mathbb{E}Z_{\text{sym}} = 0$. The following basic symmetrization result is drawn from [vW96, Lemma 2.3.1 et seq.].

Lemma 3.2. *Let $Z_1, \dots, Z_n, Z_1', \dots, Z_n'$ be independent random variables satisfying $Z_i \sim Z_i'$, and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Let $\mathscr{F}$ be a family of functions such that*

$$\sup_{f \in \mathscr{F}} \sum_{k=1}^{n} (f(Z_k) - f(Z_k'))$$

is measurable. Then

$$\mathbb{E} \sup_{f \in \mathscr{F}} \sum_{k=1}^{n} (f(Z_k) - f(Z_k')) = \mathbb{E} \sup_{f \in \mathscr{F}} \sum_{k=1}^{n} \varepsilon_k (f(Z_k) - f(Z_k')).$$

Since we work with finite-dimensional probability models and linear functions, measurability concerns can be ignored in our applications of Lemma 3.2.

While Lemma 3.2 allows us to replace certain random processes with Rademacher processes, Talagrand's Rademacher comparison theorem [LT91, Theorem 4.12 et seq.] shows that certain complicated Rademacher processes are bounded by simpler Rademacher processes. Together, these two results often allow us to reduce the analysis of complicated random processes to the analysis of simpler Rademacher processes.

Lemma 3.3. *Fix finite-dimensional vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$ and let $\boldsymbol{\varepsilon}$ be a Rademacher vector. Then*

$$\mathbb{E} \max_{\|\mathbf{u}\|_q = 1} \sum_{k=1}^{n} \varepsilon_k |\langle \mathbf{z}_k, \mathbf{u} \rangle| \leq \mathbb{E} \max_{\|\mathbf{u}\|_q = 1} \sum_{k=1}^{n} \varepsilon_k \langle \mathbf{z}_k, \mathbf{u} \rangle.$$

Lemma 3.3 involves Rademacher sums, i.e. sums of the form $\sum_k \varepsilon_k x_k$ where $\boldsymbol{\varepsilon}$ is a Rademacher vector and $\mathbf{x}$ is a fixed vector. One of the most basic tools for understanding Rademacher sums is the Khintchine inequality [Sza76], which gives information on the moments of a Rademacher sum; in particular, it tells us the expected value of the sum is equivalent with the ℓ_2 norm of the vector $\mathbf{x}$.

Lemma 3.4 (Khintchine Inequality)**.** *Let* $\mathbf{x}$ *be a real vector, and let* $\boldsymbol{\varepsilon}$ *be a Rademacher vector. Then*

$$\frac{1}{\sqrt{2}} \|\mathbf{x}\|_2 \leq \mathbb{E} \left| \sum_k \varepsilon_k x_k \right| \leq \|\mathbf{x}\|_2 .$$

In its more general form, which we do not use in this thesis, the Khintchine inequality implies that Rademacher sums are subguassian random variables.

3.3 The $\infty \to p$ norm of a random matrix

We are interested in schemes that approximate a given matrix $\mathbf{A}$ by means of a random matrix $\mathbf{X}$ in such a way that the entries of $\mathbf{X}$ are independent and $\mathbb{E}\mathbf{X} = \mathbf{A}$. It follows that the error matrix $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ has independent, zero-mean entries. Ultimately we aim to construct $\mathbf{X}$ so that it has the property that, with high probability, many of its entries are identically zero, but this property does not play a role at this stage of the analysis.

In this section, we derive a bound on the expected value of the $\infty \to p$ norm of a random matrix with independent, zero-mean entries. We also study the tails of this error. In the next two sections, we use the results of this section to reach more detailed conclusions on the $\infty \to 1$ and $\infty \to 2$ norms of $\mathbf{Z}$.

3.3.1 The expected $\infty \to p$ norm

The main tools used to derive the bound on the expected norm of $\mathbf{Z}$ are Lemma 3.2, a result on the Rademacher symmetrization of random processes, and Lemma 3.3, Talagrand's Rademacher comparison theorem.

We now state and prove the bound on the expected norm of $\mathbf{Z}$.

Theorem 3.5. *Let* $\mathbf{Z}$ *be a random matrix with independent, zero-mean entries and let* $\boldsymbol{\varepsilon}$ *be a Rademacher vector independent of* $\mathbf{Z}$*. Then*

$$\mathbb{E} \|\mathbf{Z}\|_{\infty \to p} \leq 2\mathbb{E} \left\| \sum_k \varepsilon_k \mathbf{Z}_{(k)} \right\|_p + 2 \max_{\|\mathbf{u}\|_q = 1} \mathbb{E} \sum_k \left| \sum_j \varepsilon_j Z_{jk} u_j \right|$$

where q *is the conjugate exponent of* p*.*

Proof of Theorem 3.5. By duality,

$$\mathbb{E} \|\mathbf{Z}\|_{\infty \to p} = \mathbb{E} \left\| \mathbf{Z}^T \right\|_{q \to 1} = \mathbb{E} \max_{\|\mathbf{u}\|_q = 1} \sum_k |\langle \mathbf{Z}_{(k)}, \mathbf{u} \rangle|.$$

Center the terms in the sum and apply subadditivity of the maximum to get

$$\begin{aligned} \mathbb{E} \|\mathbf{Z}\|_{\infty \to p} &\leq \mathbb{E} \max_{\|\mathbf{u}\|_q = 1} \sum_k (|\langle \mathbf{Z}_{(k)}, \mathbf{u} \rangle| - \mathbb{E}' |\langle \mathbf{Z}'_{(k)}, \mathbf{u} \rangle|) + \max_{\|\mathbf{u}\|_q = 1} \mathbb{E} \sum_k |\langle \mathbf{Z}_{(k)}, \mathbf{u} \rangle| \\ &= F + S. \end{aligned} \tag{3.3.1}$$

Begin with the first term on the right-hand side of (3.3.1). Use Jensen's inequality to draw the expectation outside of the maximum:

$$F \le \mathbb{E} \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k (|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| - |\langle \mathbf{Z}'_{(k)}, \mathbf{u}\rangle|).$$

Now apply Lemma 3.2 to symmetrize the random variable:

$$F \le \mathbb{E} \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \varepsilon_k (|\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| - |\langle \mathbf{Z}'_{(k)}, \mathbf{u}\rangle|).$$

By the subadditivity of the maximum,

$$F \le \mathbb{E}\left(\max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \varepsilon_k |\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| + \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k -\varepsilon_k |\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle| \right) = 2\mathbb{E} \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \varepsilon_k |\langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle|,$$

where we have invoked the fact that $-\varepsilon_k$ has the Rademacher distribution. Apply Lemma 3.3 to get the final estimate of F:

$$F \le 2\mathbb{E} \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \varepsilon_k \langle \mathbf{Z}_{(k)}, \mathbf{u}\rangle = 2\mathbb{E} \max_{\|\mathbf{u}\|_q=1} \left\langle \sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}, \mathbf{u} \right\rangle = 2\mathbb{E} \left\| \sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)} \right\|_p .$$

Now consider the last term on the right-hand side of (3.3.1). Use Jensen's inequality to prepare for symmetrization:

$$\begin{aligned} S &= \max_{\|\mathbf{u}\|_q=1} \mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j Z_{jk} u_j \right| = \max_{\|\mathbf{u}\|_q=1} \mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j (Z_{jk} - \mathbb{E}' Z'_{jk}) u_j \right| \\ &\le \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \mathbb{E} \left| \sum\nolimits_j (Z_{jk} - Z'_{jk}) u_j \right| . \end{aligned}$$

Apply Lemma 3.2 to the expectation of the inner sum to see

$$S \le \max_{\|\mathbf{u}\|_q=1} \sum\nolimits_k \mathbb{E} \left| \sum\nolimits_j \varepsilon_j (Z_{jk} - Z'_{jk}) u_j \right| .$$

The triangle inequality gives us the final expression:

$$S \le \max_{\|\mathbf{u}\|_q=1} 2\mathbb{E} \sum\nolimits_k \left| \sum\nolimits_j \varepsilon_j Z_{jk} u_j \right| .$$

Introduce the bounds for F and S into (3.3.1) to complete the proof. □

3.3.2 A tail bound for the $\infty \to p$ norm

We now develop a deviation bound for the $\infty \to p$ approximation error. The argument is based on Lemma 3.1, a bounded differences inequality.

To apply Lemma 3.1, we let $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ be our error matrix, $W = \|\mathbf{Z}\|_{\infty \to p}$, and $W^{jk} = \left\|\mathbf{Z}^{jk}\right\|_{\infty \to p}$, where $\mathbf{Z}^{jk}$ is a matrix obtained by replacing $a_{jk} - X_{jk}$ with an identically distributed variable $a_{jk} - X'_{jk}$ while keeping all other variables fixed. The $\infty \to p$ norms are sufficiently insensitive to each entry of the matrix that Lemma 3.1 gives us a useful deviation bound.

Theorem 3.6. *Fix an $m \times n$ matrix $\mathbf{A}$, and let $\mathbf{X}$ be a random matrix with independent entries for which $\mathbb{E}X = \mathbf{A}$. Assume $\left|X_{jk}\right| \leq \frac{D}{2}$ almost surely for all j, k. Then, for all $t > 0$,*

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbf{X}\|_{\infty\to p} > \mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p} + t\right\} \leq \mathrm{e}^{-t^2/(4D^2nm^s)}$$

where $s = \max\{0, 1-2/q\}$ and q is the conjugate exponent to p.

Proof. Let q be the conjugate exponent of p, and choose $\mathbf{u}, \mathbf{v}$ such that $W = \mathbf{u}^T\mathbf{Z}\mathbf{v}$ and $\|\mathbf{u}\|_q = 1$ and $\|\mathbf{v}\|_\infty = 1$. Then

$$(W - W^{jk})\mathbb{1}_{W>W^{jk}} \leq \mathbf{u}^T\left(\mathbf{Z}-\mathbf{Z}^{jk}\right)\mathbf{v}\,\mathbb{1}_{W>W^{jk}} = (X'_{jk} - X_{jk})u_j v_k\,\mathbb{1}_{W>W^{jk}} \leq D|u_j v_k|.$$

This implies

$$\sum\nolimits_{j,k}(W - W^{jk})^2\mathbb{1}_{W>W^{jk}} \leq D^2\sum\nolimits_{j,k}|u_j v_k|^2 \leq nD^2\|\mathbf{u}\|_2^2,$$

so we can apply Lemma 3.1 if we have an estimate for $\|\mathbf{u}\|_2^2$. We have the bounds $\|\mathbf{u}\|_2 \leq \|\mathbf{u}\|_q$ for $q \in [1,2]$ and $\|\mathbf{u}\|_2 \leq m^{1/2-1/q}\|\mathbf{u}\|_q$ for $q \in [2,\infty]$. Therefore,

$$\sum\nolimits_{j,k}(W - W^{jk})^2\mathbb{1}_{W>W^{jk}} \leq D^2\begin{cases} nm^{1-2/q}, & q \in [2,\infty] \\ n, & q \in [1,2]. \end{cases}$$

It follows from Lemma 3.1 that

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbf{X}\|_{\infty\to p} > \mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p} + t\right\} = \mathbb{P}\{W > \mathbb{E}W + t\} \leq \mathrm{e}^{-t^2/(4D^2nm^s)}$$

where $s = \max\{0, 1-2/q\}$. □

It is often convenient to measure deviations on the scale of the mean. Taking $t = \delta\mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p}$ in Theorem 3.6 gives the following result.

Corollary 3.7. *Under the conditions of Theorem 3.6, for all $\delta > 0$,*

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbf{X}\|_{\infty\to p} > (1+\delta)\mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p}\right\} \leq \mathrm{e}^{-\delta^2\left(\mathbb{E}\|\mathbf{A}-\mathbf{X}\|_{\infty\to p}\right)^2/(4D^2nm^s)}.$$

3.4 Approximation in the $\infty\to 1$ norm

In this section, we develop the $\infty\to 1$ error bound as a consequence of Theorem 3.5. We then prove that one form of the error bound is optimal, and we describe an example of its application to matrix sparsification.

3.4.1 The expected $\infty\to 1$ norm

To derive the $\infty\to 1$ error bound, we first apply Theorem 3.5 with $p = 1$.

Theorem 3.8. *Suppose that $\mathbf{Z}$ is a random matrix with independent, zero-mean entries. Then*

$$\mathbb{E}\|\mathbf{Z}\|_{\infty\to 1} \leq 2\mathbb{E}(\|\mathbf{Z}\|_{\mathrm{col}} + \left\|\mathbf{Z}^T\right\|_{\mathrm{col}}).$$

Proof. Apply Theorem 3.5 to get

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} &\le 2\mathbb{E}\left\|\sum_k \varepsilon_k \mathbf{Z}_{(k)}\right\|_1 + 2\max_{\|\mathbf{u}\|_\infty=1}\mathbb{E}\sum_k\left|\sum_j \varepsilon_j Z_{jk}u_j\right| \\ &= F + S.\end{aligned} \tag{3.4.1}$$

Use Hölder's inequality to bound the first term in (3.4.1) with a sum of squares:

$$\begin{aligned}F &= 2\mathbb{E}\sum_j\left|\sum_k \varepsilon_k Z_{jk}\right| = 2\mathbb{E}_{\mathbf{Z}}\sum_j \mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum_k \varepsilon_k Z_{jk}\right| \\ &\le 2\mathbb{E}_{\mathbf{Z}}\sum_j\left(\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2}.\end{aligned}$$

The inner expectation can be computed exactly by expanding the square and using the independence of the Rademacher variables:

$$F \le 2\mathbb{E}\sum_j\left(\sum_k Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\left\|\mathbf{Z}^T\right\|_{\text{col}}.$$

We treat the second term in the same manner. Use Hölder's inequality to replace the sum with a sum of squares and invoke the independence of the Rademacher variables to eliminate cross terms:

$$S \le 2\max_{\|\mathbf{u}\|_\infty=1}\mathbb{E}_{\mathbf{Z}}\sum_k\left(\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum_j \varepsilon_j Z_{jk}u_j\right|^2\right)^{1/2} = 2\max_{\|\mathbf{u}\|_\infty=1}\mathbb{E}\sum_k\left(\sum_j Z_{jk}^2 u_j^2\right)^{1/2}.$$

Since $\|\mathbf{u}\|_\infty = 1$, it follows that $u_j^2 \le 1$ for all j, and

$$S \le 2\mathbb{E}\sum_k\left(\sum_j Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\,\|\mathbf{Z}\|_{\text{col}}.$$

Introduce these estimates for F and S into (3.4.1) to complete the proof. □

Taking $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ in Theorem 3.8, we find

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|_{\infty\to 1} \le 2\mathbb{E}\left[\sum_k\left(\sum_j (a_{jk}-X_{jk})^2\right)^{1/2} + \sum_j\left(\sum_k (a_{jk}-X_{jk})^2\right)^{1/2}\right].$$

A simple application of Jensen's inequality gives an error bound in terms of the variances of the entries of $\mathbf{X}$.

Corollary 3.9. *Fix the matrix* $\mathbf{A}$*, and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}X_{jk} = a_{jk}$*. Then*

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|_{\infty\to 1} \le 2\left[\sum_k\left(\sum_j \operatorname{Var}(X_{jk})\right)^{1/2} + \sum_j\left(\sum_k \operatorname{Var}(X_{jk})\right)^{1/2}\right].$$

3.4.2 Optimality

A simple estimate using the Khintchine inequality shows that the bound on the expected value of the $\infty \to 1$ norm given in Theorem 3.8 is in fact optimal up to constants.

Corollary 3.10. *Suppose that* $\mathbf{Z}$ *is a random matrix with independent, zero-mean entries. Then*

$$\frac{1}{2\sqrt{2}}\mathbb{E}(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}) \leq \mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1} \leq 2\mathbb{E}(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}).$$

Proof. First we establish the inequality

$$\|\mathbf{Z}\|_{\text{col}} \leq \sqrt{2}\,\|\mathbf{Z}\|_{\infty\to 1} \tag{3.4.2}$$

as a consequence of the Khintchine inequality, Lemma 3.4. Indeed, since

$$\|\mathbf{Z}\|_{\text{col}} = \sum\nolimits_j \left\|\mathbf{Z}_{(j)}\right\|_2$$

and the Khintchine inequality gives the estimate

$$\left\|\mathbf{Z}_{(j)}\right\|_2 \leq \sqrt{2}\mathbb{E}\left|\sum\nolimits_i \varepsilon_i Z_{ij}\right|,$$

we see that

$$\begin{aligned}\|\mathbf{Z}\|_{\text{col}} &\leq \sqrt{2}\mathbb{E}\sum\nolimits_j \left|\sum\nolimits_i \varepsilon_i Z_{ij}\right| \\ &= \sqrt{2}\mathbb{E}\left\|\mathbf{Z}^T\varepsilon\right\|_1 \leq \sqrt{2}\sup_{\|\mathbf{x}\|_\infty=1}\left\|\mathbf{Z}^T\mathbf{x}\right\|_1 \\ &= \sqrt{2}\left\|\mathbf{Z}^T\right\|_{\infty\to 1} = \sqrt{2}\,\|\mathbf{Z}\|_{\infty\to 1}\,.\end{aligned}$$

Since the $\infty \to 1$ norms of $\mathbf{Z}$ and $\mathbf{Z}^T$ are equal, it also follows that

$$\left\|\mathbf{Z}^T\right\|_{\text{col}} \leq \sqrt{2}\left\|\mathbf{Z}^T\right\|_{\infty\to 1} = \sqrt{2}\,\|\mathbf{Z}\|_{\infty\to 1}\,.$$

The lower bound on $\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1}$ is now a consequence of (3.4.2),

$$\frac{1}{2\sqrt{2}}\mathbb{E}(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}) \leq \mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 1}\,,$$

while the upper bound is given by Theorem 3.8. □

Remark 3.11. Using standard arguments, one can establish that the deterministic bounds

$$\frac{1}{2\sqrt{2}}\left(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}\right) \leq \|\mathbf{Z}\|_{\infty\to 1} \leq \frac{\sqrt{n}}{2}\left(\|\mathbf{Z}\|_{\text{col}} + \left\|\mathbf{Z}^T\right\|_{\text{col}}\right)$$

hold for *any* square $n \times n$ matrix $\mathbf{Z}$. Corollary 3.10 is a refinement of this equivalence relation that holds when $\mathbf{Z}$ is a random, zero-mean matrix. In particular, the corollary tells us that when we assume this model for $\mathbf{Z}$, the equivalence relation does not depend on the dimensions of $\mathbf{Z}$, and thus if what we care about is the expected $\infty \to 1$ norm of $\mathbf{Z}$, we can work with the expected column norm of $\mathbf{Z}$ without losing any sharpness.

3.4.3 An example application

In this section we provide an example illustrating the application of Corollary 3.9 to matrix sparsification.

From Corollary 3.9 we infer that a good scheme for sparsifying a matrix $\mathbf{A}$ while minimizing the expected relative $\infty \to 1$ error is one which drastically increases the sparsity of $\mathbf{X}$ while keeping the relative error

$$\frac{\sum_k \left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2}}{\|\mathbf{A}\|_{\infty\to 1}}$$

small. Once a sparsification scheme is chosen, the hardest part of estimating this quantity is probably estimating the $\infty \to 1$ norm of $\mathbf{A}$. The example shows, for a simple family of approximation schemes, what kind of sparsification results can be obtained using Corollary 3.9 when we have a very good handle on this quantity.

Consider the case where $\mathbf{A}$ is an $n \times n$ matrix whose entries all lie within an interval bounded away from zero; for definiteness, take them to be positive. Let γ be a desired bound on the expected relative $\infty \to 1$ norm error. We choose the randomization strategy $X_{jk} \sim \frac{a_{jk}}{p}\,\mathrm{Bern}(p)$ and ask how small can p be without violating our bound on the expected error.

In this case,

$$\|\mathbf{A}\|_{\infty\to 1} = \sum_{j,k} a_{jk} = \mathrm{O}(n^2),$$

and $\mathrm{Var}(X_{jk}) = \frac{a_{jk}^2}{p} - a_{jk}^2$. Consequently, the first term in Corollary 3.9 satisfies

$$\begin{aligned}\sum_k \left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2} &= \sum_k \left(\frac{1}{p}\left\|\mathbf{a}_k\right\|_2^2 - \left\|\mathbf{a}_k\right\|_2^2\right)^{1/2} = \left(\frac{1-p}{p}\right)^{1/2} \|\mathbf{A}\|_{\mathrm{col}} \\ &= \mathrm{O}\left(\left(\frac{1-p}{p}\right)^{1/2} n\sqrt{n}\right)\end{aligned}$$

and likewise the second term satisfies

$$\sum_j \left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2} = \mathrm{O}\left(\left(\frac{1-p}{p}\right)^{1/2} n\sqrt{n}\right).$$

Therefore the relative $\infty \to 1$ norm error satisfies

$$\frac{\sum_k \left(\sum_j \mathrm{Var}(X_{jk})\right)^{1/2} + \sum_j \left(\sum_k \mathrm{Var}(X_{jk})\right)^{1/2}}{\|\mathbf{A}\|_{\infty\to 1}} = \mathrm{O}\left(\left(\frac{1-p}{pn}\right)^{1/2}\right).$$

It follows that $\mathbb{E}\,\|\mathbf{A} - \mathbf{X}\|_{\infty\to 1} < \gamma$ for p on the order of $(1 + n\gamma^2)^{-1}$ or larger. The expected number of nonzero entries in $\mathbf{X}$ is pn^2, so for matrices with this structure, we can sparsify with a relative $\infty \to 1$ norm error smaller than γ while reducing the number of expected nonzero entries to as few as $\mathrm{O}(\frac{n^2}{1+n\gamma^2}) = \mathrm{O}(\frac{n}{\gamma^2})$. Intuitively, this sparsification result is optimal in the dimension: it seems we must keep on average at least one entry per row and column if we are to faithfully approximate $\mathbf{A}$.

3.5 Approximation in the $\infty \to 2$ norm

In this section, we develop the $\infty \to 2$ error bound stated in the introduction, establish the optimality of a related bound, and provide examples of its application to matrix sparsification. To derive the error bound, we first specialize Theorem 3.5 to the case of $p = 2$.

Theorem 3.12. *Suppose that* $\mathbf{Z}$ *is a random matrix with independent, zero-mean entries. Then*

$$\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} \le 2\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} + 2\min_{\mathbf{D}} \mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty}$$

where $\mathbf{D}$ *is a positive diagonal matrix that satisfies* $\operatorname{Tr}(\mathbf{D}^2) = 1$.

Proof. Apply Theorem 3.5 to get

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} &\le 2\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \mathbf{Z}_{(k)}\right\|_2 + 2\max_{\|\mathbf{u}\|_2=1} \mathbb{E}\sum\nolimits_k \left|\sum\nolimits_j \varepsilon_j Z_{jk} u_j\right| \\ &=: F + S.\end{aligned} \tag{3.5.1}$$

Expand the first term, and use Jensen's inequality to move the expectation with respect to the Rademacher variables inside the square root:

$$F = 2\mathbb{E}\left(\sum\nolimits_j \left|\sum\nolimits_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2} \le 2\mathbb{E}_{\mathbf{Z}}\left(\sum\nolimits_j \mathbb{E}_{\boldsymbol{\varepsilon}} \left|\sum\nolimits_k \varepsilon_k Z_{jk}\right|^2\right)^{1/2}.$$

The independence of the Rademacher variables implies that the cross terms cancel, so

$$F \le 2\mathbb{E}\left(\sum\nolimits_j \sum\nolimits_k Z_{jk}^2\right)^{1/2} = 2\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}}.$$

We use the Cauchy–Schwarz inequality to replace the ℓ_1 norm with an ℓ_2 norm in the second term of (3.5.1). A direct application would introduce a possibly suboptimal factor of $\sqrt{n}$ (where n is the number of columns in $\mathbf{Z}$), so instead we choose $d_k > 0$ such that $\sum_k d_k^2 = 1$ and use the corresponding weighted ℓ_2 norm:

$$S = 2\max_{\|\mathbf{u}\|_2=1} \mathbb{E}\sum\nolimits_k \frac{\left|\sum_j \varepsilon_j Z_{jk} u_j\right|}{d_k} d_k \le 2\max_{\|\mathbf{u}\|_2=1} \mathbb{E}\left(\sum\nolimits_k \frac{\left|\sum_j \varepsilon_j Z_{jk} u_j\right|^2}{d_k^2}\right)^{1/2}.$$

Move the expectation with respect to the Rademacher variables inside the square root and observe that the cross terms cancel:

$$S \le 2\max_{\|\mathbf{u}\|_2=1} \mathbb{E}_{\mathbf{Z}}\left(\sum\nolimits_k \frac{\mathbb{E}_{\boldsymbol{\varepsilon}}\left|\sum_j \varepsilon_j Z_{jk} u_j\right|^2}{d_k^2}\right)^{1/2} = 2\max_{\|\mathbf{u}\|_2=1} \mathbb{E}\left(\sum\nolimits_{j,k} \frac{Z_{jk}^2 u_j^2}{d_k^2}\right)^{1/2}.$$

Use Jensen's inequality to pass the maximum through the expectation, and note that if $\|\mathbf{u}\|_2 = 1$ then the vector formed by elementwise squaring $\mathbf{u}$ lies on the ℓ_1 unit ball, thus

$$S \le 2\mathbb{E}\left(\max_{\|\mathbf{u}\|_1=1} \sum\nolimits_j u_j \cdot \left(\sum_k (Z_{jk}/d_k)^2\right)\right)^{1/2}.$$

Clearly this maximum is achieved when $\mathbf{u}$ is chosen so $u_j = 1$ at an index j for which $\sum_k (Z_{jk}/d_k)^2$ is maximal and $u_j = 0$ otherwise. Consequently, the maximum is the largest of the ℓ_2 norms of the rows of $\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$. Recall that this quantity is, by definition, $\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty}$. Therefore $S \le 2\mathbb{E}\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty}$. The theorem follows by optimizing our choice of $\mathbf{D}$ and introducing our estimates for F and S into (3.5.1). □

Taking $\mathbf{Z} = \mathbf{A} - \mathbf{X}$ in Theorem 3.12, we have

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|_{\infty\to 2} \le 2\mathbb{E}\left(\sum\nolimits_{j,k}(X_{jk}-a_{jk})^2\right)^{1/2} + 2\min_{\mathbf{D}}\mathbb{E}\max_j\left(\sum\nolimits_k \frac{(X_{jk}-a_{jk})^2}{d_k^2}\right)^{1/2}. \quad (3.5.2)$$

We now derive a bound which depends only on the variances of the X_{jk}.

Corollary 3.13. *Fix the $m \times n$ matrix $\mathbf{A}$ and let $\mathbf{X}$ be a random matrix with independent entries so that $\mathbb{E}X = \mathbf{A}$. Then*

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|_{\infty\to 2} \le 2\left(\sum\nolimits_{j,k}\operatorname{Var}(X_{jk})\right)^{1/2} + 2\sqrt{m}\min_{\mathbf{D}}\max_j\left(\sum\nolimits_k \frac{\operatorname{Var}(X_{jk})}{d_k^2}\right)^{1/2}$$

where $\mathbf{D}$ is a positive diagonal matrix with $\operatorname{Tr}(\mathbf{D}^2) = 1$.

Proof. Let F and S denote, respectively, the first and second term of (3.5.2). An application of Jensen's inequality shows that $F \le 2\left(\sum_{j,k}\operatorname{Var}(X_{jk})\right)^{1/2}$. A second application shows that

$$S \le 2\min_{\mathbf{D}}\left(\mathbb{E}\max_j\sum\nolimits_k \frac{(X_{jk}-a_{jk})^2}{d_k^2}\right)^{1/2}.$$

Bound the maximum with a sum:

$$S \le 2\min_{\mathbf{D}}\left(\sum\nolimits_j\mathbb{E}\sum\nolimits_k \frac{(X_{jk}-a_{jk})^2}{d_k^2}\right)^{1/2}.$$

The sum is controlled by a multiple of its largest term, so

$$S \le 2\sqrt{m}\min_{\mathbf{D}}\left(\max_j\sum\nolimits_k \frac{\operatorname{Var}(X_{jk})}{d_k^2}\right)^{1/2},$$

where m is the number of rows of $\mathbf{A}$. □

3.5.1 Optimality

We now show that Theorem 3.12 gives an optimal bound, in the sense that each of its terms is necessary. In the following, we reserve the letter $\mathbf{D}$ for a positive diagonal matrix with $\operatorname{Tr}(\mathbf{D}^2) = 1$.

First, we establish the necessity of the Frobenius term by identifying a class of random matrices whose $\infty\to 2$ norms are larger than their weighted $2\to\infty$ norms but comparable to their Frobenius norms. Let $\mathbf{Z}$ be a random $m \times \sqrt{m}$ matrix such that the entries in the first column of $\mathbf{Z}$ are equally likely to be positive or negative ones, and all other entries are zero. With this choice,

$\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} = \mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} = \sqrt{m}$. Meanwhile, $\mathbb{E}\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty} = d_{11}^{-1}$, so $\min_{\mathbf{D}} \mathbb{E}\|\mathbf{Z}\mathbf{D}^{-1}\|_{2\to\infty} = 1$, which is much smaller than $\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2}$. Clearly, the Frobenius term is necessary.

Similarly, to establish the necessity of the weighted $2\to\infty$ norm term, we consider a class of matrices whose $\infty\to 2$ norms are larger than their Frobenius norms but comparable to their weighted $2\to\infty$ norms. Consider a $\sqrt{n}\times n$ matrix $\mathbf{Z}$ whose entries are all equally likely to be positive or negative ones. It is a simple task to confirm that $\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2} \geq n$ and $\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} = n^{3/4}$; it follows that the weighted $2\to\infty$ norm term is necessary. In fact,

$$\min_{\mathbf{D}} \mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty} = \min_{\mathbf{D}} \mathbb{E} \max_{j=1,\ldots,\sqrt{n}} \left(\sum\nolimits_{k=1}^{n} \frac{Z_{jk}^2}{d_{kk}^2}\right)^{1/2} = \min_{\mathbf{D}} \left(\sum\nolimits_{k=1}^{n} \frac{1}{d_{kk}^2}\right)^{1/2} = n,$$

so we see that $\mathbb{E}\,\|\mathbf{Z}\|_{\infty\to 2}$ and the weighted $2\to\infty$ norm term are comparable.

3.5.2 An example application

From Theorem 3.12 we infer that a good scheme for sparsifying a matrix $\mathbf{A}$ while minimizing the expected relative $\infty\to 2$ norm error is one which drastically increases the sparsity of $\mathbf{X}$ while keeping the relative error

$$\frac{\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} + \min_{\mathbf{D}} \mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty}}{\|\mathbf{A}\|_{\infty\to 2}}$$

small, where $\mathbf{Z} = \mathbf{A} - \mathbf{X}$.

As before, consider the case where $\mathbf{A}$ is an $n\times n$ matrix all of whose entries are positive and in an interval bounded away from zero. Let γ be a desired bound on the expected relative $\infty\to 2$ norm error. We choose the randomization strategy $X_{jk} \sim \frac{a_{jk}}{p}\,\mathrm{Bern}(p)$ and ask how much can we sparsify while respecting our bound on the relative error. That is, how small can p be? We appeal to Theorem 3.12. In this case,

$$\|\mathbf{A}\|_{\infty\to 2} = \left(\sum\nolimits_j \sum\nolimits_k a_{jk}^2 + 2\sum\nolimits_j \sum\nolimits_{\ell<m} a_{j\ell}a_{jm}\right)^{\frac{1}{2}} = \mathrm{O}\left(\left(n^2 + n^2(n-1)\right)^{\frac{1}{2}}\right).$$

By Jensen's inequality,

$$\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} \leq \mathbb{E}\,\|\mathbf{A}\|_{\mathrm{F}} + \mathbb{E}\,\|\mathbf{X}\|_{\mathrm{F}} \leq \left(1 + \frac{1}{\sqrt{p}}\right)\|\mathbf{A}\|_{\mathrm{F}} = \mathrm{O}\left(n\left(1 + \frac{1}{\sqrt{p}}\right)\right).$$

We bound the other term in the numerator, also using Jensen's inequality:

$$\min_{\mathbf{D}} \mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty} \leq \sqrt{n}\mathbb{E}\,\|\mathbf{Z}\|_{2\to\infty} \leq \sqrt{n}\left(1 + \frac{1}{\sqrt{p}}\right)\|\mathbf{A}\|_{2\to\infty} = \mathrm{O}\left(n\left(1 + \frac{1}{\sqrt{p}}\right)\right)$$

to get

$$\frac{\mathbb{E}\,\|\mathbf{Z}\|_{\mathrm{F}} + \min_{\mathbf{D}} \mathbb{E}\left\|\mathbf{Z}\mathbf{D}^{-1}\right\|_{2\to\infty}}{\|\mathbf{A}\|_{\infty\to 2}} = \mathrm{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{pn}}\right) = \mathrm{O}\left(\frac{1}{\sqrt{pn}}\right)$$

We conclude that, for this class of matrices and this family of sparsification schemes, we can reduce the number of expected nonzero terms to $\mathrm{O}\left(\frac{n}{\gamma^2}\right)$ while maintaining an expected $\infty\to 2$ norm relative error of γ.

3.6 A spectral error bound

In this section we establish a bound on $\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|$ as an immediate consequence of Latała's result [Lat05]. We then derive a deviation inequality for the spectral approximation error using a log-Sobolev inequality from [BLM03], and use it to compare our results to those of Achlioptas and McSherry [AM07] and Arora, Hazan, and Kale [AHK06].

Theorem 3.14. *Suppose* $\mathbf{A}$ *is a fixed matrix, and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}\mathbf{X} = \mathbf{A}$*. Then*

$$\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\| \le \mathrm{C}\left[\max_j \left(\sum\nolimits_k \mathrm{Var}(X_{jk})\right)^{1/2} + \max_k \left(\sum\nolimits_j \mathrm{Var}(X_{jk})\right)^{1/2} + \left(\sum\nolimits_{jk} \mathbb{E}(X_{jk}-a_{jk})^4\right)^{1/4}\right]$$

where C *is a universal constant.*

In [Lat05], Latała considered the spectral norm of random matrices with independent, zero-mean entries, and he showed that, for any such matrix $\mathbf{Z}$,

$$\mathbb{E}\|\mathbf{Z}\| \le \mathrm{C}\left[\max_j \left(\sum\nolimits_k \mathbb{E}Z_{jk}^2\right)^{1/2} + \max_k \left(\sum\nolimits_j \mathbb{E}Z_{jk}^2\right)^{1/2} + \left(\sum\nolimits_{jk} \mathbb{E}Z_{jk}^4\right)^{1/4}\right],$$

where C is some universal constant. Unfortunately, no estimate for C is available. Theorem 3.14 follows from Latała's result, by taking $\mathbf{Z} = \mathbf{A}-\mathbf{X}$.

The bounded differences argument from Section 3.3 establishes the correct (subgaussian) tail behavior of $\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|$.

Theorem 3.15. *Fix the matrix* $\mathbf{A}$*, and let* $\mathbf{X}$ *be a random matrix with independent entries for which* $\mathbb{E}X = \mathbf{A}$*. Assume* $\left|X_{jk}\right| \le D/2$ *almost surely for all* j,k*. Then, for all* $t > 0$*,*

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbf{X}\| > \mathbb{E}\,\|\mathbf{A}-\mathbf{X}\| + t\right\} \le \mathrm{e}^{-t^2/(4D^2)}.$$

Proof. The proof is exactly that of Theorem 3.6, except now $\mathbf{u}$ and $\mathbf{v}$ are both in the ℓ_2 unit sphere. □

We find it convenient to measure deviations on the scale of the mean.

Corollary 3.16. *Under the conditions of Theorem 3.15, for all* $\delta > 0$*,*

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbf{X}\| > (1+\delta)\mathbb{E}\,\|\mathbf{A}-\mathbf{X}\|\right\} \le \mathrm{e}^{-\delta^2(\mathbb{E}\|\mathbf{A}-\mathbf{X}\|)^2/(4D^2)}.$$

3.6.1 Comparison with previous results

To demonstrate the applicability of our bound on the spectral norm error, we consider the sparsification and quantization schemes used by Achlioptas and McSherry [AM07], and the quantization scheme proposed by Arora, Hazan, and Kale [AHK06]. We show that our spectral norm error bound and the associated concentration result give results of the same order, with less effort. Throughout these comparisons, we take $\mathbf{A}$ to be a $m \times n$ matrix, with $m < n$, and we define $b = \max_{jk} |a_{jk}|$.

3.6.1.1 A matrix quantization scheme

First we consider the scheme proposed by Achlioptas and McSherry for quantization of the matrix entries:

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

With this choice $\mathrm{Var}(X_{jk}) = b^2 - a_{jk}^2 \leq b^2$, and $\mathbb{E}(X_{jk} - a_{jk})^4 = b^2 - 3a^4 + 2a^2b^2 \leq 3b^4$, so the expected spectral error satisfies

$$\mathbb{E}\|\mathbf{A} - \mathbf{X}\| \leq \mathrm{C}(\sqrt{n}b + \sqrt{m}b + b\sqrt[4]{3mn}) \leq 4\mathrm{C}b\sqrt{n}.$$

Applying Corollary 3.16, we find that the error satisfies

$$\mathbb{P}\left\{\|\mathbf{A} - \mathbf{X}\| > 4\mathrm{C}b\sqrt{n}(1+\delta)\right\} \leq \mathrm{e}^{-\delta^2\mathrm{C}^2 n}.$$

In particular, with probability at least $1 - \exp(-\mathrm{C}^2 n)$,

$$\|\mathbf{A} - \mathbf{X}\| \leq 8\mathrm{C}b\sqrt{n}.$$

Achlioptas and McSherry proved that for $n \geq n_0$, where n_0 is on the order of 10^9, with probability at least $1 - \exp(-19(\log n)^4)$,

$$\|\mathbf{A} - \mathbf{X}\| < 4b\sqrt{n}.$$

Thus, Theorem 3.15 provides a bound of the same order in n which holds with higher probability and over a larger range of n.

3.6.1.2 A nonuniform sparsification scheme

Next we consider an analog to the nonuniform sparsification scheme proposed in the same paper. Fix a number p in the range $(0, 1)$ and sparsify entries with probabilities proportional to their magnitudes:

$$X_{jk} \sim \frac{a_{jk}}{p_{jk}} \mathrm{Bern}(p_{jk}), \text{ where } p_{jk} = \max\left\{ p\left(\frac{a_{jk}}{b}\right)^2, \sqrt{p\left(\frac{a_{jk}}{b}\right)^2 \times \frac{(8\log n)^4}{n}} \right\}.$$

Achlioptas and McSherry determine that, with probability at least $1 - \exp(-19(\log n)^4)$,

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

Further, the expected number of nonzero entries in $\mathbf{X}$ is less than

$$pmn \times \mathrm{Avg}[(a_{jk}/b)^2] + m(8\log n)^4, \tag{3.6.1}$$

where the notation $\mathrm{Avg}(\cdot)$ indicates the average of a quantity over all the entries of $\mathbf{A}$.

Their choice of p_{jk}, in particular the insertion of the $(8\log n)^4/n$ factor, is an artifact of their method of proof. Instead, we consider a scheme which compares the magnitudes of a_{jk} and b to

determine p_{jk}. Introduce the quantity $R = \max_{a_{jk}\neq 0} b/|a_{jk}|$ to measure the spread of the entries in $\mathbf{A}$, and take

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

With this scheme, $\text{Var}(X_{jk}) = 0$ when $a_{jk} = 0$, otherwise $\text{Var}(X_{jk}) = b^2/p$. Likewise, $\mathbb{E}(X_{jk} - a_{jk})^4 = 0$ if $a_{jk} = 0$, otherwise

$$\mathbb{E}(X_{jk} - a_{jk})^4 \leq \text{Var}(X_{jk}) \left\|X_{jk} - a_{jk}\right\|_\infty^2 = \frac{b^2}{p} \max\left\{|a_{jk}|, |a_{jk}|\left(\frac{pa_{jk}^2 + b^2}{pa_{jk}^2} - 1\right)\right\}^2 \leq \frac{b^4}{p^2}R^2,$$

so

$$\mathbb{E}\,\|\mathbf{A} - \mathbf{X}\| \leq \text{C}\left(b\sqrt{\frac{n}{p}} + b\sqrt{\frac{m}{p}} + b\sqrt{\frac{R}{p}}\sqrt[4]{mn}\right) \leq \text{C}(2+\sqrt{R})b\sqrt{\frac{n}{p}}.$$

Applying Corollary 3.16, we find that the error satisfies

$$\mathbb{P}\left\{\|\mathbf{A} - \mathbf{X}\| > \text{C}(2+\sqrt{R})b\sqrt{\frac{n}{p}}(\epsilon + 1)\right\} \leq e^{-\epsilon^2\text{C}^2(2+\sqrt{R})^2pn/16},$$

with probability at least $1 - \exp(-\text{C}^2(2+\sqrt{R})^2pn/16)$,

$$\|\mathbf{A} - \mathbf{X}\| \leq 2\text{C}(2+\sqrt{R})b\sqrt{\frac{n}{p}}.$$

Thus, Theorem 3.14 and Achlioptas and McSherry's scheme-specific analysis yield results of the same order in n and p. As before, we see that our bound holds with higher probability and over a larger range of n. Furthermore, since the expected number of nonzero entries in $\mathbf{X}$ satisfies

$$\sum\nolimits_{jk} p_{jk} = \sum\nolimits_{jk} \frac{pa_{jk}^2}{pa_{jk}^2 + b^2} \leq pnm \times \text{Avg}\left[\left(\frac{a_{jk}}{b}\right)^2\right],$$

we have established a smaller limit on the expected number of nonzero entries.

3.6.1.3 A scheme which simultaneously sparsifies and quantizes

Finally, we use Theorem 3.15 to estimate the error of the scheme from [AHK06] which simultaneously quantizes and sparsifies. Fix $\delta > 0$ and consider

$$X_{jk} = \begin{cases} \text{sgn}(a_{jk})\frac{\delta}{\sqrt{n}} \text{ Bern}\left(\frac{|a_{jk}|\sqrt{n}}{\delta}\right), & |a_{jk}| \leq \frac{\delta}{\sqrt{n}} \\ a_{jk}, & \text{otherwise.} \end{cases}$$

Then $\text{Var}(X_{jk}) = 0$ if $|a_{jk}| \geq \delta/\sqrt{n}$, otherwise

$$\text{Var}(X_{jk}) = |a_{jk}|^3\frac{\sqrt{n}}{\delta} - 2a_{jk}^2 + |a_{jk}|\frac{\delta}{\sqrt{n}} \leq \frac{\delta^2}{n}.$$

The fourth moment term is zero when $|a_{jk}| \geq \delta/\sqrt{n}$, and when $|a_{jk}| < \delta/\sqrt{n}$,

$$\mathbb{E}(X_{jk} - a_{jk})^4 = |a_{jk}|^5 \frac{\sqrt{n}}{\delta} - 4a_{jk}^4 + 6|a_{jk}|^3 \frac{\delta}{\sqrt{n}} - 4a_{jk}^2 \frac{\delta^2}{n} + |a_{jk}| \left(\frac{\delta}{\sqrt{n}}\right)^3 \leq 8\frac{\delta^4}{n^2}.$$

This gives the estimates

$$\mathbb{E}\,\|\mathbf{A} - \mathbf{X}\| \leq C\left(\sqrt{n}\frac{\delta}{\sqrt{n}} + \sqrt{m}\frac{\delta}{\sqrt{n}} + 2\frac{\delta}{\sqrt{n}}\sqrt[4]{mn}\right) \leq 4C\delta$$

and

$$\mathbb{P}\left\{\|\mathbf{A} - \mathbf{X}\| > 4C\delta(\gamma + 1)\right\} \leq \mathrm{e}^{-\gamma^2 C^2 n}.$$

Taking $\gamma = 1$, we see that with probability at least $1 - \exp(-C^2 n)$,

$$\|\mathbf{A} - \mathbf{X}\| \leq 8C\delta.$$

Let $S = \sum_{j,k} |A_{jk}|$, then appealing to Lemma 1 in [AHK06], we find that $\mathbf{X}$ has $\mathrm{O}\left(\frac{\sqrt{n}S}{\gamma}\right)$ nonzero entries with probability at least $1 - \exp\left(-\Omega\left(\frac{\sqrt{n}S}{\gamma}\right)\right)$.

Arora, Hazan, and Kale establish that this scheme guarantees $\|\mathbf{A} - \mathbf{X}\| = \mathrm{O}(\delta)$ with probability at least $1 - \exp(-\Omega(n))$, so we see that our general bound recovers a bound of the same order.

3.7 Comparison with later bounds

The papers [NDT10, DZ11, AKL13], written after the results in this chapter were obtained, present alternative schemes for sparsification and quantization.

The scheme presented in [NDT10] sparsifies a matrix by zeroing out all sufficiently small entries of $\mathbf{A}$, keeping all sufficiently large entries, and randomly sampling the remaining entries of the matrix with a probability depending on their magnitudes. More precisely, given a parameter $s > 0$, it generates an approximation whose entries are distributed as

$$X_{jk} = \begin{cases} 0, & a_{jk}^2 \leq (\log^2(n)/n)\|\mathbf{A}\|_{\mathrm{F}}^2/s \\ a_{jk} & a_{jk}^2 \geq \|\mathbf{A}\|_{\mathrm{F}}^2/s \\ (a_{jk}/p_{jk})\mathrm{Bern}(p_{jk}), & \text{otherwise, where } p_{jk} = sa_{jk}^2/\|\mathbf{A}\|_{\mathrm{F}}^2. \end{cases}$$

The analysis offered guarantees that if $s = \Omega(\epsilon^{-2} n \log^3 n)$, then with probability at least $1 - n^{-1}$, $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$ and, in expectation, $\mathbf{X}$ has less than $2s$ nonzero entries. It is not clear whether or not this scheme can be analyzed using Theorem 3.14. It is straightforward to establish that $\mathrm{Var}(X_{jk}) \leq \epsilon^2/(n\log^3 n)$ for this scheme, but obtaining a sufficiently small upper bound on the fourth moment $\mathbb{E}(X_{jk} - a_{jk})^4$ is challenging. In particular, the estimate

$$\mathbb{E}(X_{jk} - a_{jk})^4 \leq \mathrm{Var}(X_{jk})\|X_{jk} - a_{jk}\|_\infty$$

gives an upper bound on the order of $\epsilon a_{jk}^2 n/\log^5 n$, which is sufficient only to establish a much weaker guarantee on the error $\mathbb{E}\|\mathbf{A} - \mathbf{X}\|_2$ than the guarantee given in [NDT10].

The scheme introduced in [DZ11] first zeroes out all entries of $\mathbf{A} \in \mathbb{R}^{n\times n}$ of sufficiently small magnitude, then samples elements from $\mathbf{A}$ in s i.i.d. trials with replacement. The elements are

selected with probabilities proportional to their squared magnitudes. Thus, the approximant can be written in the form

$$\mathbf{X} = \frac{1}{s} \sum_{t=1}^{s} \frac{a_{j_t k_t}}{p_{j_t k_t}} \boldsymbol{e}_{j_t} \boldsymbol{e}_{k_t}^T,$$

where (j_t, k_t) is the index of the element of $\mathbf{A}$ selected in the tth trial, $p_{jk} = a_{jk}^2 / \|\mathbf{A}\|_{\mathrm{F}}^2$ is the probability that the entry a_{jk} is selected, and $\boldsymbol{e}_j$ denotes the jth standard basis vector in n. Clearly $\mathbf{X}$ has at most s nonzero entries. Let $s = \Omega(\epsilon^{-2} n \log(n) \|\mathbf{A}\|_{\mathrm{F}}^2)$. Then the authors show that, with probability at least $1 - n^{-1}$, the error of the approximation satisfies $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$. This scheme is not easily analyzable using our Theorem 3.14. Since the approximant $\mathbf{X}$ is a sum of rank-one matrices, it is most natural to analyze its approximation error using tail bounds for sums of independent random matrices. Indeed, the authors of [DZ11] use a matrix Bernstein inequality to provide their results.

Finally, the scheme presented in [AKL13] computes an approximation of the same form as the scheme introduced in [DZ11], but samples entries of $\mathbf{A}$ with probabilities proportional their absolute values. That is,

$$\mathbf{X} = \frac{1}{s} \sum_{t=1}^{s} \frac{a_{j_t k_t}}{p_{j_t k_t}} \boldsymbol{e}_{j_t} \boldsymbol{e}_{k_t}^T,$$

where $p_{jk} = |a_{jk}| / \sum_{pq} |a_{pq}|$. Again, this scheme is not amenable to analysis using Theorem 3.14. Recall that $\mathbf{A}^{(k)}$ denotes the kth row of $\mathbf{A}$. The authors establish that, when

$$s = \Omega\left(\epsilon^{-2} \log(n/\delta) \left(\sum_{jk} |A_{jk}| \right) \max_k \|\mathbf{A}^{(k)}\|_1 \right),$$

the error bound $\|\mathbf{A} - \mathbf{X}\|_2 \leq \epsilon$ is satisfied with probability at least $1 - \delta$. The approximant $\mathbf{X}$ has, in expectation, at most $2s$ nonzero entries.

Comparing the extents to which we were able to reproduce the guarantees of the sparsification schemes introduced in [AM01, AHK06, AM07, NDT10, DZ11, AKL13], we see that Theorem 3.14 sometimes can recover competitive guarantees on the approximation errors of element-wise sparsification schemes in which X_{jk} is directly related to a_{jk} through a simple expression. When $\mathbf{X}$ is more naturally represented as a sum of rank-1 matrices, Theorem 3.14 is not easily applicable.

Chapter 4

Preliminaries for the investigation of low-rank approximation algorithms

This chapter consolidates probabilistic and linear algebraic tools used in Chapters 5 and 6. We also establish two lemmas of independent interest: the first, Lemma 4.3, is an exponential tail bound on the Frobenius-norm error incurred when approximating the product of two matrices using randomized column and row sampling without replacement; the second, Lemma 4.9, is a deterministic bound on the forward errors of column-based low-rank approximations.

4.1 Probabilistic tools

In this section, we review several tools that are used to deal with random matrices and more generally, random processes.

4.1.1 Concentration of convex functions of Rademacher variables

Rademacher random variables take the values ± 1 with equal probability. Rademacher vectors are vectors of i.i.d. Rademacher random variables. Rademacher vectors often play a crucial role in the construction of dimension reduction maps, an area where the strong measure concentration properties of Rademacher sums are often exploited. The following result states a large-deviation property of convex Lipschitz functions of Rademacher vectors: namely, these functions tend to be not much larger than their expectations.

Lemma 4.1 (A large deviation result for convex Lipschitz functions of Rademacher random variables [Corollary 1.3 ff. in [Led96]])**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function that satisfies the Lipschitz bound*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \left\|\mathbf{x} - \mathbf{y}\right\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y}.$$

Let $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ be a Rademacher vector. For all $t \geq 0$,

$$\mathbb{P}\left\{f(\boldsymbol{\varepsilon}) \geq \mathbb{E}\left[f(\boldsymbol{\varepsilon})\right] + Lt\right\} \leq \mathrm{e}^{-t^2/8}.$$

4.1.2 Chernoff bounds for sums of random matrices sampled without replacement

Classical Chernoff bounds provide tail bounds for sums of nonnegative random variables. Their matrix analogs provide tail bounds on the eigenvalues and singular values of sums of positive-

semidefinite random matrices. Matrix Chernoff bounds are particularly useful for analyzing algorithms involving randomized column-sampling. Most matrix Chernoff bounds available in the literature require the summands to be independent. Indeed, the Chernoff bounds developed in Chapter 2 bound the eigenvalues of a sum of independent random Hermitian matrices. However, occasionally one desires Chernoff bounds that do not require the summands to be independent. The following Chernoff bounds are useful in the case where the summands are drawn without replacement from a set of bounded random matrices.

Lemma 4.2 (Matrix Chernoff Bounds, Theorem 2.2 in [Tro11b])**.** *Let $\mathscr{X}$ be a finite set of positive-semidefinite matrices with dimension k, and suppose that*

$$\max_{\mathbf{X}\in\mathscr{X}} \lambda_{\max}(\mathbf{X}) \le B.$$

Sample $\{\mathbf{X}_1,\dots,\mathbf{X}_\ell\}$ uniformly at random from $\mathscr{X}$ without replacement. Compute

$$\mu_{\max} = \ell\cdot\lambda_1(\mathbb{E}\mathbf{X}_1) \quad \text{and} \quad \mu_{min} = \ell\cdot\lambda_k(\mathbb{E}\mathbf{X}_1).$$

Then

$$\mathbb{P}\left\{\lambda_1\left(\sum\nolimits_j \mathbf{X}_j\right) \ge (1+\nu)\mu_{\max}\right\} \le k\cdot\left[\frac{\mathrm{e}^{\nu}}{(1+\nu)^{1+\nu}}\right]^{\mu_{\max}/B} \quad \text{for } \nu\ge 0, \text{ and}$$

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \le (1-\nu)\mu_{\min}\right\} \le k\cdot\left[\frac{\mathrm{e}^{-\nu}}{(1-\nu)^{1-\nu}}\right]^{\mu_{\min}/B} \quad \text{for } \nu\in[0,1).$$

We also use the following standard simplification of the lower Chernoff bound, which holds under the setup of Lemma 4.2:

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \mathbf{X}_j\right) \le \varepsilon\mu_{\min}\right\} \le k\cdot\mathrm{e}^{-(1-\varepsilon)^2\mu_{\min}/(2B)} \quad \text{for } \varepsilon\in[0,1]. \tag{4.1.1}$$

4.1.3 Frobenius-norm error bounds for matrix multiplication

We now establish a tail bound on the Frobenius-norm error of a simple approximate matrix multiplication scheme based upon randomized column and row sampling. This simple approximate multiplication scheme is a staple in randomized numerical linear algebra, and variants have been analyzed multiple times [DK01, DKM06a, Sar06]. The result derived here differs in that it applies to the sampling without replacement model, and it provides bounds on the error that hold with high probability, rather than simply an estimate of the expected error.

Lemma 4.3 (Matrix Multiplication)**.** *Let $\mathbf{X}\in\mathbb{R}^{m\times n}$ and $\mathbf{Y}\in\mathbb{R}^{n\times p}$. Fix $\ell\le n$. Select uniformly at random and without replacement ℓ columns from $\mathbf{X}$ and the corresponding rows from $\mathbf{Y}$ and multiply the selected columns and rows with $\sqrt{n/\ell}$. Let $\hat{\mathbf{X}}\in\mathbb{R}^{m\times\ell}$ and $\hat{\mathbf{Y}}\in\mathbb{R}^{\ell\times p}$ contain the scaled columns and rows, respectively. Choose*

$$\sigma^2 \ge \frac{4n}{\ell}\sum\nolimits_{i=1}^{n} \|\mathbf{X}_{(i)}\|_2^2\|\mathbf{Y}^{(i)}\|_2^2 \quad \text{and} \quad B \ge \frac{2n}{\ell}\max_i \|\mathbf{X}_{(i)}\|_2\|\mathbf{Y}^{(i)}\|_2.$$

Then if $0\le t\le\sigma^2/B$,

$$\mathbb{P}\left\{\left\|\hat{\mathbf{X}}\hat{\mathbf{Y}}-\mathbf{X}\mathbf{Y}\right\|_{\mathrm{F}} \ge t+\sigma\right\} \le \exp\left(-\frac{t^2}{4\sigma^2}\right).$$

To prove Lemma 4.3, we use the following vector Bernstein inequality for sampling without replacement in Banach spaces; this result follows directly from a similar inequality for sampling with replacement established by Gross in [Gro11]. Again, vector Bernstein inequalities have been derived by multiple authors [LT91, BLM03, Rec11, Tro12, CP11, Gro11]; the value of this specific result is that it applies to the sampling without replacement model.

Lemma 4.4. *Let $\mathcal{V}$ be a collection of n vectors in a Hilbert space with norm $\|\cdot\|_2$. Choose $\mathbf{V}_1, \dots, \mathbf{V}_\ell$ from $\mathcal{V}$ uniformly at random* without *replacement. Choose $\mathbf{V}'_1, \dots, \mathbf{V}'_\ell$ from $\mathcal{V}$ uniformly at random* with *replacement. Let*

$$\mu = \mathbb{E}\left\|\sum\nolimits_{i=1}^{\ell} (\mathbf{V}'_i - \mathbb{E}\mathbf{V}'_i)\right\|_2$$

and set

$$\sigma^2 \geq 4\ell\mathbb{E}\left\|\mathbf{V}'_1\right\|_2^2 \quad \text{and} \quad B \geq 2\max_{\mathbf{V}\in\mathcal{V}} \|\mathbf{V}\|_2 .$$

If $0 \leq t \leq \sigma^2/B$, then

$$\mathbb{P}\left\{\left\|\sum\nolimits_{i=1}^{\ell} \mathbf{V}_i - \ell\mathbb{E}\mathbf{V}_1\right\|_2 \geq \mu + t\right\} \leq \exp\left(-\frac{t^2}{4\sigma^2}\right).$$

Proof. We proceed by developing a bound on the moment generating function (mgf) of

$$\left\|\sum\nolimits_{i=1}^{\ell} \mathbf{V}_i - \ell\mathbb{E}\mathbf{V}_1\right\|_2 - \mu.$$

This mgf is controlled by the mgf of a similar sum where the vectors are sampled with replacement. That is, for $\lambda \geq 0$,

$$\mathbb{E}\exp\left(\lambda \cdot \left\|\sum\nolimits_{i=1}^{\ell} \mathbf{V}_i - \ell\mathbb{E}\mathbf{V}_1\right\|_2 - \lambda\mu\right) \leq \mathbb{E}\exp\left(\lambda \cdot \left\|\sum\nolimits_{i=1}^{\ell} \mathbf{V}'_i - \ell\mathbb{E}\mathbf{V}_1\right\|_2 - \lambda\mu\right). \tag{4.1.2}$$

This follows from a classical observation due to Hoeffding [Hoe63] that for any convex real-valued function g,

$$\mathbb{E}g\left(\sum\nolimits_{i=1}^{\ell} \mathbf{V}_i\right) \leq \mathbb{E}g\left(\sum\nolimits_{i=1}^{\ell} \mathbf{V}'_i\right).$$

The paper [GN10] provides an alternate exposition of this fact. Specifically, take $g(\mathbf{V}) = \exp\left(\lambda\left\|\mathbf{V} - \ell\mathbb{E}\mathbf{V}_1\right\|_2 - \lambda\mu\right)$ to obtain the inequality of mgfs asserted in (4.1.2).

In the proof of Theorem 12 in [Gro11], Gross establishes that any random variable Z whose mgf is less than the righthand side of (4.1.2) satisfies a tail inequality of the form

$$\mathbb{P}\{Z \geq \mu + t\} \leq \exp\left(-\frac{t^2}{4s^2}\right) \tag{4.1.3}$$

when $t \leq s^2/M$, where

$$s^2 \geq \sum\nolimits_{i=1}^{\ell} \mathbb{E}\left\|\mathbf{V}'_i - \mathbb{E}\mathbf{V}'_1\right\|_2^2$$

and $\left\|\mathbf{V}'_i - \mathbb{E}\mathbf{V}'_1\right\|_2 \leq M$ almost surely for all $i = 1, \dots, \ell$. To apply this result, note that for all $i = 1, \dots, \ell$,

$$\left\|\mathbf{V}'_i - \mathbb{E}\mathbf{V}'_1\right\|_2 \leq 2\max_{\mathbf{V}\in\mathcal{V}} \|\mathbf{V}\|_2 = B.$$

Take $\mathbf{V}_1''$ to be an i.i.d. copy of $\mathbf{V}_1'$ and observe that, by Jensen's inequality,

$$\begin{aligned}\sum_{i=1}^{\ell} \mathbb{E}\left\|\mathbf{V}_i' - \mathbb{E}\mathbf{V}_1'\right\|_2^2 &= \ell\mathbb{E}\left\|\mathbf{V}_1' - \mathbb{E}\mathbf{V}_1'\right\|_2^2 \\ &\le \ell\mathbb{E}\left\|\mathbf{V}_1' - \mathbf{V}_1''\right\|_2^2 \le \ell\mathbb{E}(\left\|\mathbf{V}_1'\right\|_2 + \left\|\mathbf{V}_1''\right\|_2)^2 \\ &\le 2\ell\mathbb{E}\left\|\mathbf{V}_1'\right\|_2^2 + \left\|\mathbf{V}_1''\right\|_2^2 \\ &= 4\ell\mathbb{E}\left\|\mathbf{V}_1'\right\|_2^2 \le \sigma^2.\end{aligned}$$

The bound given in the statement of Lemma 4.4 when we take $s^2 = \sigma^2$ and $M = B$ in (4.1.3). □

With this Bernstein bound in hand, we proceed to the proof of Lemma 4.3. Let $\mathrm{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ denote the operation of vectorization, which stacks the columns of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ to form the vector $\mathrm{vec}(\mathbf{A})$.

Proof of Lemma 4.3. Let $\mathscr{V}$ be the collection of vectorized rank-one products of columns of $\sqrt{n/\ell}\cdot\mathbf{X}$ and rows of $\sqrt{n/\ell}\cdot\mathbf{Y}$. That is, take

$$\mathscr{V} = \left\{\frac{n}{\ell}\mathrm{vec}(\mathbf{X}_{(i)}\mathbf{Y}^{(i)})\right\}_{i=1}^{n}.$$

Sample $\mathbf{V}_1, \dots, \mathbf{V}_\ell$ uniformly at random from $\mathscr{V}$ without replacement, and observe that $\mathbb{E}\mathbf{V}_i = \ell^{-1}\mathrm{vec}(\mathbf{XY})$. With this notation,

$$\left\|\hat{\mathbf{X}}\hat{\mathbf{Y}} - \mathbf{XY}\right\|_{\mathrm{F}} \sim \left\|\sum_{i=1}^{\ell}(\mathbf{V}_i - \mathbb{E}\mathbf{V}_i)\right\|_2,$$

where $\sim$ refers to identical distributions. Therefore any probabilistic bound developed for the right-hand side quantity holds for the left-hand side quantity. The conclusion of the lemma follows when we apply Lemma 4.4 to bound the right-hand side quantity.

We calculate the variance-like term in Lemma 4.4,

$$4\ell\mathbb{E}\left\|\mathbf{V}_1\right\|_2^2 = 4\ell\frac{1}{n}\sum_{i=1}^{n}\frac{n^2}{\ell^2}\|\mathbf{X}_{(i)}\|_2^2\|\mathbf{Y}^{(i)}\|_2^2 = 4\frac{n}{\ell}\sum_{i=1}^{n}\|\mathbf{X}_{(i)}\|_2^2\|\mathbf{Y}^{(i)}\|_2^2 \le \sigma^2.$$

Now we consider the expectation

$$\mu = \mathbb{E}\left\|\sum_{i=1}^{\ell}(\mathbf{V}_i' - \mathbb{E}\mathbf{V}_i')\right\|_2.$$

In doing so, we will use the notation $\mathbb{E}[C \mid A, B, \dots]$ to denote the conditional expectation of a random variable C with respect to the random variables $A, B, \dots$. Recall that a Rademacher vector is a random vector whose entries are independent and take the values ± 1 with equal probability. Let $\boldsymbol{\varepsilon}$ be a Rademacher vector of length ℓ and sample $\mathbf{V}_1', \dots, \mathbf{V}_\ell'$ and $\mathbf{V}_1'', \dots, \mathbf{V}_\ell''$

uniformly at random from $\mathcal{V}$ with replacement. Now μ can be bounded as follows:

$$\begin{aligned}
\mu &= \mathbb{E}\left\|\sum_{i=1}^{\ell}(\mathbf{V}_i' - \mathbb{E}\mathbf{V}_i')\right\|_2 \\
&\le \mathbb{E}\left[\left\|\sum_{i=1}^{\ell}(\mathbf{V}_i' - \mathbf{V}_i'')\right\|_2 \,\middle|\, \{\mathbf{V}_i'\}, \{\mathbf{V}_i''\}\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^{\ell}\varepsilon_i(\mathbf{V}_i' - \mathbf{V}_i'')\right\|_2 \,\middle|\, \{\mathbf{V}_i'\}, \{\mathbf{V}_i''\}, \boldsymbol{\varepsilon}\right] \\
&\le 2\mathbb{E}\left[\left\|\sum_{i=1}^{\ell}\varepsilon_i\mathbf{V}_i'\right\|_2 \,\middle|\, \{\mathbf{V}_i'\}, \boldsymbol{\varepsilon}\right] \\
&\le 2\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{\ell}\varepsilon_i\mathbf{V}_i'\right\|_2^2 \,\middle|\, \{\mathbf{V}_i'\}, \boldsymbol{\varepsilon}\right]} \\
&= 2\sqrt{\mathbb{E}\left[\mathbb{E}\left[\sum_{i,j=1}^{\ell}\varepsilon_i\varepsilon_j{\mathbf{V}_i'}^T\mathbf{V}_j' \mid \boldsymbol{\varepsilon}\right] \,\middle|\, \{\mathbf{V}_i'\}\right]} \\
&= 2\sqrt{\mathbb{E}\sum_{i=1}^{\ell}\left\|\mathbf{V}_i'\right\|_2^2}.
\end{aligned}$$

The first inequality is Jensen's, and the following equality holds because the components of the sequence $\{\mathbf{V}_i' - \mathbf{V}_i''\}$ are symmetric and independent. The next two manipulations are the triangle inequality and Jensen's inequality. This stage of the estimate is concluded by conditioning and using the orthogonality of the Rademacher variables. Next, the triangle inequality and the fact that $\mathbb{E}\left\|\mathbf{V}_1'\right\|_2^2 = \mathbb{E}\left\|\mathbf{V}_1\right\|_2^2$ allow us to further simplify the estimate of μ :

$$\mu \le 2\sqrt{\mathbb{E}\sum_{i=1}^{\ell}\left\|\mathbf{V}_i'\right\|_2^2} = 2\sqrt{\ell\mathbb{E}\left\|\mathbf{V}_1\right\|_2^2} \le \sigma.$$

We also calculate the quantity

$$2\max_{\mathbf{V}\in\mathcal{V}}\|\mathbf{V}\|_2 = \frac{2n}{\ell}\max_i \|\mathbf{X}_{(i)}\|_2\|\mathbf{Y}^{(i)}\|_2 \le B.$$

The tail bound given in the statement of the lemma follows from applying Lemma 4.4 with our estimates for B, σ^2, and μ. □

4.2 Linear Algebra notation and results

In subsequent chapters, we use the following partitioned compact SVD to state results for rectangular matrices $\mathbf{A}$ with $\operatorname{rank}(\mathbf{A}) = \rho$:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \begin{bmatrix} \overset{k}{\mathbf{U}_1} & \overset{\rho-k}{\mathbf{U}_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}. \tag{4.2.1}$$

Here, $\boldsymbol{\Sigma}_1$ contains the k largest singular values of $\mathbf{A}$ and the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ respectively span top k-dimensional left and right singular spaces of $\mathbf{A}$. The matrix $\mathbf{A}_k = \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^T$ is the optimal rank-k approximation to $\mathbf{A}$, and $\mathbf{A}_{\rho-k} = \mathbf{A} - \mathbf{A}_k = \mathbf{U}_2\boldsymbol{\Sigma}_2\mathbf{V}_2^T$. The Moore-Penrose pseudoinverse of $\mathbf{A}$ is denoted by $\mathbf{A}^\dagger$.

When $\mathbf{A}$ is a positive-semidefinite matrix, $\mathbf{U} = \mathbf{V}$ and (4.2.1) becomes the following partitioned eigenvalue decomposition:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T = \begin{bmatrix} \overset{k}{\mathbf{U}_1} & \overset{\rho-k}{\mathbf{U}_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix}. \tag{4.2.2}$$

The eigenvalues of an $n \times n$ symmetric matrix $\mathbf{A}$ are ordered $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$.

The orthoprojector onto the column space of a matrix $\mathbf{A}$ is written $\mathbf{P}_\mathbf{A}$ and satisfies

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

Let $\mathscr{S}$ be a k-dimensional subspace of $\mathbb{R}^n$ and $\mathbf{P}_\mathscr{S}$ denote the projection onto $\mathscr{S}$. Then the *coherence* of $\mathscr{S}$ is

$$\mu(\mathscr{S}) = \frac{n}{k}\max_i(\mathbf{P}_\mathscr{S})_{ii}.$$

The coherence of a matrix $\mathbf{U} \in \mathbb{R}^{n\times k}$ with orthonormal columns is the coherence of the subspace $\mathscr{S}$ which it spans:

$$\mu(\mathbf{U}) := \mu(\mathscr{S}) = \frac{n}{k}\max_i(\mathbf{P}_\mathscr{S})_{ii} = \frac{n}{k}\max_i(\mathbf{U}\mathbf{U}^T)_{ii}.$$

The kth column of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{(k)}$; the jth row is denoted by $\mathbf{A}^{(j)}$. The vector $\boldsymbol{e}_i$ is the ith element of the standard Euclidean basis (whose dimensionality will be clear from the context).

We often compare SPSD matrices using the semidefinite ordering. In this ordering, $\mathbf{A}$ is greater than or equal to $\mathbf{B}$, written $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \preceq \mathbf{A}$, when $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Each SPSD matrix $\mathbf{A}$ has a unique square root $\mathbf{A}^{1/2}$ that is also SPSD, has the same eigenspaces as $\mathbf{A}$, and satisfies $\mathbf{A} = \left(\mathbf{A}^{1/2}\right)^2$. The eigenvalues of an SPSD matrix $\mathbf{A}$ are arranged in weakly decreasing order: $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$. Likewise, the singular values of a rectangular matrix $\mathbf{A}$ with rank ρ are ordered $\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_\rho(\mathbf{A}) = \sigma_{\min}(\mathbf{A})$. The spectral norm of a matrix $\mathbf{B}$ is written $\|\mathbf{B}\|_2$; its Frobenius norm and trace are written $\|\mathbf{B}\|_\mathrm{F}$ and $\mathrm{Tr}(\mathbf{B})$, respectively. The notation $\|\cdot\|_\xi$ indicates that an expression holds for both $\xi = 2$ and $\xi = \mathrm{F}$.

4.2.1 Column-based low-rank approximation

The remainder of this thesis concerns low-rank matrix approximation algorithms: Chapter 5 provides bounds on the approximation errors of low-rank approximations that are formed using fast orthonormal transformations, and Chapter 6 provides bounds on the approximation errors of a class of low-rank approximations to SPSD matrices.

Both of these low-rank approximation schemes are amenable to interpretation as schemes wherein a matrix is projected onto a subspace spanned by some linear combination of its columns. The problem of providing a general framework for studying the error of these projection schemes is well studied [BMD09, HMT11, BDMI11]. The authors of these works have provided a set of so-called *structural* results: deterministic bounds on the spectral and Frobenius-norm approximation errors incurred by these projection schemes. Structural results allow us to relate the errors of low-rank approximations formed using projection schemes to the optimal errors $\|\mathbf{A} - \mathbf{A}_k\|_\xi$ for $\xi = 2, \mathrm{F}$.

Before stating the specific structural results that are used in the sequel, we review the necessary background material on low-rank matrix approximations that are restricted to lie within a particular subspace.

4.2.1.1 Matrix Pythagoras and generalized least-squares regression

Lemma 4.5 is the analog of Pythagoras' theorem in the matrix setting. A proof of this lemma can be found in [BDMI11]. Lemma 4.6 is an immediate corollary that generalizes the Eckart–Young theorem.

Lemma 4.5. *If* $\mathbf{X}\mathbf{Y}^T = \mathbf{0}$ *or* $\mathbf{X}^T\mathbf{Y} = \mathbf{0}$*, then*

$$\|\mathbf{X}+\mathbf{Y}\|_{\mathrm{F}}^2 = \|\mathbf{X}\|_{\mathrm{F}}^2 + \|\mathbf{Y}\|_{\mathrm{F}}^2$$

and

$$\max\{\|\mathbf{X}\|_2^2, \|\mathbf{Y}\|_2^2\} \leq \|\mathbf{X}+\mathbf{Y}\|_2^2 \leq \|\mathbf{X}\|_2^2 + \|\mathbf{Y}\|_2^2.$$

Lemma 4.6. *Given* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{C} \in \mathbb{R}^{m\times \ell}$*, for all* $\mathbf{X} \in \mathbb{R}^{\ell\times n}$

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{C}}\mathbf{A}\|_\xi^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_\xi^2$$

for both $\xi = 2$ *and* $\xi = \mathrm{F}$*.*

Proof. Write

$$\mathbf{A} - \mathbf{C}\mathbf{X} = (\mathbf{I} - \mathbf{P}_{\mathbf{C}})\mathbf{A} + (\mathbf{P}_{\mathbf{C}}\mathbf{A} - \mathbf{C}\mathbf{X})$$

and observe that

$$((\mathbf{I} - \mathbf{P}_{\mathbf{C}})\mathbf{A})^T(\mathbf{P}_{\mathbf{C}}\mathbf{A} - \mathbf{C}\mathbf{X}) = \mathbf{0},$$

so by Lemma 4.5,

$$\|\mathbf{A} - \mathbf{C}\mathbf{X}\|_\xi^2 \geq \|(\mathbf{I} - \mathbf{P}_{\mathbf{C}})\mathbf{A}\|_\xi^2.$$

□

4.2.1.2 Low-rank approximations restricted to subspaces

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$; a target rank $k < n$; another matrix $\mathbf{Y} \in \mathbb{R}^{m\times \ell}$, where $\ell > k$; and a choice of norm ξ ($\xi = 2$ or $\xi = \mathrm{F}$), we use the notation $\mathbf{\Pi}_{\mathbf{Y},k}^{\xi}(\mathbf{A})$ to refer to the matrix that lies in the column span of $\mathbf{Y}$, has rank k or less, and minimizes the ξ-norm error in approximating $\mathbf{A}$. More concisely, $\mathbf{\Pi}_{\mathbf{Y},k}^{\xi}(\mathbf{A}) = \mathbf{Y}\mathbf{X}^{\xi}$, where

$$\mathbf{X}^{\xi} = \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}^{\ell\times n}:\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A} - \mathbf{Y}\mathbf{X}\|_\xi^2.$$

The approximation $\mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})$ can be computed using the following three-step procedure:

1: Orthonormalize the columns of $\mathbf{Y}$ to construct a matrix $\mathbf{Q} \in \mathbb{R}^{m\times \ell}$.
2: Compute $\mathbf{X}_{\mathrm{opt}} = \arg\min_{\mathbf{X}\in\mathbb{R}^{\ell\times n},\ \mathrm{rank}(\mathbf{X})\leq k} \left\|\mathbf{Q}^T\mathbf{A} - \mathbf{X}\right\|_{\mathrm{F}}$.
3: Compute and return $\mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A}) = \mathbf{Q}\mathbf{X}_{\mathrm{opt}} \in \mathbb{R}^{m\times n}$.

There does not seem to be a similarly efficient algorithm for computing $\mathbf{\Pi}_{\mathbf{Y},k}^{2}(\mathbf{A})$.

The following result, which appeared as Lemma 18 in [BDMI11], both verifies the claim that this algorithm computes $\mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})$ and shows that $\mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})$ is a constant factor approximation to $\mathbf{\Pi}_{\mathbf{Y},k}^{2}(\mathbf{A})$.

Lemma 4.7. *[Lemma 18 in [BDMI11]] Given* $\mathbf{A} \in \mathbb{R}^{m\times n}$*,* $\mathbf{Y} \in \mathbb{R}^{m\times \ell}$*, and an integer* $k \leq \ell$*, the matrix* $\mathbf{Q}\mathbf{X}_{\mathrm{opt}} \in \mathbb{R}^{m\times n}$ *described above satisfies* $\mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A}) = \mathbf{Q}\mathbf{X}_{\mathrm{opt}}$*, can be computed in* $\mathrm{O}(mn\ell + (m+n)\ell^2)$ *time, and satisfies*

$$\left\|\mathbf{A} - \mathbf{\Pi}_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})\right\|_2^2 \leq 2\left\|\mathbf{A} - \mathbf{\Pi}_{\mathbf{Y},k}^{2}(\mathbf{A})\right\|_2^2.$$

4.2.2 Structural results for low-rank approximation

The following result, which appears as Lemma 7 in [BMD09], provides an upper bound on the residual error of the low-rank matrix approximation obtained via projections onto subspaces. The paper [HMT11] also supplies an equivalent result.

Lemma 4.8. *[Lemma 7 in [BMD09]] Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ. Fix k satisfying $0 \le k \le \rho$. Given a matrix $\mathbf{S} \in \mathbb{R}^{n\times \ell}$, with $\ell \ge k$, construct $\mathbf{Y} = \mathbf{AS}$. If $\mathbf{V}_1^T\mathbf{S}$ has full row-rank, then, for $\xi = 2, \mathrm{F}$,*

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_\xi^2 \le \|\mathbf{A} - \mathbf{\Pi}_{\mathbf{Y},k}^{\xi}(\mathbf{A})\|_\xi^2 \le \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\|_\xi^2. \tag{4.2.3}$$

In addition to this bound on the residual error, we use the following novel structural bound on the forward errors of low-rank approximants.

Lemma 4.9. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ. Fix k satisfying $0 \le k \le \rho$. Given a matrix $\mathbf{S} \in \mathbb{R}^{n\times \ell}$, where $\ell \ge k$, construct $\mathbf{Y} = \mathbf{AS}$. If $\mathbf{V}_1^T\mathbf{S}$ has full row-rank, then, for $\xi = 2, \mathrm{F}$,*

$$\|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_\xi^2 \le \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\|_\xi^2. \tag{4.2.4}$$

Proof. Observe that

$$(\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}_k)^T(\mathbf{P}_{\mathbf{Y}}\mathbf{A}_{\rho-k}) = \mathbf{0},$$

so Lemma 4.5 implies that

$$\|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_\xi^2 = \|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}_{\rho-k}\|_\xi^2 \le \|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}_k\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2.$$

Applying Lemma 4.6 with $\mathbf{X} = (\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T$, we see that

$$\begin{aligned}
\|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_\xi^2 &\le \|\mathbf{A}_k - \mathbf{Y}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2 \\
&= \|\mathbf{A}_k - \mathbf{A}_k\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T + \mathbf{A}_{\rho-k}\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2 \\
&= \|\mathbf{A}_k - \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T + \mathbf{A}_{\rho-k}\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2.
\end{aligned}$$

Since $\mathbf{V}_1^T\mathbf{S}$ has full row rank, $(\mathbf{V}_1^T\mathbf{S})(\mathbf{V}_1^T\mathbf{S})^\dagger = \mathbf{I}_k$. Recall that $\mathbf{A}_k = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$ and $\mathbf{A}_{\rho-k} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$. Consequently, the above inequality reduces neatly to the desired inequality

$$\begin{aligned}
\|\mathbf{A}_k - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_\xi^2 &\le \|\mathbf{A}_k - \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T + \mathbf{A}_{\rho-k}\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2 \\
&= \|\mathbf{A}_{\rho-k}\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\mathbf{V}_1^T\|_\xi^2 + \|\mathbf{A}_{\rho-k}\|_\xi^2 \\
&= \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\|_\xi^2.
\end{aligned}$$

□

4.2.2.1 A geometric interpretation of the sampling interaction matrix

Let $\mathbf{\Omega}_1 = \mathbf{V}_1^T\mathbf{S}$ and $\mathbf{\Omega}_2 = \mathbf{V}_2^T\mathbf{S}$ denote the interaction of the sampling matrix $\mathbf{S}$ with the top and bottom right-singular spaces of $\mathbf{A}$. It is evident from Lemmas 4.8 and 4.9 that the quality of the low-rank approximations depend upon the norm of the *sampling interaction matrix*

$$\mathbf{V}_2^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger = \mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger.$$

The smaller the spectral norm of the $\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger$ the more effective $\mathbf{S}$ is as a sampling matrix. To give the sampling interaction matrix a geometric interpretation, we first recall the definition of the sine between the range spaces of two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$:

$$\sin^2(\mathbf{M}_1, \mathbf{M}_2) = \|(\mathbf{I} - \mathbf{P}_{\mathbf{M}_1})\mathbf{P}_{\mathbf{M}_2}\|_2.$$

Note that this quantity is *not* symmetric: it measures how well the range of $\mathbf{M}_1$ captures that of $\mathbf{M}_2$ [GV96, Chapter 12].

Lemma 4.10. *Fix $\mathbf{A} \in \mathbb{R}^{m\times n}$, a target rank k, and $\mathbf{S} \in \mathbb{R}^{n\times \ell}$ where $\ell > k$. Assume $\mathbf{S}$ has orthonormal columns. Define*

$$\mathbf{\Omega}_1 = \mathbf{V}_1^T\mathbf{S} \quad \text{and} \quad \mathbf{\Omega}_2 = \mathbf{V}_2^T\mathbf{S}.$$

Then, if $\mathbf{\Omega}_1$ has full row-rank,

$$\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2 = \tan^2(\mathbf{S}, \mathbf{V}_1).$$

Proof. Since $\mathbf{V}_1$ and $\mathbf{S}$ have orthonormal columns, we see that

$$\begin{aligned}
\sin^2(\mathbf{S}, \mathbf{V}_1) &= \left\|(\mathbf{I} - \mathbf{S}\mathbf{S}^T)\mathbf{V}_1\mathbf{V}_1^T\right\|_2^2 \\
&= \left\|\mathbf{V}_1^T(\mathbf{I} - \mathbf{S}\mathbf{S}^T)\mathbf{V}_1\right\|_2 \\
&= \left\|\mathbf{I} - \mathbf{V}_1^T\mathbf{S}\mathbf{S}^T\mathbf{V}_1\right\|_2 \\
&= 1 - \lambda_k(\mathbf{V}_1^T\mathbf{S}\mathbf{S}^T\mathbf{V}_1) \\
&= 1 - \|\mathbf{\Omega}_1^\dagger\|^{-2}.
\end{aligned}$$

The second to last equality holds because $\mathbf{V}_1^T\mathbf{S}$ has k rows and we assumed it has full row-rank. Accordingly,

$$\tan^2(\mathbf{S}, \mathbf{V}_1) = \frac{\sin^2(\mathbf{S}, \mathbf{V}_1)}{1 - \sin^2(\mathbf{S}, \mathbf{V}_1)} = \|\mathbf{\Omega}_1^\dagger\|_2^2 - 1.$$

Now observe that

$$\begin{aligned}
\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2^2 &= \left\|(\mathbf{S}^T\mathbf{V}_1)^\dagger\mathbf{S}^T\mathbf{V}_2\mathbf{V}_2^T\mathbf{S}(\mathbf{V}_1^T\mathbf{S})^\dagger\right\|_2 \\
&= \left\|(\mathbf{S}^T\mathbf{V}_1)^\dagger(\mathbf{I} - \mathbf{S}^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{S})(\mathbf{V}_1^T\mathbf{S})^\dagger\right\|_2 \\
&= \left\|(\mathbf{S}^T\mathbf{V}_1)^\dagger\right\|_2^2 - 1 \\
&= \tan^2(\mathbf{S}, \mathbf{V}_1).
\end{aligned}$$

The second to last equality holds because of the fact that, for any matrix $\mathbf{M}$,

$$\left\|\mathbf{M}^\dagger(\mathbf{I} - \mathbf{M}\mathbf{M}^T)(\mathbf{M}^T)^\dagger\right\|_2 = \left\|\mathbf{M}^\dagger\right\|_2^2 - 1;$$

this identity can be established with a routine SVD argument. □

Thus, when $\mathbf{S}$ has orthonormal columns and $\mathbf{V}_1^T\mathbf{S}$ has full row-rank, $\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_2$ is the tangent of the largest angle between the range of $\mathbf{S}$ and the top right singular space spanned by $\mathbf{V}_1$. If $\mathbf{V}_1^T\mathbf{S}$ does not have full row-rank, then our derivation above shows that $\sin^2(\mathbf{S}, \mathbf{V}_1) = 1$, meaning that there is a vector in the eigenspace spanned by $\mathbf{V}_1$ which has no component in the space spanned by the sketching matrix $\mathbf{S}$.

We note that $\tan(\mathbf{S}, \mathbf{V}_1)$ also arises in the classical bounds on the convergence of the orthogonal iteration algorithm for approximating the top k-dimensional singular spaces of a matrix (see, e.g. [GV96, Theorem 8.2.2]).

Chapter 5

Low-rank approximation with subsampled unitary transformations

5.1 Introduction

In this chapter, we analyze the theoretical performance of a randomized low-rank approximation algorithm introduced in [WLRT08] and analyzed in [WLRT08, HMT11, NDT09]. Our analysis often provides sharper approximation bounds than those in [WLRT08, HMT11, NDT09]. We provide bounds on the residual and forward errors of this approximation algorithm in the spectral and Frobenius norms, and provide experimental evidence that this low-rank approximation algorithm performs as well as a more expensive low-rank approximation algorithm based upon projections onto uniformly distributed random subspaces[1]. Further, we provide approximation bounds for a variant of the algorithm that returns approximations with even lower rank.

The setting is as follows: fix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a target rank $k \leq \min\{m,n\}$. We would like to approximate $\mathbf{A}$ with a matrix $\mathbf{X}$ that has rank close to k, and we would like $\|\mathbf{A}-\mathbf{X}\|_\xi$ to be within a small multiplicative factor of the smallest error achievable when approximating $\mathbf{A}_k$ with a rank-k matrix, for $\xi = 2, F$.

It is well-known that the rank-k matrix $\mathbf{A}_k$ that minimizes both the Frobenius and the spectral-norm approximation errors can be calculated using the singular value decomposition (SVD) in $\mathrm{O}(mn\min\{m,n\})$ arithmetic operations, using classical so-called *direct* algorithms such as QR iteration or Jacobi iteration [GV96]. Computing the full SVD is expensive when $\mathbf{A}$ is a large matrix. In this case, it is often more efficient to use *iterative* projection methods (e.g. Krylov subspace methods) to obtain approximations to $\mathbf{A}_k$. It is difficult to state a precise guarantee for the number of arithmetic operations carried out by Krylov methods, but one iteration of a Krylov method requires $\Omega(mnk)$ operations (assuming $\mathbf{A}$ has no special structure which can be exploited to speed up the computation of matrix–vector products). To obtain even an accurate rank-1 approximation requires $\mathrm{O}(\log n)$ iterations [KW92]. Thus, an optimistic estimate for the number of operations required to compute approximate rank-k truncated SVDs using a Krylov method is $\Omega(mnk\log n)$.

Our discussion thus far has concerned only the arithmetic cost of computing truncated SVDs, but an equally or more important issue is that of the communication costs: bandwidth costs (proportional to the amount of times storage is accessed) and latency costs (proportional to the cost of transferring the information over a network or through the levels of a hierarchical

[1] The content of this chapter is adapted from the article [GB12] co-authored with Christos Boutsidis.

Algorithm 5.1: Randomized approximate truncated SVD

Input: an $m \times n$ matrix $\mathbf{A}$ and an $n \times \ell$ matrix $\mathbf{S}$, where ℓ is an integer in $[1, n]$.
Output: matrices $\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}}$ constituting the SVD of $\mathbf{P}_{\mathbf{AS}}\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$.

1: Let $\mathbf{Y} = \mathbf{AS}$.
2: Compute the QR decomposition $\mathbf{Y} = \mathbf{QR}$.
3: Compute the SVD of $\mathbf{Q}^T\mathbf{A} = \mathbf{W}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$.
4: Set $\tilde{\mathbf{U}} = \mathbf{QW}$.

Algorithm 5.2: Rank-k randomized approximate truncated SVD

Input: an $m \times n$ matrix $\mathbf{A}$, integers ℓ and k that satisfy $\ell > k$ and $k \in [1, n]$, and an $n \times \ell$ matrix $\mathbf{S}$.
Output: matrices $\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}}$ constituting the SVD of $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A}) = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$.

1: Let $\mathbf{Y} = \mathbf{AS}$.
2: Compute the QR decomposition $\mathbf{Y} = \mathbf{QR}$.
3: Compute the rank-k truncated SVD of $\mathbf{Q}^T\mathbf{A}$ to obtain $(\mathbf{Q}^T\mathbf{A})_k = \mathbf{W}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$.
4: Set $\tilde{\mathbf{U}} = \mathbf{QW}$.

memory system) [BDHS11]. If the algorithm is to be parallelized, then the complexity of the required information interchange must also be taken into account.

The randomized algorithms considered in this chapter, Algorithms 5.1 and 5.2, are of interest because they yield low-rank approximations after $\Omega(mnk \max\{\log n, \log k\})$ arithmetic operations and have low communication costs. In particular, each element of $\mathbf{A}$ is accessed only twice, and the algorithms are simple enough that they are amenable to straightforward parallelization. The guarantees provided are probabilistic, and allow one to trade off between the operation count of the algorithms and the accuracy and failure probabilities of the algorithms.

Both of the algorithms considered in this chapter are based on the intuition that, when $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ is randomly selected and ℓ is sufficiently larger than k, the range of the matrix $\mathbf{AS}$ "captures" the top k-dimensional left singular space of $\mathbf{A}$. When this phenomenon occurs, the low-rank matrix formed by projecting $\mathbf{A}$ onto the range of $\mathbf{AS}$ should be almost as accurate an approximation of $\mathbf{A}$ as is the optimal approximation $\mathbf{A}_k$:

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_\xi \approx \|\mathbf{A} - \mathbf{A}_k\|_\xi \quad \text{for } \xi = 2, \mathrm{F}.$$

Algorithm 5.1 computes exactly this approximation, $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$. Note that this approximation may have rank up to ℓ, which may be much larger than k. Algorithm 5.2 instead returns the approximation $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})$, which is guaranteed to have rank at most k.

Unlike classical iterative methods for approximating the trucated SVD, which use as many iterations as necessary to satisfy some convergence condition, Algorithms 5.1 and 5.2 use only one matrix–matrix product $\mathbf{AS}$ to generate an approximate basis for the top left singular space of $\mathbf{A}$. Accordingly, the quality of approximations obtained using either of these algorithms is more dependent on the properties of $\mathbf{A}$ itself and the sampling matrix $\mathbf{S}$ than is the quality of approximations derived from classical iterative methods. Thus it is important to supply theoretical guarantees on the errors of the algorithms that identify which properties of $\mathbf{A}$ affect the quality of the approximations, as well as to carry to empirical studies investigating the

influence of the choice of $\mathbf{S}$.

Recent years have produced a large body of research on designing random sampling matrices $\mathbf{S}$. Some proposals for $\mathbf{S}$ include: (i) every entry of $\mathbf{S}$ takes the values $+1, -1$ with equal probability [CW09, MZ11]; (ii) the entries of $\mathbf{S}$ are i.i.d. Gaussian random variables with zero mean and unit variance [HMT11]; (iii) the columns of $\mathbf{S}$ are chosen independently from the columns of the $m \times m$ identity matrix with probabilities that are proportional to the Euclidean length of the columns of $\mathbf{A}$ [FKV98, DKM06b]; and (iv) $\mathbf{S}$ is designed carefully such that $\mathbf{AS}$ can be computed in at most $O(nnz(\mathbf{A}))$ arithmetic operations, where $nnz(\mathbf{A})$ denotes the number of non-zero entries in $\mathbf{A}$ [CW12].

In this chapter we take $\mathbf{S}$ to be a subsampled randomized Hadamard transform (SRHT) matrix, i.e. $\mathbf{S}$ comprises a subset of the columns of a randomized Hadamard matrix (see Definitions 5.1 and 5.2 below). This choice for $\mathbf{S}$ was introduced in [AC06].

Definition 5.1 (Normalized Walsh–Hadamard Matrix)**.** *Fix an integer $n = 2^p$, for $p = 1, 2, 3, \ldots$. The (non-normalized) $n \times n$ matrix of the Hadamard-Walsh transform is defined recursively as,*

$$\mathbf{H}_n = \begin{bmatrix} \mathbf{H}_{n/2} & \mathbf{H}_{n/2} \\ \mathbf{H}_{n/2} & -\mathbf{H}_{n/2} \end{bmatrix}, \qquad \textit{with} \qquad \mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

The $n \times n$ normalized matrix of the Walsh–Hadamard transform is equal to $\mathbf{H} = n^{-\frac{1}{2}}\mathbf{H}_n \in \mathbb{R}^{n\times n}$.

Definition 5.2 (Subsampled Randomized Hadamard Transform (SRHT) matrix)**.** *Fix integers ℓ and $n = 2^p$ with $\ell < n$ and $p = 1, 2, 3, \ldots$. An SRHT matrix is an $\ell \times n$ matrix of the form*

$$\boldsymbol{\Theta} = \sqrt{\frac{n}{\ell}} \cdot \mathbf{RHD};$$

- $\mathbf{D} \in \mathbb{R}^{n\times n}$ *is a random diagonal matrix whose entries are independent random signs, i.e. random variables uniformly distributed on $\{\pm 1\}$.*
- $\mathbf{H} \in \mathbb{R}^{n\times n}$ *is a normalized Walsh–Hadamard matrix.*
- $\mathbf{R} \in \mathbb{R}^{\ell\times n}$ *is a subset of ℓ rows from the $n \times n$ identity matrix, where the rows are chosen uniformly at random and without replacement.*

The choice of $\mathbf{S}$ as an SRHT matrix is particularly practical because the highly structured nature of $\mathbf{S}$ can be exploited to reduce the time of computing $\mathbf{AS}$ from $O(mn\ell)$ to $O(mn \log_2 \ell)$.

Lemma 5.3 (Fast Matrix–Vector Multiplication, Theorem 2.1 in [AL08])**.** *Given $\mathbf{x} \in \mathbb{R}^n$ and $\ell < n$, one can construct $\boldsymbol{\Theta} \in \mathbb{R}^{\ell\times n}$ and compute $\boldsymbol{\Theta}\mathbf{x}$ in at most $2n \log_2(\ell + 1))$ operations.*

Beyond the SRHT. The SRHT is defined only when the matrix dimension is a power of two. An alternative option is to use other structured orthonormal randomized transforms such as the real Fourier transform (DFT), the discrete cosine transform (DCT) or the discrete Hartley transform (DHT) [WLRT08, NDT09, RT08, AMT10], whose entries are on the order of $n^{-1/2}$. None of these transforms place restrictions on the size of the matrix being approximated. With minimal effort, the results of this chapter can be extended to encompass these transforms. Specifically, the statements of Lemma 5.5 and Lemma 5.8 in this chapter would need to be modified slightly to account for the difference in the transform; essentially, the constants present in the statements of the Lemmas would change. These two lemmas isolate the effects of the particular choice of

$\mathbf{S}$ from the remainder of the arguments used in this chapter, so the changes would propagate, *mutatis mutandis*, throughout the remaining results in this chapter.

We note, further, that Algorithms 5.1 and 5.2 can be modified to use $\mathbf{Y} = (\mathbf{A}\mathbf{A}^T)^p\mathbf{A}\mathbf{S}$, where $p \geq 1$ is an integer, as an approximate basis for the top left singular space of $\mathbf{A}$. Approximations to $\mathbf{A}_k$ generated using this choice of $\mathbf{Y}$ are more accurate than those generated using our choice of $\mathbf{A}\mathbf{S}$, but one loses the speed conferred by taking $\mathbf{S}$ to be an SRHT matrix: after the first multiplication $\mathbf{A}\mathbf{S}$, all the matrix multiplications required to form $\mathbf{Y}$ are dense and unstructured.

Outline. In Section 5.2, we present a portion of our results on the quality of SRHT low-rank approximations and compare them to prior results in the literature. Section 5.3 presents new results on the application of SRHTs to general matrices and the approximation of matrix multiplication using SRHTs under the Frobenius norm. Section 5.4 contains the statements and proofs of our main results. We conclude the chapter with an experimental evaluation of the SRHT low-rank approximation algorithms in Section 5.5.

5.2 Low-rank matrix approximation using SRHTs

Using an SRHT matrix (see Definition 5.2), one can quickly construct low-rank approximations of a given matrix $\mathbf{A}$ using Algorithms 5.1 and 5.2. Our main results, Theorems 5.13 and 5.14, given in Section 5.4, respectively provide theoretical guarantees on the spectral and Frobenius-norm residual and forward errors of these approximations. To facilitate the comparison of our results with prior work, we highlight our residual error guarantees for Algorithm 5.1.

Theorem 5.4. *Assume n is a power of 2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ and fix an integer k satisfying $2 \leq k < \rho$. Let $0 < \epsilon < 1/3$ be an accuracy parameter, $0 < \delta < 1$ be a failure probability, and $C \geq 1$ be any specified constant. If $\mathbf{\Theta} \in \mathbb{R}^{\ell\times n}$ is an SRHT matrix and ℓ satisfies*

$$6C^2\epsilon^{-1}\left[\sqrt{k} + \sqrt{8\log(n/\delta)}\right]^2 \log(k/\delta) \leq \ell \leq n, \tag{5.2.1}$$

then the approximation $\tilde{\mathbf{A}}$ generated by Algorithm 5.1 with $\mathbf{S} = \mathbf{\Theta}^T$ satisfies

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\mathbf{\Theta}^T\mathbf{A}}\right\|_{\mathrm{F}} \leq (1 + 11\epsilon)\left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}$$

with probability at least $1 - \delta^{C^2\log(k/\delta)/4} - 7\delta$, and

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\mathbf{\Theta}^T\mathbf{A}}\right\|_2 \leq \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{\ell}}\right)\cdot\left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{\ell}}\cdot\left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}$$

with probability at least $1 - 5\delta$.

The Frobenius-norm bound in this theorem is slightly stronger than the best bound appearing in prior efforts [NDT09]. The spectral-norm bound on the residual error is often much smaller than the bounds presented in prior work and sheds light on an open question mentioned in [NDT09] and [HMT11]. We do not, however, claim that the error bounds provided are the tightest possible. Certainly the specific constants (11, 4, etc.) in the error estimates are not optimized.

We now present a detailed comparison of the guarantees given in Theorem 5.4 with those available in the existing literature.

5.2.1 Detailed comparison with prior work

The subsampled randomized Fourier transform (SRFT). The algorithm in Section 5.2 of [WLRT08], which was the first to use the idea of employing subsampled randomized orthogonal transforms to compute low-rank approximations to matrices, provides a spectral-norm error bound but replaces the SRHT with an SRFT, i.e. the matrix $\mathbf{H}$ of Definition 5.2 is replaced by a matrix where the (p,q)th entry is $\mathbf{H}_{pq} = \mathrm{e}^{-2\pi \mathrm{i}(p-1)(q-1)/n}$, where $\mathrm{i} = \sqrt{-1}$, i.e. $\mathbf{H}$ is the unnormalized discrete Fourier transform. Woolfe et al. [WLRT08, Equation 190] argue that, for any $\alpha > 1$, $\beta > 1$, if

$$\ell \geq \alpha^2 \beta \, (\alpha - 1)^{-1} \, (2k)^2,$$

then with probability at least $1 - 3/\beta$ ($\omega = \max\{m, n\}$),

$$\left\| \mathbf{A} - \tilde{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^T \right\|_2 \leq 2 \left(\sqrt{2\alpha - 1} + 1 \right) \cdot \left(\sqrt{\alpha\omega + 1} + \sqrt{\alpha\omega} \right) \cdot \left\| \mathbf{A} - \mathbf{A}_k \right\|_2 .$$

Here, $\tilde{\mathbf{U}}_k \in \mathbb{R}^{m \times k}$ contains orthonormal columns, as does $\tilde{\mathbf{V}}_k \in \mathbb{R}^{n \times k}$, while $\tilde{\mathbf{\Sigma}}_k \in \mathbb{R}^{k \times k}$ is diagonal with nonegative entries. These matrices can be computed deterministically from $\mathbf{A}\mathbf{\Theta}^T$ in $\mathrm{O}\left(k^2(m+n) + k\ell^2 \log \ell\right)$ time. Also, computing $\mathbf{Y} = \mathbf{A}\mathbf{\Theta}^T$ takes $\mathrm{O}\left(mn \log \ell\right)$ time.

The analysis of [WLRT08] applies when $\ell = \Omega(k^2)$, while the spectral-norm guarantee of Theorem 5.4 applies for potentially much smaller values of $\ell = \Omega(\max\{k \log k, \log(n) \log k\})$.

Nguyen et al. [NDT09]. An analysis of the Frobenius-norm error of an SRHT-based low-rank matrix approximation algorithm appeared in Nguyen et al. [NDT09]. Let δ be a probability parameter with $0 < \delta < 1$ and ϵ be an accuracy parameter with $0 < \epsilon < 1$. Then, Nguyen et al. show that in order to get a rank-k matrix $\tilde{\mathbf{A}}_k$ satisfying

$$\left\| \mathbf{A} - \tilde{\mathbf{A}}_k \right\|_{\mathrm{F}} \leq (1 + \epsilon) \cdot \left\| \mathbf{A} - \mathbf{A}_k \right\|_{\mathrm{F}}$$

and

$$\left\| \mathbf{A} - \tilde{\mathbf{A}}_k \right\|_2 \leq \left(2 + \sqrt{2n/\ell} \right) \cdot \left\| \mathbf{A} - \mathbf{A}_k \right\|_2$$

with probability of success at least $1 - 5\delta$, one requires

$$\ell = \Omega\left(\epsilon^{-1} \max\{k, \sqrt{k} \log(2n/\delta)\} \cdot \max\{\log k, \log(3/\delta)\} \right).$$

Theorem 5.4 gives a tighter spectral-norm error bound when $\left\| \mathbf{A} - \mathbf{A}_k \right\|_{\mathrm{F}} \ll (n/\log(\rho/\delta))^{1/2} \cdot \left\| \mathbf{A} - \mathbf{A}_k \right\|_2$. It also provides an equivalent Frobenius-norm error bound with a comparable failure probability for a smaller number of samples. Specifically, if

$$\begin{aligned} \ell &\geq 528 \epsilon^{-1} [\sqrt{k} + \sqrt{8 \log(8n/\delta)}]^2 \log(8k/\delta) \\ &= \Omega\left(\epsilon^{-1} \max\{k, \log(n/\delta)\} \cdot \max\{\log k, \log(1/\delta)\} \right), \end{aligned}$$

then the Frobenius-norm bound in Theorem 5.4 ensures that, with probability at least $1 - 8\delta$, the approximation satisfies $\left\| \mathbf{A} - \tilde{\mathbf{A}} \right\|_{\mathrm{F}} \leq (1 + \epsilon) \cdot \left\| \mathbf{A} - \mathbf{A}_k \right\|_{\mathrm{F}}$.

In [HMT11] and [NDT09], the authors left as a subject for future research the explanation of a curious experimental phenomenon: when the singular values decay according to power laws, the SRHT low-rank approximation algorithm empirically achieves relative-error spectral norm approximations. Our spectral norm result provides an explanation of this phenomenon:

when the singular values of $\mathbf{A}$ decay fast enough, as in power law decay, one has $\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}} = \Theta(1)\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_2$. In this case, when ℓ is chosen to satisfy

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

our spectral-norm bound assures us that $\left\|\mathbf{A}-\tilde{\mathbf{A}}\right\|_2 \le \mathrm{O}(1)\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_2$ with probability of at least $1-8\delta$, thus predicting the observed empirical behavior of the algorithm.

The approximation scheme addressed in [NDT09] generates low-rank approximations with rank at most k, while the bounds in Theorem 5.4 are for approximations using Algorithm 5.2, which may result in an approximation with a rank larger than k. However, the comparisons above also hold, as stated, for approximations $\tilde{\mathbf{A}}$ generated using Algorithm 5.2, as can be verified directly from Theorems 5.13 and 5.14.

Halko et al. [HMT11]. Halko et al. [HMT11] consider the performance of Algorithm 5.1 when $\mathbf{S}=\mathbf{\Theta}^T$ is an SRHT matrix, and conclude that if ℓ satisfies

$$4\left[\sqrt{k}+\sqrt{8\log(kn)}\right]^2\log(k)\le\ell\le n, \tag{5.2.2}$$

then, for both $\xi = 2, \mathrm{F}$,

$$\|\mathbf{A}-\tilde{\mathbf{A}}\|_\xi \le \left(1+\sqrt{7n/\ell}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_\xi,$$

with probability at least $1-\mathrm{O}(1/k)$. Our Frobenius-norm bound is always tighter than the Frobenius-norm bound given here. To compare the spectral-norm bounds, note that our spectral-norm bound is on the order of

$$\max\left\{\sqrt{\frac{\log(\rho/\delta)\log(n/\delta)}{\ell}}\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_2, \sqrt{\frac{\log(\rho/\delta)}{\ell}}\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}\right\}. \tag{5.2.3}$$

If the residual spectrum $\mathbf{A}$ (the set of singular values smaller than $\sigma_k(\mathbf{A})$) is constant, or more generally decays slowly, then the spectral-norm result in [HMT11] is perhaps optimal. But when $\mathbf{A}$ is rank-deficient or the singular values of $\mathbf{A}$ decay fast, the spectral-norm bound in Theorem 5.4 is more useful. Specifically, if

$$\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}} \ll \sqrt{\frac{n}{\log(\rho/\delta)}}\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_2,$$

then when ℓ is chosen according to Theorem 5.4, the quantity in (5.2.3) is much smaller than $\sqrt{7n/\ell}\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_2$.

We were able to obtain this improved bound by using the results in Section 5.3.1, which allow one to take into account decay in the spectrum of $\mathbf{A}$. Finally, notice that our theorem makes explicit the intuition that the probability of failure can be driven to zero independently of the target rank k by increasing ℓ.

Two alternative approximate SVD algorithms. Instead of an SRHT matrix, one can take $\mathbf{S}$ in Algorithms 5.1 and 5.2 to be a matrix of i.i.d. standard Gaussian random variables. One gains theoretically and often empirically better worse-case tradeoffs between the number of samples taken, the failure probability, and the error guarantees. The SRHT algorithms are still faster,

since a matrix multiplication with a Gaussian matrix requires $O(mn\ell)$ time (assuming $\mathbf{A}$ is dense and unstructured). One can also take $\mathbf{S}$ to be a matrix of i.i.d. random signs (± 1 with equal probability). In many ways, this is analogous to the Gaussian algorithm: in both cases $\mathbf{S}$ is a matrix of i.i.d. subgaussian random variables. Consequently, we expect this algorithm to have the same advantages and disadvantages relative to the SRHT algorithm. We now compare the best available performance bounds for these schemes to our SRHT performance bounds.

We use the notion of the stable rank of a matrix,

$$\mathrm{sr}(\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{F}}^2 / \|\mathbf{A}\|_2^2,$$

to capture the decay of the singular values of $\mathbf{A}$. As can be seen by considering a matrix with a flat singular spectrum, in general the stable rank is no smaller than the rank.

When $\ell > k + 4$, Theorem 10.7 and Corollary 10.9 in [HMT11] imply that, when using Gaussian sampling in Algorithm 5.1, with probability at least $1 - 2 \cdot 32^{-(\ell-k)} - \mathrm{e}^{\frac{-(\ell-k+1)}{2}}$,

$$\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_{\mathrm{F}} \leq \left(1 + 32\frac{\sqrt{3k} + \mathrm{e}\sqrt{\ell}}{\sqrt{\ell - k + 1}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}$$

and with probability at least $1 - 3e^{-(\ell-k)}$,

$$\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2 \leq \left(1 + 16\sqrt{1 + \frac{k}{\ell - k}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \frac{8\sqrt{\ell}}{\ell - k + 1} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

Comparing to the guarantees of Theorem 5.4 we see that the two bounds just stated suggest that with the same number of samples, Gaussian low-rank approximations outperform SRHT-based low-rank approximations. In particular, the spectral-norm bound guarantees that if $\mathrm{sr}\left(\mathbf{A} - \mathbf{A}_k\right) \leq k$, that is, $\left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \leq \sqrt{k}\left\|\mathbf{A} - \mathbf{A}_k\right\|_2$, then the Gaussian version of Algorithm 5.1 requires $O(k/\epsilon^2)$ samples to return a $17 + \epsilon$ constant factor spectral-norm error approximation with high probability. Similarly, the Frobenius-norm bound guarantees that the same number of samples returns a $1 + 32\epsilon$ constant factor Frobenius-norm error approximation with high probability. Neither the spectral nor Frobenius bounds given in Theorem 5.4 for SRHT-based low-rank approximations apply for this few samples.

The paper [MZ11] does not consider the Frobenius-norm error of the random sign low-rank approximation algorithm, but Remark 4 in [MZ11] shows that when $\ell = O(k/\epsilon^4 \log(1/\delta))$, for $0 < \delta < 1$, and $\mathrm{sr}\left(\mathbf{A} - \mathbf{A}_k\right) \leq k$, Algorithm 5.1 ensures that with probability at least $1 - \delta$,

$$\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2 \leq (1 + \epsilon)\left\|\mathbf{A} - \mathbf{A}_k\right\|_2.$$

To compare our results with those stated in [HMT11, MZ11] we assume that $k \gg \log(n/\delta)$ so that $\ell > k \log k$ suffices for Theorem 5.4 to apply. Then, in order to acquire a $4 + \epsilon$ relative error bound from Theorem 5.4, it suffices that

$$\ell \geq \mathrm{C}'\epsilon^{-2} k \log(\rho/\delta) \quad \text{and} \quad \mathrm{sr}\left(\mathbf{A} - \mathbf{A}_k\right) \leq \mathrm{C}'k,$$

where C' is an explicit constant no larger than 6.

We see, therefore, that the Gaussian and random sign approximation versions of Algorithm 5.1 return $17 + \epsilon$ and $1 + \epsilon$ relative spectral error approximations, respectively, when ℓ is on the order of k and the relatively weak spectral decay condition $\mathrm{sr}\left(\mathbf{A} - \mathbf{A}_k\right) \leq k$ is satisfied,

while our bound for the SRHT version of Algorithm 5.1 requires $\ell > k\log(\rho/\delta)$ and the spectral decay condition

$$\mathrm{sr}\left(\mathbf{A}-\mathbf{A}_k\right) \leq \mathrm{C}'k$$

to ensure a $6+\epsilon$ relative spectral error approximation. We note that the SRHT algorithm can be used to obtain relative spectral error approximations of matrices with arbitrary stable rank at the cost of increasing ℓ; the same is of course true for the Gaussian and random sign algorithms.

The bounds for the SRHT, Gaussian, and random sign low-rank approximation algorithms differ in two significant ways. First, there are logarithmic factors in the spectral-norm error bound for the SRHT algorithm that are absent from the corresponding bounds for the Gaussian and random sign algorithms. Second, the spectral-norm bound for the SRHT algorithm applies only when $\ell > k\log(\rho/\delta)$, while the corresponding bounds for the Gaussian and random sign algorithms apply when ℓ is on the order of k. These disparities may reflect a fundamental tradeoff between the structure and randomness of the sampling matrix $\mathbf{S}$. Assuming $\mathbf{A}$ is dense and unstructured, the highly structured nature of SRHT matrices makes it possible to calculate $\mathbf{AS}$ much faster than when Gaussian or random sign sampling matrices are used, but this moves us away from the very nice isotropic randomness present in the Gaussian $\mathbf{S}$ and the similarly nice properties of a matrix of i.i.d subgaussian random variables, thus resulting in slacker bounds which require more samples.

5.3 Matrix computations with SRHT matrices

An important ingredient in analyzing the performance of Algorithms 5.1 and 5.2 is understanding how an SRHT changes the spectrum of a matrix after postmultiplication: given a matrix $\mathbf{A}$ and an SRHT matrix $\mathbf{\Theta}$, how are the singular values of $\mathbf{A}$ and $\mathbf{A\Theta}^T$ related?

To be more precise, Lemma 4.8 suggests that one path towards establishing the efficacy of SRHT-based low-rank approximations lies in understanding how the SRHT perturbs the singular values of matrices. To see this, we repeat the statement of the lemma here. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ and recall the following partitioning of its SVD:

$$\mathbf{A} = \mathbf{U\Sigma V}^T = \begin{array}{c} \begin{matrix} k & \rho-k \end{matrix} \\ \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \end{array} \begin{array}{c} \begin{matrix} k & \rho-k \end{matrix} \\ \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \end{array} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}. \tag{5.3.1}$$

Fix k satisfying $0 \leq k \leq \rho$. Given a matrix $\mathbf{S} \in \mathbb{R}^{n\times\ell}$, with $\ell \geq k$, Lemma 4.8 states that if $\mathbf{V}_1^T\mathbf{S}$ has full row-rank, then for $\xi = 2, \mathrm{F}$,

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

Note that $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$ is exactly the low-rank approximation of $\mathbf{A}$ returned by Algorithm 5.1.

Now take $\mathbf{S} = \mathbf{\Theta}^T$ to be an SRHT matrix and observe that if the product $\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\left(\mathbf{V}_1^T\mathbf{\Theta}^T\right)^\dagger$ has small norm, then the residual error of the approximant $\mathbf{P}_{\mathbf{A\Theta}^T}\mathbf{A}$ is small. The norm of this product is small when the norm of the perturbed matrix $\left(\mathbf{V}_1^T\mathbf{\Theta}^T\right)^\dagger$ is small, because

$$\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\|_\xi^2 \leq \|\mathbf{\Sigma}_2\|_\xi^2\|\mathbf{V}_2^T\mathbf{\Theta}^T\|_2^2\left\|(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\right\|_2^2. \tag{5.3.2}$$

The matrix $(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger$ has small norm precisely when the singular values of $\mathbf{V}_1^T\mathbf{\Theta}^T$ are close to those of $\mathbf{V}_1$. This strategy is developed in [HMT11] to supply bounds for the SRHT low-rank

approximation algorithm, and relies upon knowledge of how multiplication with SRHT matrices perturbs the singular values of matrices with orthonormal rows.

The main contribution of this chapter is the realization that one can take advantage of the decay in the singular values of $\mathbf{A}$ encoded in $\mathbf{\Sigma}_2$ to obtain sharper results. In view of the fact that

$$\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\|_\xi^2 \leq \|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\|_\xi^2 \left\|(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\right\|_2^2, \tag{5.3.3}$$

we can additionally consider the behavior of the singular values of $\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T$. It is clear that (5.3.3) provides a tighter bound than (5.3.2): for example, when $\xi = \mathrm{F}$,

$$\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\|_\mathrm{F}^2 \leq \|\mathbf{\Sigma}_2\|_\mathrm{F}^2\|\mathbf{V}_2^T\mathbf{\Theta}^T\|_2^2,$$

and the quantity on the left-hand side is potentially much smaller than that on the right-hand side.

In the remainder of this section, we refer to matrices with more columns than rows as "fat"; similarly, we refer to matrices with more rows than columns as "tall."

5.3.1 SRHTs applied to orthonormal matrices

In this subsection, we collect known results on how the singular values of a matrix with orthonormal rows are affected by postmultiplication by an SRHT matrix. These results will be used to control the quantity $\left\|(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\right\|_2^2$ in (5.3.3).

It has recently been shown by Tropp [Tro11b] that, if the SRHT matrix is of sufficiently large dimensions, post-multiplying a fat matrix with orthonormal rows by an SRHT matrix preserves the singular values of the orthonormal matrix, with high probability, up to a small multiplicative factor. The following lemma is essentially a reparametrization of Theorem 3.1 in [Tro11b], but we include a full proof for completeness.

Lemma 5.5 (The SRHT preserves geometry)**.** *Assume n is a power of 2. Let $\mathbf{V}^T \in \mathbb{R}^{k\times n}$ have orthonormal rows. Choose parameters $0 < \epsilon < 1/3$ and $0 < \delta < 1$. Construct an SRHT matrix $\mathbf{\Theta} \in \mathbb{R}^{\ell\times n}$ with ℓ satisfying*

$$6\epsilon^{-1}\left[\sqrt{k} + \sqrt{8\log(n/\delta)}\right]^2 \log(k/\delta) \leq \ell \leq n. \tag{5.3.4}$$

Then, with probability at least $1 - 3\delta$, for all $i = 1, \ldots, k$,

$$\sqrt{1-\sqrt{\epsilon}} \leq \sigma_i(\mathbf{V}^T\mathbf{\Theta}^T) \leq \sqrt{1+\sqrt{\epsilon}}$$

and

$$\left\|(\mathbf{V}^T\mathbf{\Theta}^T)^\dagger - (\mathbf{V}^T\mathbf{\Theta}^T)^T\right\|_2 \leq 1.54\sqrt{\epsilon}.$$

By the definition of an SRHT matrix, $\mathbf{V}^T\mathbf{\Theta}^T = \sqrt{n/\ell}\cdot(\mathbf{RHDV})^T$. Tropp [Tro11b] argues that the above lemma follows from a more fundamental fact: if $\mathbf{V}$ has orthonormal columns, then the rows of the product $\mathbf{HDV}$ all have roughly the same norm. That is, premultiplication by $\mathbf{HD}$ equalizes the row norms of an orthonormal matrix. See also [AC06].

Lemma 5.6 (Row norms, Lemma 3.3 in [Tro11b])**.** *Assume n is a power of 2. Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ have orthonormal columns; let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a normalized Hadamard matrix; and let $\mathbf{D} \in \mathbb{R}^{n\times n}$*

be a diagonal matrix of independent Rademacher random variables. Choose a failure probability $0 < \delta < 1$. *Then, with probability at least* $1 - \delta$,

$$\max_{i=1,\dots,n} \left\| (\mathbf{HDV})^{(i)} \right\|_2 \leq \sqrt{\frac{k}{n}} + \sqrt{\frac{8\log(n/\delta)}{n}}.$$

Recall that $(\mathbf{HDV})^{(i)}$ *denotes the* i*th row of the matrix* $\mathbf{HDV} \in \mathbb{R}^{n\times k}$.

To prove Lemma 5.5 we need one more result on uniform random sampling (without replacement) of rows from thin matrices with orthonormal columns. The following lemma is a reparameterization of Lemma 3.4 of [Tro11b].

Lemma 5.7 (Uniform sampling without replacement from an orthonormal matrix)**.** *Let* $\mathbf{W} \in \mathbb{R}^{n\times k}$ *have orthonormal columns. Let* $0 < \epsilon < 1$ *and* $0 < \delta < 1$. *Let* $M := n \cdot \max_{i=1,\dots,n} \left\| \mathbf{W}^{(i)} \right\|_2^2$. *Let* ℓ *be an integer such that*

$$6\epsilon^{-2} M \log(k/\delta) \leq \ell \leq n\,. \tag{5.3.5}$$

Let $\mathbf{R} \in \mathbb{R}^{\ell\times n}$ *be a matrix which consists of a subset of* ℓ *rows from* $\mathbf{I}_n$ *where the rows are chosen uniformly at random and without replacement. Then, with probability at least* $1 - 2\delta$, *for* $i \in [k]$:

$$\sqrt{\frac{\ell}{n}} \cdot \sqrt{1-\epsilon} \leq \sigma_i(\mathbf{RW}) \leq \sqrt{1+\epsilon} \cdot \sqrt{\frac{\ell}{n}}.$$

Proof. Apply Lemma 3.4 of [Tro11b] with the following choice of parameters: $\ell = \alpha M \log(k/\delta)$, $\alpha = 6/\epsilon^2$, and $\delta_{\text{tropp}} = \eta = \epsilon$. Here, ℓ, α, M, k, η are the variables named in Lemma 3.4 of [Tro11b], and δ_{tropp} plays the role of an error parameter named δ in Lemma 3.4 of [Tro11b]. The variables ϵ and δ are from our Lemma. The choice of ℓ proportional to $\log(k/\delta)$ rather than proportional to $\log(k)$, as in the original statement of Lemma 3.4, is what results in a probability proportional to δ instead of k; this can easily be seen by tracing the modified choice of ℓ through the proof of Lemma 3.4. □

Proof of Lemma 5.5. To obtain the bounds on the singular values, we combine Lemmas 5.6 and 5.7. More specifically, apply Lemma 5.7 with $\mathbf{W} = \mathbf{HDV}$ and use the bound for M from Lemma 5.6. Then, the bound on ℓ in (5.3.5), the bound on the singular values in Lemma 5.7, and a union bound together establish that, with probability at least $1 - 3\delta$,

$$\sqrt{\frac{\ell}{n}} \cdot \sqrt{1-\epsilon} \leq \sigma_i(\mathbf{RHDV}) \leq \sqrt{1+\epsilon} \cdot \sqrt{\frac{\ell}{n}} \quad \text{for all } i \in [k].$$

Now, multiply this inequality with $\sqrt{n/\ell}$ and recall the definition $\mathbf{\Theta} = \sqrt{n/\ell} \cdot \mathbf{RHD}$ to obtain

$$\sqrt{1-\epsilon} \leq \sigma_i(\mathbf{\Theta V}) \leq \sqrt{1+\epsilon} \quad \text{for all } i \in [k]. \tag{5.3.6}$$

Replacing ϵ with $\sqrt{\epsilon}$ and using the bound on ℓ given in (5.3.4) concludes the proof of the first inequality in Lemma 5.5.

The second bound in the lemma follows from the first bound after a simple algebraic manipulation. Let $\mathbf{X} = \mathbf{V}^T\mathbf{\Theta}^T \in \mathbb{R}^{k\times\ell}$ have SVD $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^T$ where $\mathbf{\Sigma}_\mathbf{X} \in \mathbb{R}^{k\times k}$ is invertible, then

$$\left\| (\mathbf{V}^T\mathbf{\Theta}^T)^\dagger - (\mathbf{V}^T\mathbf{\Theta}^T)^T \right\|_2 = \left\| \mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-1}\mathbf{U}_\mathbf{X}^T - \mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{U}_\mathbf{X}^T \right\|_2 = \left\| \mathbf{V}_\mathbf{X}(\mathbf{\Sigma}_\mathbf{X}^{-1} - \mathbf{\Sigma}_\mathbf{X})\mathbf{U}_\mathbf{X}^T \right\|_2 = \left\| \mathbf{\Sigma}_\mathbf{X}^{-1} - \mathbf{\Sigma}_\mathbf{X} \right\|_2,$$

by unitary invariance of the spectral norm. Let $\mathbf{Y} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{k\times k}$. Then, for all $i = 1, \ldots, k$, $\mathbf{Y}_{ii} = (1 - \sigma_i^2(\mathbf{X}))/\sigma_i(\mathbf{X})$. We conclude the proof with the series of estimates

$$\begin{aligned}\|\mathbf{Y}\|_2 &= \max_{1\le i\le k} \left|\mathbf{Y}_{ii}\right| = \max_{1\le i\le k} \left|\frac{1-\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}\right| \\ &= \max\left\{\frac{\left|1-\sigma_1^2(\mathbf{X})\right|}{\sigma_1(\mathbf{X})}, \frac{\left|1-\sigma_k^2(\mathbf{X})\right|}{\sigma_k(\mathbf{X})}\right\} \\ &\le \frac{\sqrt{\epsilon}}{\sqrt{1-\sqrt{\epsilon}}} \le 1.54\sqrt{\epsilon}.\end{aligned}$$

The second to last inequality follows from the lower bound in (5.3.6). □

5.3.2 SRHTs applied to general matrices

Recall our observation that the inequality

$$\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T (\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger\|_\xi^2 \le \|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T\|_\xi^2 \left\|(\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger\right\|_2^2, \tag{5.3.7}$$

together with Lemma 4.8, allows us to bound the errors of SRHT-based low-rank approximations. In the previous subsection, we collected the results necessary to bound the term $\left\|(\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger\right\|_2^2$. In this subsection, we present new results on the perturbative effects of SRHT multiplication. These allow us to bound the quantities $\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T\|_\xi^2$.

Our main tool is a generalization of Lemma 5.6 that states that the maximum column norm of a matrix to which an SRHT has been applied is, with high probability, not much larger than the root mean square of the column norms of the original matrix.

Lemma 5.8 (SRHT equalization of column norms)**.** *Suppose that* $\mathbf{A}$ *is a matrix with* n *columns, where* n *is a power of 2. Let* $\mathbf{H} \in \mathbb{R}^{n\times n}$ *be a normalized Walsh–Hadamard matrix, and* $\mathbf{D} \in \mathbb{R}^{n\times n}$ *a diagonal matrix of Rademacher random variables. Then for every* $t \ge 0$*,*

$$\mathbb{P}\left\{\max_{j=1,\ldots,n} \|(\mathbf{A}\mathbf{D}\mathbf{H}^T)_{(j)}\|_2 \le \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\|\mathbf{A}\|_2\right\} \ge 1 - n\cdot \mathrm{e}^{-t^2/8}.$$

Recall that $(\mathbf{A}\mathbf{D}\mathbf{H}^T)_{(j)}$ *denotes the* j*th column of* $\mathbf{A}\mathbf{D}\mathbf{H}^T$*.*

Proof. Our proof of Lemma 5.8 is essentially that of Lemma 5.6 in [Tro11b], with attention paid to the fact that $\mathbf{A}$ is no longer assumed to have orthonormal columns.

Lemma 5.8 follows immediately from the observation that the norm of any one column of $\mathbf{A}\mathbf{D}\mathbf{H}^T$ is a convex Lipschitz function of a Rademacher vector. Consider the norm of the jth column of $\mathbf{A}\mathbf{D}\mathbf{H}^T$ as a function of $\boldsymbol{\varepsilon}$, where $\mathbf{D} = \operatorname{diag}(\boldsymbol{\varepsilon})$:

$$f_j(\boldsymbol{\varepsilon}) = \|\mathbf{A}\mathbf{D}\mathbf{H}^T\mathbf{e}_j\| = \left\|\mathbf{A}\operatorname{diag}(\boldsymbol{\varepsilon})\mathbf{h}_j\right\|_2 = \left\|\mathbf{A}\operatorname{diag}(\mathbf{h}_j)\boldsymbol{\varepsilon}\right\|_2,$$

where $\mathbf{h}_j$ denotes the jth column of $\mathbf{H}^T$. Evidently f_j is convex. Furthermore,

$$|f_j(\mathbf{x}) - f_j(\mathbf{y})| \le \left\|\mathbf{A}\operatorname{diag}(\mathbf{h}_j)(\mathbf{x}-\mathbf{y})\right\|_2 \le \|\mathbf{A}\|_2 \left\|\operatorname{diag}(\mathbf{h}_j)\right\|_2 \left\|\mathbf{x}-\mathbf{y}\right\|_2 = \frac{1}{\sqrt{n}}\|\mathbf{A}\|_2 \left\|\mathbf{x}-\mathbf{y}\right\|_2,$$

where we used the triangle inequality and the fact that $\left\|\operatorname{diag}(\mathbf{h}_j)\right\|_2 = \|\mathbf{h}_j\|_\infty = n^{-1/2}$. Thus f_j is convex and Lipschitz with Lipschitz constant of at most $n^{-1/2}\|\mathbf{A}\|_2$.

We calculate

$$\begin{aligned}\mathbb{E}\left[f_j(\boldsymbol{\varepsilon})\right] \le \mathbb{E}\left[f_j(\boldsymbol{\varepsilon})^2\right]^{1/2} &= \left[\mathrm{Tr}\left(\mathbf{A}\,\mathrm{diag}(\mathbf{h}_j)\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^*\right]\mathrm{diag}(\mathbf{h}_j)\,\mathbf{A}^T\right)\right]^{1/2} \\ &= \left[\mathrm{Tr}\left(\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right]^{1/2} \\ &= \frac{1}{\sqrt{n}}\,\|\mathbf{A}\|_{\mathrm{F}}\,.\end{aligned}$$

It now follows from Lemma 4.1 that, for all $j = 1, 2, \ldots, n$, the norm of the jth column of $\mathbf{ADH}^T$ satisfies the tail bound

$$\mathbb{P}\left\{\left\|\mathbf{ADH}^T\mathbf{e}_j\right\|_2 \ge \frac{1}{\sqrt{n}}\,\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\,\|\mathbf{A}\|_2\right\} \le \mathrm{e}^{-t^2/8}.$$

Taking a union bound over all columns of $\mathbf{ADH}^T$, we conclude that

$$\mathbb{P}\left\{\max_{j=1,\ldots,n}\left\|(\mathbf{ADH}^T)_{(j)}\right\|_2 \ge \frac{1}{\sqrt{n}}\,\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\,\|\mathbf{A}\|_2\right\} \le n\cdot\mathrm{e}^{-t^2/8}.$$

□

Our next lemma shows that the SRHT does not substantially increase the spectral norm of a matrix.

Lemma 5.9 (SRHT-based subsampling in the spectral norm)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ, and assume that n is a power of 2. For some $r < n$, let $\boldsymbol{\Theta} \in \mathbb{R}^{\ell\times n}$ be an SRHT matrix. Fix a failure probability $0 < \delta < 1$. Then,*

$$\mathbb{P}\left\{\left\|\mathbf{A}\boldsymbol{\Theta}^T\right\|_2^2 \le 5\,\|\mathbf{A}\|_2^2 + \frac{\log(\rho/\delta)}{\ell}\left(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{A}\|_2\right)^2\right\} \ge 1 - 2\delta.$$

To establish Lemma 5.9, we use the upper Chernoff bound for sums of matrices stated in Lemma 4.2.

Proof of Lemma 5.9. Write the SVD of $\mathbf{A}$ as $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{\rho\times\rho}$, and observe that the spectral norm of $\mathbf{A}\boldsymbol{\Theta}^T$ is the same as that of $\sqrt{n/\ell}\cdot\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Theta}^T$.

We control the norm of $\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Theta}^T$ by considering the maximum singular value of its Gram matrix. Define $\mathbf{M} = \boldsymbol{\Sigma}\mathbf{V}^T\mathbf{DH}^T$, so that $\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Theta}^T = \sqrt{n/\ell}\cdot\mathbf{MR}^T$, and let $\mathbf{G}$ be the $\rho\times\rho$ Gram matrix of $\mathbf{MR}^T$:

$$\mathbf{G} = \mathbf{MR}^T(\mathbf{MR}^T)^T.$$

Evidently

$$\lambda_{\max}(\mathbf{G}) = \frac{\ell}{n}\left\|\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Theta}^T\right\|_2^2. \tag{5.3.8}$$

Recall that $\mathbf{M}_{(j)}$ denotes the jth column of $\mathbf{M}$. If we denote by C the random set of r coordinates to which $\mathbf{R}$ restricts, then

$$\mathbf{G} = \sum\nolimits_{j\in C}\mathbf{M}_{(j)}\mathbf{M}_{(j)}^T.$$

Thus $\mathbf{G}$ is a sum of r random matrices $\mathbf{X}_1, \ldots, \mathbf{X}_r$ sampled without replacement from the set $\mathscr{X} = \{\mathbf{M}_{(j)}\mathbf{M}_{(j)}^T : j = 1, 2, \ldots, n\}$. There are two sources of randomness in $\mathbf{G}$: the subsampling matrix $\mathbf{R}$ and the Rademacher random variables on the diagonal of $\mathbf{D}$.

Set

$$B = \frac{1}{n}\left(\|\mathbf{\Sigma}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{\Sigma}\|_2\right)^2$$

and let E be the event

$$\max_{j=1,\dots,n} \|\mathbf{M}_{(j)}\|_2^2 \le B.$$

By Lemma 5.8, E occurs with probability at least $1-\delta$. When E holds, for all $j = 1, 2, \dots, n$,

$$\lambda_{\max}\left(\mathbf{M}_{(j)}\mathbf{M}_{(j)}^T\right) = \|\mathbf{M}_{(j)}\|_2^2 \le B,$$

so $\mathbf{G}$ is a sum of random positive-semidefinite matrices each of whose norms is bounded by B. Note that the event E is entirely determined by the random matrix $\mathbf{D}$; in particular E is independent of $\mathbf{R}$.

Conditioning on E, the randomness in $\mathbf{R}$ allows us to use the upper matrix Chernoff bound of Lemma 4.2 to control the maximum eigenvalue of $\mathbf{G}$. We observe that

$$\mu_{\max} = \ell \cdot \lambda_{\max}\left(\mathbb{E}\left[\mathbf{X}_1\right]\right) = \frac{\ell}{n}\lambda_{\max}\left(\sum\nolimits_{j=1}^{n} \mathbf{M}_{(j)}\mathbf{M}_{(j)}^T\right) = \frac{\ell}{n}\,\|\mathbf{\Sigma}\|_2^2.$$

Take the parameter ν in Lemma 4.2 to be

$$\nu = 4 + \frac{B}{\mu_{\max}}\log(\rho/\delta)$$

to obtain the relation

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}(\mathbf{G}) \ge 5\mu_{\max} + B\log(\rho/\delta) \mid E\right\} &\le (\rho - k)\cdot \mathrm{e}^{[\delta-(1+\nu)\log(1+\nu)](\mu_{\max}/B)} \\ &\le \rho \cdot \mathrm{e}^{(1-(5/4)\log 5)\delta(\mu_{\max}/B)} \\ &\le \rho \cdot \mathrm{e}^{(-(5/4)\log 5 - 1)\log(\rho/\delta)} < \delta. \end{aligned} \tag{5.3.9}$$

The second inequality holds because $\nu \ge 4$ implies that $(1+\nu)\log(1+\nu) \ge \nu \cdot (5/4)\log 5$.

We have conditioned on E, the event that the squared norms of the columns of $\mathbf{M}$ are all smaller than B. Thus, substituting the values of B and $\mu_{\max}$ into (5.3.9), we find that

$$\mathbb{P}\left\{\lambda_{\max}(\mathbf{G}) \ge \frac{\ell}{n}\left(5\,\|\mathbf{\Sigma}\|_2^2 + \frac{\log(\rho/\delta)}{\ell}\left(\|\mathbf{\Sigma}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{\Sigma}\|_2\right)^2\right)\right\} \le 2\delta.$$

Use equation (5.3.8) to wrap up. □

Similarly, the SRHT is unlikely to substantially increase the Frobenius norm of a matrix.

Lemma 5.10 (SRHT-based subsampling in the Frobenius norm)**.** *Assume n is a power of 2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, and let $\mathbf{\Theta} \in \mathbb{R}^{\ell\times n}$ be an SRHT matrix for some $\ell < n$. Fix a failure probability $0 < \delta < 1$. Then, for any $\eta \ge 0$,*

$$\mathbb{P}\left\{\left\|\mathbf{A}\mathbf{\Theta}^T\right\|_{\mathrm{F}}^2 \le (1+\eta)\,\|\mathbf{A}\|_{\mathrm{F}}^2\right\} \ge 1 - \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{1+\eta}}\right]^{\ell/\left(1+\sqrt{8\log(n/\delta)}\right)^2} - \delta.$$

Proof. Let $c_j = (n/\ell) \cdot \left\|(\mathbf{ADH}^T)_{(j)}\right\|_2^2$ denote the squared norm of the jth column of $\sqrt{n/\ell} \cdot \mathbf{ADH}^T$. Then, since right multiplication by $\mathbf{R}^T$ samples columns uniformly at random without replacement,

$$\left\|\mathbf{A\Theta}^T\right\|_{\mathrm{F}}^2 = \frac{n}{\ell}\left\|\mathbf{ADH}^T\mathbf{R}^T\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{\ell} X_i \tag{5.3.10}$$

where the random variables X_i are chosen randomly without replacement from the set $\{c_j\}_{j=1}^n$. There are two independent sources of randomness in this sum: the choice of summands, which is determined by $\mathbf{R}$, and the magnitudes of the $\{c_j\}$, which are determined by $\mathbf{D}$.

To bound this sum, we first condition on the event E that each c_j is bounded by a quantity B which depends only on the random matrix $\mathbf{D}$. Then

$$\mathbb{P}\left\{\sum_{i=1}^{\ell} X_i \geq (1+\eta)\sum_{i=1}^{\ell}\mathbb{E}X_i\right\} \leq \mathbb{P}\left\{\sum_{i=1}^{\ell} X_i \geq (1+\eta)\sum_{i=1}^{\ell}\mathbb{E}X_i \mid E\right\} + \mathbb{P}(E^c).$$

To select B, we observe that Lemma 5.8 implies that with probability $1-\delta$, the entries of $\mathbf{D}$ are such that

$$\max_j c_j \leq \frac{n}{\ell}\cdot\frac{1}{n}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{A}\|_2)^2 \leq \frac{1}{\ell}(1+\sqrt{8\log(n/\delta)})^2\,\|\mathbf{A}\|_{\mathrm{F}}^2.$$

Accordingly, we take

$$B = \frac{1}{\ell}(1+\sqrt{8\log(n/\delta)})^2\,\|\mathbf{A}\|_{\mathrm{F}}^2,$$

thereby arriving at the bound

$$\mathbb{P}\left\{\sum_{i=1}^{\ell} X_i \geq (1+\eta)\sum_{i=1}^{\ell}\mathbb{E}X_i\right\} \leq \mathbb{P}\left\{\sum_{i=1}^{\ell} X_i \geq (1+\eta)\sum_{i=1}^{\ell}\mathbb{E}X_i \mid E\right\} + \delta. \tag{5.3.11}$$

After conditioning on $\mathbf{D}$, we observe that the randomness remaining on the right-hand side of (5.3.11) is due to the choice of the summands X_i, which is determined by $\mathbf{R}$. We address this randomness by applying a scalar Chernoff bound (Lemma 4.2 with $k=1$). To do so, we need $\mu_{\max}$, the expected value of the sum; this is an elementary calculation:

$$\mathbb{E}X_1 = n^{-1}\sum_{j=1}^{n} c_j = \frac{1}{\ell}\,\|\mathbf{A}\|_{\mathrm{F}}^2,$$

so $\mu_{\max} = \ell\mathbb{E}X_1 = \|\mathbf{A}\|_{\mathrm{F}}^2$.

Applying Lemma 4.2 conditioned on E, we conclude that

$$\mathbb{P}\left\{\left\|\mathbf{A\Theta}^T\right\|_{\mathrm{F}}^2 \geq (1+\eta)\,\|\mathbf{A}\|_{\mathrm{F}}^2 \mid E\right\} \leq \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{1+\eta}}\right]^{\ell/(1+\sqrt{8\log(n/\delta)})^2} + \delta$$

for $\eta \geq 0$. □

Finally, we prove a novel result on approximate matrix multiplication involving SRHT matrices.

Lemma 5.11 (SRHT for approximate matrix multiplication)**.** *Assume n is a power of 2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$. For some $\ell < n$, let $\mathbf{\Theta} \in \mathbb{R}^{\ell\times n}$ be an SRHT matrix. Fix a failure probability $0 < \delta < 1$. Assume R satisfies $0 \leq R \leq \sqrt{\ell}/(1+\sqrt{8\log(n/\delta)})$. Then,*

$$\mathbb{P}\left\{\left\|\mathbf{A\Theta}^T\mathbf{\Theta B} - \mathbf{AB}\right\|_{\mathrm{F}} \leq 2(R+1)\frac{\|\mathbf{A}\|_{\mathrm{F}}\,\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{A}\|_{\mathrm{F}}\,\|\mathbf{B}\|_2}{\sqrt{\ell}}\right\} \geq 1 - \mathrm{e}^{-R^2/4} - 2\delta.$$

Remark 5.12. Recall that the stable rank $\mathrm{sr}(\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{F}}^2 / \|\mathbf{A}\|_2^2$ reflects the decay of the spectrum of the matrix $\mathbf{A}$. The event under consideration in Lemma 5.11 can be rewritten as a bound on the relative error of the approximation $\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B}$ to the product $\mathbf{AB}$:

$$\frac{\left\|\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B} - \mathbf{AB}\right\|_{\mathrm{F}}}{\|\mathbf{AB}\|_{\mathrm{F}}} \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{AB}\|_{\mathrm{F}}} \cdot \frac{R+2}{\sqrt{\ell}} \cdot \left(1 + \frac{\sqrt{8\log(n/\delta)}}{\mathrm{sr}(\mathbf{B})}\right).$$

In this form, we see that the relative error is controlled by the deterministic condition number for the matrix multiplication problem as well as the stable rank of $\mathbf{B}$ and the number of column samples ℓ. Since the roles of $\mathbf{A}$ and $\mathbf{B}$ in this bound can be interchanged, in fact we have the bound

$$\frac{\left\|\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B} - \mathbf{AB}\right\|_{\mathrm{F}}}{\|\mathbf{AB}\|_{\mathrm{F}}} \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{AB}\|_{\mathrm{F}}} \cdot \frac{R+2}{\sqrt{\ell}} \cdot \left(1 + \frac{\sqrt{8\log(n/\delta)}}{\max(\mathrm{sr}(\mathbf{B}), \mathrm{sr}(\mathbf{A}))}\right).$$

To prove the lemma, we use Lemma 4.3, our result for approximate matrix multiplication via uniform sampling (without replacement) of the columns and the rows of the two matrices involved in the product. Lemma 5.11 is simply a specific application of this generic result. We mention that Lemma 3.2.8 in [Dri02] gives a similar result for approximate matrix multiplication which, however, gives a bound on the expected value of the error term, while our Lemma 5.11 gives a comparable bound that holds with high probability.

Proof of Lemma 5.11. Let $\mathbf{X} = \mathbf{ADH}^T$ and $\mathbf{Y} = \mathbf{HDB}$ and form $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ according to Lemma 4.3. Then, $\mathbf{XY} = \mathbf{AB}$ and

$$\left\|\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B} - \mathbf{AB}\right\|_{\mathrm{F}} = \left\|\hat{\mathbf{X}}\hat{\mathbf{Y}} - \mathbf{XY}\right\|_{\mathrm{F}}.$$

To apply Lemma 4.3, we first condition on the event that the SRHT equalizes the column norms of our matrices. Namely, we observe that, from Lemma 5.8, with probability at least $1 - 2\delta$,

$$\max_i \|\mathbf{X}_{(i)}\|_2 \leq \frac{1}{\sqrt{n}}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{A}\|_2), \text{ and} \tag{5.3.12}$$
$$\max_i \|\mathbf{Y}^{(i)}\|_2 \leq \frac{1}{\sqrt{n}}(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{B}\|_2).$$

We choose the parameters σ and B in Lemma 4.3. Set

$$\sigma^2 = \frac{4}{\ell}(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{B}\|_2)^2\,\|\mathbf{A}\|_{\mathrm{F}}^2. \tag{5.3.13}$$

In view of (5.3.12),

$$\sigma^2 = 4\frac{n}{\ell} \cdot \frac{(\|\mathbf{Y}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{Y}\|_2)^2}{n}\,\|\mathbf{X}\|_{\mathrm{F}}^2 \geq 4\frac{n}{\ell}\sum_{i=1}^n \|\mathbf{X}_{(i)}\|_2^2\|\mathbf{Y}^{(i)}\|_2^2$$

so this choice of σ satisfies the inequality condition of Lemma 4.3. Next we choose

$$B = \frac{2}{\ell}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{A}\|_2)(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\,\|\mathbf{B}\|_2).$$

Again, because of (5.3.12), B satisfies the requirement $B \geq \frac{2n}{\ell}\max_i \|\mathbf{X}_{(i)}\|_2\|\mathbf{Y}^{(i)}\|_2$.

For simplicity, abbreviate $\gamma = 8\log(n/\delta)$. With these choices for σ^2 and B,

$$\begin{aligned}\frac{\sigma^2}{B} &= \frac{2\,\|\mathbf{A}\|_{\mathrm{F}}^2\,(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{B}\|_2)^2}{(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{A}\|_2)(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{B}\|_2)} \\ &\geq \frac{2\,\|\mathbf{A}\|_{\mathrm{F}}^2\,(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{B}\|_2)^2}{(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{A}\|_{\mathrm{F}})(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{B}\|_2)} \\ &= \frac{2\,\|\mathbf{A}\|_{\mathrm{F}}\,(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{\gamma}\,\|\mathbf{B}\|_2)}{1 + \sqrt{\gamma}}.\end{aligned}$$

Now, referring to (5.3.13), identify the numerator as $\sqrt{\ell}\sigma$ to see that

$$\frac{\sigma^2}{B} \geq \frac{\sqrt{\ell}\sigma}{1 + \sqrt{8\log(n/\delta)}}.$$

Apply Lemma 4.3 to see that, when (5.3.12) holds and $0 \leq R\sigma \leq \sigma^2/B$,

$$\mathbb{P}\left\{\left\|\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\right\|_{\mathrm{F}} \geq (R+1)\sigma\right\} \leq \exp\left(-\frac{R^2}{4}\right).$$

From our lower bound on σ^2/B, we know that the condition $R\sigma \leq \sigma^2/B$ is satisfied when

$$R \leq \sqrt{\ell}/(1 + \sqrt{8\log(n/\delta)}).$$

We established above that (5.3.12) holds with probability at least $1 - 2\delta$. From these two facts, it follows that when $0 \leq R \leq \sqrt{\ell}/(1 + \sqrt{8\log(n/\delta)})$,

$$\mathbb{P}\left\{\left\|\mathbf{A}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\right\|_{\mathrm{F}} \geq (R+1)\sigma\right\} \leq \exp\left(-\frac{R^2}{4}\right) + 2\delta.$$

The tail bound given in the statement of Lemma 5.11 follows when we substitute our estimate of σ. □

5.4 Proof of the quality of approximation guarantees

With the necessary preliminaries in hand, we now proceed to the proof of our main results: bounds on the spectral and Frobenius-norm residual and forward errors of low-rank approximations generated using Algorithms 5.1 and 5.2 with SRHT sampling matrices. We note that prior works have provided only residual error bounds [WLRT08, HMT11, NDT09].

Our first result bounds the spectral-norm errors.

Theorem 5.13. *Assume n is a power of 2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ and fix an integer k satisfying $2 \leq k < \rho$. Let $0 < \epsilon < 1/3$ be an accuracy parameter, $0 < \delta < 1$ be a failure probability, and $C \geq 1$ be any specified constant. If $\boldsymbol{\Theta} \in \mathbb{R}^{\ell\times n}$ is an SRHT matrix and ℓ satisfies*

$$6C^2\epsilon^{-1}\left[\sqrt{k} + \sqrt{8\log(n/\delta)}\right]^2 \log(k/\delta) \leq \ell \leq n, \tag{5.4.1}$$

then the approximation $\mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}$ *generated by Algorithm 5.1 with* $\mathbf{S} = \boldsymbol{\Theta}^T$ *satisfies the residual error bound*

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2 \leq \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \quad \text{(i)}$$

and the forward error bound

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2 \leq \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \quad \text{(ii)}$$

simultaneously, with probability at least $1 - 5\delta$.

Likewise, the approximation $\boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})$ *generated by Algorithm 5.2 with* $\mathbf{S} = \boldsymbol{\Theta}^T$ *satisfies the residual error bound*

$$\left\|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\right\|_2 \leq \left(6 + \sqrt{\frac{6\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{6\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \quad \text{(iii)}$$

and the forward error bound

$$\left\|\mathbf{A}_k - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\right\|_2 \leq \left(7 + \sqrt{\frac{6\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{6\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \quad \text{(iv)}$$

simultaneously, with probability at least $1 - 5\delta$.

Proof. First we derive the residual error bounds. Lemma 5.5 implies that, when ℓ satisfies (5.4.1),

$$\left\|(\mathbf{V}_1^T\boldsymbol{\Theta}^T)^\dagger\right\|_2^2 \leq (1 - \sqrt{\epsilon})^{-1}$$

with probability at least $1 - 3\delta$. Consequently, $\mathbf{V}_1^T\boldsymbol{\Theta}^T$ has full row-rank and Lemma 4.8 with $\mathbf{S} = \boldsymbol{\Theta}^T \in \mathbb{R}^{n\times\ell}$ and $\xi = 2$ applies with the same probability, yielding a bound on the residual error of the approximation $\mathbf{P}_{\mathbf{AS}}\mathbf{A}$:

$$\begin{aligned}\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2^2 &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + \|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T(\mathbf{V}_1^T\boldsymbol{\Theta}^T)^\dagger\|_2^2 \\ &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + \|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T\|_2^2\|(\mathbf{V}_1^T\boldsymbol{\Theta}^T)^\dagger\|_2^2 \\ &\leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + (1 - \sqrt{\epsilon})^{-1}\left\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T\right\|_2^2. \end{aligned} \quad (5.4.2)$$

We now provide an upper bound for the quantity

$$Z = \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + (1 - \sqrt{\epsilon})^{-1}\left\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T\right\|_2^2.$$

After applying Lemma 5.9 to estimate the term $\left\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T\right\|_2^2$, we see that the estimate

$$Z \leq \left(1 + \frac{5}{1 - \sqrt{\epsilon}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + \frac{\log(\rho/\delta)}{(1 - \sqrt{\epsilon})\ell}\left(\left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)}\left\|\mathbf{A} - \mathbf{A}_k\right\|_2\right)^2$$

holds with probability at least $1 - 5\delta$. Our assumption that $\epsilon < 1/3$ ensures that $(1 - \sqrt{\epsilon})^{-1} < 3$, so

$$Z \leq 16 \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + \frac{3\log(\rho/\delta)}{\ell} \left(\left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} + \sqrt{8\log(n/\delta)} \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 \right)^2. \tag{5.4.3}$$

Introduce this estimate for Z into (5.4.2), use the subadditivity of the square-root function, and rearrange the spectral and Frobenius norm terms to arrive at Eqn. (i) in the theorem:

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}}\mathbf{A}\right\|_2 \leq \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{\ell}} \right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

We now establish the residual error bound for the approximation $\boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})$. We begin by recalling that Lemma 4.7 states that

$$\|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2^2 \leq 2\|\mathbf{A} - \boldsymbol{\Pi}^{2}_{\mathbf{AS},k}(\mathbf{A})\|_2^2.$$

Lemma 4.8 can be used to bound the right-hand side quantity:

$$2\|\mathbf{A} - \boldsymbol{\Pi}^{2}_{\mathbf{AS},k}(\mathbf{A})\|_2^2 \leq 2\left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + 2\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T(\mathbf{V}_1^T\boldsymbol{\Theta}^T)^\dagger\|_2^2.$$

We have already encountered the right-hand side of this expression, without the factor of two, in (5.4.2). It follows that

$$\|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2^2 \leq 2Z.$$

Introduce our earlier estimate for Z, given in (5.4.3), into this inequality; apply the submultiplicativity of the square-root function; and rearrange terms to obtain Eqn. (iii) in the theorem:

$$\left\|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\right\|_2 \leq \left(6 + \sqrt{\frac{6\log(n/\delta)\log(\rho/\delta)}{\ell}} \right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{6\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

Once again, this bound holds with probability at least $1 - 5\delta$.

The forward error bounds follow in a similar manner. To establish Eqn. (ii) in the theorem, observe that Lemma 4.9 gives the bound

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2^2 \leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_2^2 + \|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\boldsymbol{\Theta}^T(\mathbf{V}_1^T\boldsymbol{\Theta}^T)^\dagger\|_2^2.$$

Once again, we observe that we encountered the right-hand side of this expression in (5.4.2), where we argued that it is bounded by Z, so

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2^2 \leq Z.$$

Introduce into this inequality the estimate for Z given in (5.4.3), apply the submultiplicativity of the square-root function, and rearrange terms to obtain Eqn. (ii) in the theorem:

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}\right\|_2^2 \leq \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{\ell}} \right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{3\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

To establish Eqn. (iv) in the theorem, observe that

$$\begin{aligned}
\|\mathbf{A}_k - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2 &\leq \|\mathbf{A}_k + \mathbf{A}_{\rho-k} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A}) + \mathbf{A}_{\rho-k}\|_2 \\
&\leq \|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2 + \|\mathbf{A}_{\rho-k}\|_2 \\
&= \|\mathbf{A} - \boldsymbol{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2 + \|\mathbf{A} - \mathbf{A}_{\rho-k}\|_2.
\end{aligned}$$

The first term on the right-hand side of this inequality is simply the forward error of the approximation $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})$. Introduce our bound on this error, given by Eqn. (iii) of the theorem, into this inequality to obtain the desired bound:

$$\|\mathbf{A}_k - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\|_2 \leq \left(7 + \sqrt{\frac{6\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_2 + \sqrt{\frac{6\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

This bound holds with probability at least $1 - 5\delta$. □

Our second result bounds the Frobenius-norm errors of the SRHT-based low-rank approximation algorithms.

Theorem 5.14. *Assume n is a power of 2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ and fix an integer k satisfying $2 \leq k < \rho$. Let $0 < \epsilon < 1/3$ be an accuracy parameter, $0 < \delta < 1$ be a failure probability, and $C \geq 1$ be any specified constant. If $\mathbf{\Theta} \in \mathbb{R}^{\ell\times n}$ is an SRHT matrix and ℓ satisfies*

$$6C^2\epsilon^{-1}\left[\sqrt{k} + \sqrt{8\log(n/\delta)}\right]^2 \log(k/\delta) \leq \ell \leq n, \tag{5.4.4}$$

then the approximation $\mathbf{P}_{\mathbf{A\Theta}^T}\mathbf{A}$ generated by Algorithm 5.1 with $\mathbf{S} = \mathbf{\Theta}^T$ satisfies the residual error bound

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A\Theta}^T}\mathbf{A}\right\|_{\mathrm{F}} \leq (1 + 11\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \tag{i}$$

and the forward error bound

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{A\Theta}^T}\mathbf{A}\right\|_{\mathrm{F}} \leq (1 + 11\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \tag{ii}$$

simultaneously, with probability at least $1 - \delta^{C^2\log(k/\delta)/4} - 7\delta$.

Likewise, the approximation $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})$ generated by Algorithm 5.2 with $\mathbf{S} = \mathbf{\Theta}^T$ satisfies the residual error bound

$$\left\|\mathbf{A} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\right\|_{\mathrm{F}} \leq (1 + 11\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \tag{iii}$$

and the forward error bound

$$\left\|\mathbf{A}_k - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{AS},k}(\mathbf{A})\right\|_{\mathrm{F}} \leq (2 + 11\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}} \tag{iv}$$

simultaneously, with probability at least $1 - \delta^{C^2\log(k/\delta)/4} - 7\delta$.

Proof. We first establish the residual error bounds. Because ℓ satisfies (5.4.4), Lemma 5.5 implies that with probability at least $1 - 3\delta$,

$$\operatorname{rank}(\mathbf{V}_1^T\mathbf{\Theta}^T) = k;$$

so, Lemma 4.8 applies with the same probability, yielding

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{A\Theta}^T}\mathbf{A}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{A} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{A\Theta}^T,k}(\mathbf{A})\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}^2 + \left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\right\|_{\mathrm{F}}^2. \tag{5.4.5}$$

We complete the estimate by bounding the second term in the right hand side of the above inequality. Justifications appear below.

$$\begin{aligned} S &:= \left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T (\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger\right\|_{\mathrm{F}}^2 \\ &\le 2\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{V}_1\right\|_{\mathrm{F}}^2 + 2\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T ((\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger - (\mathbf{V}_1^T \boldsymbol{\Theta}^T)^T)\right\|_{\mathrm{F}}^2 \qquad (5.4.6) \\ &\le 2\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{V}_1\right\|_{\mathrm{F}}^2 + 2\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T\right\|_{\mathrm{F}}^2 \left\|(\mathbf{V}_1^T \boldsymbol{\Theta}^T)^\dagger - (\mathbf{V}_1^T \boldsymbol{\Theta}^T)^T\right\|_2^2 \\ &\le 8\epsilon \cdot \left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T\right\|_{\mathrm{F}}^2 + 2 \cdot \left(\frac{11}{4}\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right) \cdot (2.38\epsilon) \qquad (5.4.7) \\ &\le 22\epsilon \cdot \left\|\boldsymbol{\Sigma}_2\right\|_{\mathrm{F}}^2. \qquad (5.4.8) \end{aligned}$$

In (5.4.6) we used the fact that $\|\mathbf{X}+\mathbf{Y}\|_{\mathrm{F}}^2 \le 2\|\mathbf{X}\|_{\mathrm{F}}^2 + 2\|\mathbf{Y}\|_{\mathrm{F}}^2$ for any two matrices $\mathbf{X}$ and $\mathbf{Y}$. The first estimate in (5.4.7) is justified by an application of Lemma 5.11 on SRHT-based approximate matrix multiplication; the second estimate is an application of Lemma 5.10, which predicts the effect that postmultiplication by an SRHT matrix has upon the Frobenius norm of a matrix; and the third estimate is an application of Lemma 5.5, which predicts the effect of postmultiplication by an SRHT matrix on the singular values of a matrix with orthonormal rows. We postpone a detailed discussion of the calculations involved in these estimations until after we have established the residual error bounds.

Combining (5.4.5) with the bound on S, we obtain

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\right\|_{\mathrm{F}}^2 \le \|\mathbf{A} - \boldsymbol{\Pi}_{\mathbf{AS},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}}^2 \le (1+22\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}^2. \qquad (5.4.9)$$

Since $1+2x \le (1+x)^2$ when x is positive, it follows that $\sqrt{1+2x} \le 1+x$ when x is positive. In particular, $\sqrt{1+22\epsilon} \le 1+11\epsilon$. Introduce this observation into (5.4.9) to conclude that

$$\left\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\right\|_{\mathrm{F}} \le \|\mathbf{A} - \boldsymbol{\Pi}_{\mathbf{AS},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}} \le (1+11\epsilon) \cdot \left\|\mathbf{A} - \mathbf{A}_k\right\|_{\mathrm{F}}.$$

Thus we have established Eqns. (i) and (iii) in the theorem.

We now supply the details of the manipulations in (5.4.7). To justify the first estimate, notice that $\mathbf{V}_2^T \mathbf{V}_1 = \mathbf{0}$. Next use Lemma 5.11 with $R = C\sqrt{\log(k/\delta)}$. From the lower bound (5.4.4) on ℓ, we have that

$$\frac{\sqrt{\ell}}{1+\sqrt{8\log(n/\delta)}} \ge \sqrt{6\epsilon^{-1}} \cdot \frac{\sqrt{k}+\sqrt{8\log(n/\delta)}}{1+\sqrt{8\log(n/\delta)}} \cdot C\sqrt{\log(k/\delta)} > R > 0,$$

so this choice of R satisfies the requirements of Lemma 5.11. Apply Lemma 5.11 to obtain

$$\mathbb{P}\left\{\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{V}_1\right\|_{\mathrm{F}}^2 \le 4(R+1)^2 \frac{(\sqrt{k}+\sqrt{8\log(n/\delta)})^2}{\ell}\left\|\boldsymbol{\Sigma}_2 \mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right\} \ge 1 - \mathrm{e}^{-R^2/4} - 2\delta.$$

Recall that $R = C\sqrt{\log(k/\delta)}$. Use the lower bound (5.4.4) on ℓ to justify the estimate

$$\begin{aligned}
4(R+1)^2 \frac{\left[\sqrt{k}+\sqrt{8\log(n/\delta)}\right]^2}{\ell} &\le 4(R+1)^2 \frac{\left[\sqrt{k}+\sqrt{8\log(n/\delta)}\right]^2}{6C^2\epsilon^{-1}\left[\sqrt{k}+\sqrt{8\log(n/\delta)}\right]^2 \log(k/\delta)} \\
&= \frac{2\epsilon}{3} \cdot \frac{(C\sqrt{\log(k/\delta)}+1)^2}{C^2\log(k/\delta)} \\
&\le \frac{2\epsilon}{3}\left(1+\frac{1}{C\sqrt{\log(k/\delta)}}\right)^2.
\end{aligned}$$

This estimate implies that

$$\mathbb{P}\left\{\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\mathbf{\Theta}\mathbf{V}_1\right\|_{\mathrm{F}}^2 \le \frac{2\epsilon}{3}\left(1+\frac{1}{C\sqrt{\log(k/\delta)}}\right)^2 \left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right\} \ge 1-\delta^{C^2\log(k/\delta)/4}-2\delta.$$

Since $C > 1$ and $k \ge 2$, a simple numerical bound allows us to state that, more simply,

$$\mathbb{P}\left\{\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\mathbf{\Theta}\mathbf{V}_1\right\|_{\mathrm{F}}^2 \le 4\epsilon\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right\} \ge 1-\delta^{C^2\log(k/\delta)/4}-2\delta.$$

This bound on $\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\mathbf{\Theta}\mathbf{V}_1\right\|_{\mathrm{F}}^2$ is used to estimate the first term in (5.4.7). The remaining estimates in (5.4.7) follow from applying Lemma 5.5 (keeping in mind the lower bound (5.4.4) on ℓ) to obtain

$$\mathbb{P}\left\{\left\|(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger - (\mathbf{V}_1^T\mathbf{\Theta}^T)^T\right\|_2^2 \le 2.38\epsilon\right\} \ge 1-3\delta.$$

and Lemma 5.10 with $\eta = 7/4$ to obtain

$$\mathbb{P}\left\{\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\right\|_{\mathrm{F}}^2 \le \frac{11}{4}\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right\} \ge 1-\left(\frac{\mathrm{e}^{7/4}}{(1+7/4)^{1+7/4}}\right)^{\ell/(1+\sqrt{8\log(n/\delta)})^2}-\delta.$$

We have the estimate

$$\frac{\mathrm{e}^{7/4}}{(1+7/4)^{1+7/4}} < \frac{1}{\mathrm{e}},$$

so in fact

$$\begin{aligned}
\mathbb{P}\left\{\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T\right\|_{\mathrm{F}}^2 \le \frac{11}{4}\left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\right\|_{\mathrm{F}}^2\right\} &\ge 1-\mathrm{e}^{-\ell/(1+\sqrt{8\log(n/\delta)})^2}-\delta \\
&\ge 1-\mathrm{e}^{-6C^2\epsilon^{-1}\log(k/\delta)}-\delta \\
&\ge 1-\mathrm{e}^{-\log(k/\delta)}-\delta \\
&\ge 1-2\delta.
\end{aligned}$$

Adding up the failure probabilities of the three estimates used in (5.4.7), we conclude that the bound on S given in (5.4.8) holds with probability at least $1-\delta^{C^2\log(k/\delta)/4}-7\delta$. Thus Eqns. (i) and (iii) hold with this probability.

Next we establish the forward error bounds. Lemma 4.9 with $\mathbf{S} = \mathbf{\Theta}^T \in \mathbb{R}^{n\times\ell}$ gives

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\right\|_{\mathrm{F}}^2 \le \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}^2 + \left\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{\Theta}^T(\mathbf{V}_1^T\mathbf{\Theta}^T)^\dagger\right\|_{\mathrm{F}}^2.$$

Identify the second term on the right-hand side as S and introduce the estimate for S given in (5.4.8) to see that

$$\left\|\mathbf{A}_k - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}^2 + 22\epsilon \cdot \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}^2 = (1+22\epsilon)\cdot\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}^2.$$

This bound holds with probability at least $1-\delta^{C^2\log(k/\delta)/4}-7\delta$. Taking the square-roots of both sides and using the fact that $\sqrt{1+22\epsilon}\leq 1+22\epsilon$ gives Eqn. (iii).

Finally, we prove Eqn. (iv):

$$\begin{aligned}\left\|\mathbf{A}_k - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{A}\boldsymbol{\Theta}^T,k}(\mathbf{A})\right\|_{\mathrm{F}} &= \left\|\mathbf{A}-\mathbf{A}_k-(\mathbf{A}-\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{A}\boldsymbol{\Theta}^T,k}(\mathbf{A}))\right\|_{\mathrm{F}}\\ &\leq \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}} + \left\|\mathbf{A}-\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{A}\boldsymbol{\Theta}^T,k}(\mathbf{A})\right\|_{\mathrm{F}} \leq (2+11\epsilon)\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}},\end{aligned}$$

where the first inequality follows by the triangle inequality and the second follows from Eqn. (ii) in the theorem. This bound holds with probability at least $1-\delta^{C^2\log(k/\delta)/4}-7\delta$.

□

5.5 Experiments

In this section, we experimentally investigate the tightness of the residual and forward error bounds provided in Theorems 5.13 and 5.14 for the spectral and Frobenius-norm approximation errors of SRHT low-rank approximations of the forms $\mathbf{P}_{\mathbf{A}\boldsymbol{\Theta}^T}\mathbf{A}$ and $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{A}\boldsymbol{\Theta}^T,k}(\mathbf{A})$. Additionally, we experimentally verify that the SRHT algorithms are not significantly less accurate than the Gaussian low-rank approximation algorithms.

5.5.1 The test matrices

Let $n = 1024$, and consider the following three test matrices:

1. Matrix $\mathbf{A}\in\mathbb{R}^{(n+1)\times n}$ is given by

$$\mathbf{A} = [100\mathbf{e}_1+\mathbf{e}_2, 100\mathbf{e}_1+\mathbf{e}_3, \ldots, 100\mathbf{e}_1+\mathbf{e}_{n+1}],$$

   where $\mathbf{e}_i\in\mathbb{R}^{n+1}$ are the standard basis vectors.

2. Matrix $\mathbf{B}\in\mathbb{R}^{n\times n}$ is diagonal with entries $(\mathbf{B})_{ii} = 100(1-(i-1)/n)$.

3. Matrix $\mathbf{C}\in\mathbb{R}^{n\times n}$ has the same singular values as $\mathbf{B}$, but its singular spaces are sampled from the uniform measure on the set of orthogonal matrices. More precisely, $\mathbf{C} = \mathbf{U}\mathbf{B}\mathbf{V}^T$, where $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD of an $n\times n$ matrix whose entries are standard Gaussian random variables.

These three matrices exhibit properties that, judging from the bounds in Theorems 5.13 and 5.14, could challenge the SRHT approximation algorithm. Matrix $\mathbf{A}$ is approximately rank one—there is a large spectral gap after the first singular value—but the residual spectrum is flat, so for $k\geq 1$, the $\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}$ terms in the spectral norm bounds of Theorem 5.13 are quite large compared to the $\left\|\mathbf{A}-\mathbf{A}_k\right\|_2$ terms. Matrices $\mathbf{B}$ and $\mathbf{C}$ both have slowly decaying spectrums, so one again has a large Frobenius term present in the spectral norm error bound.

Matrices $\mathbf{B}$ and $\mathbf{C}$ were chosen to have the same singular values but different singular spaces to reveal any effect that the structure of the singular spaces of the matrix has on the quality

of SRHT approximations. The coherence of their right singular spaces provides a summary of the relevant difference in the singular spaces of **B** and **C**. Recall that the coherence of a k-dimensional subspace $\mathscr{S}$ is defined as

$$\mu(\mathscr{S}) = \frac{n}{k} \max_i \mathbf{P}_{ii},$$

where **P** is the projection onto $\mathscr{S}$; the coherence of $\mathscr{S}$ is always between 1 and n/k [CR09]. It is clear that all the right singular spaces of **B** are maximally coherent, and it is known that with high probability the dominant right k-dimensional singular space of **C** is quite incoherent, with coherence on the order of $k \log n$ [CR09].

To gain an intuition for the potential significance of this difference in coherence, consider a randomized column sampling approach to forming low-rank approximants; that is, consider approximating $\mathbf{M}_k$ with a matrix $\mathbf{P}_\mathbf{Y}\mathbf{M}$ where **Y** comprises randomly sampled columns of **M**. It is known that such approximations are quite inaccurate unless the dominant k-dimensional right singular space of **M** is incoherent (see, e.g., Chapter 6 or [TR10]). One could interpret SRHT approximation algorithms as consisting of a rotation of the right singular spaces of **M** by multiplying from the right with $\mathbf{DH}^T$ followed by forming a column sample-based approximation. The rotation lowers the coherence of the right singular spaces and thereby increases the probability of obtaining an accurate low-rank approximation. One expects that if **M** has highly coherent right singular spaces then the right singular spaces of $\mathbf{MDH}^T$ will be less coherent. Thus we compare the performance of the SRHT approximations on **B**, which has maximally coherent right singular spaces, to their performance on **C**, which has almost maximally incoherent right singular spaces.

5.5.2 Empirical comparison of the SRHT and Gaussian algorithms

Figure 5.1 depicts the relative residual errors of the Gaussian and SRHT algorithms for approximations generated using Algorithms 5.1 and 5.2: $\mathbf{P}_{\mathbf{MS}}\mathbf{M}$ and $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{MS},k}(\mathbf{M})$, which we shall hereafter refer to respectively as the non-rank-restricted and rank-restricted approximations. Here the matrix **M** is used to refer interchangeably to **A**, **B**, and **C**. The relative residual errors ($\|\mathbf{M} - \mathbf{P}_{\mathbf{MS}}\mathbf{M}\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ and $\|\mathbf{M} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{MS},k}(\mathbf{M})\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ for $\xi = 2, \mathrm{F}$) shown in this figure for each value of k were obtained by taking the average of the relative residual errors observed over 30 trials of low-rank approximations, each formed using $\ell = \lceil 2k \log n \rceil$ samples.

With the exception of the residual spectral errors on **A**, which range from between two and nine times the size of the optimal rank-k spectral residual error for $k < 20$, we see that the residual errors for all three matrices are less than 1.1 times the residual error of $\mathbf{M}_k$, if not significantly smaller. Specifically, the relative residual errors of the restricted-rank approximations remain less than 1.1 over the entire range of k while the relative residual errors of the non-rank-restricted approximations actually decrease as k increases. Note that, because $\ell > k$, the relative errors of the non-rank-restricted approximations are often smaller than 1, while those of the restricted-rank approximations are never smaller than 1.

Since the matrices **B** and **C** have the same singular values, but the singular spaces of **C** are less coherent, the difference in the residual errors of the approximations of **B** and **C** is evidence that the spectral-norm accuracy of the SRHT approximations is increased on less coherent datasets; the same is true for the Frobenius norm accuracy to a lesser extent. The Gaussian approximations seem insensitive to the level of coherence. Only on the highly coherent matrix **B** do we see a notable decrease in the residual errors when Gaussian sampling is used rather than an SRHT; however, even in this case the residual errors of the SRHT approximations are

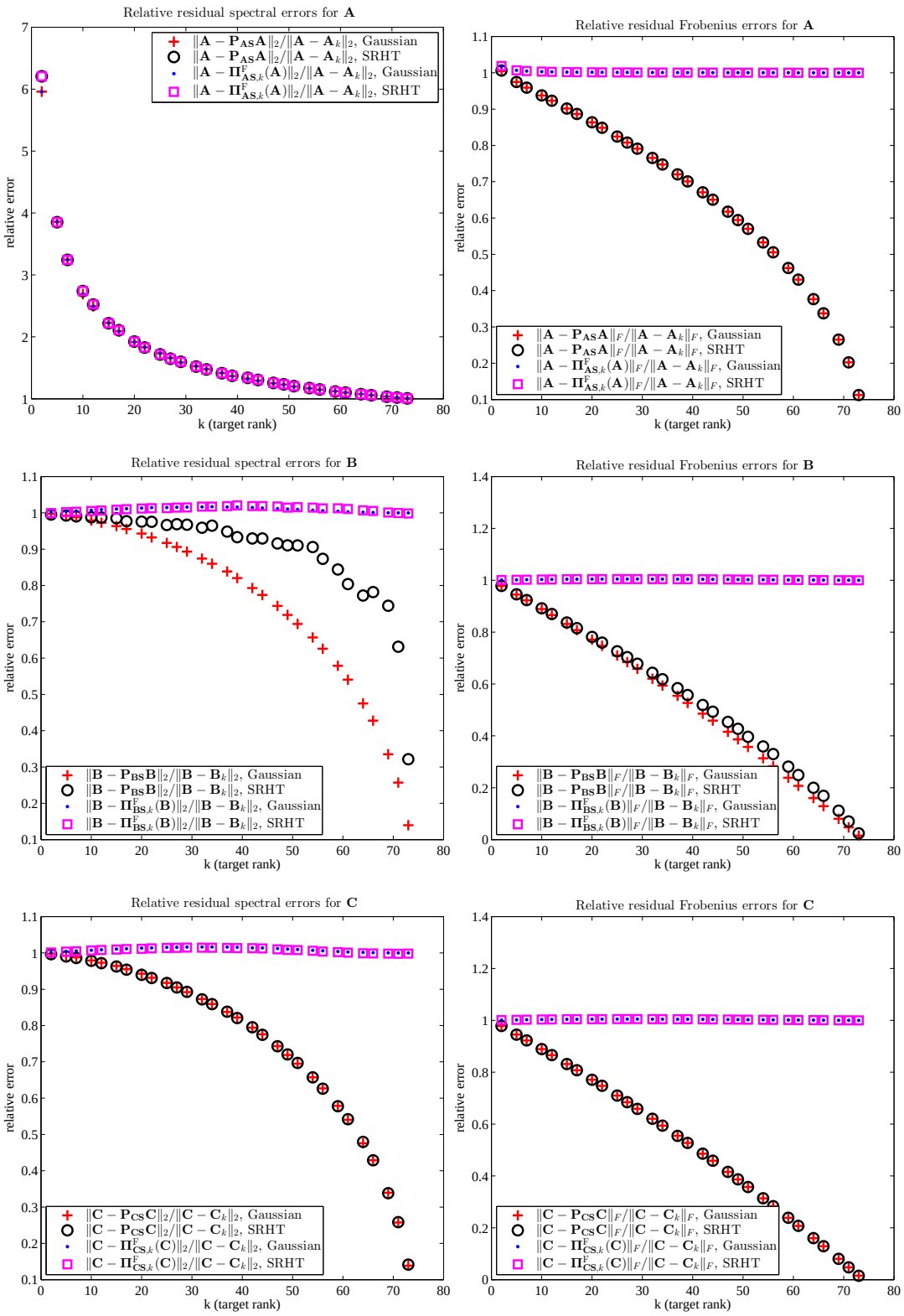


Figure 5.1: RESIDUAL ERRORS OF LOW-RANK APPROXIMATION ALGORITHMS. Relative spectral and Frobenius-norm residual errors of the SRHT and Gaussian low-rank approximation algorithms ($\|\mathbf{M} - \mathbf{P}_{\mathbf{MS}}\mathbf{M}\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ and $\|\mathbf{M} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{MS},k}(\mathbf{M})\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ for $\xi = 2, \mathrm{F}$) as a function of the target rank k for the three matrices $\mathbf{M} = \mathbf{A}, \mathbf{B}, \mathbf{C}$. Each point is the average error observed over 30 trials. In each trial, $\ell = \lceil 2k \log n \rceil$ column samples were used.

comparable with that of $\mathbf{B}_k$. In all, Figure 5.1 suggests that the gain in computational efficiency provided by the SRHT does not come at the cost of a significant loss in accuracy and that taking $\ell = \lceil 2k \log n \rceil$ samples suffices to obtain approximations with small residual errors relative to those of the optimal rank-k approximations. Up to the specific value of the constant, this latter observation coincides with the conclusions of Theorems 5.13 and 5.14.

Figure 5.2 depicts the relative forward errors of the Gaussian and SRHT algorithms ($\|\mathbf{M}_k - \mathbf{P}_{\mathbf{MS}}\mathbf{M}\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ and $\|\mathbf{M}_k - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{MS},k}(\mathbf{M})\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ for $\xi = 2, \mathrm{F}$) for the non-rank-restricted and rank-restricted approximations. The error shown for each k is the average relative forward error observed over 30 trials of low-rank approximations each formed using $\ell = \lceil 2k \log n \rceil$ samples. We observe that the forward errors of both algorithms for both choices of sampling matrices are on the scale of the norm of $\mathbf{M}_k$. By looking at the relative spectral-norm forward errors we see that in this norm, perhaps contrary to intuition, the rank-restricted approximation does not provide a more accurate approximation to $\mathbf{M}_k$ than does the non-rank-restricted approximation. However the rank-restricted approximation clearly provides a more accurate approximation to $\mathbf{M}_k$ than the non-rank-restricted approximation in the Frobenius norm. A rather unexpected observation is that the rank-restricted approximations are more accurate in the spectral norm for highly coherent matrices (**B**) than they are for matrices which are almost minimally coherent (**C**). Overall, Figure 5.2 suggests that the SRHT low-rank approximation algorithms provide accurate approximations to $\mathbf{M}_k$ when ℓ is in the regime suggested by Theorems 5.13 and 5.14.

5.5.3 Empirical evaluation of our error bounds

Figures 5.1 and 5.2 show that when $\ell = \lceil 2k \log n \rceil$ samples are taken, the SRHT low-rank approximation algorithms both provide approximations to $\mathbf{M}$ that are within a factor of $1+\epsilon$ as accurate in the Frobenius norm as $\mathbf{M}_k$, as Theorem 5.14 suggests should be the case. More precisely, Theorem 5.14 assures us that $528\epsilon^{-1}[\sqrt{k} + \sqrt{8\log(8n/\delta)}]^2 \log(8k/\delta)$ column samples are sufficient to ensure that, with at least probability $1-\delta$, $\mathbf{P}_{\mathbf{M\Theta}^T}\mathbf{M}$ and $\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{M\Theta}^T,k}(\mathbf{M})$ have Frobenius norm residual and forward error within $1+\epsilon$ of that of $\mathbf{M}_k$. The factor 528 can certainly be reduced by optimizing the numerical constants given in Theorem 5.14. But what is the smallest ℓ that ensures the Frobenius norm residual error bounds $\left\|\mathbf{M} - \mathbf{P}_{\mathbf{M\Theta}^T}\mathbf{M}\right\|_{\mathrm{F}} \leq (1+\epsilon)\left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}}$ and $\left\|\mathbf{M} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{M\Theta}^T,k}(\mathbf{M})\right\|_{\mathrm{F}} \leq (1+\epsilon)\left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}}$ are satisfied with some fixed probability? To investigate, in Figure 5.3 we plot the values of ℓ determined empirically to be sufficient to obtain $(1+\epsilon)$ Frobenius norm residual errors relative to the optimal rank-k approximation; we fix the failure probability $\delta = 1/2$ and vary ϵ. Specifically, the ℓ plotted for each k is the smallest number of samples for which $\left\|\mathbf{M} - \mathbf{P}_{\mathbf{M\Theta}^T}\mathbf{M}\right\|_{\mathrm{F}} \leq (1+\epsilon)\left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}}$ or $\left\|\mathbf{M} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{M\Theta}^T,k}(\mathbf{M})\right\|_{\mathrm{F}} \leq (1+\epsilon)\left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}}$ in at least 15 out of 30 trials.

It is clear that, for fixed k and ϵ, the number of samples ℓ required to form a non-rank-restricted approximation to $\mathbf{M}$ with $1+\epsilon$ relative residual error is smaller than the ℓ required to form a rank-restricted approximation with $1+\epsilon$ relative residual error. Note that for small values of k, the ℓ necessary for relative residual error to be achieved is actually smaller than k for all three datasets. This is a reflection of the fact that when $k_1 < k_2$ are small, the ratio $\left\|\mathbf{M} - \mathbf{M}_{k_2}\right\|_{\mathrm{F}} / \left\|\mathbf{M} - \mathbf{M}_{k_1}\right\|_{\mathrm{F}}$ is very close to one. Outside of the initial flat regions, the empirically determined value of r seems to grow linearly with k; this matches with the observation of Woolfe et al. that taking $\ell = k + 8$ suffices to consistently form accurate low-rank approximations using the SRFT scheme, which is very similar to the SRHT scheme [WLRT08]. We also note that this matches with Theorem 5.14, which predicts that the necessary ℓ grows at most linearly with k

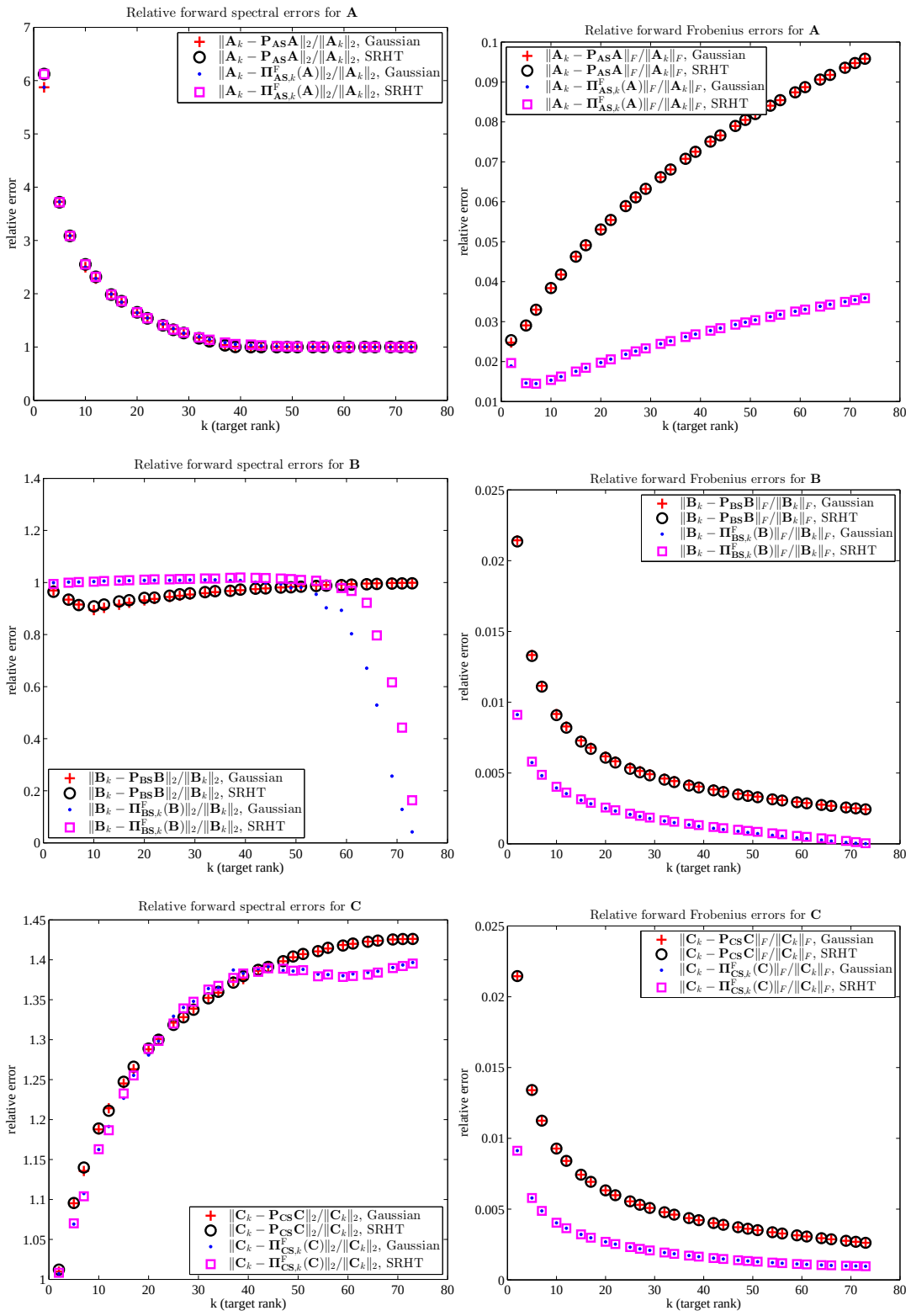


Figure 5.2: FORWARD ERRORS OF LOW-RANK APPROXIMATION ALGORITHMS. The relative spectral and Frobenius-norm forward errors of the SRHT and Gaussian low-rank approximation algorithms ($\|\mathbf{M}_k - \mathbf{P}_{\mathbf{MS}}\mathbf{M}\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ and $\|\mathbf{M}_k - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{MS},k}(\mathbf{M})\|_\xi / \|\mathbf{M} - \mathbf{M}_k\|_\xi$ for $\xi = 2, \mathrm{F}$) as a function of the target rank k for the three matrices $\mathbf{M} = \mathbf{A}, \mathbf{B}, \mathbf{C}$. Each point is the average of the errors observed over 30 trials. In each trial, $\ell = \lceil 2k \log n \rceil$ column samples were used.

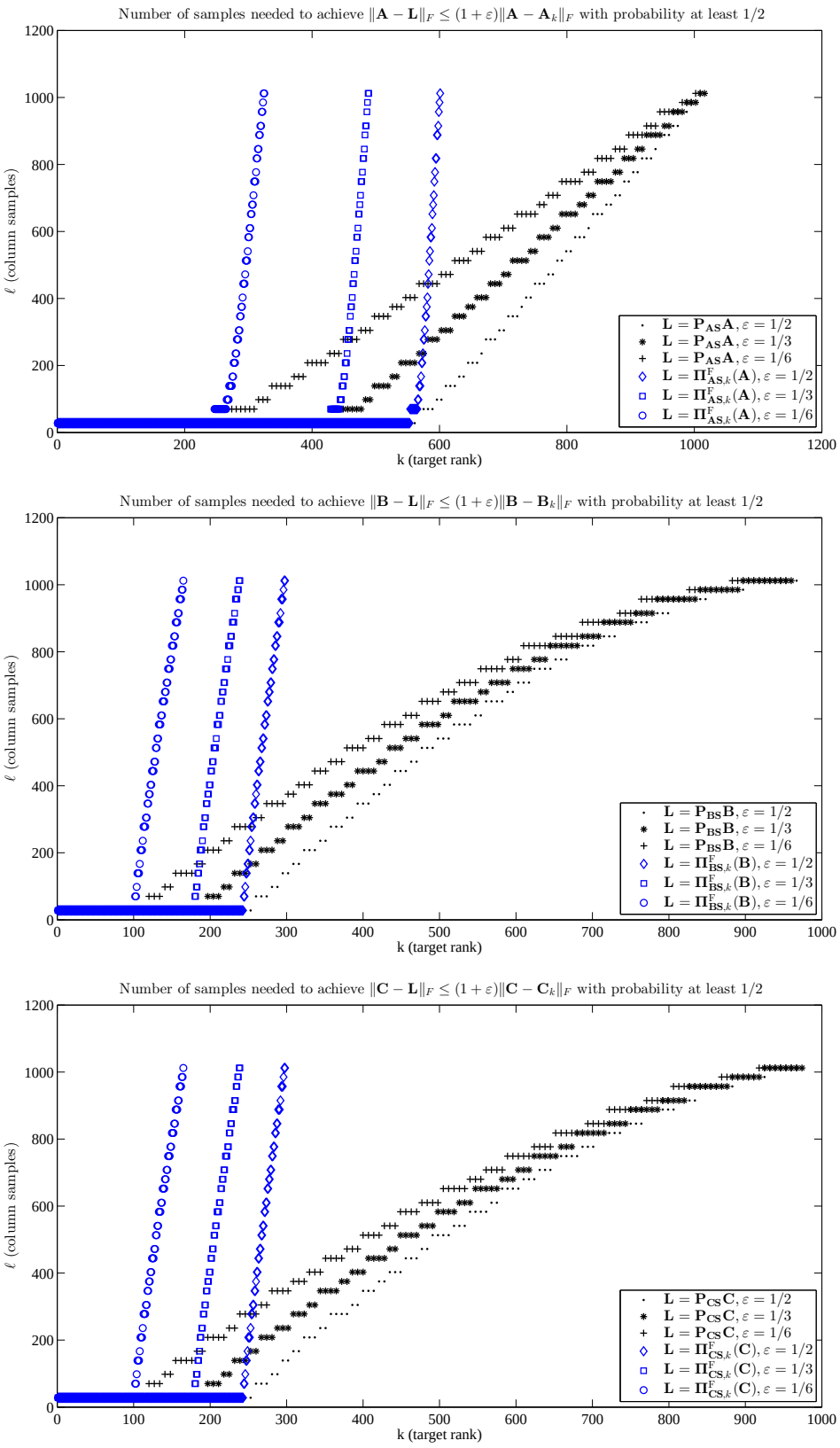


Figure 5.3: THE NUMBER OF COLUMN SAMPLES REQUIRED FOR RELATIVE ERROR FROBENIUS-NORM APPROXIMATIONS. The value of ℓ empirically necessary to ensure that, with probability at least one-half, approximations generated by the SRHT algorithms satisfy $\left\|\mathbf{M}-\mathbf{P}_{\mathbf{M}\mathbf{\Theta}^T}\mathbf{M}\right\|_{\mathrm{F}} \le (1+\epsilon)\left\|\mathbf{M}-\mathbf{M}_k\right\|_{\mathrm{F}}$ and $\left\|\mathbf{M}-\mathbf{\Pi}^{\mathrm{F}}_{\mathbf{M}\mathbf{\Theta}^T,k}(\mathbf{M})\right\|_{\mathrm{F}} \le (1+\epsilon)\left\|\mathbf{M}-\mathbf{M}_k\right\|_{\mathrm{F}}$ (for $\mathbf{M}=\mathbf{A},\mathbf{B},\mathbf{C}$).

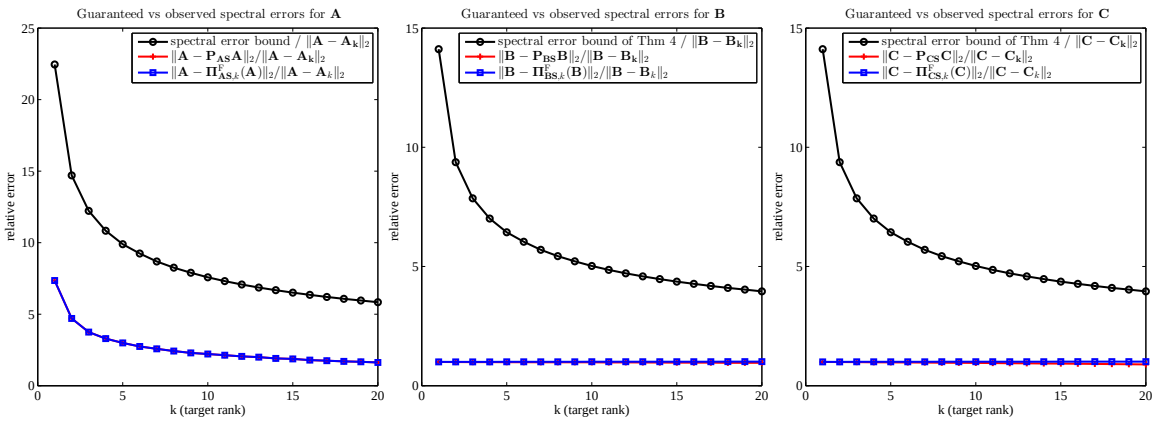


Figure 5.4: EMPIRICAL VERSUS PREDICTED SPECTRAL-NORM RESIDUAL ERRORS OF LOW-RANK APPROXIMATIONS. The empirical spectral-norm residual errors relative to those of the optimal rank-k approximants ($\left\|\mathbf{M} - \mathbf{P}_{\mathbf{M}\Theta^T}\mathbf{M}\right\|_2 / \|\mathbf{M} - \mathbf{M}_k\|_2$ and $\|\mathbf{M} - \mathbf{\Pi}^{\mathrm{F}}_{\mathbf{M}\Theta^T,k}(\mathbf{M})\|_2 / \|\mathbf{M} - \mathbf{M}_k\|_2$) plotted alongside the same ratio for the bounds given in Theorem 5.13, when $\ell = \lceil 2[\sqrt{k} + \sqrt{\log(2n)}]^2 \log(2k) \rceil$ (for $\mathbf{M} = \mathbf{A}, \mathbf{B}, \mathbf{C}$). On the scale shown, the errors of the two SRHT-based approximation algorithms are essentially identical.

with a slope like $\log n$.

Finally, Theorem 5.13 does *not* guarantee that $1+\epsilon$ spectral-norm relative residual errors can be achieved. Instead, it provides bounds on the spectral-norm residual errors achieved in terms of $\left\|\mathbf{M} - \mathbf{M}_k\right\|_2$ and $\left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}}$ that are guaranteed to hold when ℓ is sufficiently large. In Figure 5.4 we compare the spectral-norm residual error guarantees of Theorem 5.13 to what is achieved in practice. To do so, we take the optimistic viewpoint that the constants in Theorem 5.13 can be optimized to unity. Under this view, if more columns than $\ell_2 = \epsilon^{-1}[\sqrt{k} + \sqrt{\log(n/\delta)}]^2 \log(k/\delta)$ are used to construct the SRHT approximations, then the spectral-norm residual error is no larger than

$$b_2 = \left(1 + \sqrt{\frac{\log(n/\delta)\log(\rho/\delta)}{\ell}}\right) \cdot \left\|\mathbf{M} - \mathbf{M}_k\right\|_2 + \sqrt{\frac{\log(\rho/\delta)}{\ell}} \cdot \left\|\mathbf{M} - \mathbf{M}_k\right\|_{\mathrm{F}},$$

where ρ is the rank of $\mathbf{M}$, with probability greater than $1-\delta$. Our comparison consists of using ℓ_2 samples to construct the SRHT approximations and then comparing the predicted upper bound on the spectral-norm residual error, b_2, to the empirically observed spectral-norm residual errors. Figure 5.4 shows, for several values of k, the upper bound b_2 and the observed relative spectral-norm residual errors, with precision parameter $\epsilon = 1/2$ and failure parameter $\delta = 1/2$. For each value of k, the empirical spectral-norm residual error plotted is the average of the errors over 30 trials of low-rank approximations. Note from Figure 5.4 that with this choice of ℓ, the spectral-norm residual errors of the rank-restricted and non-rank-restricted SRHT approximations are essentially the same.

Judging from Figures 5.3 and 5.4, even when we assume the constants present can be optimized away, the bounds given in Theorems 5.13 and 5.14 are pessimistic: it seems that in fact approximations with Frobenius-norm residual error within $1+\epsilon$ of the error of the optimal rank-k approximation can be obtained with ℓ linear in k, and the spectral-norm residual errors are smaller than the supplied upper bounds. Thus there is still room for improvement in our

understanding of the SRHT low-rank approximation algorithm, but as explained in Section 5.2.1, ignoring constants, the bounds of Theorem 5.13 are often tighter than those obtained in earlier works.

To bring perspective to this discussion, consider that even if one limits consideration to deterministic algorithms, the known error bounds for the Gu–Eisenstat rank-revealing QR—a popular and widely used algorithm for low-rank approximation—are quite pessimistic and do not reflect the excellent accuracy that is seen in practice [GE96]. Regardless, we do not advocate using these approximation schemes for applications in which highly accurate low-rank approximations are needed. Rather, Theorems 5.13 and 5.14 and our numerical experiments suggest that they are appropriate in situations where one is willing to trade some accuracy for a gain in computational efficiency.

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 $\mathbf{A}$).

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 $\mathbf{P}_{\mathbf{AS}}\mathbf{AP}_{\mathbf{AS}}$, 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 $\mathbf{P}_{\mathbf{AS}}\mathbf{AP}_{\mathbf{AS}}$.

[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.

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:

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

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

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

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

$$\mathrm{Tr}\left(\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right) \leq \left(1 + \frac{1}{\delta^2(1-\epsilon)}\right)\mathrm{Tr}\left(\mathbf{A} - \mathbf{A}_k\right)$$

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, $\mathbf{S} = \mathbf{R}\mathbf{D}$. 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 $\mathbf{S} = \sqrt{n/\ell}\mathbf{D}\mathbf{F}\mathbf{R}$, 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

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

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 $\ell \geq (1+\epsilon^2)k$. Theorem 6.17 implies the following:

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

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*

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

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

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

□

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*

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

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

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

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 $\mathbf{\Pi A \Pi}^T = \mathbf{B}^T\mathbf{B}$ where $\mathbf{\Pi}$ is a permutation matrix, then

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

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, $\mathbf{P}_{\mathbf{U}\mathbf{M}} = \mathbf{U}\mathbf{P}_{\mathbf{M}}\mathbf{U}^T$. 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 $\mathbf{\Sigma}^{p-1/2}\mathbf{U}^T\mathbf{S}$,

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

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

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

This implies that

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

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

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

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

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

We proceed to bound T_2 by using the estimate

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

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

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

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

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

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

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

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

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

□

Remark. The quality of approximation guarantee provided by Theorem 6.6 depends on two quantities, $\|\mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\|_2$ and $\|\mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}$, 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

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

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

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

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

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

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 $\mathscr{X} = \{\mathbf{u}_i\mathbf{u}_i^T\}_{i=1,\dots,n}$. Then

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

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, $k \geq \operatorname{rank}(\mathbf{A})$, 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 $\mathrm{O}(k \log(k/\delta))$ 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:

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

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

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

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

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

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

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

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

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

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 $\boldsymbol{\Omega}_1$ 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*

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

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

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

each hold, individually, with probability at least $1 - \delta$. When either estimate holds, $\mathbf{\Omega}_1$ 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:

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

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

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

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

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

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*

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

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 $u, t \ge 1$, then

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

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

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

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

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

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 $\left\|\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_{\mathrm{F}}$,

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

The following estimates hold for the coefficients in this inequality:

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

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

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

□

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.

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.

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 $\tilde{\mathbf{A}}$.

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.

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 $\tilde{\mathbf{A}}$.

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

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

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*

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

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.

6.8.2 Dependence on coherence

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

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

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 $[\mathbf{U}_1 \quad \mathbf{U}_2]$ 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.

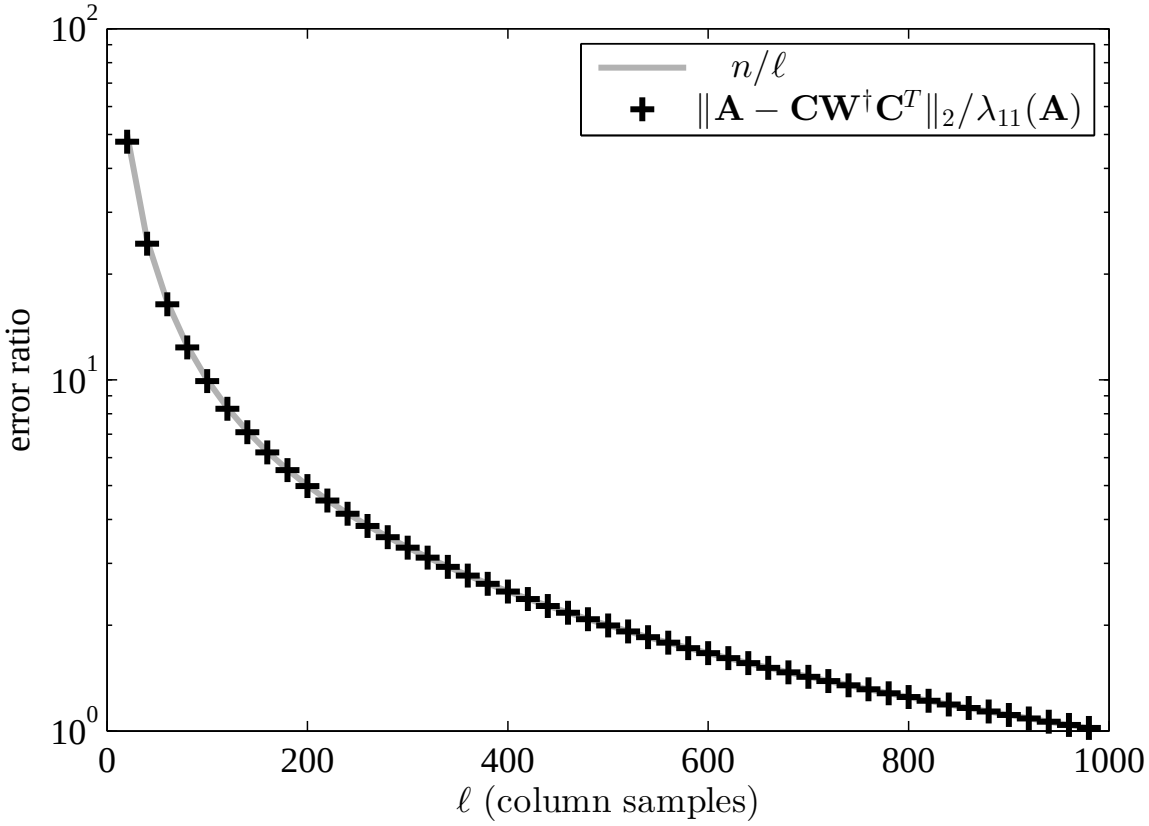
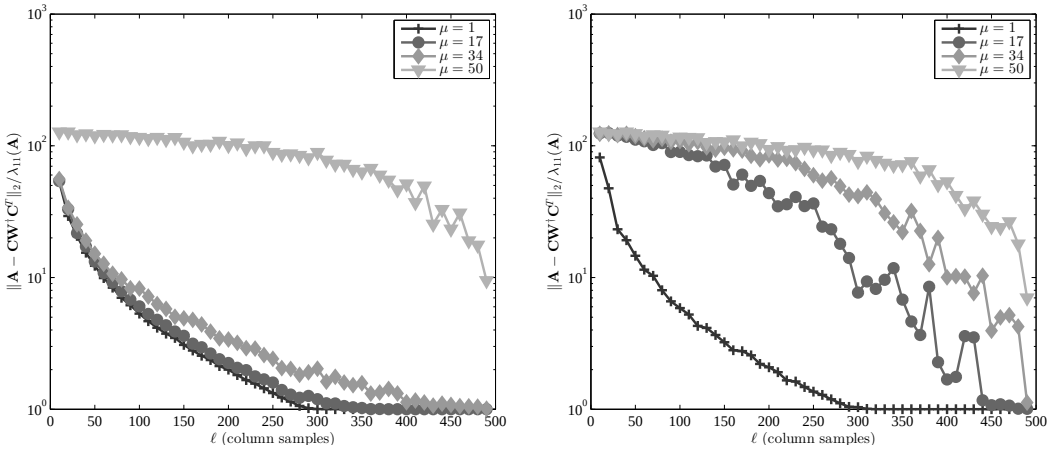


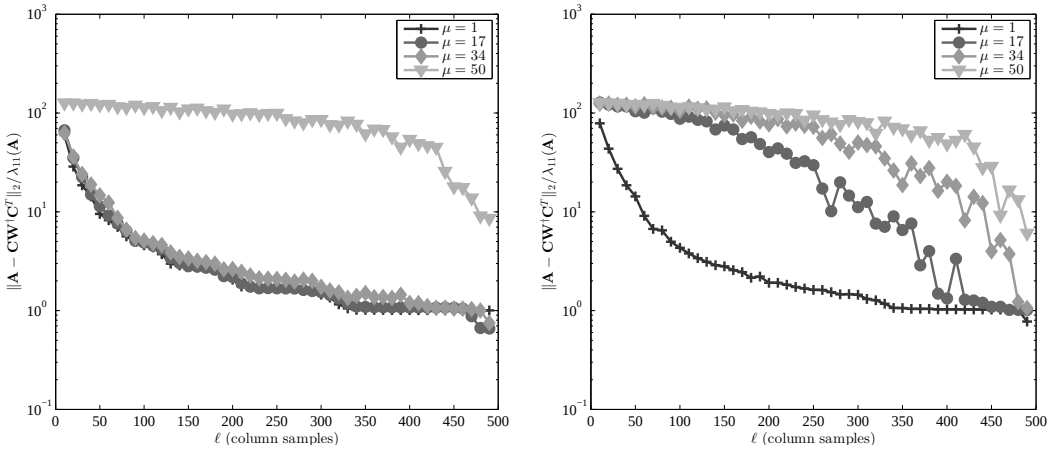
Figure 6.1: EMPIRICAL DEMONSTRATION OF THE OPTIMALITY OF THEOREM 6.9. The empirical spectral-norm error of Nyström extensions of $\mathbf{A}$, the matrix defined in (6.8.1), relative to the spectral-norm error of the optimal rank-10 approximation of $\mathbf{A}$. Each point is the worst relative error 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.

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.



(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.

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 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 $\begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}$ 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.

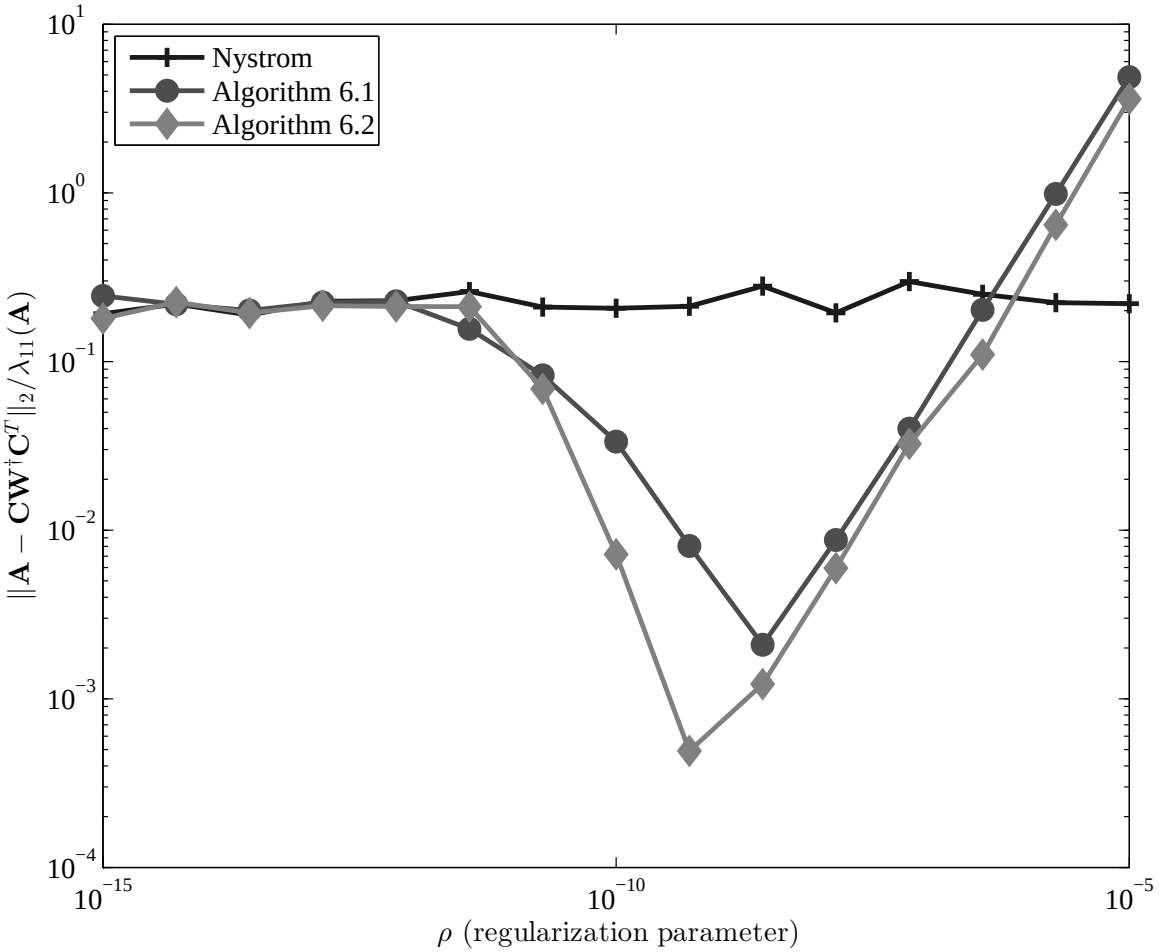


Figure 6.3: SPECTRAL-NORM ERROR OF REGULARIZED NYSTRÖM EXTENSIONS AS REGULARIZATION PARAMETER VARIES. For the matrix $\mathbf{A}$ defined in (6.8.3), the spectral-norm errors of the Nyström extension and the extensions generated using Algorithms 6.1 and 6.2, as a function of the regularization parameter 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.

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.

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.

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
- 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 $A^{(\sigma,\nu,C)}_{ij}$.

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 $\lceil \|\mathbf{A}\|_{\mathrm{F}}^2 / \|\mathbf{A}\|_2^2 \rceil$ (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.

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.

- 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.

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

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

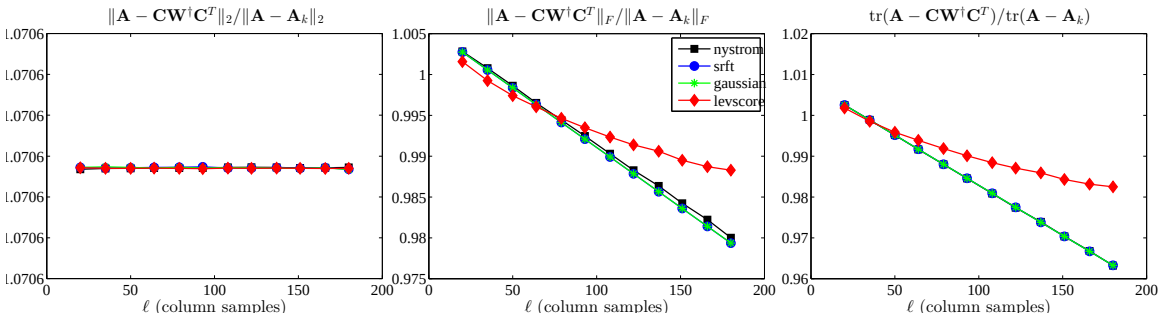
for the non-rank-restricted sketches, and the errors

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

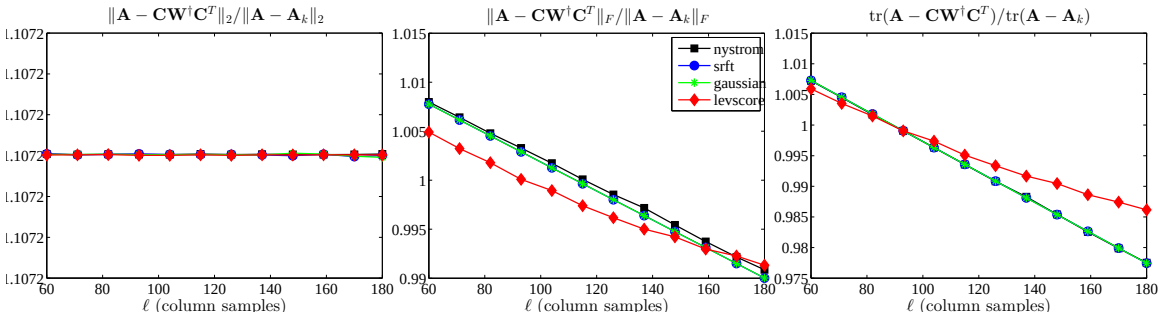
for the rank-restricted sketches.

6.9.3.1 Graph Laplacians

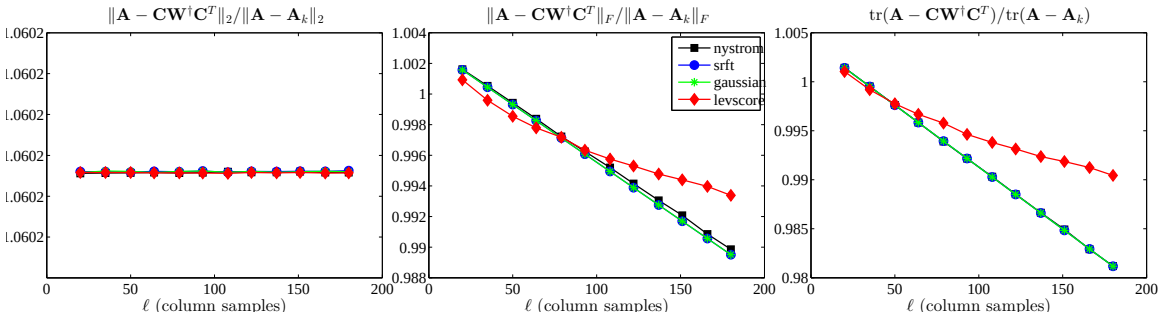
Figures 6.4–6.7 show the reconstruction error results for sampling and mixture methods applied to several normalized graph Laplacians. Figures 6.4 and 6.6 show GR and HEP, each for two values of the rank parameter. The remaining two show Enron and Gnutella, again each for two values of the rank parameter. The first two figures show the ratios of the spectral, Frobenius,



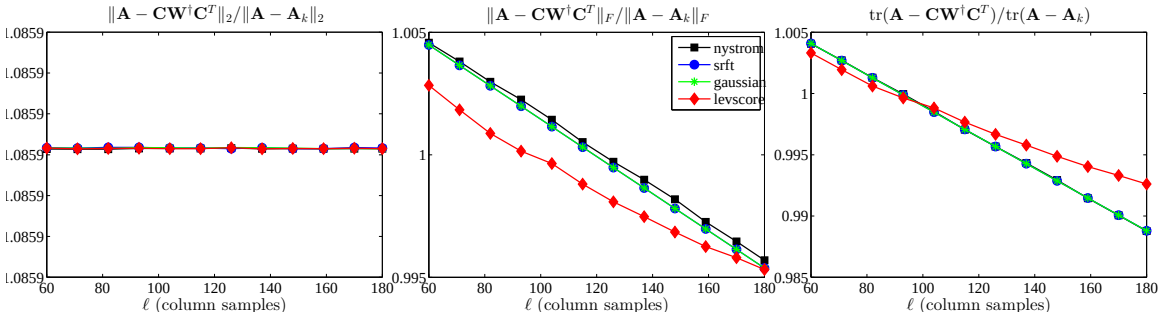
(a) GR, $k = 20$



(b) GR, $k = 60$

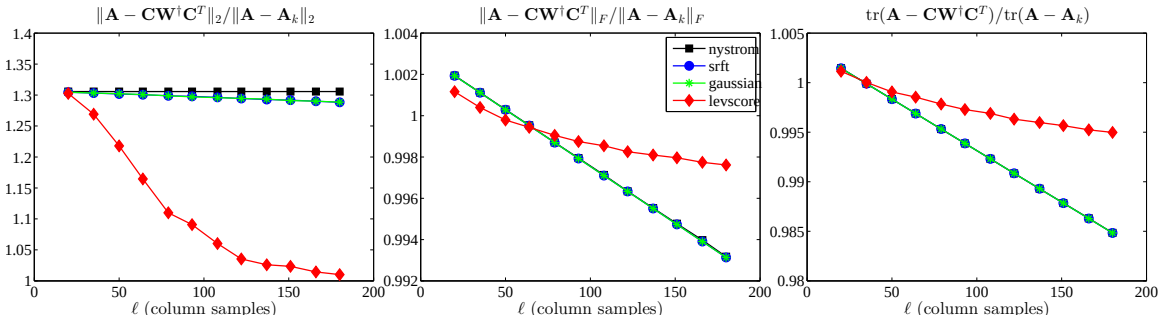


(c) HEP, $k = 20$

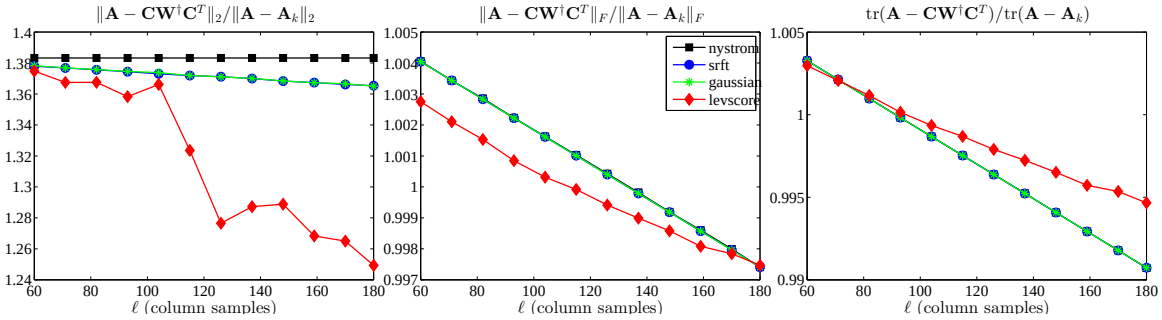


(d) HEP, $k = 60$

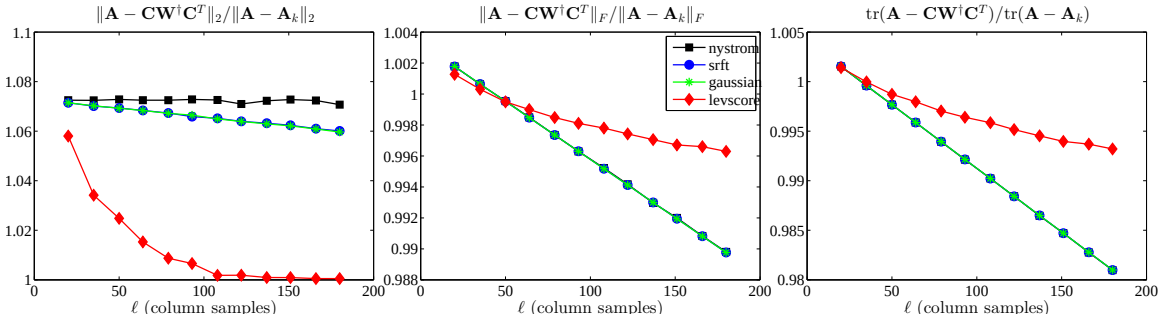
Figure 6.4: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE GR AND HEP LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the GR and HEP Laplacian matrices,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



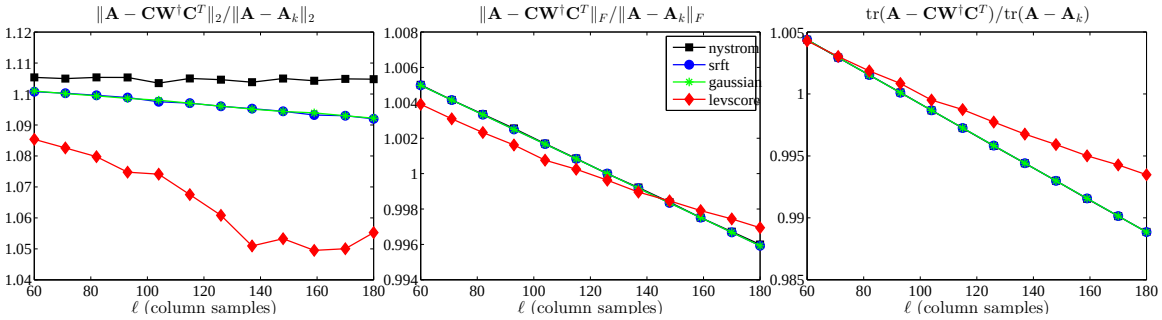
(a) Enron, $k = 20$



(b) Enron, $k = 60$

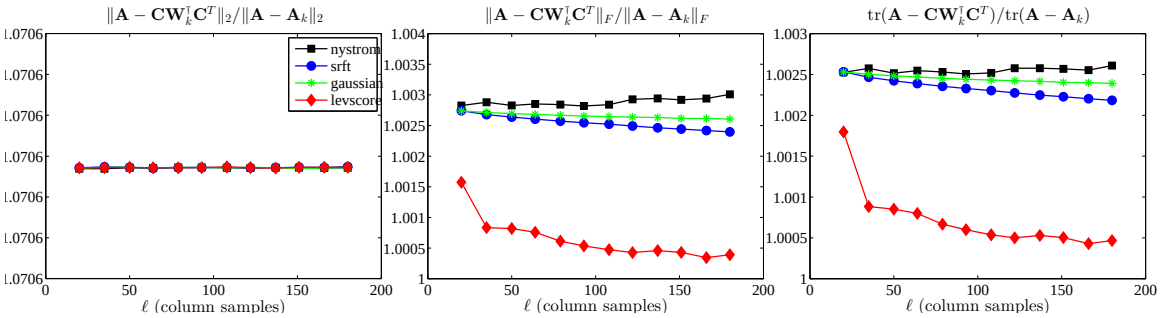


(c) Gnutella, $k = 20$

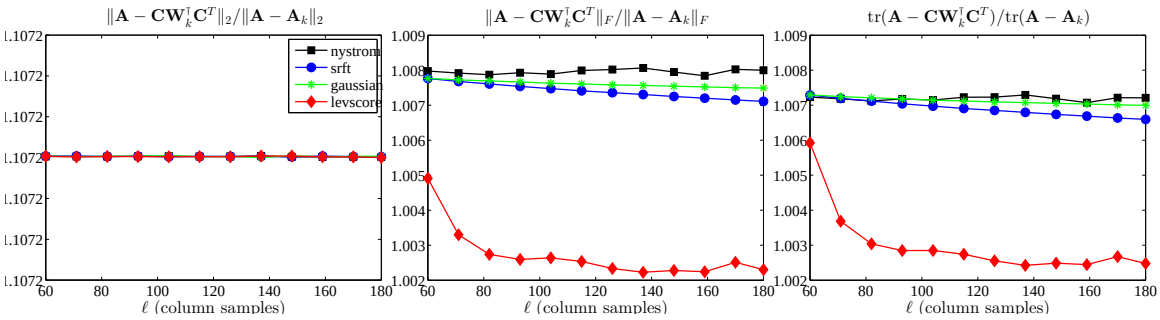


(d) Gnutella, $k = 60$

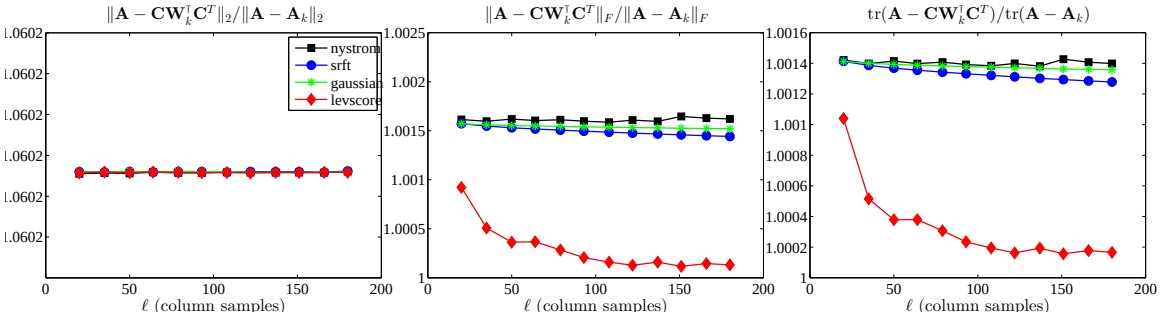
Figure 6.5: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE ENRON AND GNUTELLA LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the Enron and Gnutella Laplacian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



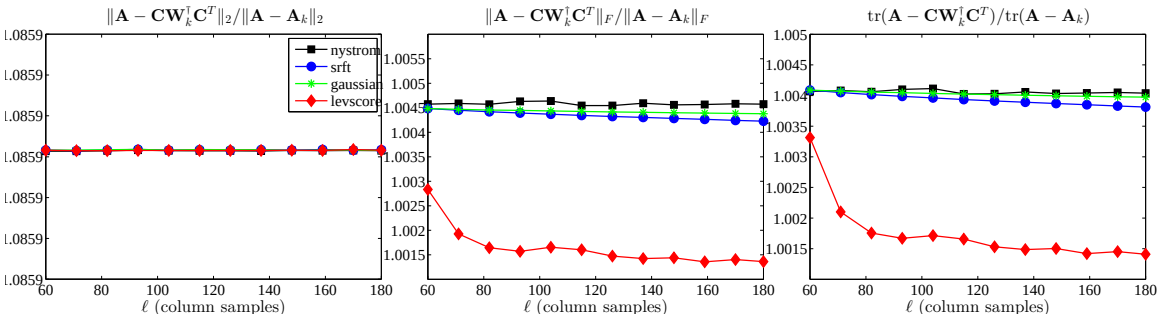
(a) GR, $k = 20$



(b) GR, $k = 60$

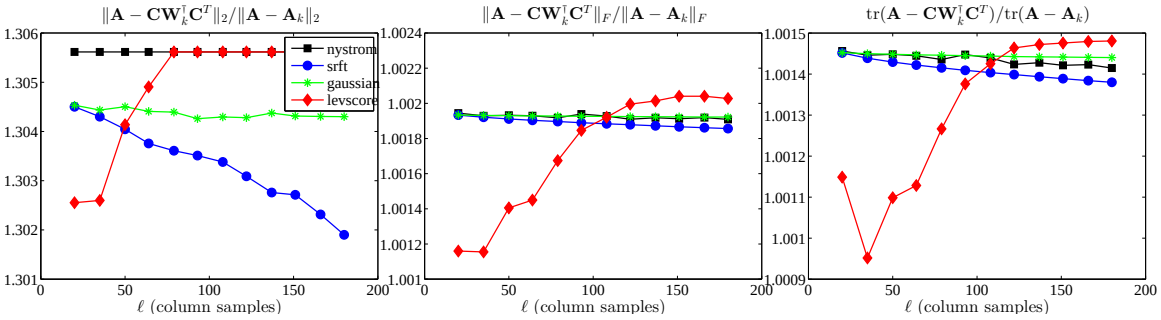


(c) HEP, $k = 20$

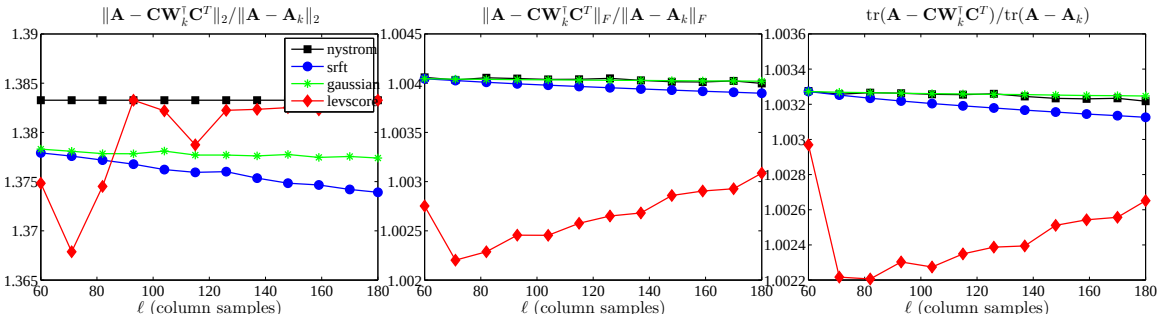


(d) HEP, $k = 60$

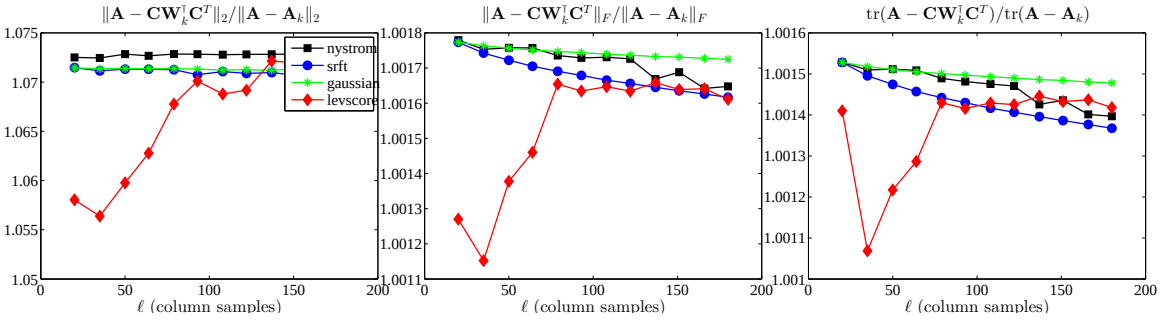
Figure 6.6: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE GR AND HEP LAPLACIAN MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the GR and HEP Laplacian matrices, with two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



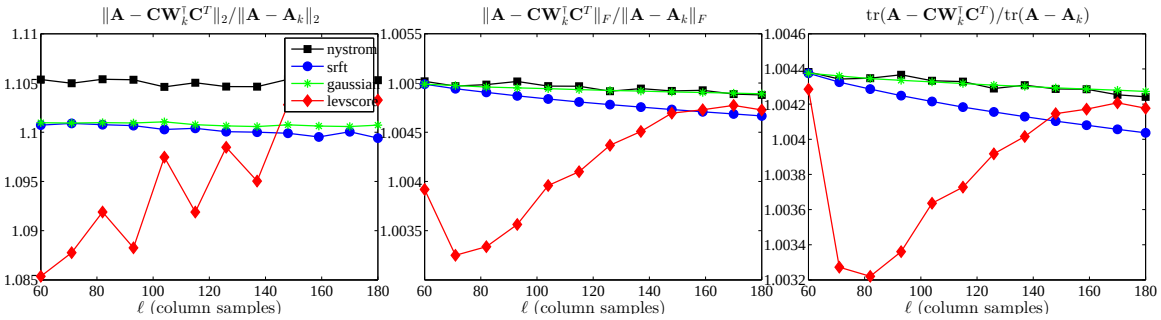
(a) Enron, $k = 20$



(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.

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.

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 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.

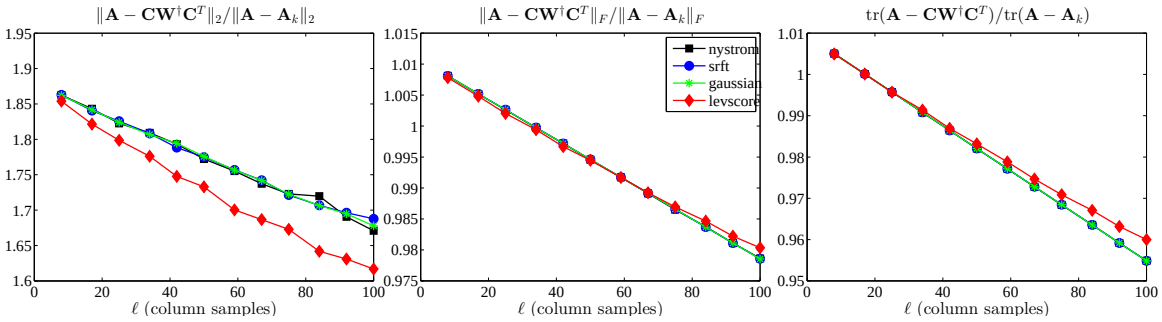
- The scale of the vertical axes is much larger than for the Laplacian matrices, mostly since the matrices are much better approximated by low-rank matrices, although the scale decreases as one goes from spectral to Frobenius to trace reconstruction error, as before.

- 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.

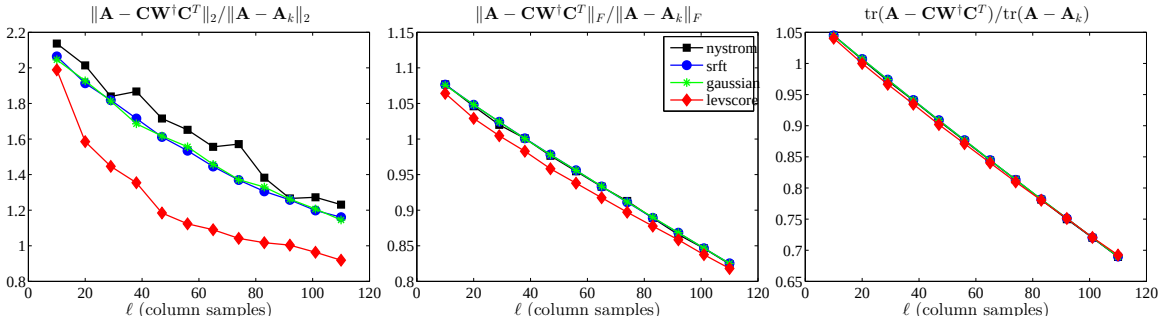
These linear kernels (and also to some extent the dense RBF kernels below that have larger σ parameter) are examples of relatively "nice" machine learning matrices that are similar to matrices where uniform sampling has been shown to perform well previously [TKR08, KMT09a, KMT09b, KMT12]; and for these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

6.9.3.3 Dense and sparse RBF kernels

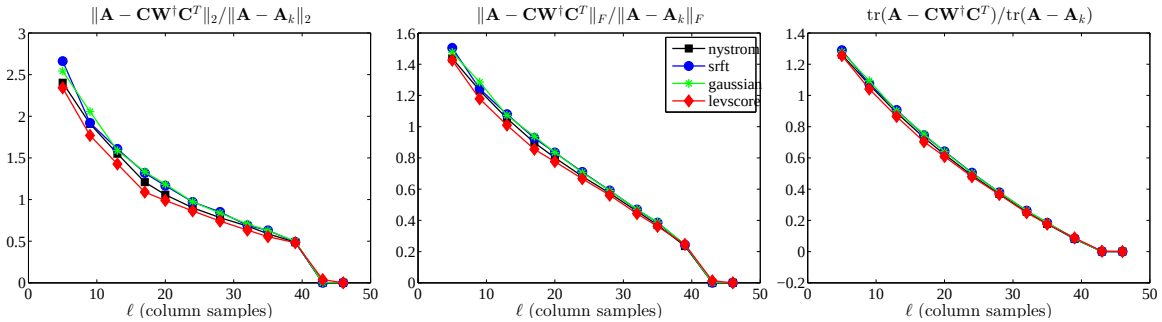
Figures 6.10–6.13 present the reconstruction error results for sampling and mixture methods applied to several dense RBF and sparse RBF kernels. Several observations are worth making about the results presented in these figures.



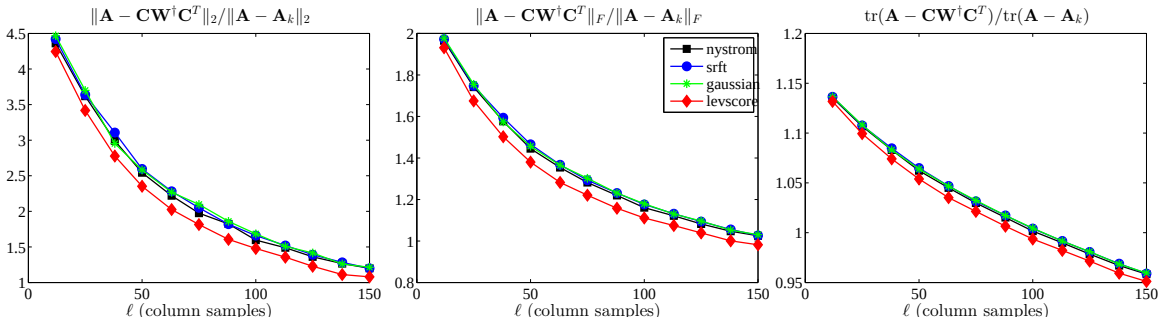
(a) Dexter, $k = 8$



(b) Protein, $k = 10$

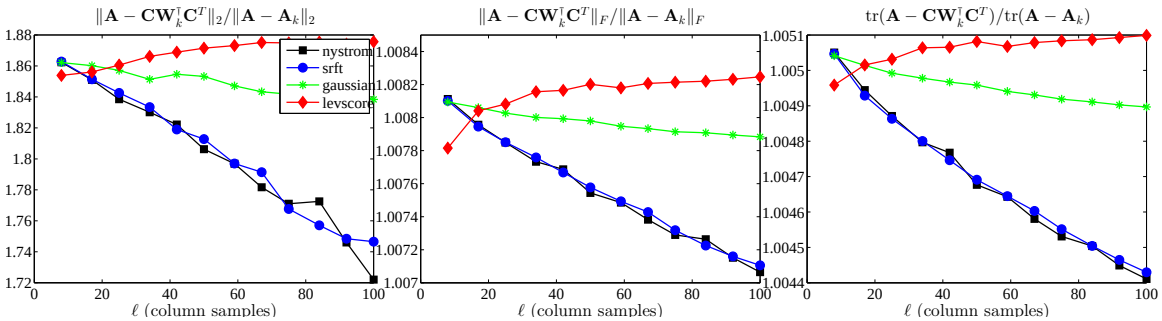


(c) SNPs, $k = 5$

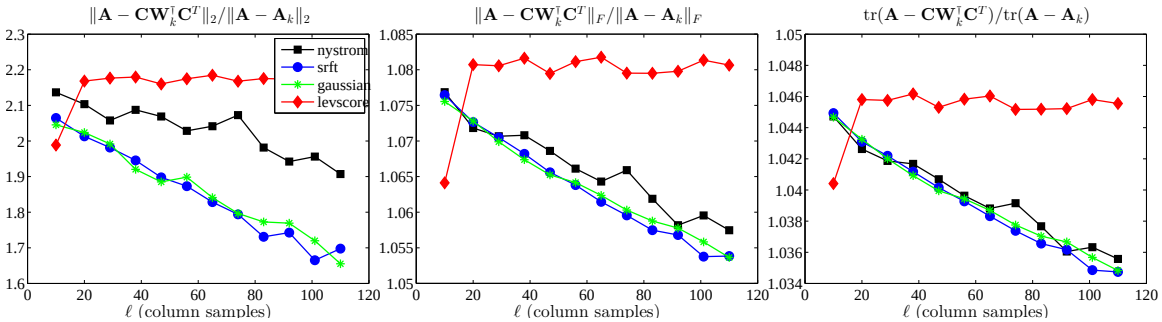


(d) Gisette, $k = 12$

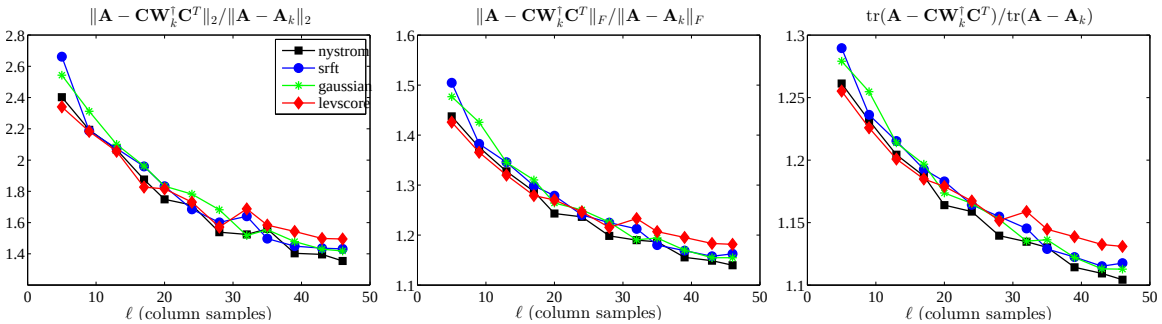
Figure 6.8: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE LINEAR KERNEL MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the linear kernel matrices.



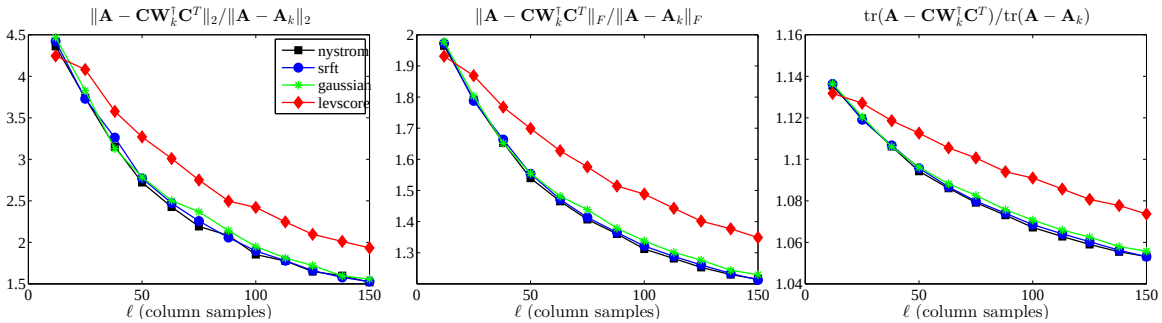
(a) Dexter, $k = 8$



(b) Protein, $k = 10$

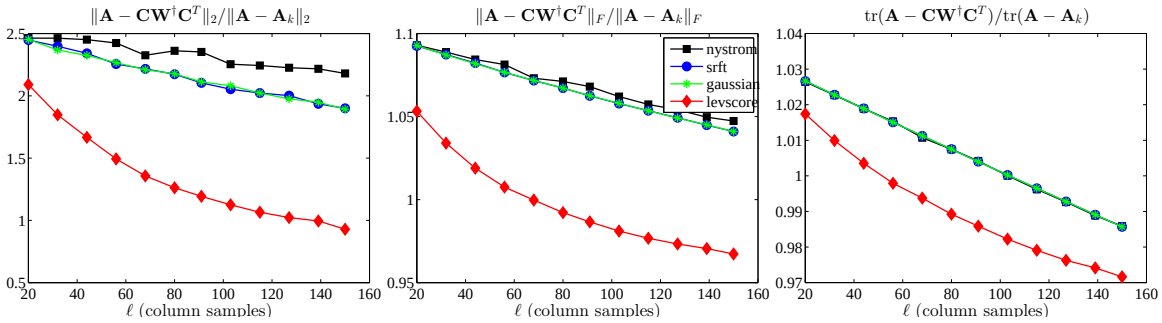


(c) SNPs, $k = 5$

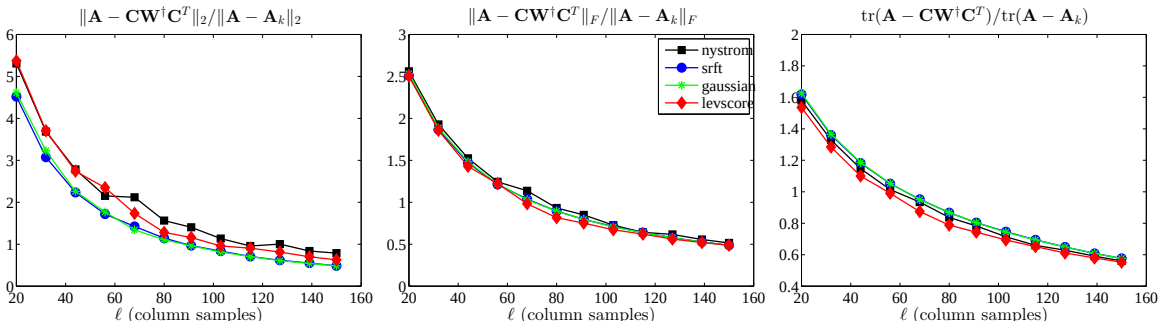


(d) Gisette, $k = 12$

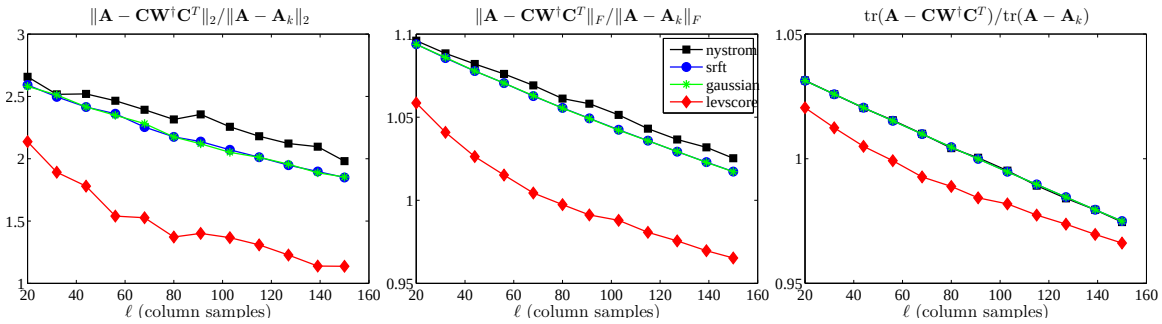
Figure 6.9: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE LINEAR KERNEL MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the linear kernel matrices.



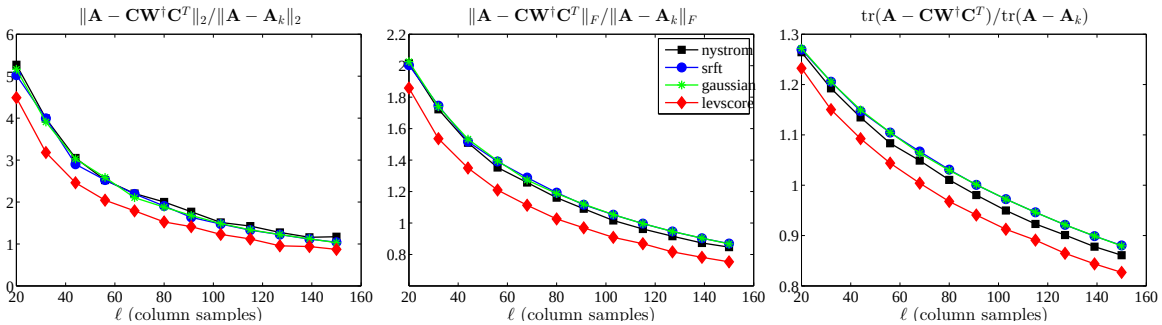
(a) AbaloneD, $\sigma = .15, k = 20$



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

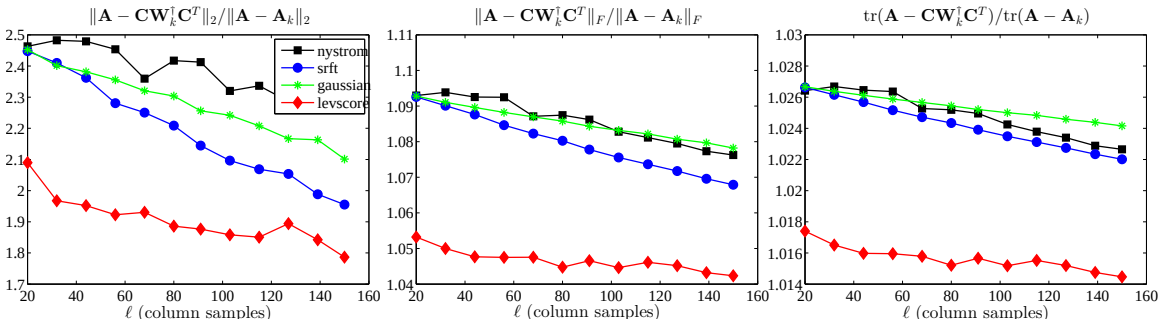


(c) WineD, $\sigma = 1, k = 20$

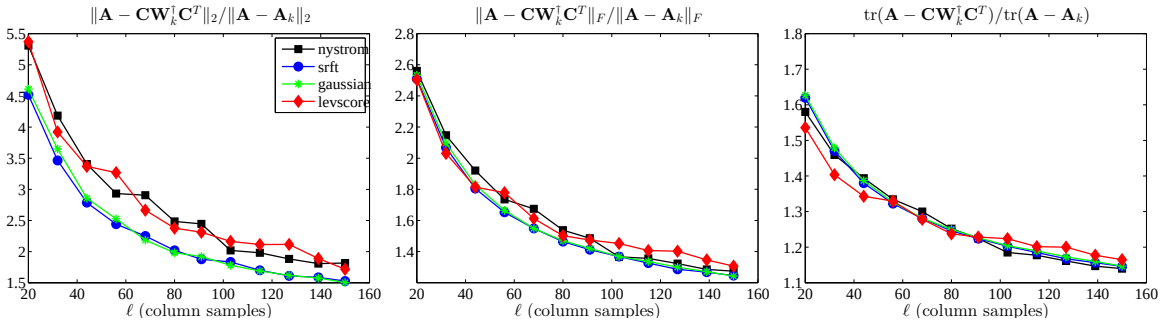


(d) WineD, $\sigma = 2.1, k = 20$

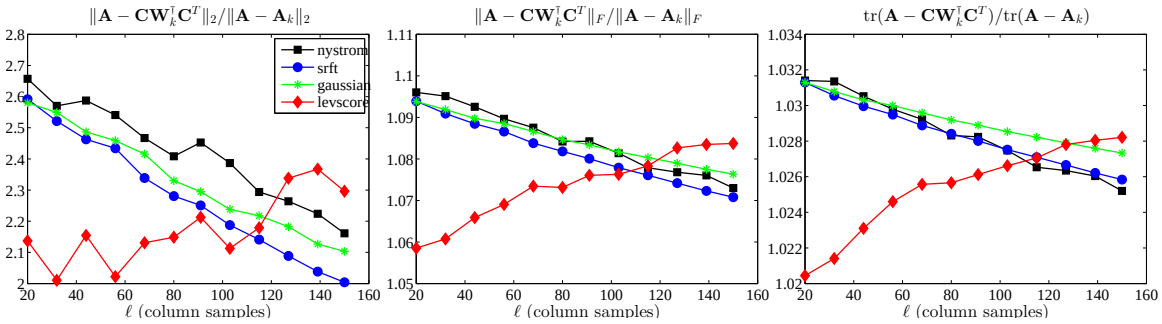
Figure 6.10: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE DENSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the dense RBFK matrices.



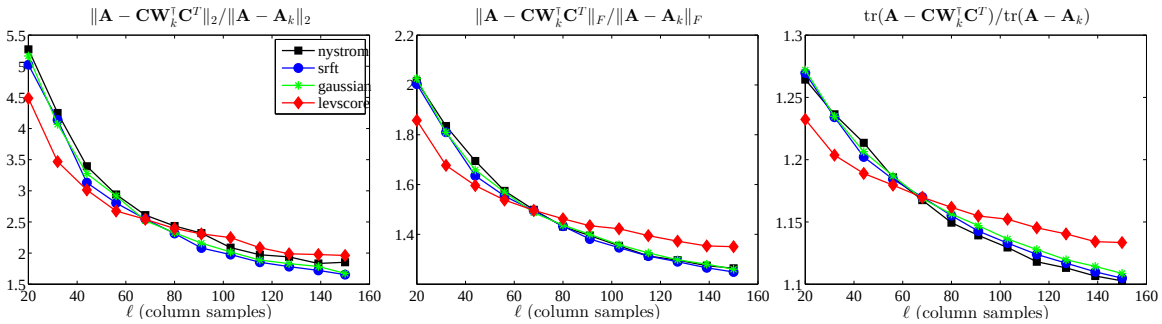
(a) AbaloneD, $\sigma = .15, k = 20$



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

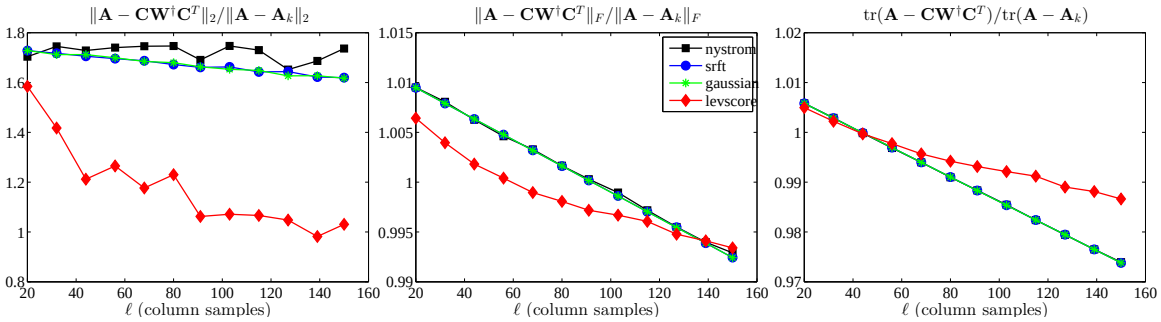


(c) WineD, $\sigma = 1, k = 20$

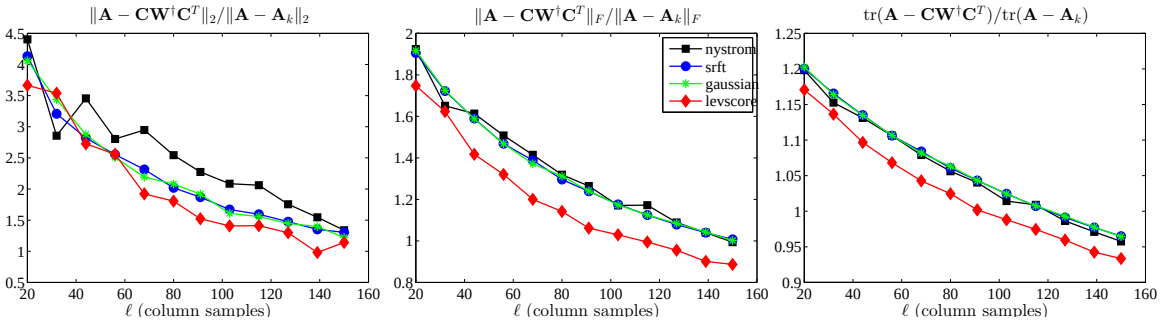


(d) WineD, $\sigma = 2.1, k = 20$

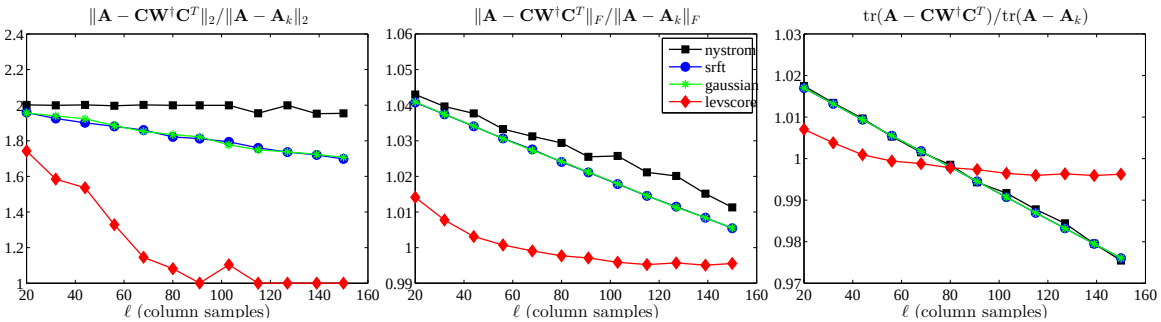
Figure 6.11: RELATIVE ERRORS OF RANK-RESTRICTED SPSD SKETCHES OF THE DENSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.2) of several rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the dense RBFK matrices.



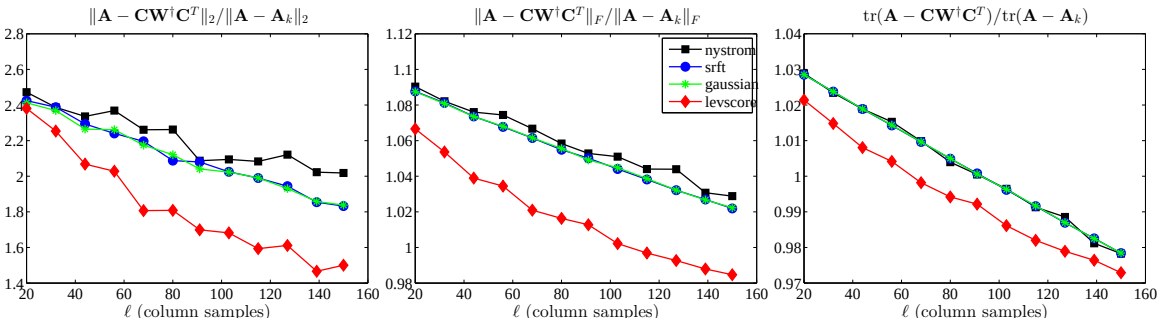
(a) AbaloneS, $\sigma = .15, k = 20$



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

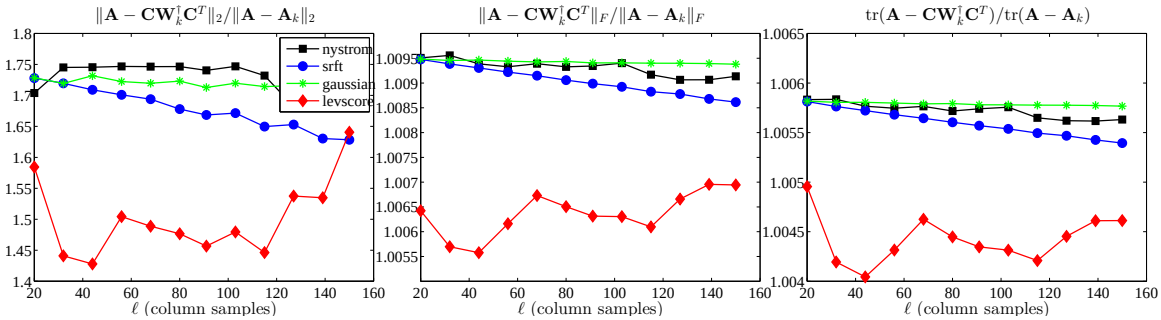


(c) WineS, $\sigma = 1, k = 20$

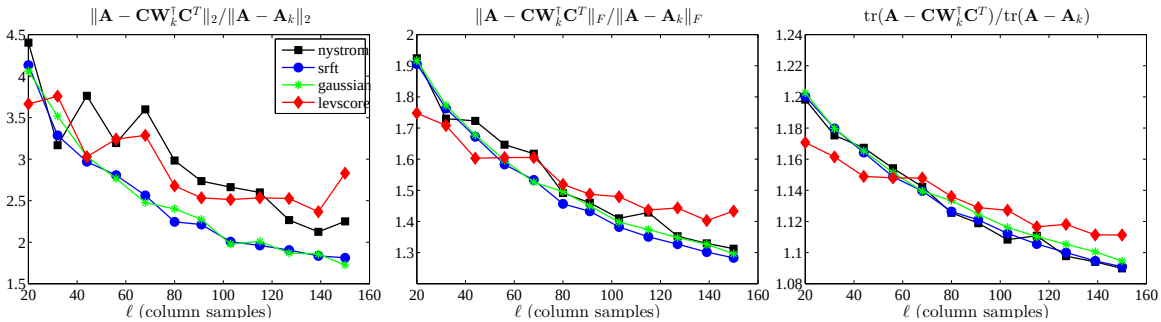


(d) WineS, $\sigma = 2.1, k = 20$

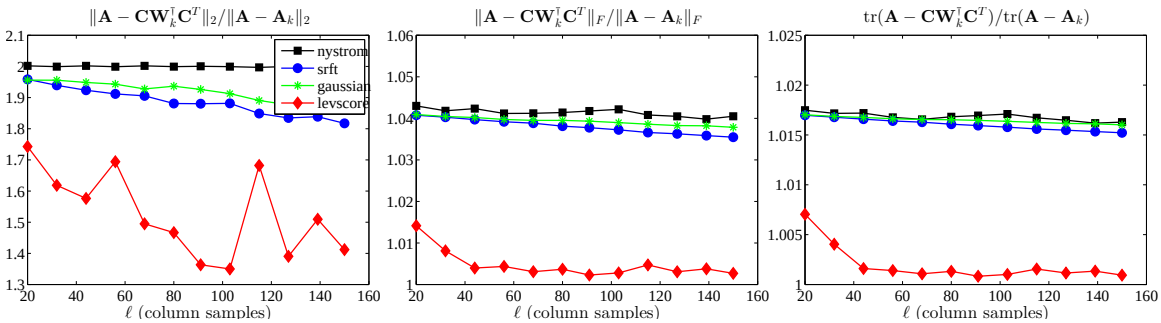
Figure 6.12: RELATIVE ERRORS OF NON-RANK-RESTRICTED SPSD SKETCHES OF THE SPARSE RBFK MATRICES. The relative spectral, Frobenius, and trace-norm errors (6.9.1) of several non-rank-restricted SPSD sketches, as a function of the number of columns samples ℓ, for the sparse RBFK matrices.



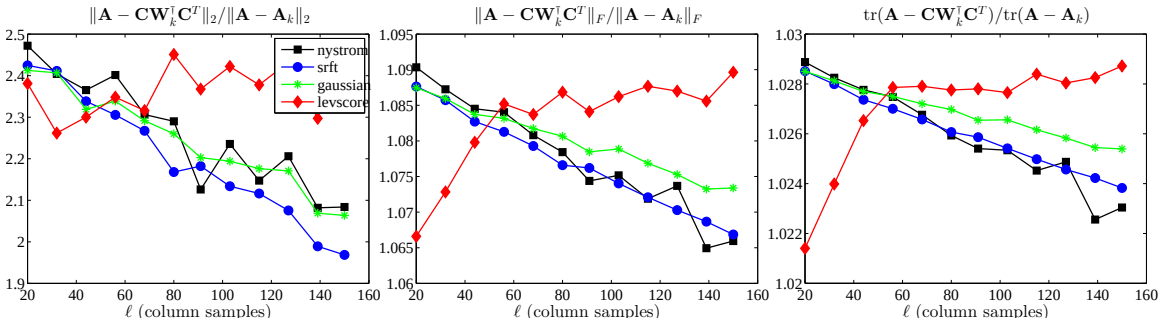
(a) AbaloneS, $\sigma = .15, k = 20$



(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 $\mathbf{C} = \mathbf{AX}$ and $\mathbf{W} = \mathbf{X}^T\mathbf{AX}$,

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

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

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

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.

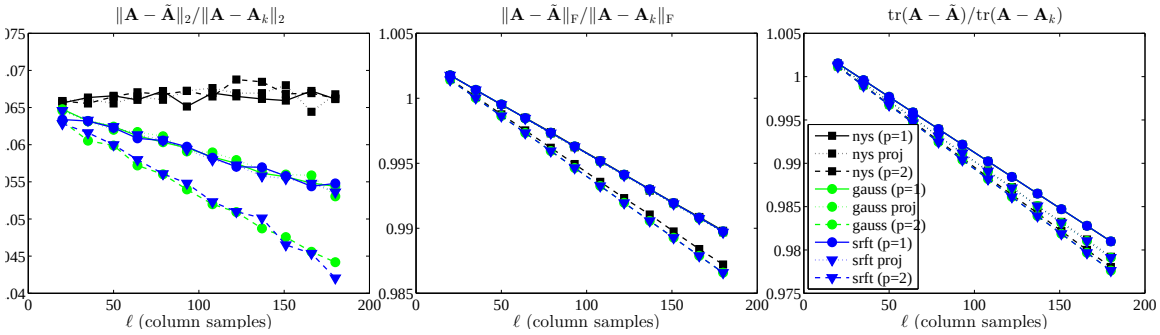
Some trends are clear from Figure 6.14.

- 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.

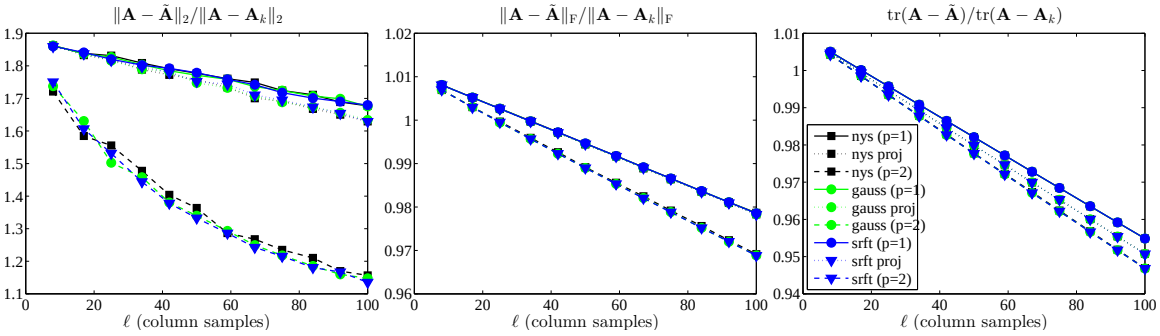
- After the prolonged approximants, the pinched Gaussian and SRFT column-mixture approximants have the smallest spectral, Frobenius, and trace-norm errors. Again however, we see that the pinched uniformly column-sampled approximants are considerably less accurate than the pinched Gaussian and SRFT column-mixture 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.

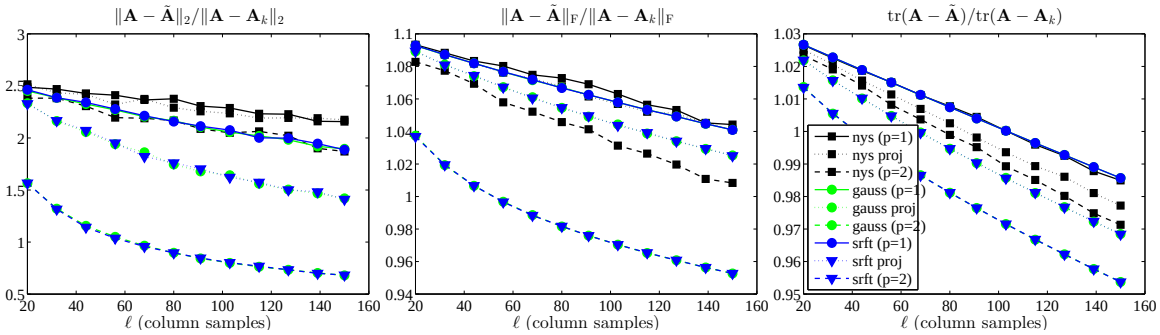
It is also evident that the pinched approximants often yield a much slighter increase in accuracy over the one-pass sketches than do the prolonged approximants. Recall that the prolonged approximants are simply two-pass sketches. Our investigations point to the conclusion that two-pass sketches are significantly 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.



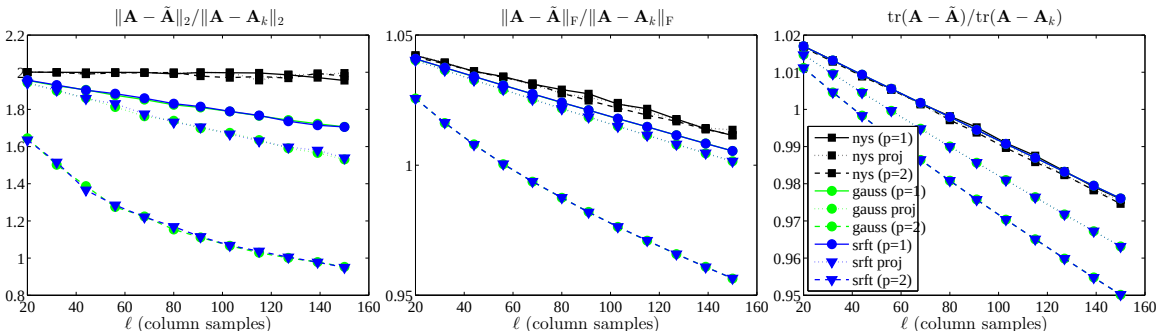
(a) Gnutella, $k = 20$



(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.

Bibliography

[AC06] N. Ailon and B. Chazelle, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*, Proc. 38th ACM Symposium on Theory of Computing, 2006, pp. 557–563. 1.2, 5.1, 5.3.1

[AHK05] S. Arora, E. Hazan, and S. Kale, *Fast algorithms for approximate semidefinite programming using the multiplicative weights update method*, 46th IEEE Symposium on Foundations of Computer Science, 2005. 1.1

[AHK06] S. Arora, E. Hazan, and S. Kale, *A Fast Random Sampling Algorithm for Sparsifying Matrices*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer Berlin, 2006, pp. 272–279. 1.1, 1.3, 1.4.2, 1.4.2, 3.6, 3.6.1, 3.6.1.3, 3.7

[AKL13] D. Achlioptas, Z. Karnin, and E. Liberty, *Matrix entry-wise sampling: Simple is best*, Submitted to KDD 2013, 2013. 1.1, 1.4.2, 3.7

[AKV02] N. Alon, M. Krivelevich, and V. H. Vu, *On the concentration of eigenvalues of random symmetric matrices*, Israel J. Math. **131** (2002), 259–267. 1.3

[AL08] N. Ailon and E. Liberty, *Fast dimension reduction using Rademacher series on dual BCH codes*, Proc. 19th ACM-SIAM Symposium on Discrete Algorithms, 2008, pp. 1–9. 5.3

[ALPTJ11] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, *Sharp bounds on the rate of convergence of the empirical covariance matrix*, C. R. Math. Acad. Sci. Paris **349** (2011), 195–200. 1.3, 2.8, 2.8.2

[AM01] D. Achlioptas and F. McSherry, *Fast Computation of Low Rank Matrix Approximations*, Proc. 33rd ACM Symposium on Theory of Computing, 2001, pp. 611–618. 1.1, 1.3, 1.4.2, 1.4.2, 3.7

[AM07] ———, *Fast Computation of Low Rank Matrix Approximations*, J. ACM **54** (2007), no. 2. 1.1, 1.3, 1.4.2, 1.4.2, 3.6, 3.6.1, 3.7

[AMT10] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging LAPACK's least-squares solver*, SIAM J. Sci. Comput. **32** (2010), no. 3, 1217–1236. 5.1

[AN04] N. Alon and A. Naor, *Approximating the Cut-Norm via Grothendieck's inequality*, Proc. 36th ACM symposium on Theory of Computing, 2004, pp. 72–80. 3.1.0.1

[AW02] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579. 1.3, 1.4.1, 2.5

[BDHS11] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz, *Minimizing Communication in Numerical Linear Algebra*, SIAM J. Matrix Anal. Appl. **32** (2011), 866–901. 5.1

[BDMI11] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near optimal column-based matrix reconstruction*, Proc. 52nd IEEE Symposium on Foundations of Computer Science (FOCS), 2011, IEEE, 2011, pp. 305–314. 1.1, 1.2, 4.2.1, 4.2.1.1, 4.2.1.2, 4.7, 6.11, 6.8.1

[BF12] A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, Multiscale Model. Simul. **10** (2012), 1090–1118. 6.7

[BL13] K. Bache and M. Lichman, *UCI Machine Learning Repository*, 2013, http://www.ics.uci.edu/~mlearn/MLRepository.html, Retrieved June 10, 2013. 6.2

[BLM03] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities using the entropy method*, Ann. Probab. **31** (2003), 1583–1614. 3.1, 3.6, 4.1.3

[BMD09] C. Boutsidis, M. W. Mahoney, and P. Drineas, *An Improved Approximation Algorithm for the Column Selection Problem*, Proc. 20th ACM-SIAM Symposium on Discrete Algorithms (SODA 2009), 2009. 4.2.1, 4.2.2, 4.8

[BSS09] J. Batson, D. Spielman, and N. Srivastava, *Twice-Ramanujan Sparsifiers*, Proc. 41st ACM Symposium on Theory of Computing, 2009, pp. 255–262. 3, 3.1.0.1

[BW09] M. Belabbas and P. J. Wolfe, *Spectral methods in machine learning and new strategies for very large datasets*, Proc. Nat. Acad. Sci. **106** (2009), 369–374. 6.3, 6.9.2

[CC00] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, 2nd ed., Chapman and Hall/CRC, 2000. 1

[CD05] Z. Chen and J. J. Dongarra, *Condition numbers of Gaussian random matrices*, SIAM J. Matrix Anal. Appl. **27** (2005), 603–620. 1.3

[CD11] J. Chiu and L. Demanet, *Sublinear randomized algorithms for skeleton decompositions*, Preprint, arXiv:1110.4193, 2011. 1.4.4, 6.1, 6.7, 6.2

[CH92] T. F. Chan and P. C. Hansen, *Some Applications of the Rank Revealing QR Factorization*, SIAM J. Sci. Comput. **13** (1992), 727–741. 1

[Cha07] S. Chatterjee, *Stein's method for concentration inequalities*, Probab. Theory Relat. Fields **138** (2007), 305–321. 1.3

[CM08] D. Christofides and K. Markström, *Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales*, Random Structures Algorithms **32** (2008), 88–100. 1.3

[Cor96] P. I. Corke, *A Robotics Toolbox for MATLAB*, IEEE Robotics and Automation Magazine **3** (1996), 24–32. 6.2

[CP11] E. Candès and Y. Plan, *A probabilistic and RIPless theory of compressed sensing*, IEEE Trans. Inf. Theory **57** (2011), 7235–7254. 4.1.3

[CR09] E. Candès and B. Recht, *Exact Matrix Completion via Convex Optimization*, Found. Comput. Math. **9** (2009), 717–772. 1.4.4, 5.5.1

[CW09] K. L. Clarkson and D. P. Woodruff, *Numerical Linear Algebra in the Streaming Model*, Proc. 41st ACM Symposium on Theory of Computing, 2009, pp. 205–214. 1, 5.1

[CW12] _____, *Low Rank Approximation and Regression in Input Sparsity Time*, Preprint, arXiv:1207.6365, 2012. 5.1

[d'A11] A. d'Aspremont, *Subsampling Algorithm for Semidefinite Programming*, Stoch. Syst. **2** (2011), no. 1, 274–305. 1.1

[DDF+90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, J. Am. Soc. Inf. Sci. Technol. **41** (1990), 391–407. 1

[DG98] V. De la Peña and E. Giné, *Decoupling: From Dependence to Independence*, Probability and its Applications, Springer, 1998. 2.6

[DK01] P. Drineas and R. Kannan, *Fast Monte-Carlo Algorithms for Approximate Matrix Multiplication*, Proc. 42nd IEEE Symposium on the Foundations of Computer Science, 2001, pp. 452–459. 1.1, 4.1.3

[DK03] _____, *Pass efficient algorithms for approximating large matrices*, Symposium on Discrete Algorithms, 2003, pp. 223–232. 1.1

[DKM06a] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication*, SIAM J. Comput. **36** (2006), no. 1, 132–157. 1.1, 4.1.3

[DKM06b] _____, *Fast Monte Carlo Algorithms for Matrices II: Computing Low-Rank Approximations to a Matrix*, SIAM J. Comput. **36** (2006), no. 1, 158–183. 1.1, 5.1

[DKM06c] _____, *Fast Monte Carlo Algorithms for Matrices III: Computing an Efficient Approximate Decomposition of a Matrix*, SIAM J. Comput. **36** (2006), no. 1, 184–206. 1.1

[DM05] P. Drineas and M. W. Mahoney, *On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning*, J. Mach. Learn. Res. **6** (2005), 2153–2175. 6.3, 6.1, 6.13, 6.9.3.4

[DM09] _____, *CUR matrix decompositions for improved data analysis*, Proc. Nat. Acad. Sci. USA **106** (2009), 697–702. 6.2, 6.6.1, 6.15

[DM10] _____, *Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving*, Preprint, arXiv:1005.3097, 2010. 1.1, 6.2

[DMIMW12] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, *Fast approximation of matrix coherence and statistical leverage*, J. Mach. Learn. Res. **13** (2012), 3475–3506. 6.2, 6.14

[DMM08] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Relative-Error CUR Matrix Decompositions*, SIAM J. Matrix Anal. Appl. **30** (2008), 844–881. 1.1, 6.15

[DMMS11] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math. **117** (2011), 219–249. 1.3

[Dri02] P. Drineas, *Randomized Algorithms for Matrix Operations*, Ph.D. thesis, Yale University, 2002. 5.3.2

[DRVW06] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, *Matrix Approximation and Projective Clustering via Volume Sampling*, Theory Comput. **2** (2006), 225–247. 1.1

[DZ11] P. Drineas and A. Zouzias, *A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality*, Inform. Process. Lett. **111** (2011), 385–389. 1.1, 1.3, 1.4.2, 3.7

[Far10] B. Farrell, *Limiting Empirical Singular Value Distribution of Restrictions of Discrete Fourier Transform Matrices*, J. Fourier Anal. Appl. (2010), 1–21. 2.7

[FBCM04] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, *Spectral Grouping Using the Nyström Method*, IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), 214–225. 6.7

[FKV98] A. Frieze, R. Kannan, and S. Vempala, *Fast Monte-Carlo Algorithms for finding low-rank approximations*, Proc. 39th Symposium on Foundations of Computer Science, 1998, pp. 378–390. 1.1, 5.1

[FKV04] ———, *Fast Monte-Carlo algorithms for finding low-rank approximations*, J. ACM **51** (2004), 1025–1041. 1.1

[FNL+09] P. Freeman, J. Newman, A. Lee, J. Richards, and C. Schafer, *Photometric redshift estimation using spectral connectivity analysis*, Mon. Not. R. Astron. Soc. **398** (2009), 2012–2021. 6.7

[GB12] A. Gittens and C. Boutsidis, *Improved matrix algorithms via the Subsampled Randomized Hadamard Transform*, SIAM J. Matrix Anal. Appl., to appear. Preprint available at arXiv:1204.0062, 2012. 1

[GE96] M. Gu and S. C. Eisenstat, *Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization*, SIAM J. Sci. Comput. **17** (1996), 848–869. 5.5.3

[Gen02] M. Genton, *Classes of Kernels for Machine Learning: A Statistics Perspective*, J. Mach. Learn. Res. **2** (2002), 299–312. 6.9.1

[GGBHD05] I. Guyon, S. R. Gunn, A. Ben-Hur, and G. Dror, *Result analysis of the NIPS 2003 feature selection challenge*, Advances in Neural Information Processing Systems 17, MIT Press, 2005. 6.2

[GH08] S. Gurevich and R. Hadani, *The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries*, Preprint, arXiv:0812.2602, 2008. 2.7

[Git11] A. Gittens, *The spectral norm error of the naive Nystrom extension*, Preprint, arXiv:1110.5305. Submitted to SIAM J. Matrix Anal. Appl., 2011. 1

[GM13a] A. Gittens and Mahoney M., *Revisiting the Nyström Method for Improved Large-Scale Machine Learning*, Preprint, arXiv:1303.1849, 2013. 1

[GM13b] A. Gittens and M. Mahoney, *Revisiting the Nyström method for improved large-scale machine learning*, Proc. 30th International Conference on Machine Learning, 2013. 1

[GN10] D. Gross and V. Nesme, *Note on sampling without replacing from a finite collection of matrices*, Preprint, arXiv:1001.2738, January 2010. 4.1.3

[Gro11] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory **57** (2011), 1548–1566. 1.3, 4.1.3, 4.1.3

[GSP+06] A. M. Gustafson, E. S. Snitkin, S. C. J. Parker, C. DeLisi, and S. Kasif, *Towards the identification of essential genes using targeted genome sequencing and comparative analysis*, BMC Genomics **7** (2006), 265. 6.2

[GT09] A. Gittens and J. A. Tropp, *Error bounds for random matrix approximation schemes*, Preprint, arXiv:0911.4108, 2009. 1

[GT11] ———, *Tail bounds for all eigenvalues of a sum of random matrices*, Preprint, arXiv:1104.4513, 2011. 1

[GV96] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, 1996. 4.2.2.1, 4.2.2.1, 5.1, 6.7

[Han90] P. C. Hansen, *Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank*, SIAM J. Sci. Comput. **11** (1990), 503–518. 1

[HJ85] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985. 2.1

[HLMS04] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, *A kernel view of the dimensionality reduction of manifolds*, Proc. 21th International Conference on Machine Learning, 2004. 1

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev. **53** (2011), 217–288. 1.2, 1.4.3, 4.2.1, 4.2.2, 5.1, 5.1, 5.2, 5.2.1, 5.2.1, 5.2.1, 5.2.1, 5.3, 5.4, 6.1, 6.1.1, 6.4.1, 6.4.2, 6.4.2, 6.6.2, 6.6.3, 6.10

[Hoe63] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30. 4.1.3

[HP06] S. Har-Peled, *Low rank matrix approximation in linear time*, Manuscript, 2006. 1.1

[HTF08] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2008. 1

[IW12] I. F. Ipsen and T. Wentworth, *The Effect of Coherence on Sampling from Matrices with Orthonormal Columns, and Preconditioned Least Squares Problems*, Preprint, arXiv:1203.4809, March 2012. 6.8.2

[Kar94a] D. R. Karger, *Random sampling in cut, flow, and network design problems*, Proc. 26th ACM Symposium on Theory of Computing, May 1994, pp. 648–657. 3.1.0.1

[Kar94b] _____, *Using randomized sparsification to approximate minimum cuts*, Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms, January 1994, pp. 424–432. 3.1.0.1

[Kar95] _____, *Random Sampling in Graph Optimization Problems*, Ph.D. thesis, Stanford University, 1995. 3.1.0.1

[Kar96] _____, *Approximating s–t minimum cuts in* $\tilde{O}(n^2)$ *time*, Proc. 28th ACM Symposium on Theory of Computing, May 1996, pp. 47–55. 3.1.0.1

[KMT09a] S. Kumar, M. Mohri, and A. Talwalkar, *On sampling-based approximate spectral decomposition*, Proc. 26th International Conference on Machine Learning, 2009. 6.9.3.2

[KMT09b] _____, *Sampling Techniques for the Nyström Method*, Proc. 12th International Workshop on Artificial Intelligence and Statistics, 2009, pp. 304–311. 6.9.3.2

[KMT12] _____, *Sampling Methods for the Nyström Method*, J. Mach. Learn. Res. **13** (2012), 981–1006. 6.3, 6.9.2, 6.9.3.2, 6.9.3.4

[KW92] J. Kuczynski and H. Wozniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl. **13** (1992), 1094–1122. 5.1

[KY04] B. Klimt and Y. Yang, *The Enron Corpus: A New Dataset for Email Classification Research*, Proc. 15th European Conference on Machine Learning, 2004, pp. 217–226. 6.2

[Lat05] R. Latała, *Some estimates of norms of random matrices*, Proc. Amer. Math. Soc. **133** (2005), 1273–1282. 1.3, 1.4.2, 3, 3.6, 3.6

[Led96] M. Ledoux, *On Talagrand's deviation inequalities for product measures*, ESAIM Probab. Stat. **1** (1996), 63–87. 4.1

[Lie73] E. H. Lieb, *Convex trace functions and the Wigner–Yanase–Dyson conjecture*, Adv. Math. **11** (1973), no. 3, 267–288. 2.4

[LKF07] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph Evolution: Densification and Shrinking Diameters*, ACM Transactions on Knowledge Discovery from Data **1** (2007). 6.2

[LPP91] F. Lust-Piquard and G. Pisier, *Noncommutative Khintchine and Paley Inequalities*, Ark. Mat. **29** (1991), 241–260. 1.3

[LS05] E. H. Lieb and R. Seiringer, *Stronger subadditivity of entropy*, Phys. Rev. A **71** (2005), no. 6. 2.4, 2.4

[LT91] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, Springer-Verlag, 1991. 3.2, 4.1.3

[Mah11] M. Mahoney, *Randomized algorithms for matrices and data*, Foundations and Trends in Machine Learning **3** (2011), 123–224, Preprint, arXiv:1104.5557, 2011. 6.2, 6.4, 6.6.1

[Mah12] _____, *Combinatorial scientific computing*, ch. Algorithmic and Statistical Perspectives on Large-Scale Data Analysis, Chapman and Hall/CRC, 2012. 6.4

[Mec04] M. W. Meckes, *Concentration of norms and eigenvalues of random matrices*, J. Funct. Anal. **211** (2004), 508–524. 1.3

[MJC+12] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp, *Matrix Concentration Inequalities via the Method of Exchangeable Pairs*, Preprint, arXiv:1201.6002., 2012. 1.3

[MP06] S. Mendelson and A. Pajor, *On singular values of matrices with independent rows*, Bernoulli **12** (2006), 761–773. 1.3

[MRT06] P.-G. Martinsson, V. Rohklin, and M. Tygert, *A Randomized Algorithm for the Approximation of Matrices*, Tech. report, Yale, 2006. 1.1, 1.2

[MRT11] P.-G. Martinsson, V. Rokhlin, and M. Tygert, *A randomized algorithm for the decomposition of matrices*, Appl. Comput. Harmon. Anal. **30** (2011), 47–68. 1.1, 1.2

[MTJ12] L. Mackey, A. Talwalkar, and M. I. Jordan, *Divide-and-Conquer Matrix Factorization*, Advances in Neural Information Processing Systems (NIPS) 24, 2012, Technical arXiV:1107.0789, August 2011. 6.6.1

[MZ11] A. Magen and A. Zouzias, *Low Rank Matrix-valued Chernoff Bounds and Approximate Matrix Multiplication*, ACM-SIAM Symposium on Discrete Algorithms (SODA), 2011. 5.1, 5.2.1

[NDT09] N. H. Nguyen, T. T. Do, and T. D. Tran, *A fast and efficient algorithm for low-rank approximation of a matrix*, Proc. 41st ACM Symposium on Theory of Computing, 2009, pp. 215–224. 1.2, 5.1, 5.1, 5.2, 5.2.1, 5.4

[NDT10] N. H. Nguyen, P. Drineas, and T. D. Tran, *Tensor sparsification via a bound on the spectral norm of random tensors*, Preprint, arXiv:1005.4732, 2010. 1.1, 1.3, 1.4.2, 3.7

[NWL+02] T. O. Nielsen, R. B. West, S. C. Linn, O. Alter, M. A. Knowling, J. X. O'Connell, S. Zhu, M. Fero, G. Sherlock, J. R. Pollack, P. O. Brown, D. Botstein, and M. van de Rijn, *Molecular characterisation of soft tissue tumours: a gene expression study*, The Lancet **359** (2002), 1301–1307. 6.2

[Oli09] R. I. Oliveira, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, Preprint, arXiv:0911.0600, 2009. 1.3

[Oli10] _____, *Sums of random Hermitian matrices and an inequality due to Rudelson*, Elect. Comm. in Probab. **15** (2010), 203–212. 1.3

[PMT13] D. Paulin, L. Mackey, and J. A. Tropp, *Deriving Matrix Concentration Inequalities from Kernel Couplings*, Preprint, arXiv:1305.0612, 2013. 1.3

[PRTV00] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, *Latent semantic indexing: a probabilistic analysis*, J. Comput. System Sci. **61** (2000), 217–235. 1.2

[PZB+07] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, *PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations*, PLoS Genetics **3** (2007), 1672–1686. 6.2, 6.6.1

[Rec11] B. Recht, *A Simpler Approach to Matrix Completion*, J. Mach. Learn. Res. **12** (2011), 3413–3430, . 1.3, 4.1.3

[Roh00] J. Rohn, *Computing the Norm $\|A\|_{\infty\to 1}$ is NP-Hard*, Linear Multilinear Algebra **47** (2000), 195–204. 3.1

[RT08] V. Rokhlin and M. Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Natl. Acad. Sci. USA **105** (2008), no. 36, 13212–13217. 5.1

[Rud99] M. Rudelson, *Random Vectors in the Isotropic Position*, J. Funct. Anal. **164** (1999), no. 1, 60–72. 1.3, 2.8

[RV07] M. Rudelson and R. Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, J. ACM **54** (2007), no. 4. 1.1, 1.3, 2.7, 3

[RV08] ———, *The least singular value of a random square matrix is $O(n^{-1/2})$*, C. R. Math. Acad. Sci. Paris **346** (2008), 893–896. 1.3

[SAJ10] N. Srebro, N. Alon, and T. S. Jaakkola, *Generalization Error Bounds for Collaborative Prediction with Low-Rank Matrices*, Proc. 13th International Conference on Artificial Intelligence and Statistics, 2010. 1

[Sar06] T. Sarlós, *Improved Approximation Algorithms for Large Matrices via Random Projections*, Proc. 47th IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, 2006, pp. 143–152. 1.2, 4.1.3

[SS08] D. A. Spielman and N. Srivastava, *Graph Sparsification by Effective Resistances*, Proc. 40th ACM Symposium on Theory of Computing, 2008. 3.1.0.1

[Ste77] G. W. Stewart, *On the Perturbation of Pseudo-inverses, Projections and Linear Least Squares Problems*, SIAM Rev. **19** (1977), 634–662. 6.4.2

[SV] N. Srivastava and R. Vershynin, *Covariance estimation for distributions with $2+\varepsilon$ moments*, Ann. Probab., to appear. Preprint, arXiv:1106.2775. 2.8

[Sza76] S. J. Szarek, *On the best constants in the Khintchin inequality*, Studia Math **58** (1976), 197–208. 3.2

[Sza90] S. J. Szarek, *Spaces with Large Distance to ℓ_∞^n and Random Matrices*, Amer. J. Math. **112** (1990), 899–942. 1.3

[Tal95] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. **81** (1995), 73–205. 1.3

[Tal05] ———, *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*, Springer, 2005. 1.3

[TKR08] A. Talwalkar, S. Kumar, and H. Rowley, *Large-scale manifold learning*, 2008 Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008. 6.9.3.2

[TR10] A. Talwalkar and A. Rostamizadeh, *Matrix Coherence and the Nyström Method*, Proc. 26th Conference on Uncertainty in Artifical Intelligence (UAI 2010), 2010. 1.4.4, 5.5.1, 6.3, 6.10, 6.9.2

[Tro08] J. A. Tropp, *On the conditioning of random subdictionaries*, Appl. Comput. Harmon. Anal. **25** (2008), no. 1, 1–24. 2.7

[Tro09] _____, *Column subset selection, matrix factorization, and eigenvalue optimization*, Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 978–986. 3

[Tro11a] _____, *Freedman's inequality for matrix martingales*, Electron. Commun. Probab. **16** (2011), 262–270. 1.3

[Tro11b] _____, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images" **3** (2011), 115–126. 4.2, 5.3.1, 5.3.1, 5.6, 5.3.1, 5.3.1, 5.3.2, 6.6.2

[Tro11c] _____, *User-Friendly Tail Bounds for Matrix Martingales*, Tech. report, California Institute of Technology, 2011, http://www.acm.caltech.edu/~jtropp/reports/Tro10-User-Friendly-Martingale-TR.pdf, retrieved June 13, 2011. 1.3

[Tro12] _____, *User-Friendly Tail Bounds for Sums of Random Matrices*, Found. Comput. Math. **12** (2012), 389–434. 1.3, 1.4.1, 2.1, 2.4, 2.4, 2.4, 2.4, 2.5, 2.5, 2.5, 2.6, 2.6, 2.6, 4.1.3

[Ver11a] R. Vershynin, *Compressed sensing: Theory and applications*, ch. Introduction to the non-asymptotic analysis of random matrices, Cambridge University Press, 2011. 2.8, 2.8

[Ver11b] _____, *How close is the sample covariance matrix to the actual covariance matrix?*, J. Theoret. Probab. (2011), 1–32. 2.8, 2.8.2

[vW96] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, 1996. 3.2

[WLRT08] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Appl. Comput. Harmon. Anal. **25** (2008), no. 3, 335–366. 1.1, 1.2, 5.1, 5.1, 5.2.1, 5.4, 5.5.3

[WS01] C. K. I. Williams and M. Seeger, *Using the Nyström Method to Speed Up Kernel Machines*, Annual Advances in Neural Information Processing Systems 13, 2001, pp. 682–688. 1.1, 1.4.4, 6.1, 6.7, 6.1

[YMS+13] C.-W. Yip, M. W. Mahoney, A. S. Szalay, I. Csabai, T. Budavári, R. F. G. Wyse, and L. Dobos, *Objective Identification of Informative Wavelength Regions in Galaxy Spectra*, Manuscript submitted for publication. (2013). 6.2, 6.6.1