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 **C** and **R** and showed that approximations formed with $O(k \log k)$ columns and rows in this manner achieve Frobenius norm errors close to the optimal rank-k approximation error: $\|\mathbf{A} - \mathbf{CUR}\|_F \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F$. The *leverage scores* of the columns of **A** are used to generate the probability distribution used for column sampling: given **P**, a projection onto the dominant k-dimensional right singular space of **A**, the leverage score of the jth column of **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 **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 **A** are represented as linear combinations of some small subset of the columns of **A**. The algorithm of [MRT06] is accelerated in [WLRT08]. With high probability, it constructs matrices **B** and **Π** such that **B** consists of k columns sampled from **A**, some subset of the columns of **Π** make up the $k \times k$ identity matrix, and $\|\mathbf{A} - \mathbf{B\Pi}\|_2 = O(\sqrt{kmn}\|\mathbf{A} - \mathbf{A}_k\|_2)$.

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 $O(mnk)$ 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 $O(mn\log(k)+k^4(m+n))$ 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 $O(k^2)$ 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 $O(k \log k)$ 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:

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

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:

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

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

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

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

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

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

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

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

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.