SIAM J. MATRIX ANAL. APPL.
Vol. 34, No. 3, pp. 1301–1340

IMPROVED MATRIX ALGORITHMS VIA THE SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM*

CHRISTOS BOUTSIDIS† AND ALEX GITTENS‡

Abstract. Several recent randomized linear algebra algorithms rely upon fast dimension reduction methods. A popular choice is the subsampled randomized Hadamard transform (SRHT). In this article, we address the efficacy, in the Frobenius and spectral norms, of an SRHT-based low-rank matrix approximation technique introduced by Woolfe, Liberty, Rohklin, and Tygert. We establish a slightly better Frobenius norm error bound than is currently available, and a much sharper spectral norm error bound (in the presence of reasonable decay of the singular values). Along the way, we produce several results on matrix operations with SRHTs (such as approximate matrix multiplication) that may be of independent interest. Our approach builds upon Tropp's in "Improved Analysis of the Subsampled Randomized Hadamard Transform" [*Adv. Adaptive Data Anal.*, 3 (2011), pp. 115–126].

Key words. low-rank approximation, least-squares regression, Hadamard transform, sampling, randomized algorithms

AMS subject classifications. 15B52, 15A18, 11K45

DOI. 10.1137/120874540

1. Introduction.

Numerical linear algebra algorithms are traditionally deterministic. For example, given a full-rank matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ and a vector $\mathbf{b} \in \mathbb{R}^m$, Gaussian elimination requires at most $2m^3/3$ arithmetic operations to compute a vector $\mathbf{x} \in \mathbb{R}^n$ that satisfies $\mathbf{Ax} = \mathbf{b}$, while the matrix-matrix multiplication $\mathbf{AA}^{\mathrm{T}}$ requires at most $(2m-1)m^2$ operations, assuming that the matrix multiplication exponent equals 3. Another important problem is eigenvalue computation: current state-of-the-art solvers compute all m eigenvalues of $\mathbf{AA}^{\mathrm{T}}$ in $\mathrm{O}(m^3)$ arithmetic operations. All these computations are deterministic, i.e., ensure that the solution of the underlying problem is returned after the corresponding operation count.

Although these algorithms are numerically stable and run in polynomial time, $\mathrm{O}(m^3)$ arithmetic operations can be prohibitive for many applications when the size of the matrix is large, e.g., on the order of millions or billions [32, 31]. One way to speed up these algorithms is to reduce the size of $\mathbf{A}$, and then apply standard deterministic procedures to the resulting matrix. In more detail, for a matrix $\mathbf{\Omega} \in \mathbb{R}^{m\times r}$ $(m > r = \mathrm{o}(m))$, let $\mathbf{Y} = \mathbf{A\Omega} \in \mathbb{R}^{m\times r}$. $\mathbf{\Omega}$ is a so-called dimension reduction matrix and $\mathbf{Y}$ contains as much information of $\mathbf{A}$ as possible. Consider, for example, the matrix-matrix multiplication operation mentioned above. In this setting, one can compute $\mathbf{YY}^{\mathrm{T}}$ instead of $\mathbf{AA}^{\mathrm{T}}$. If $\mathbf{\Omega}$ is chosen carefully, then

$$\mathbf{YY}^{\mathrm{T}} \approx \mathbf{AA}^{\mathrm{T}},$$

*Received by the editors April 23, 2012; accepted for publication (in revised form) by I. C. F. Ipsen June 4, 2013; published electronically September 12, 2013.
http://www.siam.org/journals/simax/34-3/87454.html

†Mathematical Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 (cboutsi@us.ibm.com). This author was supported by the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323.

‡Applied and Computational Mathematics Department, California Institute of Technology, Pasadena, CA 91125 (gittens@caltech.edu). This author was supported by ONR awards N00014-08-1-0883 and N00014-11-1002, AFOSR award FA9550-09-1-0643, DARPA award N66001-08-1-2065, and a Sloan Research Fellowship awarded to Joel Tropp.

and the number of operations needed to compute $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ is at most $o(m^3)$ [15, 16].

Recent years have produced a large body of research on designing random matrices $\mathbf{\Omega}$ with which many popular problems in numerical linear algebra (e.g., low-rank matrix approximation [17, 18, 30, 34, 37], least-squares regression [40, 6, 11], k-means clustering [8]) can be solved approximately in $o(m^3)$ arithmetic operations. We refer the reader to a recent comprehensive survey of the topic [26], which has now emerged as *randomized numerical linear algebra*.

Some proposed choices for $\mathbf{\Omega}$ include the following: (i) every entry of $\mathbf{\Omega}$ takes the values $+1, -1$ with equal probability [12, 33]; (ii) the entries of $\mathbf{\Omega}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and unit variance [26]; (iii) the columns of $\mathbf{\Omega}$ 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 $\mathbf{A}$ [20, 17]; (vi) the columns of $\mathbf{\Omega}$ are chosen independently from the columns of the $m \times m$ identity matrix uniformly at random [22]; (v) $\mathbf{\Omega}$ is designed carefully such that $\mathbf{A}\mathbf{\Omega}$ can be computed in at most $O(\mathrm{nnz}(\mathbf{A}))$ arithmetic operations, where $\mathrm{nnz}(\mathbf{A})$ denotes the number of nonzero entries in $\mathbf{A}$ [13].

In this article we focus on the so-called subsampled randomized Hadamard transform (SRHT), i.e., the matrix $\mathbf{\Omega}$ contains a subset of the columns of a randomized Hadamard matrix (see Definitions 1.1 and 1.2 below). This form of dimension reduction was introduced in [1]. It is of particular interest because the highly structured nature of $\mathbf{\Omega}$ can be exploited to reduce the time of computing $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$ from $O(m^2 r)$ to $O(m^2 \log_2 r)$ (see Lemma 1.3 below).

DEFINITION 1.1 (normalized Walsh–Hadamard matrix). *Fix an integer $n = 2^p$ for $p = 1, 2, 3, \ldots$. The (nonnormalized) $n \times n$ matrix of the Walsh–Hadamard transform is defined recursively as*

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

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 1.2 (subsampled randomized Hadamard transform (SRHT) matrix). *Fix integers r and $n = 2^p$ with $r < n$ and $p = 1, 2, 3, \ldots$. An SRHT matrix is an $r \times n$ matrix of the form*

$$\mathbf{\Theta} = \sqrt{\frac{n}{r}} \cdot \mathbf{R}\mathbf{H}\mathbf{D};$$

- $\mathbf{D} \in \mathbb{R}^{n \times n}$ *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;*
- $\mathbf{R} \in \mathbb{R}^{r \times n}$ *is a subset of r rows from the $n \times n$ identity matrix, where the rows are chosen uniformly at random and without replacement.*

LEMMA 1.3 (fast matrix-vector multiplication; see Theorem 2.1 in [2]). *Given $\mathbf{x} \in \mathbb{R}^n$ and $r < n$, one can construct $\mathbf{\Theta} \in \mathbb{R}^{r \times n}$ and compute $\mathbf{\Theta}\mathbf{x}$ in at most $2n \log_2(r+1))$ operations.*

The purpose of this article is to analyze the theoretical performance of an SRHT-based randomized low-rank approximation algorithm introduced in [43] and analyzed in [43, 26, 35]. Our analysis (see Theorem 2.1) provides sharper approximation bounds than those in [43, 26, 35].

Our study should also be viewed as a followup to the work of Drineas et al. [19] and [38, 3] on designing fast approximation algorithms for solving least-squares re-

gression problems. One of the two algorithms presented in [19] employs the SRHT to quickly reduce the dimension of the least-squares problem and then solves the smaller problem with a direct least-squares solver, while [38, 3] use the SRHT to design a good preconditioner for an iterative method, which is then used to solve the regression problem. The results in this article along with the work in [41] have implications in all these studies [38, 19, 3]. We discuss these implications in section 3.

1.1. Beyond the SRHT. Finally, notice that the SRHT is defined only when the matrix dimension is a power of 2. An alternative option is to use other structured orthonormal randomized transforms such as the discrete cosine transform (DCT) or the discrete Hartley transform (DHT) [43, 35, 38, 3], whose entries are on the order of $n^{-1/2}$. All these transforms do not place any restrictions on the size of the matrix. The results of this paper—with minimal effort—can be extended *unchanged* to encompass these transforms. To see this, notice that Lemma 3.3 in [41] remains unchanged for all these orthogonal transforms. Thus Lemma 4.1 in our work as well as all other results presented in this article are true for these orthogonal transforms as well.

1.2. Outline. This article is structured as follows. Section 1.3 introduces the notation. In section 2, we present our main results on the quality of SRHT low-rank approximations and compare them to prior results in the literature. In section 3, we discuss two approaches to least-squares regression involving SRHT dimensionality reduction. Section 4 first recalls known facts on the application of SRHTs to orthogonal matrices and then 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 contains the proofs of our two main theorems presented in sections 2 and 3. We conclude the paper with an experimental evaluation of the SRHT low-rank approximation algorithm in section 6.

1.3. Preliminaries. We use $\mathbf{A}, \mathbf{B}, \ldots$ to denote real matrices and $\mathbf{a}, \mathbf{b}, \ldots$ to denote real column vectors. $\mathbf{I}_n$ is the $n \times n$ identity matrix; $\mathbf{0}_{m\times n}$ is the $m \times n$ matrix of zeros; $\boldsymbol{e}_i$ is the standard basis (whose dimensionality will be clear from the context). $\mathbf{A}_{(i)}$ denotes the ith row of $\mathbf{A}$; $\mathbf{A}^{(j)}$ denotes the jth column of $\mathbf{A}$; $\mathbf{A}_{ij}$ denotes the (i,j)th element of $\mathbf{A}$. We use the Frobenius and the spectral norm of a matrix, $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$ and $\|\mathbf{A}\|_2 = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$, respectively. The notation $\|\mathbf{A}\|_\xi$ indicates that an expression holds for both $\xi = 2$ and $\xi = \mathrm{F}$.

A (compact) singular value decomposition (SVD) of the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\mathbf{A}) = \rho$ is a decomposition of the form

$$\mathbf{A} = \underbrace{\begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{\rho-k} \end{pmatrix}}_{\mathbf{U}_\mathbf{A}\in\mathbb{R}^{m\times\rho}} \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_{\rho-k} \end{pmatrix}}_{\boldsymbol{\Sigma}_\mathbf{A}\in\mathbb{R}^{\rho\times\rho}} \underbrace{\begin{pmatrix} \mathbf{V}_k^{\mathrm{T}} \\ \mathbf{V}_{\rho-k}^{\mathrm{T}} \end{pmatrix}}_{\mathbf{V}_\mathbf{A}^{\mathrm{T}}\in\mathbb{R}^{\rho\times n}},$$

where the singular values of $\mathbf{A}$ are ordered $\sigma_1 \geq \cdots \geq \sigma_k \geq \sigma_{k+1} \geq \cdots \geq \sigma_\rho > 0$. Here k is a parameter in the interval $1 \leq k \leq \rho$, and the above formula corresponds to a partition of the SVD in block form using k. We denote the ith singular value of $\mathbf{A}$ by $\sigma_i(\mathbf{A})$ and sometimes refer to σ_1 as $\sigma_{\max}$ and σ_ρ as $\sigma_{\min}$. The matrices $\mathbf{U}_k \in \mathbb{R}^{m\times k}$ and $\mathbf{U}_{\rho-k} \in \mathbb{R}^{m\times(\rho-k)}$ contain the left singular vectors of $\mathbf{A}$; similarly, the matrices $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ and $\mathbf{V}_{\rho-k} \in \mathbb{R}^{n\times(\rho-k)}$ contain the right singular vectors of $\mathbf{A}$. We denote $\mathbf{A}_k = \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^{\mathrm{T}} \in \mathbb{R}^{m\times n}$. $\mathbf{A}_k$ minimizes $\|\mathbf{A} - \mathbf{X}\|_\xi$ over all $m \times n$ matrices $\mathbf{X}$ of rank at most k. $\mathbf{A}^\dagger = \mathbf{V}_\mathbf{A}\boldsymbol{\Sigma}_\mathbf{A}^{-1}\mathbf{U}_\mathbf{A}^{\mathrm{T}} \in \mathbb{R}^{n\times m}$ denotes the Moore–Penrose pseudoinverse of

$\mathbf{A} \in \mathbb{R}^{m\times n}$. Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ $(n \geq m)$ and $\mathbf{B} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \in \mathbb{R}^{m\times m}$; any matrix that can be written in this form is called a symmetric positive semidefinite (SPSD) matrix. For all $i = 1, \ldots, m$, $\lambda_i(\mathbf{B}) = \sigma_i^2(\mathbf{X})$ denotes the ith eigenvalue of $\mathbf{B}$. We sometimes use $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ to denote the smallest (nonzero) and largest eigenvalues of $\mathbf{B}$, respectively.

2. Low-rank matrix approximation using SRHTs. Using an SRHT matrix (see Definition 1.2), one can quickly construct a low-rank approximation to a given matrix $\mathbf{A}$. Our main result, Theorem 2.1 below, provides theoretical guarantees on the spectral and Frobenius norm accuracy of these approximations.

THEOREM 2.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ with rank ρ and n a power of 2. Fix an integer k satisfying $2 \leq k < \rho$. Let $0 < \varepsilon < 1/3$ be an accuracy parameter, let $0 < \delta < 1$ be a failure probability, and let $\mathrm{C} \geq 1$ be any specified constant. Let $\mathbf{Y} = \mathbf{A}\mathbf{\Theta}^{\mathrm{T}}$, where $\mathbf{\Theta} \in \mathbb{R}^{r\times n}$ is an SRHT with r satisfying*

$$6\mathrm{C}^2\varepsilon^{-1}\left[\sqrt{k} + \sqrt{8\ln(n/\delta)}\right]^2 \ln(k/\delta) \leq r \leq n. \tag{2.1}$$

Let $\ell = \min\{m, r\}$. Furthermore, let $\mathbf{Q} \in \mathbb{R}^{m\times\ell}$ satisfy $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_\ell$ and be such that the column space of $\mathbf{Y}$ is contained in the range of $\mathbf{Q}$ (e.g., such a $\mathbf{Q}$ can be computed with the QR factorization of $\mathbf{Y}$ in $\mathrm{O}(m\ell^2)$ arithmetic operations), and let $\tilde{\mathbf{A}}_k = \mathbf{Q}\mathbf{X}_{opt} \in \mathbb{R}^{m\times n}$, where $\mathbf{X}_{opt}$ is computed via the SVD of $\mathbf{Q}^{\mathrm{T}}\mathbf{A}$ as follows:

$$\mathbf{X}_{opt} = \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{\ell\times n},\ \mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{Q}^{\mathrm{T}}\mathbf{A} - \mathbf{X}\|_{\mathrm{F}}.$$

Given this setup, with probability at least $1 - \delta^{\mathrm{C}^2\ln(k/\delta)/4} - 7\delta$, the following Frobenius norm bounds hold simultaneously:

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}} \leq (1 + 22\varepsilon)\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}, \tag{i}$$

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (1 + 22\varepsilon)\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}, \tag{ii}$$

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}} \leq (1 + 22\varepsilon)\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}, \tag{iii}$$

$$\|\mathbf{A}_k - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (2 + 22\varepsilon)\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}. \tag{iv}$$

Similarly, the same setup ensures that with probability at least $1 - 5\delta$, the following spectral norm bounds hold simultaneously:

(v)
$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \leq \left(4 + \sqrt{\frac{3\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{3\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}},$$

(vi)
$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \left(6 + \sqrt{\frac{6\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{6\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}},$$

(vii)
$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \leq \left(4 + \sqrt{\frac{3\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{3\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}},$$

(viii)
$$\|\mathbf{A}_k - \tilde{\mathbf{A}}_k\|_2 \leq \left(7 + \sqrt{\frac{12\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{6\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

Recall that $\ell = \min\{m, r\}$. *The matrix* $\mathbf{Y}$ *can be constructed using* $2mn\log_2(r+1)$ *arithmetic operations and, given* $\mathbf{Y}$, *the matrices* $\mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}$ *and* $\tilde{\mathbf{A}}_k$ *can be formed using* $\mathrm{O}(mn\ell + mr\ell)$ *and* $\mathrm{O}(mn\ell + \ell^2 n)$ *additional arithmetic operations, respectively.*

We prove this theorem in section 5.2. Notice that the theorem provides residual and forward error bounds for two low-rank matrices in the spectral and Frobenius norms. The matrix $\mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}$ has rank at most $r > k$, while the matrix $\tilde{\mathbf{A}}_k$ has rank at most k. Previous works have provided only residual error bounds [43, 26, 35].

The first two Frobenius norm bounds in this theorem (residual error analysis) are slightly stronger than the best bounds appearing in prior efforts [35]. The spectral norm bounds on the residual error are significantly better than the bounds presented in prior work and shed light on an open question mentioned in [35]. We do not, however, claim that the error bounds provided are the tightest possible. Certainly the specific constants $(22, 6,$ etc.$)$ in the error estimates are not optimized.

We now present a detailed comparison of the guarantees given in Theorem 2.1 with those available in the existing literature.

2.1. Detailed comparison to prior work.

2.1.1. Halko, Martinson, and Tropp [26]. To put our result into perspective, we compare it to prior efforts at analyzing the SRHT algorithm introduced above. Halko, Martinson, and Tropp [26] argue that if r satisfies

$$4\left[\sqrt{k} + \sqrt{8\ln(kn)}\right]^2 \ln(k) \leq r \leq n, \tag{2.2}$$

then, for both $\xi = 2, \mathrm{F}$,

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}\|_\xi \leq \left(1 + \sqrt{7n/r}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_\xi,$$

with probability at least $1 - \mathrm{O}(1/k)$. Our first Frobenius norm bound is always tighter than the Frobenius norm bound given here. To compare the spectral norm bounds, note that our first spectral norm bound is on the order of

$$\max\left\{\sqrt{\frac{\ln(\rho/\delta)\ln(n/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_2,\ \sqrt{\frac{\ln(\rho/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_\mathrm{F}\right\}. \tag{2.3}$$

If the singular values of $\mathbf{A}$ are flat and $\mathbf{A}$ has close to full rank, then the spectral norm result in [26] is perhaps optimal. But in the cases where it makes most sense to ask for low-rank approximations—viz., $\mathbf{A}$ is rank-deficient or the singular values of $\mathbf{A}$ decay fast—the spectral error norm bound in Theorem 2.1 is more useful. Specifically, if

$$\|\mathbf{A} - \mathbf{A}_k\|_\mathrm{F} \ll \sqrt{\frac{n}{\ln(\rho/\delta)}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_2,$$

then when r is chosen according to Theorem 2.1, the quantity in (2.3) is much smaller than

$$\sqrt{7n/r} \cdot \|\mathbf{A} - \mathbf{A}_k\|_2.$$

We were able to obtain this improved bound by using the results in section 4.1, which allow one to take into account decays 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 the number of samples r.

2.1.2. Nguyen, Do, and Tran [35]. A tighter analysis of the Frobenius norm error term of the SRHT low-rank matrix approximation algorithm appeared in Nguyen, Do, and Tran [35]. Let δ be a probability parameter with $0 < \delta < 1$ and let ε be an accuracy parameter with $0 < \varepsilon < 1$. Then Nguyen et al. show that in order to get a rank-k matrix $\tilde{\mathbf{A}}_k$ satisfying

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (1+\varepsilon) \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$$

and

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \left(2 + \sqrt{2n/r}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$$

with probability of success at least $1 - 5\delta$, one requires

$$r = \Omega\left(\varepsilon^{-1} \max\{k, \sqrt{k}\ln(2n/\delta)\} \cdot \max\{\ln k, \ln(3/\delta)\}\right).$$

Theorem 2.1 gives a tighter spectral norm error bound in the cases of most interest, where $\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}} \ll \sqrt{\frac{n}{\ln(\rho/\delta)}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$. It also provides an equivalent Frobenius norm error bound with a comparable failure probability for a smaller number of samples. Specifically, if

$$r \geq 528\varepsilon^{-1}[\sqrt{k} + \sqrt{8\ln(8n/\delta)}]^2 \ln(8k/\delta) = \Omega\left(\varepsilon^{-1} \max\{k, \ln(n/\delta)\} \cdot \max\{\ln k, \ln(1/\delta)\}\right),$$

then the second Frobenius norm bound in Theorem 2.1 ensures $\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (1+\varepsilon) \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$, with probability at least $1 - 8\delta$.

In the conclusion of [35], 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 $\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}} = \Theta(1) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$. In this case, by choosing r

$$24\varepsilon^{-1}\left[\sqrt{k} + \sqrt{8\ln(n/\delta)}\right]^2 \ln(k/\delta)\ln(n/\delta) \leq r \leq n,$$

our second spectral norm bound ensures $\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \mathrm{O}(1) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$ with probability of at least $1 - 8\delta$, thus predicting the observed empirical behavior of the algorithm.

2.1.3. The subsampled randomized Fourier transform. The algorithm in section 5.2 of [43], 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 the subsampled randomized Fourier transform (SRFT), i.e., the matrix $\mathbf{H}$ of Definition 1.2 is replaced by a matrix where the (j,h)th entry is $\mathbf{H}_{jh} = e^{-2\pi i(j-1)(h-1)/n}$, where $i = \sqrt{-1}$, i.e., $\mathbf{H}$ is the unnormalized discrete Fourier transform. Woolfe et al. [43] (see eqn. (190)) argue that, for any $\alpha > 1$, $\beta > 1$, if

$$r \geq \alpha^2\beta\,(\alpha - 1)^{-1}\,(2k)^2,$$

then with probability at least $1 - 3/\beta$ ($\omega = \max\{m, n\}$),

$$\|\mathbf{A} - \tilde{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_k\tilde{\mathbf{V}}_k^{\mathrm{T}}\|_2 \leq 2\left(\sqrt{2\alpha - 1} + 1\right) \cdot \left(\sqrt{\alpha\omega + 1} + \sqrt{\alpha\omega}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2.$$

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 nonnegative entries. These matrices can be computed deterministically from $\mathbf{A}\mathbf{\Theta}^{\mathrm{T}}$ in $\mathrm{O}(k^2(m+n)+kr^2\ln r)$ time. Also, computing $\mathbf{Y}=\mathbf{A}\mathbf{\Theta}^{\mathrm{T}}$ takes $O(mn\ln r)$ time.

2.1.4. Two alternative dimensionality-reduction algorithms. Instead of using an SRHT matrix, one can take $\mathbf{\Theta}^{\mathrm{T}}$ in Theorem 2.1 to be a matrix of i.i.d. standard Gaussian random variables. One gains theoretically and often empirically better worse-case trade-offs between the number of samples taken, the failure probability, and the error guarantees. The SRHT algorithm is still faster, though, since matrix multiplications with Gaussian matrices require $\mathrm{O}(mnr)$ time. One can also take $\mathbf{\Theta}^{\mathrm{T}}$ 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 $\mathbf{\Theta}$ is a matrix of i.i.d. sub-Gaussian random variables—so 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 spectrum of $\mathbf{A}$ (spectrum here refers to the singular values of $\mathbf{A}$). As can be seen by considering a matrix with a flat spectrum, in general the stable rank is no smaller than the rank; the smaller the stable rank, the more pronounced the decay in the spectrum of $\mathbf{A}$.

When $r > k+4$, Theorem 10.7 and Corollary 10.9 in [26] imply that, when using Gaussian sampling, with probability at least $1-2\cdot 32^{-(r-k)}-e^{\frac{-(r-k+1)}{2}}$,

$$\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}} \le \left(1+32\frac{\sqrt{3k}+\mathrm{e}\sqrt{r}}{\sqrt{r-k+1}}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}},$$

and with probability at least $1-3e^{-(r-k)}$,

$$\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \le \left(1+16\sqrt{1+\frac{k}{r-k}}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_2+\frac{8\sqrt{r}}{r-k+1}\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

Comparing to the guarantees of Theorem 2.1 we see that these bounds suggest that with the same number of samples, Gaussian low-rank approximations outperform SRHT low-rank approximations. In particular, the spectral norm bound guarantees that if $\mathrm{sr}\,(\mathbf{A}-\mathbf{A}_k)\le k$, i.e., $\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}\le\sqrt{k}\|\mathbf{A}-\mathbf{A}_k\|_2$, then the Gaussian low-rank approximation algorithm requires $\mathrm{O}(k/\varepsilon^2)$ samples to return a $(17+\varepsilon)$ constant factor spectral norm error approximation with high probability. Similarly, the Frobenius norm bound guarantees that the same number of samples returns a $(1+32\varepsilon)$ constant factor Frobenius norm error approximation with high probability. Neither the spectral nor Frobenius bounds given in Theorem 2.1 for SRHT low-rank approximations apply for this few samples.

Although [33] does not consider the Frobenius norm error of the random sign low-rank approximation algorithm, Remark 4 in [33] shows that when $r=\mathrm{O}(k/\varepsilon^4\ln(1/\delta))$ for $1<\delta<0$, and $\mathrm{sr}\,(\mathbf{A}-\mathbf{A}_k)\le k$, this algorithm ensures that with high probability of at least $1-\delta$,

$$\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \le (1+\varepsilon)\|\mathbf{A}-\mathbf{A}_k\|_2.$$

To compare our results to those stated in [26, 33] we assume that $k \gg \ln(n/\delta)$, so that $r > k \ln k$ suffices for Theorem 2.1 to apply. Then, in order to acquire a $(4+\varepsilon)$ relative error bound from Theorem 2.1, it suffices that (here C' is an explicit constant no larger than 6)

$$r \geq \mathrm{C}' \varepsilon^{-2} k \ln(\rho/\delta) \quad \text{and} \quad \mathrm{sr}\,(\mathbf{A} - \mathbf{A}_k) \leq \mathrm{C}' k.$$

We see that the Gaussian and random sign approximation algorithms return $(17+\varepsilon)$ and $(1+\varepsilon)$ relative spectral error approximations, respectively, when r is on the order of k, and the relatively weak spectral decay condition $\mathrm{sr}\,(\mathbf{A} - \mathbf{A}_k) \leq k$ is satisfied, while our bounds for the SRHT algorithm require $r > k \ln(\rho/\delta)$ and the spectral decay condition

$$\mathrm{sr}\,(\mathbf{A} - \mathbf{A}_k) \leq \mathrm{C}' k$$

to ensure a $(6+\varepsilon)$ 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 r (the same is of course true for the Gaussian and random sign algorithms).

The disparity in the bounds for these three schemes—the presence of the logarithmic factors in the SRHT bounds and the fact that these bounds apply only when $r > k \ln(\rho/\delta)$—may reflect a fundamental trade-off between the structure and randomness of $\mathbf{\Theta}^{\mathrm{T}}$. The highly structured nature of SRHT matrices makes it possible to calculate $\mathbf{Y}$ 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 $\mathbf{\Theta}^{\mathrm{T}}$ and the similarly nice properties of a matrix of i.i.d. sub-Gaussian random variables, thus resulting in slacker bounds which require more samples.

3. Least squares regression. We now show how one can use the SRHT to solve least-squares problems of the form

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2.$$

Here $\mathbf{A}$ is an $m \times n$ matrix with $m \gg n$ and $\mathrm{rank}(\mathbf{A}) = n$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{x} \in \mathbb{R}^n$. One approach to solve this optimization problem is via the SVD of $\mathbf{A}$, $\mathbf{x}_{opt} = \mathbf{A}^{\dagger}\mathbf{b}$, while an example of an iterative algorithm is the LSQR algorithm in [36].

During the last decade, researchers have developed several randomized algorithms that (approximately) solve the regression problem in less running time than the approaches mentioned above [40, 38, 33, 3, 19]. We refer the reader to section 3.3 in [4] for a survey of these methods. The fastest noniterative method is in [19], while the fastest iterative algorithm is in [38, 3]. Both approaches employ the SRHT.

3.1. Least squares via the SRHT and the SVD. The idea in the SRHT algorithm of Drineas et al. [19] is to reduce the dimensions of $\mathbf{A}$ and $\mathbf{b}$ by pre-multiplication with an SRHT matrix $\mathbf{\Theta} \in \mathbb{R}^{r \times m}$ (the matrix $\mathbf{R}$ in this case is constructed by uniform sampling without replacement) and then solve quickly the smaller problem,

$$\min_{\mathbf{x}} \|\mathbf{\Theta}\mathbf{A}\mathbf{x} - \mathbf{\Theta}\mathbf{b}\|_2.$$

Let $\tilde{\mathbf{x}}_{opt} = (\mathbf{\Theta}\mathbf{A})^{\dagger}\mathbf{\Theta}\mathbf{b}$; then, assuming r satisfies ($\varepsilon > 0$ is an accuracy parameter)

$$r = \max\{48^2 n \ln(40mn) \ln(10^4 n \ln(40mn)), 40n \ln(40mn)/\varepsilon\},$$

[19] shows that with probability at least 0.8,

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2 \leq (1+\varepsilon) \cdot \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2.$$

Furthermore, assume that there exists a $\gamma \in (0,1]$ such that $\|\mathbf{U_A}\mathbf{U_A}^{\mathrm{T}}\mathbf{b}\|_2 = \gamma\|\mathbf{b}\|_2$. Then, with the same probability,

$$\|\mathbf{x}_{opt} - \tilde{\mathbf{x}}_{opt}\|_2 \leq \sqrt{\varepsilon}\left(\kappa(\mathbf{A})\sqrt{\gamma^{-2}-1}\right)\|\mathbf{x}_{opt}\|_2.$$

Here, $\kappa(\mathbf{A})$ is the two-norm condition number of $\mathbf{A}$:

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{\dagger}\|_2.$$

The running time of this approximation algorithm is $\mathrm{O}(mn\log_2 r + rn^2)$, since the SRHT multiplication takes $\mathrm{O}(mn\log_2 r)$ time and the solution of the small regression problem another $\mathrm{O}(rn^2)$.

Below, we provide a novel analysis of this SRHT least-squares algorithm which shows that one needs asymptotically fewer samples r. This immediately implies an improvement on the running time of the algorithm. Additionally, we show logarithmic dependence on the failure probability.

THEOREM 3.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ ($m \gg n$) have rank $\rho = n$, with n a power of 2, and let $\mathbf{b} \in \mathbb{R}^m$. Let $0 < \varepsilon < 1/3$ denote an accuracy parameter, $0 < \delta < 1$ a failure probability, and $\mathrm{C} \geq 1$ a constant. Let $\mathbf{\Theta}$ be an $r \times m$ SRHT matrix with r satisfying*

$$6\mathrm{C}^2\varepsilon^{-1}\left[\sqrt{n} + \sqrt{8\ln(m/\delta)}\right]^2 \ln(n/\delta) \leq r \leq m.$$

Then, with probability at least $1 - \delta^{\mathrm{C}^2\ln(n/\delta)/4} - 7\delta$,

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2 \leq (1+22\varepsilon) \cdot \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2.$$

Furthermore, assume that there exists a $\gamma \in (0,1]$ such that $\|\mathbf{U_A}\mathbf{U_A}^{\mathrm{T}}\mathbf{b}\|_2 = \gamma\|\mathbf{b}\|_2$. Then, with the same probability,

$$\|\mathbf{x}_{opt} - \tilde{\mathbf{x}}_{opt}\|_2 \leq \left(\frac{1-\sqrt{\varepsilon}}{4\varepsilon}\right)^{\frac{1}{2}}\left(\kappa(\mathbf{A})\sqrt{\gamma^{-2}-1}\right)\|\mathbf{x}_{opt}\|_2.$$

We prove this theorem in section 5.3. Another possibility to obtain a better analysis of the method of Drineas et al. is to use Lemma 4.5 in this article, which was proved in [28] and presents bounds for sampling without replacement. This analysis is not straightforward and is beyond the scope of this paper.

3.2. Iterative methods. The key idea of an iterative algorithm such as the LSQR method of [36] is *preconditioning*. Blendenpik in [3] constructs such a preconditioner by using the SRHT (the matrix $\mathbf{R}$ in this case is constructed by uniform sampling without replacement) as follows. First, an SRHT matrix $\mathbf{\Theta} \in \mathbb{R}^{r\times m}$ is constructed. Then one forms a QR factorization $\mathbf{\Theta A} = \mathbf{QR_A}$, with $\mathbf{Q} \in \mathbb{R}^{r\times n}$ and $\mathbf{R_A} \in \mathbb{R}^{n\times n}$. Finally, $\mathbf{A}$ and $\mathbf{R_A}$ are given as inputs to LSQR to find a solution to the least-squares problem. We refer the reader to [3] (see also [28]) for a detailed discussion of this approach. The purpose of our discussion here is to comment on the first step of the above procedure and show that a preconditioner of the same quality can be constructed with a smaller r. Avron, Maymounkov, and Toledo [3] argue that if the

number of samples is sufficiently large, then the two-norm condition number of $\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}$ is small. A small condition number is desirable because the number of iterations required for convergence of the LSQR method is proportional to the condition number. More specifically, Theorem 3.2 in [3] argues that with $r = \Omega\left(n \ln(m) \ln(n \ln(m))\right)$, and with constant probability (e.g., 0.9),

$$\kappa\big(\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}\big) = \mathrm{O}(1).$$

The analysis of Blendenpik was recently improved in [28]. More specifically, Corollary 3.11 in [28], along with Lemma 4.2 in our manuscript, which gives a bound on the coherence, shows that if

$$\frac{8}{3}\varepsilon^{-2}\left[\sqrt{n} + \sqrt{8\ln(m/\delta)}\right]^2 \ln(2n/\delta) \leq r \leq m,$$

then, with probability at least $1 - 2\delta$,

$$\kappa\big(\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}\big) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}.$$

We now provide a similar bound in the case where the SRHT is constructed via sampling without replacement. This bound is a simple combination of results in prior work. More specifically, Theorem 1 in [38] argues that the two-norm condition number of $\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}$ equals the two-norm condition number of $\mathbf{U}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{m\times n}$ contains the top n left singular vectors of $\mathbf{A}$. Combine this fact with the bounds on the singular values of $\mathbf{U}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$ from Lemma 4.1 to obtain the following observation.

Remark. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \gg n)$ have rank $\rho = n$, and let n be a power of 2. Fix $0 < \delta < 1$ and $0 < \varepsilon < 1/3$. Construct the upper triangular matrix $\mathbf{R}_{\mathbf{A}} \in \mathbb{R}^{n\times n}$ via the QR factorization $\mathbf{\Theta}\mathbf{A} = \mathbf{Q}\mathbf{R}_{\mathbf{A}}$, where $\mathbf{\Theta}$ is an $r \times m$ SRHT matrix with r satisfying

$$6\varepsilon^{-2}\left[\sqrt{n} + \sqrt{8\ln(m/\delta)}\right]^2 \ln(2n/\delta) \leq r \leq m.$$

Then, with probability at least $1 - 2\delta$,

$$\kappa\big(\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}\big) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}.$$

Finally, notice that we form the SRHT by uniform sampling without replacement, while Blendenpik samples the columns of the randomized Hadamard matrix *with* replacement. A different sampling scheme—Bernoulli sampling—was analyzed in Theorem 6.1 in [23] and section 4 in [28].

3.2.1. The subsampled randomized Fourier transform. Finally, we mention the work of Rokhlin and Tygert [38], which was the first to use the idea of employing subsampled randomized orthogonal transforms to precondition iterative solvers for least-squares regression problems. In [38] the SRHT is replaced with the SRFT; notice, though, that one still needs $O\left(mn \ln r\right)$ time to compute the product $\mathbf{\Theta}\mathbf{A}$. In this case, for any $\alpha > 1$, $0 < \delta < 1$, if

$$r \geq \left(\frac{\alpha^2+1}{\alpha^2-1}\right)^2 \frac{n^2}{\delta},$$

then, with probability at least $1 - \delta$,

$$\kappa\left(\mathbf{A}\mathbf{R}_{\mathbf{A}}^{-1}\right) \leq \alpha.$$

4. Matrix computations with SRHT matrices.

4.1. SRHTs applied to orthonormal matrices. An important ingredient in analyzing the low-rank approximation algorithm of Theorem 2.1 is understanding how an SRHT changes the spectrum of a matrix after postmultiplication: given a matrix $\mathbf{X}$ and an SRHT matrix $\mathbf{\Theta}$, how are the singular values of $\mathbf{X}$ and $\mathbf{X}\mathbf{\Theta}^{\mathrm{T}}$ related? To be more precise, Lemma 5.4 in section 5.1.2 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 orthonormal matrices. To see this, we informally repeat the statement of the lemma here. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ. Fix k satisfying $0 \le k \le \rho$. Given a matrix $\mathbf{\Omega} \in \mathbb{R}^{n\times r}$, with $r \ge k$, construct $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$. If $\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}$ has full row-rank, then, for $\xi = 2, \mathrm{F}$,

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_\xi^2 \le \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \left\|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Omega}\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}\right)^{\dagger}\right\|_\xi^2. \tag{4.1}$$

Now take $\mathbf{\Omega} = \mathbf{\Theta}^{\mathrm{T}}$ and observe that if the product $\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}$ has small norm, then the residual error of the approximant $\mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}$ is small. The norm of this product is small when the norms of the perturbed orthonormal matrices $\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$ and $\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}$ are in turn small, because

$$\left\|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}\right\|_\xi^2 \le \|\mathbf{\Sigma}_{\rho-k}\|_\xi^2\|\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\|_\xi^2\left\|\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}\right\|_\xi^2. \tag{4.2}$$

These perturbed orthogonal matrices have small norm precisely when their singular values are close to those of the original orthogonal matrices.

4.1.1. SRHTs by uniform sampling without replacement. In this section, we collect known results on how the singular values of a matrix with orthonormal rows are affected by postmultiplication by an SRHT matrix.

It has recently been shown by Tropp [41] that if the SRHT matrix is of sufficiently large dimensions, postmultiplying a short-fat matrix with orthonormal rows with 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 restatement of Theorem 3.1 in [41], but we include a full proof (later in this subsection) for completeness.

LEMMA 4.1 (the SRHT preserves geometry). *Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ have orthonormal columns, and let n be a power of 2. Let $0 < \varepsilon < 1/3$ and $0 < \delta < 1$. Construct an SRHT matrix $\mathbf{\Theta} \in \mathbb{R}^{r\times n}$ with r satisfying*

$$6\varepsilon^{-1}\left[\sqrt{k} + \sqrt{8\ln(n/\delta)}\right]^2 \ln(k/\delta) \le r \le n. \tag{4.3}$$

Then, with probability at least $1 - 3\delta$, for all $i = 1, \ldots, k$,

$$\sqrt{1-\sqrt{\varepsilon}} \le \sigma_i(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}) \le \sqrt{1+\sqrt{\varepsilon}}$$

and

$$\|(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\dagger} - (\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\mathrm{T}}\|_2 \le 1.54\sqrt{\varepsilon}.$$

Tropp [41] (see also [1]) 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.

LEMMA 4.2 (row norms; see Lemma 3.3 in [41]). *Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ have orthonormal columns (n is a power of 2), let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a normalized Hadamard matrix, let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix of independent random signs, and let $0 < \delta < 1$ be a failure probability. Recall that $(\mathbf{HDV})_{(i)}$ denotes the ith row of the matrix $\mathbf{HDV} \in \mathbb{R}^{n\times k}$. Then, with probability at least $1-\delta$,*

$$\max_{i=1,\dots,n} \| (\mathbf{HDV})_{(i)} \|_2 \leq \sqrt{\frac{k}{n}} + \sqrt{\frac{8\ln(n/\delta)}{n}}.$$

To prove Lemma 4.1 we need one more result on uniform random sampling (without replacement) of rows from tall-thin matrices with orthonormal columns.

LEMMA 4.3 (uniform sampling without replacement from an orthonormal matrix; corollary to Lemma 3.4 of [41]). *Let $\mathbf{W} \in \mathbb{R}^{n\times k}$ have orthonormal columns. Let $0 < \varepsilon < 1$ and $0 < \delta < 1$. Let $M := n \cdot \max_{i=1,\dots,n} \|\mathbf{W}_{(i)}\|_2^2$. Let r be an integer such that*

$$6\varepsilon^{-2} M \ln(k/\delta) \leq r \leq n\,. \tag{4.4}$$

Let $\mathbf{R} \in \mathbb{R}^{r\times n}$ be a matrix which consists of a subset of r 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 $i \in [k]$,

$$\sqrt{\frac{r}{n}} \cdot \sqrt{1-\varepsilon} \leq \sigma_i(\mathbf{RW}) \leq \sqrt{1+\varepsilon} \cdot \sqrt{\frac{r}{n}}.$$

Proof. Apply Lemma 3.4 from [41] with the following choice of parameters: $\ell = \alpha M \ln(k/\delta)$, $\alpha = 6/\varepsilon^2$, and $\delta_{tropp} = \eta = \varepsilon$. Here, ℓ, α, M, k, η are the variables of Lemma 3.4 from [41] (we also use M and k), and δ_{tropp} plays the role of δ, an error parameter, of Lemma 3.4 from [41]. The variables ε and δ are from our lemma. The choice of ℓ proportional to $\ln(k/\delta)$ rather than proportional to $\ln(k)$, 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 4.1. To obtain the bounds on the singular values, we combine Lemmas 4.2 and 4.3. More specifically, apply Lemma 4.3 with $\mathbf{W} = \mathbf{HDV}$ and use the bound for M from Lemma 4.2. Then the bound on r in (4.4), the bound on the singular values in Lemma 4.3, and the union bound establish that, with probability at least $1-3\delta$,

$$\sqrt{\frac{r}{n}} \cdot \sqrt{1-\varepsilon} \leq \sigma_i(\mathbf{RHDV}) \leq \sqrt{1+\varepsilon} \cdot \sqrt{\frac{r}{n}}.$$

Now, multiply this inequality with $\sqrt{n/r}$ and recall the definition $\mathbf{\Theta} = \sqrt{\frac{n}{r}} \cdot \mathbf{RHD}$ to obtain

$$\sqrt{1-\varepsilon} \leq \sigma_i(\mathbf{\Theta V}) \leq \sqrt{1+\varepsilon}.$$

Replacing ε with $\sqrt{\varepsilon}$ and using the bound on r in (4.3) concludes the proof.

The second bound in the lemma follows from the first bound after a simple algebraic manipulation. Let $\mathbf{X} = \mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}} \in \mathbb{R}^{k\times r}$ with SVD $\mathbf{X} = \mathbf{U_X}\mathbf{\Sigma_X}\mathbf{V_X}^{\mathrm{T}}$. Here,

$\mathbf{U_X} \in \mathbb{R}^{k\times k}$, $\mathbf{\Sigma_X} \in \mathbb{R}^{k\times k}$, and $\mathbf{V_X} \in \mathbb{R}^{r\times k}$, since $r > k$. Consider taking the SVDs of $(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\dagger}$ and $(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\mathrm{T}}$,

$$\begin{aligned}\|(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\dagger} - (\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\mathrm{T}}\|_2 &= \|\mathbf{V_X}\mathbf{\Sigma_X}^{-1}\mathbf{U_X}^{\mathrm{T}} - \mathbf{V_X}\mathbf{\Sigma_X}\mathbf{U_X}^{\mathrm{T}}\|_2 = \|\mathbf{V_X}(\mathbf{\Sigma_X}^{-1} - \mathbf{\Sigma_X})\mathbf{U_X}^{\mathrm{T}}\|_2 \\ &= \|\mathbf{\Sigma_X}^{-1} - \mathbf{\Sigma_X}\|_2,\end{aligned}$$

since $\mathbf{V_X}$ and $\mathbf{U_X}^{\mathrm{T}}$ can be dropped without changing the spectral norm. Let $\mathbf{Y} = \mathbf{\Sigma_X}^{-1} - \mathbf{\Sigma_X} \in \mathbb{R}^{k\times k}$. Then, for all $i = 1, \ldots, k$, $\mathbf{Y}_{ii} = \frac{1-\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}$. We conclude the proof as follows:

$$\begin{aligned}\|\mathbf{Y}\|_2 &= \max_{1\le i\le k} |\mathbf{Y}_{ii}| = \max_{1\le i\le k} \left|\frac{1-\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}\right| \\ &= \max_{1\le i\le k} \frac{\left|1-\sigma_i^2(\mathbf{X})\right|}{\sigma_i(\mathbf{X})} \\ &\le \frac{\sqrt{\varepsilon}}{\sqrt{1-\sqrt{\varepsilon}}} \le 1.54\sqrt{\varepsilon}. \qquad \square\end{aligned}$$

4.1.2. SRHTs by uniform sampling with replacement. Lemmas 4.1 and 4.3 analyze uniform random sampling without replacement. Below, we present the analogues of these two lemmas for uniform random sampling with replacement. Lemma 4.4 is essentially a restatement of Algorithm 2 (with the probabilities set to $1/m$) along with the third point in Remark 3.9 and Lemma 2.1 (with $\alpha = \sqrt{n/r}$) in [28].

LEMMA 4.4 (uniform sampling with replacement from an orthonormal matrix [28]). *Let $\mathbf{W} \in \mathbb{R}^{n\times k}$ have orthonormal columns. Let $0 < \varepsilon < 1$ and $0 < \delta < 1$. Let $M := n \cdot \max_{i=1,\ldots,n} \|\mathbf{W}_{(i)}\|_2^2$. Let r be an integer such that*

$$\frac{8}{3}\varepsilon^{-2} M \ln(k/\delta) \le r \le n\,. \tag{4.5}$$

Let $\mathbf{R} \in \mathbb{R}^{r\times n}$ be a matrix which consists of a subset of r rows from $\mathbf{I}_n$ where the rows are chosen uniformly at random and with replacement. Then, with probability of at least $1 - 2\delta$, for $i \in [k]$,

$$\sqrt{1-\varepsilon} \le \sigma_i\left(\sqrt{\frac{n}{r}}\mathbf{RW}\right) \le \sqrt{1+\varepsilon}.$$

LEMMA 4.5 (the SRHT preserves geometry). *Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ have orthonormal columns, and let n be a power of 2. Let $0 < \varepsilon < 1$ and $0 < \delta < 1$. Construct an SRHT matrix $\mathbf{\Theta} \in \mathbb{R}^{r\times n}$ ($\mathbf{R}$ is constructed as in Lemma 4.4, i.e., via uniform random sampling with replacement) with r satisfying*

$$\frac{8}{3}\varepsilon^{-1}\left[\sqrt{k} + \sqrt{8\ln(n/\delta)}\right]^2 \ln(k/\delta) \le r \le n. \tag{4.6}$$

Then, with probability at least $1 - 3\delta$, for all $i = 1, \ldots, k$,

$$\sqrt{1-\sqrt{\varepsilon}} \le \sigma_i\left(\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right) \le \sqrt{1+\sqrt{\varepsilon}}.$$

Proof. To obtain the bounds on the singular values, we combine Lemmas 4.2 and 4.4. More specifically, apply Lemma 4.4 with $\mathbf{W} = \mathbf{HDV}$ and use the bound for

M from Lemma 4.2. Then the bound on r in (4.4), the bound on the singular values in Lemma 4.4, and the union bound establish that, with probability of at least $1-3\delta$,

$$\sqrt{1-\varepsilon} \leq \sigma_i\left(\sqrt{\frac{n}{r}} \cdot \mathbf{RHDV}\right) \leq \sqrt{1+\varepsilon}.$$

Replacing ε with $\sqrt{\varepsilon}$ and using the bound on r in (4.4) concludes the proof. ∎

4.2. SRHTs applied to general matrices. The structural result in Lemma 5.4, Lemma 4.1 on the perturbative effects of SRHTs on the singular values of orthonormal matrices, and the basic estimate in (4.2) are enough to reproduce the results on the approximation error of SRHT-based low-rank approximation in [26]. The main contribution of this paper is the realization that one can take advantage of the decay in the singular values of $\mathbf{A}$ encoded in $\mathbf{\Sigma}_{\rho-k}$ to obtain sharper results. In view of the fact that

$$\left\|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}\right\|_{\xi}^2 \leq \|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\|_{\xi}^2\left\|\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\right)^{\dagger}\right\|_{\xi}^2, \tag{4.7}$$

we should consider the behavior of the singular values of $\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$ instead of those of $\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$. Accordingly, in this section we extend the analysis of [41] to include the application of SRHTs to general matrices.

Our main tool is a generalization of Lemma 4.2 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-squared average of the column norms of the original matrix.

4.2.1. SRHT equalizes column norms.

LEMMA 4.6 (SRHT equalization of column norms). *Suppose that* $\mathbf{A}$ *is a matrix with* n *columns and* 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 independent random signs. Then for every* $t \geq 0$,

$$\mathbb{P}\left[\max_{i=1,\ldots,n}\left\|\left(\mathbf{ADH}^{\mathrm{T}}\right)^{(i)}\right\|_2 \leq \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\|\mathbf{A}\|_2\right] \geq 1 - n\cdot \mathrm{e}^{-t^2/8}.$$

Proof. Our proof of Lemma 4.6 is essentially that of Lemma 4.2 in [41], with attention paid to the fact that $\mathbf{A}$ is no longer assumed to have orthonormal columns. In particular, the following concentration result for Lipschitz functions of Rademacher vectors is central to establishing the result. Recall that a Rademacher vector is a random vector whose entries are independent and take the values ± 1 with equal probability.

LEMMA 4.7 (concentration of convex Lipschitz functions of Rademacher random variables (Corollary 1.3 ff. in [29])). *Suppose* f *is a convex function on vectors that satisfies the Lipschitz bound*

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

Let $\boldsymbol{\varepsilon}$ *be a Rademacher vector. For all* $t \geq 0$,

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

Lemma 4.6 follows immediately from the observation that the norm of any one column of $\mathbf{ADH}^{\mathrm{T}}$ is a convex Lipschitz function of a Rademacher vector. Consider the norm of the jth column of $\mathbf{ADH}^{\mathrm{T}}$ as a function of $\boldsymbol{\varepsilon}$, where $\mathbf{D} = \mathrm{diag}(\boldsymbol{\varepsilon})$:

$$f_j(\boldsymbol{\varepsilon}) = \|\mathbf{ADH}^{\mathrm{T}}\mathbf{e}_j\| = \|\mathbf{A}\,\mathrm{diag}(\boldsymbol{\varepsilon})\mathbf{h}_j\|_2 = \|\mathbf{A}\,\mathrm{diag}(\mathbf{h}_j)\boldsymbol{\varepsilon}\|_2,$$

where $\mathbf{h}_j$ denotes the jth column of $\mathbf{H}^{\mathrm{T}}$. Evidently f_j is convex. Furthermore,

$$|f_j(\mathbf{x})-f_j(\mathbf{y})| \le \|\mathbf{A}\,\mathrm{diag}(\mathbf{h}_j)(\mathbf{x}-\mathbf{y})\|_2 \le \|\mathbf{A}\|_2\|\,\mathrm{diag}(\mathbf{h}_j)\|_2\|\mathbf{x}-\mathbf{y}\|_2 = \frac{1}{\sqrt{n}}\|\mathbf{A}\|_2\|\mathbf{x}-\mathbf{y}\|_2,$$

where we used the triangle inequality and the fact that $\|\mathrm{diag}(\mathbf{h}_j)\|_2 = \|\mathbf{h}_j\|_\infty = \frac{1}{\sqrt{n}}$. Thus f_j is convex and Lipschitz with Lipschitz constant at most $\|\mathbf{A}\|_2/\sqrt{n}$.

We calculate

$$\begin{aligned}\mathbb{E}\left[f_j(\varepsilon)\right] \le \left[\mathbb{E}f_j(\varepsilon)^2\right]^{1/2} &= \left[\mathrm{Tr}\left(\mathbf{A}\,\mathrm{diag}(\mathbf{h}_j)\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\star\right]\mathrm{diag}(\mathbf{h}_j)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\right)\right]^{1/2} \\ &= \left[\mathrm{Tr}\left(\frac{1}{n}\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)\right]^{1/2} \\ &= \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}}.\end{aligned}$$

It now follows from Lemma 4.7 that, for all $j = 1, 2, \ldots, n$, the norm of the jth column of $\mathbf{ADH}^{\mathrm{T}}$ satisfies the tail bound

$$\mathbb{P}\left[\|\mathbf{ADH}^{\mathrm{T}}\mathbf{e}_j\|_2 \ge \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\|\mathbf{A}\|_2\right] \le \mathrm{e}^{-t^2/8}.$$

Taking a union bound over all columns of $\mathbf{ADH}^{\mathrm{T}}$, we conclude that

$$\mathbb{P}\left[\max_{j=1,\ldots,n}\|(\mathbf{ADH}^{\mathrm{T}})^{(j)}\|_2 \ge \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}} + \frac{t}{\sqrt{n}}\|\mathbf{A}\|_2\right] \le n\cdot\mathrm{e}^{-t^2/8}. \qquad \square$$

As an interesting aside, we note that just as Lemma 4.1—which states that the SRHT essentially preserves the singular value of matrices with orthonormal rows and an aspect ratio of k/n—follows from Lemma 4.2, Lemma 4.6 implies that the SRHT essentially preserves the singular values of general rectangular matrices with the same aspect ratio. This can be shown using, e.g., the results on the effects of column sampling on the singular values of matrices from [23, section 6].

4.2.2. SRHT preserves the spectral norm. The following lemma shows that even if the aspect ratio is larger than k/n, the SRHT does not substantially increase the spectral norm of a matrix.

LEMMA 4.8 (SRHT-based subsampling in the spectral norm). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ, and let n be a power of 2. For some $r < n$, let $\boldsymbol{\Theta} \in \mathbb{R}^{r\times n}$ be an SRHT matrix. Fix a failure probability $0 < \delta < 1$. Then*

$$\mathbb{P}\left[\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2 \le 5\|\mathbf{A}\|_2^2 + \frac{\ln(\rho/\delta)}{r}\left(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}\|_2\right)^2\right] \ge 1 - 2\delta.$$

To establish Lemma 4.8, we use the following Chernoff bound for sampling matrices without replacement.

LEMMA 4.9 (matrix Chernoff bound; Theorem 2.2 in [41]; see also the corollary in [42]). *Let $\mathcal{X}$ be a finite set of positive semidefinite matrices with dimension k, and suppose that*

$$\max_{\mathbf{X}\in\mathcal{X}} \lambda_{\max}(\mathbf{X}) \leq B.$$

Sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_r\}$ uniformly at random from $\mathcal{X}$ without replacement. Compute

$$\mu_{\max} = r \cdot \lambda_{\max}(\mathbb{E}[\mathbf{X}_1]).$$

Then

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_j \mathbf{X}_j\right) \geq (1+\nu)\mu_{\max}\right] \leq k \cdot \left[\frac{\mathrm{e}^{\nu}}{(1+\nu)^{1+\nu}}\right]^{\mu_{\max}/B} \quad \text{for } \nu \geq 0.$$

Proof of Lemma 4.8. Write the SVD of $\mathbf{A}$ as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ where $\mathbf{\Sigma} \in \mathbb{R}^{\rho\times\rho}$ and observe that the spectral norm of $\mathbf{A}\mathbf{\Theta}^{\mathrm{T}}$ is the same as that of $\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$.

We control the norm of $\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}$ by considering the maximum singular value of its Gram matrix. Define $\mathbf{M} = \mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{D}\mathbf{H}^{\mathrm{T}}$ and let $\mathbf{G}$ be the Gram matrix of $\mathbf{M}\mathbf{R}^{\mathrm{T}}$:

$$\mathbf{G} = \mathbf{M}\mathbf{R}^{\mathrm{T}}(\mathbf{M}\mathbf{R}^{\mathrm{T}})^{\mathrm{T}}.$$

Evidently

$$\lambda_{\max}(\mathbf{G}) = \frac{r}{n}\|\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\|_2^2. \tag{4.8}$$

Recall that $\mathbf{M}^{(j)}$ denotes the jth column of $\mathbf{M}$. If we denote the random set of r coordinates to which $\mathbf{R}$ restricts by T, then

$$\mathbf{G} = \sum\nolimits_{j\in T} \mathbf{M}^{(j)}\left(\mathbf{M}^{(j)}\right)^{\mathrm{T}}.$$

Thus $\mathbf{G}$ is a sum of r random matrices $\mathbf{X}_1, \ldots, \mathbf{X}_r$ sampled without replacement from the set $\mathcal{X} = \{\mathbf{M}^{(j)}\left(\mathbf{M}^{(j)}\right)^{\mathrm{T}} : j = 1, 2, \ldots, n\}$. There are two sources of randomness in $\mathbf{G}$: $\mathbf{R}$ and the Rademacher random variables on the diagonal of $\mathbf{D}$.

Set

$$B = \frac{1}{n}\left(\|\mathbf{\Sigma}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{\Sigma}\|_2\right)^2$$

and let E be the event

$$\max_{j=1,\ldots,n} \|\mathbf{M}^{(j)}\|_2^2 \leq B.$$

When E holds, for all $j = 1, 2, \ldots, n$,

$$\lambda_{\max}\left(\mathbf{M}^{(j)}\left(\mathbf{M}^{(j)}\right)^{\mathrm{T}}\right) = \|\mathbf{M}^{(j)}\|_2^2 \leq B,$$

so $\mathbf{G}$ is a sum of random positive semidefinite matrices, each of whose norms is bounded by B. Note that whether or not E holds is determined by $\mathbf{D}$ and independent of $\mathbf{R}$.

Conditioning on E, the randomness in $\mathbf{R}$ allows us to use the matrix Chernoff bound of Lemma 4.9 to control the maximum eigenvalue of $\mathbf{G}$. We observe that

$$\mu_{\max} = r \cdot \lambda_{\max}(\mathbb{E}[\mathbf{X}_1]) = \frac{r}{n}\lambda_{\max}\left(\sum\nolimits_{j=1}^{n} \mathbf{M}^{(j)}\left(\mathbf{M}^{(j)}\right)^{\mathrm{T}}\right) = \frac{r}{n}\|\mathbf{\Sigma}\|_2^2.$$

Take the parameter ν in Lemma 4.9 to be

$$\nu = 4 + \frac{B}{\mu_{\max}} \ln(\rho/\delta)$$

to obtain the relation

$$\begin{aligned}\mathbb{P}\left[\lambda_{\max}(\mathbf{G}) \geq 5\mu_{\max} + B\ln(\rho/\delta) \mid E\right] &\leq (\rho - k) \cdot \mathrm{e}^{[\delta - (1+\nu)\ln(1+\nu)]\frac{\mu_{\max}}{B}} \\ &\leq \rho \cdot \mathrm{e}^{\left(1 - \frac{5}{4}\ln 5\right)\delta\frac{\mu_{\max}}{B}} \\ &\leq \rho \cdot \mathrm{e}^{-\left(\frac{5}{4}\ln 5 - 1\right)\ln(\rho/\delta)} < \delta.\end{aligned}$$

The second inequality holds because $\nu \geq 4$ implies that $(1+\nu)\ln(1+\nu) \geq \nu \cdot \frac{5}{4}\ln 5$.

We have conditioned on E the event that the squared norms of the columns of $\mathbf{M}$ are all smaller than B. By Lemma 4.6, E occurs with probability at least $1 - \delta$. Thus, substituting the values of B and $\mu_{\max}$, we find that

$$\mathbb{P}\left[\lambda_{\max}(\mathbf{G}) \geq \frac{r}{n}\left(5\|\mathbf{\Sigma}\|_2^2 + \frac{\ln(\rho/\delta)}{r}\left(\|\mathbf{\Sigma}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{\Sigma}\|_2\right)^2\right)\right] \leq 2\delta.$$

Use (4.8) to wrap up. ∎

4.2.3. SRHT preserves the Frobenius norm. Similarly, the SRHT is unlikely to substantially increase the Frobenius norm of a matrix.

LEMMA 4.10 (SRHT-based subsampling in the Frobenius norm). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ (n is a power of 2), and let $\mathbf{\Theta} \in \mathbb{R}^{r\times n}$ be an SRHT matrix for some $r < n$. Fix a failure probability $0 < \delta < 1$. Then, for any $\eta \geq 0$,*

$$\mathbb{P}\left[\|\mathbf{A}\mathbf{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2 \leq (1+\eta)\|\mathbf{A}\|_{\mathrm{F}}^2\right] \geq 1 - \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{1+\eta}}\right]^{r/\left(1+\sqrt{8\ln(n/\delta)}\right)^2} - \delta.$$

Proof. Let $c_j = \frac{n}{r}\|(\mathbf{A}\mathbf{D}\mathbf{H}^{\mathrm{T}})_j\|_2^2$ denote the squared norm of the jth column of $\sqrt{n/r}\cdot\mathbf{A}\mathbf{D}\mathbf{H}^{\mathrm{T}}$. Then, since right multiplication by $\mathbf{R}^{\mathrm{T}}$ samples columns uniformly at random without replacement,

$$\|\mathbf{A}\mathbf{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \frac{n}{r}\|\mathbf{A}\mathbf{D}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \sum\nolimits_{i=1}^{r} X_i, \tag{4.9}$$

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 is determined by $\mathbf{D}$.

To bound this sum, we first condition on $\mathbf{D}$ being such that each c_j is bounded by a quantity B. Call this event E. Then

$$\mathbb{P}\left[\sum\nolimits_{i=1}^{r} X_i \geq (1+\eta)\sum\nolimits_{i=1}^{r}\mathbf{E}[X_i]\right] \leq \mathbb{P}\left[\sum\nolimits_{i=1}^{r} X_i \leq (1+\eta)\sum\nolimits_{i=1}^{r}\mathbf{E}[X_i] \mid E\right] + \mathbb{P}[E^c].$$

To select B, we observe that Lemma 4.6 implies that, with probability $1-\delta$, the entries of $\mathbf{D}$ are such that

$$\max_j c_j \leq \frac{n}{r}\cdot\frac{1}{n}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}\|_2)^2 \leq \frac{1}{r}(1+\sqrt{8\ln(n/\delta)})^2\|\mathbf{A}\|_{\mathrm{F}}^2.$$

Accordingly, we take

$$B = \frac{1}{r}(1 + \sqrt{8\ln(n/\delta)})^2 \|\mathbf{A}\|_{\mathrm{F}}^2,$$

thereby arriving at the bound

(4.10)
$$\mathbb{P}\left[\sum_{i=1}^r X_i \geq (1+\eta)\sum_{i=1}^r \mathbf{E}\left[X_i\right]\right] \leq \mathbb{P}\left[\sum_{i=1}^r X_i \leq (1+\eta)\sum_{i=1}^r \mathbf{E}\left[X_i\right] \mid E\right] + \delta.$$

After conditioning on $\mathbf{D}$, we observe that the randomness remaining on the right-hand side of (4.10) is the choice of the summands X_i, which is determined by $\mathbf{R}$. We address this randomness by applying a scalar Chernoff bound (Lemma 4.9 with $k = 1$). To do so, we need $\mu_{\max}$, the expected value of the sum; this is an elementary calculation:

$$\mathbf{E}\left[X_1\right] = n^{-1}\sum_{j=1}^n c_j = \frac{1}{r}\|\mathbf{A}\|_{\mathrm{F}}^2,$$

so $\mu_{\max} = r\mathbf{E}\left[X_1\right] = \|\mathbf{A}\|_{\mathrm{F}}^2$.

Applying Lemma 4.9 conditioned on E, we conclude that

$$\mathbb{P}\left[\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2 \geq (1+\eta)\|\mathbf{A}\|_{\mathrm{F}}^2 \mid E\right] \leq \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{1+\eta}}\right]^{r/(1+\sqrt{8\ln(n/\delta)})^2} + \delta$$

for $\eta \geq 0$. ∎

4.2.4. SRHT preserves matrix multiplication. Finally, we prove a novel result on approximate matrix multiplication involving SRHT matrices.

LEMMA 4.11 (SRHT for approximate matrix multiplication). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$, *let* $\mathbf{B} \in \mathbb{R}^{n\times p}$, *and let* n *be a power of* 2. *For some* $r < n$, *let* $\boldsymbol{\Theta} \in \mathbb{R}^{r\times n}$ *be an SRHT matrix. Fix a failure probability* $0 < \delta < 1$. *Assume* R *satisfies* $0 \leq \mathrm{R} \leq \frac{\sqrt{r}}{1+\sqrt{8\ln(n/\delta)}}$. *Then*

$$\mathbb{P}\left[\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}} \leq 2(\mathrm{R}+1)\frac{\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_2}{\sqrt{r}}\right] \geq 1 - \mathrm{e}^{-\mathrm{R}^2/4} - 2\delta.$$

Remark. Recall that the stable rank $\mathrm{sr}\left(\mathbf{A}\right) = \|\mathbf{A}\|_{\mathrm{F}}^2/\|\mathbf{A}\|_2^2$ reflects the decay of the spectrum of the matrix $\mathbf{A}$. Lemma 4.11 can be rewritten as a bound on the relative error of the approximation $\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B}$ to the product $\mathbf{A}\mathbf{B}$:

$$\frac{\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}} \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}} \cdot \frac{R+2}{\sqrt{r}} \cdot \left(1 + \frac{\sqrt{8\ln(n/\delta)}}{\mathrm{sr}\left(\mathbf{B}\right)}\right).$$

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 r. Since the roles of $\mathbf{A}$ and $\mathbf{B}$ in this bound can be interchanged, in fact we have the bound

$$\frac{\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}} \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}}{\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}} \cdot \frac{R+2}{\sqrt{r}} \cdot \left(1 + \frac{\sqrt{8\ln(n/\delta)}}{\max(\mathrm{sr}\left(\mathbf{B}\right), \mathrm{sr}\left(\mathbf{A}\right))}\right).$$

Proof of Lemma 4.11. To prove the lemma, we first develop a generic result for approximate matrix multiplication via uniform sampling (without replacement) of the columns and the rows of the two matrices involved in the product (see Lemma 4.13 below). Lemma 4.11 is a simple instance of this generic result. We mention that Lemma 3.2.8 in [14] gives a similar result for approximate matrix multiplication, which, however, gives a bound for the expected value of the error term, while our Lemma 4.11 gives a comparable bound which holds with high probability. To prove Lemma 4.13, 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 [24].

LEMMA 4.12. *Let $\mathcal{V}$ be a collection of n vectors in a normed space with norm $|\cdot|$. Choose $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_r$ from $\mathcal{V}$ uniformly at random* without *replacement. Also choose $\boldsymbol{V}_1', \ldots, \boldsymbol{V}_r'$ from $\mathcal{V}$ uniformly at random* with *replacement. Let*

$$\mu = \mathbf{E}\left[\left|\sum_{i=1}^{r}(\boldsymbol{V}_i' - \mathbf{E}\left[\boldsymbol{V}_i'\right])\right|\right]$$

and set

$$\sigma^2 \geq 4r\mathbf{E}\left[\left|\boldsymbol{V}_1'\right|^2\right] \quad \text{and} \quad B \geq 2\max_{\boldsymbol{V}\in\mathcal{V}}|\boldsymbol{V}|.$$

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

$$\mathbb{P}\left[\left|\sum_{i=1}^{r}\boldsymbol{V}_i - r\mathbf{E}\left[\boldsymbol{V}_1\right]\right| \geq \mu + t\right] \leq \exp\left(-\frac{t^2}{4\sigma^2}\right).$$

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

$$\left|\sum_{i=1}^{r}\boldsymbol{V}_i - r\mathbf{E}\left[\boldsymbol{V}_1\right]\right| - \mu.$$

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

(4.11)
$$\mathbf{E}\left[\exp\left(\lambda\cdot\left|\sum_{i=1}^{r}\boldsymbol{V}_i - r\mathbf{E}\left[\boldsymbol{V}_1\right]\right| - \lambda\mu\right)\right] \leq \mathbf{E}\left[\exp\left(\lambda\cdot\left|\sum_{i=1}^{r}\boldsymbol{V}_i' - r\mathbf{E}\left[\boldsymbol{V}_1\right]\right| - \lambda\mu\right)\right].$$

This follows from a classical observation due to Hoeffding [27] (see also [25] for a more modern exposition) that for any convex $\mathbb{R}$-valued function g,

$$\mathbf{E}\left[g\left(\sum_{i=1}^{r}\boldsymbol{V}_i\right)\right] \leq \mathbf{E}\left[g\left(\sum_{i=1}^{r}\boldsymbol{V}_i'\right)\right].$$

Specifically, take $g(\boldsymbol{V}) = \exp\left(\lambda\left|\boldsymbol{V} - r\mathbf{E}\left[\boldsymbol{V}_1\right]\right| - \lambda\mu\right)$ to obtain the asserted inequality of mgfs.

In the proof of Theorem 12 in [24], Gross establishes that any random variable Z whose mgf is less than the right-hand side of (4.11) satisfies a tail inequality of the form

$$\mathbb{P}\left[Z \geq \mu + t\right] \leq \exp\left(-\frac{t^2}{4s^2}\right) \tag{4.12}$$

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

$$s^2 \geq \sum_{i=1}^{r}\mathbf{E}\left[\left|\boldsymbol{V}_i' - \mathbf{E}\left[\boldsymbol{V}_1'\right]\right|^2\right]$$

and M almost surely bounds $|\boldsymbol{V}_i' - \mathbf{E}[\boldsymbol{V}_1']|$ for all $i = 1, \dots, r$. To apply this result, note that for all $i = 1, \dots, r$,

$$|\boldsymbol{V}_i' - \mathbf{E}[\boldsymbol{V}_1']| \leq 2 \max_{\boldsymbol{V} \in \mathcal{V}} |\boldsymbol{V}| = B.$$

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

$$\begin{aligned}
\sum_{i=1}^{r} \mathbf{E}\left[|\boldsymbol{V}_i' - \mathbf{E}[\boldsymbol{V}_1']|^2 \right] &= r\mathbf{E}\left[|\boldsymbol{V}_1' - \mathbf{E}[\boldsymbol{V}_1']|^2 \right] \\
&\leq r\mathbf{E}\left[|\boldsymbol{V}_1' - \boldsymbol{V}_1''|^2 \right] \leq r\mathbf{E}\left[(|\boldsymbol{V}_1'| + |\boldsymbol{V}_1''|)^2 \right] \\
&\leq 2r\mathbf{E}\left[|\boldsymbol{V}_1'|^2 + |\boldsymbol{V}_1''|^2 \right] \\
&= 4r\mathbf{E}\left[|\boldsymbol{V}_1'|^2 \right] \leq \sigma^2.
\end{aligned}$$

The bound given in the statement of Lemma 4.12 follows from taking $s^2 = \sigma^2$ and $M = B$ in (4.12). ∎

This vector Bernstein inequality gives us a tail bound on the Frobenius error of a simple approximate matrix multiplication scheme based upon column and row sampling.

LEMMA 4.13 (matrix multiplication). *Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{n \times \ell}$. Fix $r \leq n$. Select uniformly at random and without replacement r columns from $\mathbf{X}$ and the corresponding rows from $\mathbf{Y}$ and multiply the selected columns and rows by $\sqrt{n/r}$. Let $\hat{\mathbf{X}} \in \mathbb{R}^{m \times r}$ and $\hat{\mathbf{Y}} \in \mathbb{R}^{r \times \ell}$ contain the selected columns and rows, respectively. Choose*

$$\sigma^2 \geq \frac{4n}{r} \sum_{i=1}^{n} \|\boldsymbol{X}^{(i)}\|_2^2 \|\boldsymbol{Y}_{(i)}\|_2^2 \quad \text{and} \quad B \geq \frac{2n}{r} \max_i \|\boldsymbol{X}^{(i)}\|_2 \|\boldsymbol{Y}_{(i)}\|_2.$$

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

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

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

$$\mathcal{V} = \left\{ \frac{n}{r} \mathrm{vec}(\boldsymbol{X}^{(i)} \boldsymbol{Y}_{(i)}) \right\}_{i=1}^{n}.$$

Sample $\boldsymbol{V}_1, \dots, \boldsymbol{V}_r$ uniformly at random from $\mathcal{V}$ without replacement, and observe that $\mathbf{E}[\boldsymbol{V}_i] = \frac{1}{r}\mathrm{vec}(\mathbf{X}\mathbf{Y})$. With this notation, the quantities $\|\hat{\mathbf{X}}\hat{\mathbf{Y}} - \mathbf{X}\mathbf{Y}\|_{\mathrm{F}}$ and

$$\left\| \sum\nolimits_{i=1}^{r} (\boldsymbol{V}_i - \mathbf{E}[\boldsymbol{V}_i]) \right\|_2$$

have the same distribution; therefore any probabilistic bound developed for the latter holds for the former. The conclusion of the lemma follows from applying Lemma 4.12 to bound the second quantity.

We calculate the variance-like term in Lemma 4.12, $4r\mathbf{E}\left[\|\boldsymbol{V}_1\|_2^2\right]$:

$$4r\mathbf{E}\left[\|\boldsymbol{V}_1\|_2^2\right] = 4r\frac{1}{n}\sum_{i=1}^{n} \frac{n^2}{r^2} \|\boldsymbol{X}^{(i)}\|_2^2 \|\boldsymbol{Y}_{(i)}\|_2^2 = 4\frac{n}{r}\sum_{i=1}^{n} \|\boldsymbol{X}^{(i)}\|_2^2 \|\boldsymbol{Y}_{(i)}\|_2^2 \leq \sigma^2.$$

Now we consider the expectation

$$\mu = \mathbf{E}\left[\left\|\sum_{i=1}^{r}(\boldsymbol{V}_i' - \mathbf{E}\,[\boldsymbol{V}_i'])\right\|_2\right].$$

In doing so, we will use the notation $\mathbf{E}_{A,B,\ldots}[C]$ to denote the conditional expectation of a random variable C with respect to the random variables $A, B, \ldots$. 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 r and sample $\boldsymbol{V}_1', \ldots, \boldsymbol{V}_r'$ and $\boldsymbol{V}_1'', \ldots, \boldsymbol{V}_r''$ uniformly at random from $\mathcal{V}$ with replacement. Now μ can be bounded as follows:

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

The first inequality is Jensen's, and the following equality holds because the components of the sequence $\{\boldsymbol{V}_i' - \boldsymbol{V}_i''\}$ 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 $\mathbf{E}\left[\|\boldsymbol{V}_1'\|_2^2\right] = \mathbf{E}\left[\|\boldsymbol{V}_1\|_2^2\right]$ allow us to further simplify the estimate of μ:

$$\mu \le 2\sqrt{\mathbf{E}\left[\sum_{i=1}^{r}\|\boldsymbol{V}_i\|_2^2\right]} = 2\sqrt{r\mathbf{E}\left[\|\boldsymbol{V}_1\|_2^2\right]} \le \sigma.$$

We also calculate the quantity

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

The stipulated tail bound follows from applying Lemma 4.12 with our estimates for B, σ^2, and μ. ∎

Lemma 4.11 now follows from this result on matrix multiplication.

Proof of Lemma 4.11. Let $\mathbf{X} = \mathbf{A}\mathbf{D}\mathbf{H}^{\mathrm{T}}$ and $\mathbf{Y} = \mathbf{H}\mathbf{D}\mathbf{B}$, and form $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ according to Lemma 4.13. Then $\mathbf{X}\mathbf{Y} = \mathbf{A}\mathbf{B}$ and

$$\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}} = \|\hat{\mathbf{X}}\hat{\mathbf{Y}} - \mathbf{X}\mathbf{Y}\|_{\mathrm{F}}.$$

To apply Lemma 4.13, we first condition on the event that the SRHT equalizes the column norms of our matrices. Namely, we observe that, from Lemma 4.6, with

probability at least $1-2\delta$,

$$\max_i \|\mathbf{X}^{(i)}\|_2 \leq \frac{1}{\sqrt{n}}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}\|_2) \quad \text{and} \tag{4.13}$$
$$\max_i \|\mathbf{Y}_{(i)}\|_2 \leq \frac{1}{\sqrt{n}}(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{B}\|_2).$$

Conditioning on these nice interactions, we choose the parameters σ and B in Lemma 4.13. We first take

$$\sigma^2 = \frac{4}{r}(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{B}\|_2)^2\|\mathbf{A}\|_{\mathrm{F}}^2. \tag{4.14}$$

Observe that because of (4.13),

$$\sigma^2 = 4\frac{n}{r}\cdot\frac{(\|\mathbf{Y}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{Y}\|_2)^2}{n}\|\mathbf{X}\|_{\mathrm{F}}^2 \geq 4\frac{n}{r}\sum\nolimits_{i=1}^{n}\|\mathbf{X}^{(i)}\|_2^2\|\mathbf{Y}_{(i)}\|_2^2,$$

so this choice of σ satisfies the inequality stipulated in Lemma 4.13. Next we choose

$$B = \frac{2}{r}(\|\mathbf{A}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}\|_2)(\|\mathbf{B}\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{B}\|_2).$$

Again, because of (4.13), B satisfies the stipulation $B \geq \frac{2n}{r}\max_i \|\mathbf{X}^{(i)}\|_2\|\mathbf{Y}_{(i)}\|_2$.

For simplicity, let $\gamma = 8\ln(n/\delta)$. 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 (4.14), identify the numerator as $\sqrt{r}\sigma$ to see that

$$\frac{\sigma^2}{B} \geq \frac{\sqrt{r}\sigma}{1+\sqrt{8\ln(n/\delta)}}.$$

Apply Lemma 4.13 to see that, when (4.13) hold and $0 \leq R\sigma \leq \sigma^2/B$,

$$\mathbb{P}\left[\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}} \geq (R+1)\sigma\right] \leq \exp\left(-\frac{R^2}{4}\right).$$

From our lower bound on σ^2/B, we know that the condition $R\sigma \leq \sigma^2/B$ is satisfied when

$$R \leq \sqrt{r}/(1+\sqrt{8\ln(n/\delta)}).$$

Also, we established above that (4.13) hold with probability at least $1-2\delta$. From these two facts, it follows that when $0 \leq R \leq \sqrt{r}/(1+\sqrt{8\ln(n/\delta)})$,

$$\mathbb{P}\left[\|\mathbf{A}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B} - \mathbf{A}\mathbf{B}\|_{\mathrm{F}} \geq (R+1)\sigma\right] \leq \exp\left(-\frac{R^2}{4}\right) + 2\delta.$$

The tail bound given in the statement of Lemma 4.11 follows from substituting our estimate of σ. ∎

5. Proofs of our main theorems.

5.1. Preliminaries. To prove Theorem 2.1 we first need some background on restricted (within a subspace) low-rank matrix approximations. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, let $k < n$ be an integer, and let $\mathbf{Y} \in \mathbb{R}^{m\times r}$ with $r > k$ (the case $m = r$ corresponds to the standard unrestricted low-rank approximation problem which can be addressed via the SVD). We call $\Pi^{\xi}_{\mathbf{Y},k}(\mathbf{A}) \in \mathbb{R}^{m\times n}$ the best rank-k approximation to $\mathbf{A}$ in the column space of $\mathbf{Y}$, with respect to the ξ norm ($\xi = 2$ or $\xi = \mathrm{F}$). Formally, for fixed ξ, we can write $\Pi^{\xi}_{\mathbf{Y},k}(\mathbf{A}) = \mathbf{Y}\mathbf{X}^{\xi}$, where

$$\mathbf{X}^{\xi} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^{r\times n}:\mathrm{rank}(\mathbf{x})\le k} \|\mathbf{A} - \mathbf{Y}\mathbf{X}\|_{\xi}^2.$$

In order to compute (or approximate) $\Pi^{\xi}_{\mathbf{Y},k}(\mathbf{A})$ we will use the following 3-step procedure:

1. Let $\ell = \min\{m, r\}$. Use an SVD to construct a matrix $\mathbf{Q} \in \mathbb{R}^{m\times\ell}$ that satisfies $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_{\ell}$ and spans the range of $\mathbf{Y}$. This construction takes $\mathrm{O}(mr\ell)$ time.
2. Compute $\mathbf{X}_{opt} = \mathrm{argmin}_{\mathbf{X}\in\mathbb{R}^{\ell\times n},\,\mathrm{rank}(\mathbf{X})\le k} \|\mathbf{Q}^{\mathrm{T}}\mathbf{A} - \mathbf{X}\|_{\mathrm{F}}$ in $\mathrm{O}(mn\ell + n\ell^2)$ time. In fact, since $\ell \le m$, we see that $\mathbf{X}_{opt}$ can be computed in $\mathrm{O}(mn\ell)$ time.
3. Return $\mathbf{Q}\mathbf{X}_{opt} \in \mathbb{R}^{m\times n}$ in $\mathrm{O}(mn\ell)$ time.

$\mathbf{Q}\mathbf{X}_{opt}$ is a matrix of rank at most k that lies within the column span of $\mathbf{Y}$. Note that though $\Pi^{\xi}_{\mathbf{Y},k}(\mathbf{A})$ can depend on ξ, the algorithm above computes the same matrix, independent of ξ. The following result, which appeared as Lemma 18 in [5], proves that this algorithm computes $\Pi^{\mathrm{F}}_{\mathbf{Y},k}(\mathbf{A})$ and a constant factor approximation to $\Pi^{2}_{\mathbf{Y},k}(\mathbf{A})$.

LEMMA 5.1 (see [5, Lemma 18]). *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{Y} \in \mathbb{R}^{m\times r}$, and an integer $k \le r$, the matrix $\mathbf{Q}\mathbf{X}_{opt} \in \mathbb{R}^{m\times n}$ described above satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{X}_{opt}\|_{\mathrm{F}}^2 = \|\mathbf{A} - \Pi^{\mathrm{F}}_{\mathbf{Y},k}(\mathbf{A})\|_{\mathrm{F}}^2,$$
$$\|\mathbf{A} - \mathbf{Q}\mathbf{X}_{opt}\|_{2}^2 \le 2\|\mathbf{A} - \Pi^{2}_{\mathbf{Y},k}(\mathbf{A})\|_{2}^2.$$

The discussion above the lemma shows that $\mathbf{Q}\mathbf{X}_{opt}$ can be computed in $\mathrm{O}(mn\ell + mr\ell)$ time.

5.1.1. Matrix-Pythagoras and generalized least-squares regression. Lemma 5.2 is the analogue of the Pythagoras theorem in the matrix setting. A proof of this lemma can be found in [5]. Lemma 5.3 is an immediate corollary of matrix-Pythagoras.

LEMMA 5.2. *If $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ and $\mathbf{X}\mathbf{Y}^{\mathrm{T}} = \mathbf{0}_{m\times m}$ or $\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{0}_{n\times n}$, then for both $\xi = 2, \mathrm{F}$*

$$\|\mathbf{X} + \mathbf{Y}\|_{\xi}^2 \le \|\mathbf{X}\|_{\xi}^2 + \|\mathbf{Y}\|_{\xi}^2.$$

LEMMA 5.3. *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{C} \in \mathbb{R}^{m\times r}$, and for all $\mathbf{X} \in \mathbb{R}^{r\times n}$ and for both $\xi = 2, \mathrm{F}$*

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{+}\mathbf{A}\|_{\xi}^2 \le \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_{\xi}^2.$$

Proof. Write $\mathbf{A} - \mathbf{C}\mathbf{X} = (\mathbf{I} - \mathbf{C}\mathbf{C}^{+})\mathbf{A} + \mathbf{C}(\mathbf{C}^{+}\mathbf{A} - \mathbf{X})$. Observe that $((\mathbf{I} - \mathbf{C}\mathbf{C}^{+})\mathbf{A})^{\mathrm{T}}\mathbf{C}(\mathbf{C}^{+}\mathbf{A}) = \mathbf{0}_{n\times n}$. By Lemma 5.2, $\|\mathbf{A} - \mathbf{C}\mathbf{X}\|_{\xi}^2 \ge \|(\mathbf{I} - \mathbf{C}\mathbf{C}^{+})\mathbf{A}\|_{\xi}^2 + \|\mathbf{C}_1(\mathbf{C}^{+}\mathbf{A} - \mathbf{X})\|_{\xi}^2 \ge \|(\mathbf{I} - \mathbf{C}\mathbf{C}^{+})\mathbf{A}\|_{\xi}^2$. ∎

5.1.2. Low-rank matrix approximation based on projections. The low-rank matrix approximation algorithm investigated in this paper is an instance of a wider class of low-rank approximation 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 such projection schemes is well studied [9, 26, 5]. The following result appeared as Lemma 7 in [9] (see also Theorem 9.1 in [26]).

LEMMA 5.4 (see [9, Lemma 7]). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ. Fix k satisfying $0 \le k \le \rho$. Given a matrix $\mathbf{\Omega} \in \mathbb{R}^{n\times r}$, with $r \ge k$, construct $\mathbf{Y} = \mathbf{A\Omega}$. If $\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}$ has full row-rank, then, for $\xi = 2, \mathrm{F}$,*

$$\|\mathbf{A} - \mathbf{YY}^{\dagger}\mathbf{A}\|_\xi^2 \le \|\mathbf{A} - \Pi_{\mathbf{Y},k}^{\xi}(\mathbf{A})\|_\xi^2 \le \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \left\|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Omega}\Big(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}\Big)^{\dagger}\right\|_\xi^2. \tag{5.1}$$

This lemma provides an upper bound for the residual error of the low-rank matrix approximation obtained via projections. We now prove a new result for the forward error.

LEMMA 5.5. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ. Fix k satisfying $0 \le k \le \rho$. Given a matrix $\mathbf{\Omega} \in \mathbb{R}^{n\times r}$, with $r \ge k$, construct $\mathbf{Y} = \mathbf{A\Omega}$. If $\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}$ has full row-rank, then, for $\xi = 2, \mathrm{F}$,*

$$\|\mathbf{A}_k - \mathbf{YY}^{\dagger}\mathbf{A}\|_\xi^2 \le \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \left\|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Omega}\Big(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}\Big)^{\dagger}\right\|_\xi^2. \tag{5.2}$$

Proof. For both $\xi = 2, \mathrm{F}$,

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

In the above, in the first inequality we used Lemma 5.2 ($(\mathbf{A}_k - \mathbf{YY}^{\dagger}\mathbf{A}_k)(-\mathbf{YY}^{\dagger}\mathbf{A}_{\rho-k})^{\mathrm{T}} = \mathbf{0}_{m\times m}$ because $\mathbf{A}_k\mathbf{A}_{\rho-k}^{\mathrm{T}} = \mathbf{0}_{m\times m}$). In the second inequality we used Lemma 5.3 (with $\mathbf{X} = (\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega})^{\dagger}\mathbf{V}_k^{\mathrm{T}}$). In the third equality, we used the fact that $(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega})(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega})^{+} = \mathbf{I}_k$, since, by assumption, $\mathrm{rank}(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}) = k$. In the last inequality we used the submultiplicativity property of the spectral and Frobenius norms, i.e., for any three matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,

$$\|\mathbf{XYZ}\|_\xi^2 \le \|\mathbf{X}\|_2^2\|\mathbf{YZ}\|_\xi^2 \le \|\mathbf{X}\|_2^2\|\mathbf{Y}\|_\xi^2\|\mathbf{Z}\|_2^2. \qquad \square$$

5.1.3. Least-squares regression based on projections. Similarly, one of the two SRHT least-squares regression algorithms analyzed in this article is an instance of a wider class of approximation algorithms where the dimensions of the input matrix and the vector of the regression problem are reduced via premultiplication with a

random matrix. Lemma 9 in [7] provides a general framework for the analysis of such projection algorithms.

LEMMA 5.6. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *(*$m \geq n$*) of rank* ρ *and* $\mathbf{b} \in \mathbb{R}^m$ *be inputs to the least-squares problem* $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$*. Let* $\mathbf{U} \in \mathbb{R}^{m\times\rho}$ *contain the top* ρ *left singular vectors of* $\mathbf{A}$*, and let* $\mathbf{\Omega} \in \mathbb{R}^{m\times r}$ *(*$\rho \leq r \leq m$*) be a matrix such that* $\mathrm{rank}(\mathbf{\Omega}^{\mathrm{T}}\mathbf{U}) = \mathrm{rank}(\mathbf{U})$*. Then*

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 + \left\|\left(\mathbf{\Omega}^{\mathrm{T}}\mathbf{U}\right)^{\dagger}\mathbf{\Omega}^{\mathrm{T}}\left(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\right)\right\|_2^2.$$

In the above, $\mathbf{x}_{opt} = \mathbf{A}^{\dagger}\mathbf{b}$ *and* $\tilde{\mathbf{x}}_{opt} = \left(\mathbf{\Omega}^{\mathrm{T}}\mathbf{A}\right)^{\dagger}\mathbf{\Omega}^{\mathrm{T}}\mathbf{b}$*.*

The following lemma is a restatement of [19, Lemma 2], along with [19, (9) and (11)]. It gives a bound on the forward error of the approximation of a least-squares problem that is obtained via projections. In [19] the parameters α and β are fixed to $\alpha = 1/\sqrt{2}$ and $\beta = \varepsilon/2$ for some parameter $0 < \varepsilon < 1$. Showing the result for general $\alpha > 0$ and $\beta > 0$ is straightforward, and hence a detailed proof is omitted.

LEMMA 5.7 (see [19, Lemma 2]). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *(*$m \geq n$*) of rank* $\rho = n$ *and* $\mathbf{b} \in \mathbb{R}^m$ *be inputs to the least-squares problem* $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$*. Let* $\mathbf{U} \in \mathbb{R}^{m\times\rho}$ *contain the top* ρ *left singular vectors of* $\mathbf{A}$*, and let* $\mathbf{\Omega} \in \mathbb{R}^{m\times r}$ *(*$\rho \leq r \leq m$*). For some* $\alpha > 0$ *and* $\beta > 0$*, assume that*

$$\sigma_{\min}\left(\mathbf{U}^{\mathrm{T}}\mathbf{\Omega}\right) \geq \alpha^{\frac{1}{2}} \tag{5.3}$$

and

$$\|\mathbf{U}^{\mathrm{T}}\mathbf{\Omega}\mathbf{\Omega}^{\mathrm{T}}\left(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\right)\|_2^2 \leq \beta\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2. \tag{5.4}$$

Furthermore, assume that there exists a $\gamma \in (0, 1]$ *such that* $\|\mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^{\mathrm{T}}\mathbf{b}\|_2 \geq \gamma\|\mathbf{b}\|_2$*. Then*

$$\|\mathbf{x}_{opt} - \tilde{\mathbf{x}}_{opt}\|_2 \leq \left(\frac{\alpha}{\beta}\right)^{\frac{1}{2}} \cdot \left(\kappa(\mathbf{A})\sqrt{\gamma^{-2} - 1}\right)\|\mathbf{x}_{opt}\|_2. \tag{5.5}$$

In the above, $\mathbf{x}_{opt} = \mathbf{A}^{\dagger}\mathbf{b}$ *and* $\tilde{\mathbf{x}}_{opt} = \left(\mathbf{\Omega}^{\mathrm{T}}\mathbf{A}\right)^{\dagger}\mathbf{\Omega}^{\mathrm{T}}\mathbf{b}$*.*

5.2. Proof of Theorem 2.1.

5.2.1. Frobenius norm bounds. We first prove the Frobenius norm bounds in the theorem (i.e., (i), (ii), (iii), and (v)). We would like to apply Lemma 5.4 with $\mathbf{\Omega} = \mathbf{\Theta}^{\mathrm{T}} \in \mathbb{R}^{n\times r}$ and $\xi = \mathrm{F}$. Notice that because of our assumption that

$$r \geq 6\mathrm{C}^2\varepsilon^{-1}\left[\sqrt{k} + \sqrt{8\ln(n/\delta)}\right]^2\ln(k/\delta),$$

where $\mathrm{C} > 1$, Lemma 4.1 implies that, with probability at least $1 - 3\delta$,

$$\mathrm{rank}(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}) = k;$$

so, for $\xi = \mathrm{F}$, Lemma 5.4 applies with the same probability, yielding

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \Pi_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2 + \|\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}})^{\dagger}\|_{\mathrm{F}}^2. \tag{5.6}$$

We continue by bounding the second term on the right-hand side of the above inequality,

$$\begin{aligned}
S &:= \|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger}\|_{\mathrm{F}}^2 \\
&\leq 2\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{V}_k\|_{\mathrm{F}}^2 + 2\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}((\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\mathrm{T}})\|_{\mathrm{F}}^2 \\
&\leq 2\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{V}_k\|_{\mathrm{F}}^2 + 2\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2\|(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\mathrm{T}}\|_2^2 \\
&\leq 8\varepsilon\cdot\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2 + 2\cdot\left(\frac{11}{4}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right)\cdot(2.38\varepsilon) \\
&\leq 22\varepsilon\cdot\|\boldsymbol{\Sigma}_{\rho-k}\|_{\mathrm{F}}^2.
\end{aligned}$$

In the above, in the first inequality we used the fact that for any two matrices $\mathbf{X}, \mathbf{Y}$, $\|\mathbf{X}+\mathbf{Y}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{X}\|_{\mathrm{F}}^2 + 2\|\mathbf{Y}\|_{\mathrm{F}}^2$. To justify the first estimate in the third inequality, first notice that $\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{V}_k = \mathbf{0}_{n\times k}$. Next use Lemma 4.11 with $\mathrm{R} = \mathrm{C}\sqrt{\ln(k/\delta)}$. From the lower bound on r, we have that

$$\frac{\sqrt{r}}{1+\sqrt{8\ln(n/\delta)}} \geq \sqrt{6\varepsilon^{-1}}\cdot\frac{\sqrt{k}+\sqrt{8\ln(n/\delta)}}{1+\sqrt{8\ln(n/\delta)}}\cdot\mathrm{C}\sqrt{\ln(k/\delta)} > \mathrm{R} > 0,$$

so this choice of R satisfies the requirements of Lemma 4.11. Apply Lemma 4.11 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{V}_k\|_{\mathrm{F}}^2 \leq 4(\mathrm{R}+1)^2\frac{(\sqrt{k}+\sqrt{8\ln(n/\delta)})^2}{r}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right] \\
&\geq 1 - e^{-\mathrm{R}^2/4} - 2\delta.
\end{aligned}$$

Use the lower bound on r to justify the estimate

$$\begin{aligned}
4(\mathrm{R}+1)^2\frac{\left[\sqrt{k}+\sqrt{8\ln(n/\delta)}\right]^2}{r} &\leq 4(\mathrm{R}+1)^2\frac{\left[\sqrt{k}+\sqrt{8\ln(n/\delta)}\right]^2}{6\mathrm{C}^2\varepsilon^{-1}\left[\sqrt{k}+\sqrt{8\ln(n/\delta)}\right]^2\ln(k/\delta)} \\
&= \frac{2\varepsilon}{3}\cdot\frac{(\mathrm{C}\sqrt{\ln(k/\delta)}+1)^2}{\mathrm{C}^2\ln(k/\delta)} \\
&\leq \frac{2\varepsilon}{3}\left(1+\frac{1}{\mathrm{C}\sqrt{\ln(k/\delta)}}\right)^2.
\end{aligned}$$

This estimate implies that

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{V}_k\|_{\mathrm{F}}^2 \leq \frac{2\varepsilon}{3}\left(1+\frac{1}{\mathrm{C}\sqrt{\ln(k/\delta)}}\right)^2\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right] \\
&\geq 1 - \delta^{C^2\ln(k/\delta)/4} - 2\delta.
\end{aligned}$$

Since $\mathrm{C} > 1$ and $k \geq 2$, a simple numerical estimation allows us to state that, more simply,

$$\mathbb{P}\left[\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{V}_k\|_{\mathrm{F}}^2 \leq 4\varepsilon\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right] \geq 1 - \delta^{C^2\ln(k/\delta)/4} - 2\delta.$$

The remaining estimates in the third inequality follow from applying Lemma 4.1 (keeping in mind our lower bound on r) to obtain

$$\mathbb{P}\left[\|(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\mathrm{T}}\|_2^2 \leq 2.38\varepsilon\right] \geq 1 - 3\delta.$$

Applying Lemma 4.10 with $\eta = 7/4$, we obtain

$$\mathbb{P}\left[\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2 \leq \frac{11}{4}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right] \geq 1 - \left[\frac{\mathrm{e}^{7/4}}{(1+7/4)^{1+7/4}}\right]^{r/(1+\sqrt{8\ln(n/\delta)})^2} - \delta.$$

We have the estimate

$$\frac{\mathrm{e}^{7/4}}{(1+7/4)^{1+7/4}} < \frac{1}{\mathrm{e}},$$

so in fact

$$\begin{aligned}\mathbb{P}\left[\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_{\mathrm{F}}^2 \leq \frac{11}{4}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\|_{\mathrm{F}}^2\right] &\geq 1 - \mathrm{e}^{-r/(1+\sqrt{8\ln(n/\delta)})^2} - \delta \\ &\geq 1 - \mathrm{e}^{-6\mathrm{C}^2\varepsilon^{-1}\ln(k/\delta)} - \delta \\ &\geq 1 - \mathrm{e}^{-\ln(k/\delta)} - \delta \\ &\geq 1 - 2\delta.\end{aligned}$$

Combining (5.6) with the bound on S, we obtain

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \Pi_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}}^2 \leq (1+22\varepsilon)\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2.$$

Taking the square roots of both sides and using the fact that $\sqrt{1+22\varepsilon} \leq 1+22\varepsilon$ gives the bound

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}} \leq \|\mathbf{A} - \Pi_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}} \leq \sqrt{1+22\varepsilon}\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} \leq (1+22\varepsilon)\,\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

Equation (i) in the theorem follows directly from this:

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}} \leq (1+22\varepsilon)\,\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

To derive (ii), recall the equality $\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} = \|\mathbf{A} - \Pi_{\mathbf{Y},k}^{\mathrm{F}}(\mathbf{A})\|_{\mathrm{F}}$, established in Lemma 5.1. From this it follows that

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (1+22\varepsilon)\,\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}$$

also.

We now prove (iii) in the theorem. Equation (5.2) with $\xi = \mathrm{F}$ and $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{\mathrm{T}} \in \mathbb{R}^{n\times r}$ gives

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_{\mathrm{F}}^2 \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + \|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\right)^{\dagger}\|_{\mathrm{F}}^2.$$

Now recall the bound for S:

$$\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\right)^{\dagger}\|_{\mathrm{F}}^2 \leq 22\varepsilon\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2.$$

So,

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}\|_{\mathrm{F}}^2 \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 + 22\varepsilon\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 = (1+22\varepsilon)\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2.$$

Taking the square roots of both sides and using the fact that $\sqrt{1+22\varepsilon} \leq 1+22\varepsilon$ gives (iii).

Finally, we prove (iv):

$$\|\mathbf{A}_k-\tilde{\mathbf{A}}_k\|_{\mathrm{F}} = \|\mathbf{A}-\mathbf{A}_k-(\mathbf{A}-\tilde{\mathbf{A}}_k)\|_{\mathrm{F}} \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}+\|\mathbf{A}-\tilde{\mathbf{A}}_k\|_{\mathrm{F}} \leq (2+22\varepsilon)\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}},$$

where the first inequality follows by the triangle inequality, and the second by using the bound obtained in (ii) in the theorem.

The failure probability in the theorem follows from a union bound on all the probabilistic events involved in bounding S.

5.2.2. Spectral norm bounds. We now prove the spectral norm bounds in Theorem 2.1 (i.e., (v), (vi), (vii), and (viii)). Lemma 4.1 implies that, with this choice of r,

$$\|(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^\dagger\|_2^2 \leq (1-\sqrt{\varepsilon})^{-1},$$

with probability at least $1-3\delta$. Consequently, $\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}$ has full row-rank, and Lemma 5.4 with $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{\mathrm{T}} \in \mathbb{R}^{n\times r}$ and $\xi = 2$ applies with the same probability, yielding

$$\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}\|_2^2 \leq \|\mathbf{A}-\mathbf{A}_k\|_2^2 + (1-\sqrt{\varepsilon})^{-1}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2. \tag{5.7}$$

Also, the spectral norm bound in Lemma 5.1 implies

$$\|\mathbf{A}-\tilde{\mathbf{A}}_k\|_2^2 \leq 2\|\mathbf{A}-\Pi_{\mathbf{Y},k}^2(\mathbf{A})\|_2^2 \leq 2\left(\|\mathbf{A}-\mathbf{A}_k\|_2^2 + (1-\sqrt{\varepsilon})^{-1}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2\right). \tag{5.8}$$

We now provide an upper bound for $\sqrt{Z}$ where Z is the scalar

$$Z := \|\mathbf{A}-\mathbf{A}_k\|_2^2 + (1-\sqrt{\varepsilon})^{-1}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2.$$

From Lemma 4.8 we obtain

$$Z \leq \left(1+\frac{5}{1-\sqrt{\varepsilon}}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_2^2 + \frac{\ln(\rho/\delta)}{(1-\sqrt{\varepsilon})r}\left(\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}-\mathbf{A}_k\|_2\right)^2$$

with probability at least $1-5\delta$. Using that $\varepsilon < 1/3$, we see that $(1-\sqrt{\varepsilon})^{-1} < 3$, so

$$Z \leq 16\cdot\|\mathbf{A}-\mathbf{A}_k\|_2^2 + \frac{3\ln(\rho/\delta)}{r}\left(\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + \sqrt{8\ln(n/\delta)}\|\mathbf{A}-\mathbf{A}_k\|_2\right)^2.$$

Use the subadditivity of the square-root function and rearrange the spectral and Frobenius norm terms to obtain that

$$\sqrt{Z} \leq \left(4+\sqrt{\frac{3\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_2 + \sqrt{\frac{3\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

Apply (5.7) to arrive at (v) in the theorem,

$$\|\mathbf{A}-\mathbf{Y}\mathbf{Y}^\dagger\mathbf{A}\|_2 \leq \left(4+\sqrt{\frac{3\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{A}-\mathbf{A}_k\|_2 + \sqrt{\frac{3\ln(\rho/\delta)}{r}}\cdot\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

Take the square root of both sides of (5.8), use the subadditivity of the square root function, and use the bound for $\sqrt{Z}$ to find (vi):

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \left(6 + \sqrt{\frac{6\ln(n/\delta)\ln(\rho/\delta)}{r}}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{6\ln(\rho/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

We now derive the spectral norm bounds on the forward errors. Equation (5.2) with $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{\mathrm{T}} \in \mathbb{R}^{n\times r}$ gives

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2 + \|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\left(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\right)^{\dagger}\|_2^2.$$

Use the inequality $\|(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger}\|_2^2 \leq (1-\sqrt{\varepsilon})^{-1}$ to obtain

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2^2 + (1-\sqrt{\varepsilon})^{-1}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2.$$

Take the square root of both sides of this inequality to obtain

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A} - \mathbf{A}_k\|_2^2 + (1-\sqrt{\varepsilon})^{-1}\|\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2}$$

and identify the right-hand side as $\sqrt{Z}$. Use the bound on $\sqrt{Z}$ to arrive at (vii):

$$\|\mathbf{A}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{A}\|_2 \leq \left(4 + \sqrt{\frac{3\ln(n/\delta)\ln(\rho/\delta)}{r}}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{3\ln(\rho/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

We now prove (viii). First, recall that $\tilde{\mathbf{A}}_k = \mathbf{Q}\mathbf{X}_{opt}$ and observe that

$$\|\mathbf{A}_k - \tilde{\mathbf{A}}_k\|_2 = \|\mathbf{A}_k + \mathbf{A}_{\rho-k} - \tilde{\mathbf{A}}_k - \mathbf{A}_{\rho-k}\|_2 \leq \|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 + \|\mathbf{A}_{\rho-k}\|_2.$$

Now, recall (vi):

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \left(6 + \sqrt{\frac{6\ln(n/\delta)\ln(\rho/\delta)}{r}}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{6\ln(\rho/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

In conjunction with the previous inequality, this gives us the desired bound:

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_2 \leq \left(7 + \sqrt{\frac{12\ln(n/\delta)\ln(\rho/\delta)}{r}}\right) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2 + \sqrt{\frac{6\ln(\rho/\delta)}{r}} \cdot \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

Finally, we recall that the two probabilistic events we used in our derivations—that $(\mathbf{V}_k^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})^{\dagger}$ is bounded and that our application of Lemma 4.8 succeeds—hold with probabilities at least $1-3\delta$ and $1-2\delta$, respectively, so the failure probability for each of these four spectral error bounds is no more than 5δ.

5.2.3. Running time analysis. The matrix $\mathbf{Y}$ can be constructed in at most $2mn\log_2(r+1)$ arithmetic operations (see Lemma 1.3).

Given $\mathbf{Y} \in \mathbb{R}^{m\times r}$, the matrix $\mathbf{Y}(\mathbf{Y}^{\dagger}\mathbf{A}) \in \mathbb{R}^{m\times n}$ can be constructed in $\mathrm{O}(mr\ell + mn\ell)$ arithmetic operations as follows. Observe that $\mathbf{Y}(\mathbf{Y}^{\dagger}\mathbf{A}) = \mathbf{Q}(\mathbf{Q}^{\mathrm{T}}\mathbf{A})$ and $\mathbf{Q} \in \mathbb{R}^{m\times\ell}$ can be computed in $\mathrm{O}(mr\ell)$ time. Recall that $\ell = \min\{m, r\}$. Computing $\mathbf{Q}^{\mathrm{T}}\mathbf{A}$ requires $\mathrm{O}(mn\ell)$ operations, as does the subsequent computation of $\mathbf{Q}(\mathbf{Q}^{\mathrm{T}}\mathbf{A})$. Thus,

in total, $\mathrm{O}(mr\ell + mn\ell)$ operations are required. If $\ell = r$, this is $\mathrm{O}(mr^2 + mnr)$, but if $r > m$, the total operation count becomes $\mathrm{O}(m^2(r+n))$.

Finally, given $\mathbf{Y}$, the matrix $\tilde{\mathbf{A}}_k$ can be constructed in $\mathrm{O}(mn\ell + \ell^2 n)$ arithmetic operations as follows. As argued above, $\mathbf{Q}$ can be constructed in $\mathrm{O}(mr\ell)$ operations; then the product $\mathbf{Q}^{\mathrm{T}}\mathbf{A} \in \mathbb{R}^{\ell \times n}$ can be computed in $\mathrm{O}(mn\ell)$ operations. The SVD of $\mathbf{Q}^{\mathrm{T}}\mathbf{A}$ requires $\mathrm{O}(\ell n \min\{\ell, n\})$ operations, which is proportional to $\mathrm{O}(\ell^2 n)$ because $\min\{\ell, n\} = \min\{m, r, n\} = \min\{m, r\} = \ell$, since $r < n$. The final matrix multiplication $\mathbf{Q}\mathbf{X}_{opt}$ again requires $\mathrm{O}(mn\ell)$ arithmetic operations. The total operation count is therefore $\mathrm{O}(mn\ell + \ell^2 n)$. If $m > r$, the operation count is $\mathrm{O}(mnr + r^2 n)$, while if $m < r$, the operation count becomes $\mathrm{O}(m^2 n)$.

5.3. Proof of Theorem 3.1. To prove the first bound in the theorem (residual error analysis), we will use Lemma 5.6, which is the analogue of Lemma 5.4 but for linear regression. Using this lemma, the proof of the first bound in Theorem 3.1 is similar to the proof of the first Frobenius norm bound of Theorem 2.1.

We would like to apply Lemma 5.6 with $\mathbf{\Omega} = \mathbf{\Theta}^{\mathrm{T}} \in \mathbb{R}^{m \times r}$. For convenience, we take $\mathbf{U} = \mathbf{U}_{\mathbf{A}}$. Notice that Lemma 4.1 implies that, with probability at least $1 - 3\delta$, $\mathrm{rank}(\mathbf{\Theta U}) = \rho = n$; so, Lemma 5.6 applies with the same probability, yielding

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 + \|(\mathbf{\Theta U})^{\dagger}\mathbf{\Theta}\left(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\right)\|_2^2. \tag{5.9}$$

We continue by bounding the second term on the right-hand side of the above inequality (for notational convenience, let $\mathbf{z}_{opt} = \mathbf{A}\mathbf{x}_{opt} - \mathbf{b}$),

$$\begin{aligned}
S &:= \|(\mathbf{\Theta U})^{\dagger}\mathbf{\Theta}\left(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\right)\|_2^2 \\
&\leq 2\|\mathbf{U}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Theta}\mathbf{z}_{opt}\|_2^2 + 2\|((\mathbf{\Theta U})^{\dagger} - (\mathbf{\Theta U})^{\mathrm{T}})\mathbf{\Theta}\mathbf{z}_{opt}\|_2^2 \\
&\leq 2\|\mathbf{U}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Theta}\mathbf{z}_{opt}\|_2^2 + 2\|((\mathbf{\Theta U})^{\dagger} - (\mathbf{\Theta U})^{\mathrm{T}})\|_2^2\|\mathbf{\Theta}\mathbf{z}_{opt}\|_2^2 \\
&= 2\|\mathbf{z}_{opt}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Theta U}\|_2^2 + 2\|((\mathbf{\Theta U})^{\dagger} - (\mathbf{\Theta U})^{\mathrm{T}})\|_2^2\|\mathbf{z}_{opt}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\|_2^2 \\
&\leq 8\varepsilon \cdot \|\mathbf{z}_{opt}\|_2^2 + 2 \cdot (2.38\varepsilon) \cdot \left(\frac{11}{4}\|\mathbf{z}_{opt}\|_2^2\right) \\
&\leq 22\varepsilon \cdot \|\mathbf{z}_{opt}\|_2^2.
\end{aligned}$$

In the above, in the first inequality we used the fact that for any two matrices $\mathbf{X}, \mathbf{Y}$, $\|\mathbf{X} + \mathbf{Y}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{X}\|_{\mathrm{F}}^2 + 2\|\mathbf{Y}\|_{\mathrm{F}}^2$. To justify the first estimate in the third inequality, first notice that $\mathbf{z}_{opt}^{\mathrm{T}}\mathbf{U} = \mathbf{0}_{1\times n}$ since

$$\mathbf{z}_{opt}^{\mathrm{T}}\mathbf{U} = \left(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\right)^{\mathrm{T}}\mathbf{U} = \left(\mathbf{A}\mathbf{A}^{+}\mathbf{b} - \mathbf{b}\right)^{\mathrm{T}}\mathbf{U} = \left(\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{b} - \mathbf{b}\right)^{\mathrm{T}}\mathbf{U} = \mathbf{0}_{1\times n}.$$

Next, use Lemma 4.11 with $\mathrm{R} = \mathrm{C}\sqrt{\ln(n/\delta)}$. Recall that

$$r \geq 6\mathrm{C}^2\varepsilon^{-1}\left[\sqrt{n} + \sqrt{8\ln(m/\delta)}\right]^2 \ln(n/\delta),$$

where $\mathrm{C} \geq 1$, so

$$\frac{\sqrt{r}}{1 + \sqrt{8\ln(m/\delta)}} \geq \sqrt{6\varepsilon^{-1}} \cdot \frac{\sqrt{n} + \sqrt{8\ln(m/\delta)}}{1 + \sqrt{8\ln(m/\delta)}} \cdot \mathrm{C}\sqrt{\ln(n/\delta)} > \mathrm{R} > 0,$$

and this choice of R satisfies the requirements of Lemma 4.11. Apply Lemma 4.11 to obtain

$$\mathbb{P}\left[\|\mathbf{z}_{opt}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Theta U}\|_2^2 \leq 4(\mathrm{R}+1)^2\frac{(\sqrt{n} + \sqrt{8\ln(m/\delta)})^2}{r}\|\mathbf{z}_{opt}^{\mathrm{T}}\|_2^2\right] \geq 1 - e^{-\mathrm{R}^2/4} - 2\delta.$$

Manipulations similar to those used in proving the first Frobenius norm bound in Theorem 2.1 show that the lower bound on r implies

$$\mathbb{P}\left[\|\mathbf{z}_{opt}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{U}\|_2^2 \le 4\varepsilon\|\mathbf{z}_{opt}^{\mathrm{T}}\|_2^2\right] \ge 1-\delta^{C^2\ln(n/\delta)/4}-2\delta.$$

The remaining estimates in the third inequality follow from applying Lemma 4.1 to obtain

$$\mathbb{P}\left[\|((\boldsymbol{\Theta}\mathbf{U})^{\dagger}-(\boldsymbol{\Theta}\mathbf{U})^{\mathrm{T}})\|_2^2 \le 2.38\varepsilon\right] \ge 1-3\delta.$$

Applying Lemma 4.10 with $\eta = 7/4$, we obtain

$$\mathbb{P}\left[\|\mathbf{z}_{opt}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2 \le \frac{11}{4}\|\mathbf{z}_{opt}^{\mathrm{T}}\|_2^2\right] \ge 1-\left[\frac{e^{7/4}}{(1+7/4)^{1+7/4}}\right]^{r/(1+\sqrt{8\ln(m/\delta)})^2}-\delta.$$

Manipulations similar to those used in proving the first Frobenius norm bound in Theorem 2.1 show that the latter bound implies

$$\mathbb{P}\left[\|\mathbf{z}_{opt}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\|_2^2 \le \frac{11}{4}\|\mathbf{z}_{opt}^{\mathrm{T}}\|_2^2\right] \ge 1-2\delta.$$

Combining (5.9) with the bound on S, we obtain

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt}-\mathbf{b}\|_2^2 \le (1+22\varepsilon)\cdot\|\mathbf{A}\mathbf{x}_{opt}-\mathbf{b}\|_2^2.$$

Taking the square root to both sides and using the fact that $\sqrt{1+22\varepsilon} \le 1+22\varepsilon$ gives the bound in the theorem,

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt}-\mathbf{b}\|_2 \le (1+22\varepsilon)\cdot\|\mathbf{A}\mathbf{x}_{opt}-\mathbf{b}\|_2.$$

The failure probability in the theorem follows by a union bound on all the probabilistic events involved in the proof.

We now prove the forward error bound in the theorem. Towards this end, we will use Lemma 5.7 with $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{\mathrm{T}}$. Recall that (5.5) in the lemma,

$$\|\mathbf{x}_{opt}-\tilde{\mathbf{x}}_{opt}\|_2 \le \left(\frac{\alpha}{\beta}\right)^{\frac{1}{2}}\cdot\left(\kappa(\mathbf{A})\sqrt{\gamma^{-2}-1}\right)\|\mathbf{x}_{opt}\|_2,$$

is satisfied if α and β satisfy (5.3) and (5.4), respectively, and $\gamma \in (0,1]$ satisfies $\|\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{b}\|_2 \ge \gamma\|\mathbf{b}\|_2$. By hypothesis, such a γ exists. We now show that appropriate α and β exist.

Lemma 4.1 implies that

$$\sigma_{\min}\left(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\right) \ge \left(1-\sqrt{\varepsilon}\right)^{\frac{1}{2}},$$

so $\alpha = 1-\sqrt{\varepsilon}$ satisfies (5.3). In the proof of the residual error bound in this theorem, we showed that

$$\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\left(\mathbf{A}\mathbf{x}_{opt}-\mathbf{b}\right)\|_2^2 \le 4\varepsilon\|\mathbf{A}\mathbf{x}_{opt}-\mathbf{b}\|_2^2,$$

so $\beta = 4\varepsilon$ satisfies (5.4). With these choices of α and β, (5.5) in Lemma 5.7 gives the claimed forward error bound.

6. Experiments. In this section, we experimentally investigate the tightness of the residual and forward error bounds provided in Theorem 2.1 for the spectral and Frobenius norm approximation errors of SRHT low-rank approximations of the forms $\mathbf{YY}^\dagger\mathbf{A}$ and $\tilde{\mathbf{A}}_k = \mathbf{QX}_{opt}$. Additionally, we experimentally verify that the SRHT algorithm is not significantly less accurate than the Gaussian low-rank approximation algorithm.

6.1. 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
$$\mathbf{A} = [100\boldsymbol{e}_1 + \boldsymbol{e}_2, 100\boldsymbol{e}_1 + \boldsymbol{e}_3, \ldots, 100\boldsymbol{e}_1 + \boldsymbol{e}_{n+1}],$$
where $\boldsymbol{e}_i \in \mathbb{R}^{n+1}$ 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, $\mathbf{C} = \mathbf{UBV}^{\mathrm{T}}$, where $\mathbf{G} = \mathbf{U\Sigma V}^{\mathrm{T}}$ 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 Theorem 2.1, 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 $\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$ terms in the spectral norm bound of Theorem 2.1 are quite large compared to the $\|\mathbf{A} - \mathbf{A}_k\|_2$ 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.

$\mathbf{B}$ 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 $\mathbf{B}$ and $\mathbf{C}$. Let $\mathcal{S}$ be a k-dimensional subspace; then its coherence is defined as
$$\mu(\mathcal{S}) = \max_i \mathbf{P}_{ii},$$
where $\mathbf{P}$ is the projection onto $\mathcal{S}$; the coherence of $\mathcal{S}$ is always between k/n and 1 [10]. It is clear that all the right singular spaces of $\mathbf{B}$ are maximally coherent, and it is known that with high probability the dominant right k-dimensional singular space of $\mathbf{C}$ is quite incoherent, with coherence on the order of $\max\{k, \log n\}/n$ [10].

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 $\mathbf{YY}^\dagger\mathbf{M}$ where $\mathbf{Y}$ comprises randomly sampled columns of $\mathbf{M}$. Here and elsewhere we use the matrix $\mathbf{M}$ to refer interchangeably to $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. It is known that such approximations are quite inaccurate unless the dominant k-dimensional right singular space of $\mathbf{M}$ is incoherent [39, 22]. One could interpret SRHT approximation algorithms as consisting of a rotation of the right singular spaces of $\mathbf{M}$ by multiplying from the right with $\mathbf{DH}^{\mathrm{T}}$ 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 $\mathbf{M}$ has highly coherent right singular spaces, then the right singular spaces of $\mathbf{MDH}^{\mathrm{T}}$ will be less coherent but possibly still far from incoherent. Thus we compare the performance of the SRHT

approximations on $\mathbf{B}$, which has maximally coherent right singular spaces, to their performance on $\mathbf{C}$, which has almost maximally incoherent right singular spaces.

6.2. Empirical comparison of the SRHT and Gaussian algorithms. Figure 6.1 depicts the relative residual errors of the Gaussian and SRHT algorithms for both approximations addressed in Theorem 2.1, $\mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{M}$ and $\mathbf{Q}\mathbf{X}_{opt}$, which we shall hereafter refer to, respectively, as the non-rank-restricted and rank-restricted approximations. The relative residual errors ($\|\mathbf{M} - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{M}\|_{\xi}/\|\mathbf{M} - \mathbf{M}_k\|_{\xi}$ and $\|\mathbf{M} - \mathbf{Q}\mathbf{X}_{opt}\|_{\xi}/\|\mathbf{M} - \mathbf{M}_k\|_{\xi}$ for $\xi = 2, \mathrm{F}$) shown in this figure for each value of k were obtained by taking the largest of the relative residual errors observed over 10 trials of low-rank approximations, each formed using $r = \lceil 2k \ln n \rceil$ samples.

With the exception of the residual spectral errors on dataset $\mathbf{A}$, which range between 2 and 9 times greater than the optimal rank-k spectral residual error for $k < 20$, we see that the residual errors for all three datasets 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.

By comparing the residual errors for datasets $\mathbf{B}$ and $\mathbf{C}$, which has the same singular values as $\mathbf{B}$ but is less coherent, we see evidence that the spectral norm accuracy of the SRHT approximations is increased on less coherent datasets; the same is true to a lesser extent for the Frobenius norm accuracy. The Gaussian approximations seem insensitive to the level of coherence. Only on the highly coherent dataset $\mathbf{B}$ 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 comparable to that of $\mathbf{B}_k$. In all, Figure 6.1 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 $r = \lceil 2k \ln n \rceil$ 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 conclusion of Theorem 2.1.

Figure 6.2 depicts the relative forward errors of the Gaussian and SRHT algorithms ($\|\mathbf{M}_k - \mathbf{Y}\mathbf{Y}^{\dagger}\mathbf{M}\|_{\xi}/\|\mathbf{M}_k\|_{\xi}$ and $\|\mathbf{M}_k - \mathbf{Q}\mathbf{X}_{opt}\|_{\xi}/\|\mathbf{M}_k\|_{\xi}$ for $\xi = 2, \mathrm{F}$) for the non-rank-restricted and rank-restricted approximations. The error shown for each k is the largest relative forward error observed among 10 trials of low-rank approximations, each formed using $r = \lceil 2k \ln n \rceil$ 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 ($\mathbf{B}$) than they are for matrices which are almost minimally coherent ($\mathbf{C}$). Overall, Figure 6.2 suggests that the SRHT low-rank approximation algorithms provide accurate approximations to $\mathbf{M}_k$ when r is in the regime suggested by Theorem 2.1.

6.3. Empirical evaluation of our error bounds. Figures 6.1 and 6.2 show that when $r = \lceil 2k \ln n \rceil$ samples are taken, the SRHT low-rank approximation algorithms both provide approximations to $\mathbf{M}$ that are within a factor of $(1+\varepsilon)$ as accurate

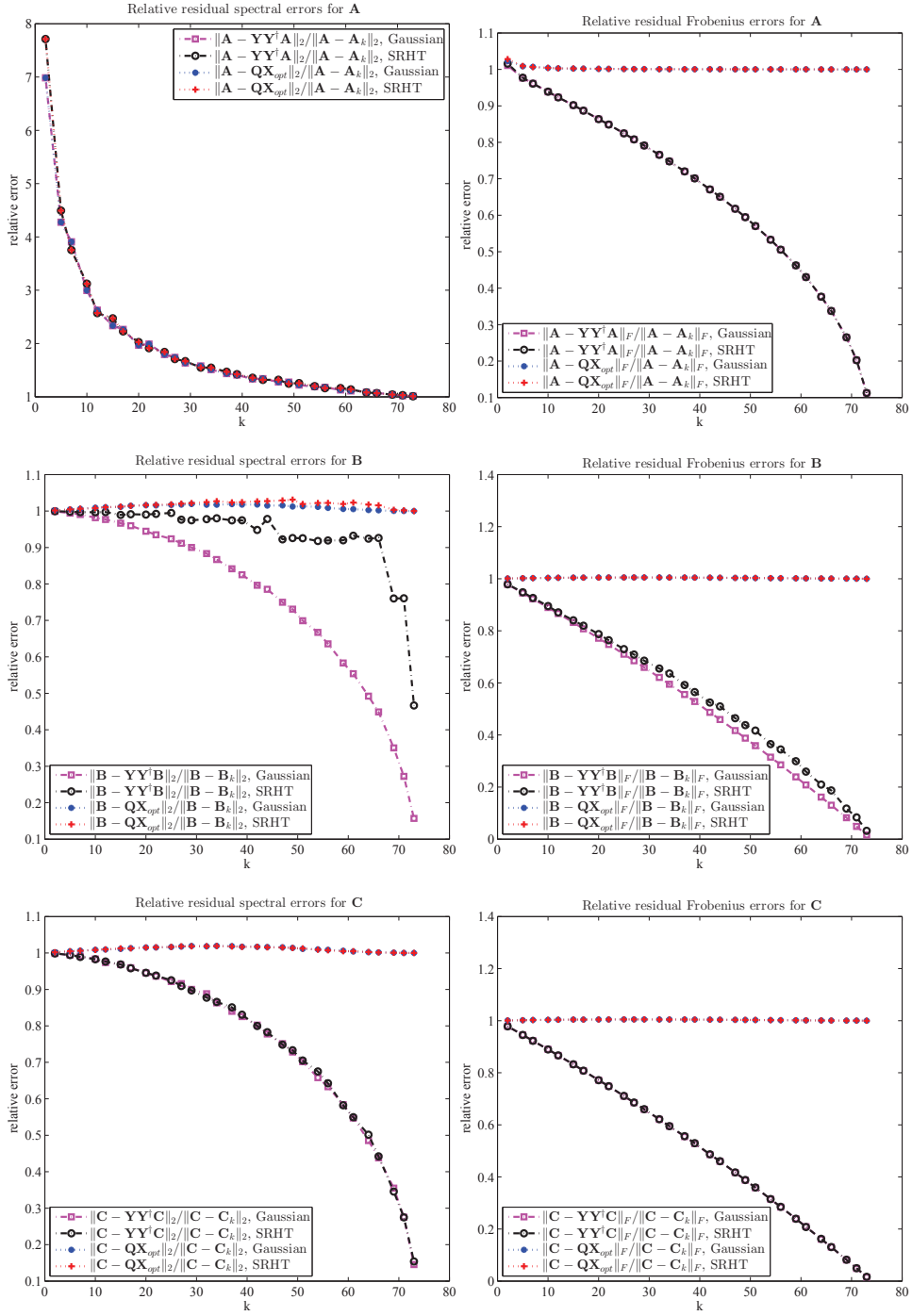

FIG. 6.1. *Relative spectral and Frobenius norm residual errors of the SRHT and Gaussian low-rank approximation algorithms* ($\|\mathbf{M}-\mathbf{YY}^{\dagger}\mathbf{M}\|_\xi/\|\mathbf{M}-\mathbf{M}_k\|_\xi$ *and* $\|\mathbf{M}-\mathbf{QX}_{opt}\|_\xi/\|\mathbf{M}-\mathbf{M}_k\|_\xi$ *for* $\xi=2,\mathrm{F}$*) as a function of* k *for the three datasets* $\mathbf{M}=\mathbf{A},\mathbf{B},\mathbf{C}$*. Each point is the worst of the errors observed over* 10 *trials.* $r=\lceil 2k\ln n\rceil$ *column samples were used in each trial.*

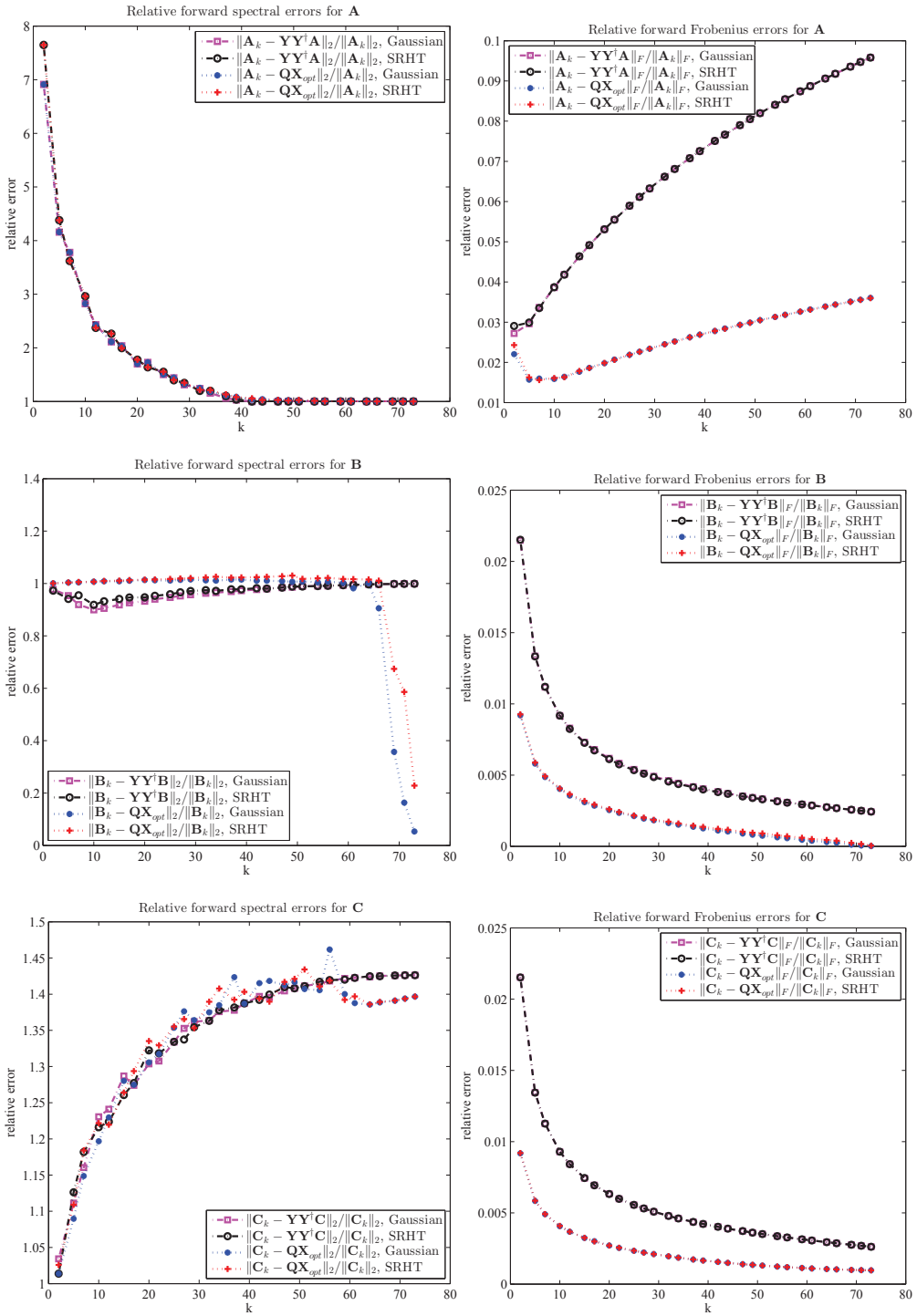

FIG. 6.2. *The relative spectral and Frobenius norm forward errors of the SRHT and Gaussian low-rank approximation algorithms* ($\|\mathbf{M}_k - \mathbf{Y}\mathbf{Y}^\dagger\mathbf{M}\|_\xi/\|\mathbf{M}_k\|_\xi$ *and* $\|\mathbf{M}_k - \mathbf{Q}\mathbf{X}_{opt}\|_\xi/\|\mathbf{M}_k\|_\xi$ *for* $\xi = 2, \mathrm{F}$) *as a function of* k *for the three datasets* $\mathbf{M} = \mathbf{A}, \mathbf{B}, \mathbf{C}$. *Each point is the worst of the errors observed over* 10 *trials.* $r = \lceil 2k \ln n \rceil$ *column samples were used in each trial.*

in the Frobenius norm as $\mathbf{M}_k$, as Theorem 2.1 suggests should be the case. More precisely, Theorem 2.1 assures us that $528\varepsilon^{-1}[\sqrt{k}+\sqrt{8\ln(8n/\delta)}]^2\ln(8k/\delta)$ column samples are sufficient to ensure that, with at least probability $1-\delta$, $\mathbf{Y}\mathbf{Y}^\dagger\mathbf{M}$ and $\mathbf{Q}\mathbf{X}_{opt}$ have Frobenius norm residual and forward error within $(1+\varepsilon)$ of that of $\mathbf{M}_k$. The factor 528 can certainly be reduced by optimizing the numerical constants given in Theorem 2.1 (as noted after the statement of the theorem). But what is the smallest r that ensures the Frobenius norm residual error bounds $\|\mathbf{M}-\mathbf{Y}\mathbf{Y}^\dagger\mathbf{M}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ and $\|\mathbf{M}-\mathbf{Q}\mathbf{X}_{opt}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ are satisfied with some fixed probability? To investigate, in Figure 6.3 we plot the values of r determined empirically to be sufficient to obtain $(1+\varepsilon)$ Frobenius norm residual errors relative to the optimal rank-k approximation; we fix the failure probability $\delta = 1/2$ and vary ε. Specifically, the r plotted for each k is the smallest number of samples for which $\|\mathbf{M}-\mathbf{Y}\mathbf{Y}^\dagger\mathbf{M}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ (or $\|\mathbf{M}-\mathbf{Q}\mathbf{X}_{opt}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$) in at least 5 out of 10 trials.

It is clear that, for fixed k and ε, the number of samples r required to form a non-rank-restricted approximation to $\mathbf{M}$ with $(1+\varepsilon)$ relative residual error is smaller than the r required to form a rank-restricted approximation with $(1+\varepsilon)$ relative residual error. Note that for small values of k, the r necessary for the 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 $\|\mathbf{M}-\mathbf{M}_{k_2}\|_{\mathrm{F}}/\|\mathbf{M}-\mathbf{M}_{k_1}\|_{\mathrm{F}}$ 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 the observation of Woolfe et al. that taking $r = k+8$ suffices to consistently form accurate low-rank approximations using the SRFT scheme, which is very similar to the SRHT scheme [43]. We also note that this matches Theorem 2.1, which predicts that the necessary r grows at most linearly with k with a slope like $\ln n$.

Finally, Theorem 2.1 does *not* guarantee that $(1+\varepsilon)$ spectral norm relative residual errors can be achieved. Instead, it provides bounds on the spectral norm residual errors achieved in terms of $\|\mathbf{M}-\mathbf{M}_k\|_2$ and $\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ that are guaranteed to hold when r is sufficiently large. In Figure 6.4 we compare the spectral norm residual error guarantees of Theorem 2.1 to what is achieved in practice. To do so, we take the optimistic viewpoint that the constants in Theorem 2.1 can be optimized to unity. Under this view, if more columns than

$$r_2 = \varepsilon^{-1}[\sqrt{k}+\sqrt{\ln(n/\delta)}]^2\ln(k/\delta)$$

are used to construct the SRHT approximations, then the spectral norm residual error is no larger than

$$b_2 = \left(1+\sqrt{\frac{\ln(n/\delta)\ln(\rho/\delta)}{r}}\right)\cdot\|\mathbf{M}-\mathbf{M}_k\|_2 + \sqrt{\frac{\ln(\rho/\delta)}{r}}\cdot\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}},$$

where ρ is the rank of $\mathbf{M}$, with probability greater than $1-\delta$. Our comparison consists of using r_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 6.4 shows, for several values of k, the upper bound b_2 and the observed relative spectral norm residual errors, with precision parameter $\varepsilon = 1/2$ and failure parameter $\delta = 1/2$. For each value of k, the empirical spectral norm residual error plotted is the largest of the errors from among 10 trials of low-rank approximations. Note from Figure 6.4 that with this choice of r, the

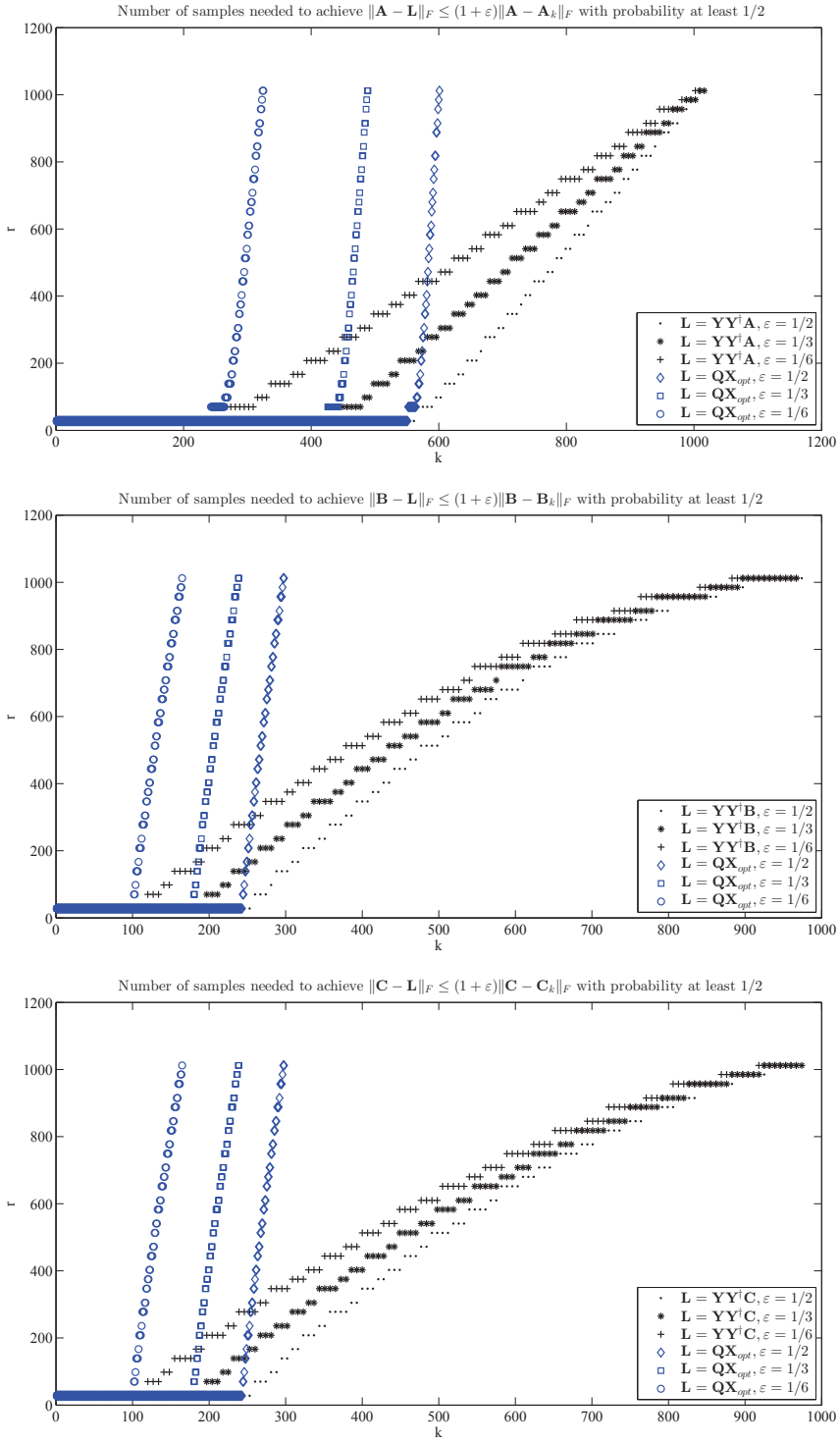

FIG. 6.3. *The value of r empirically necessary to ensure that, with probability at least 1/2, approximations generated by the SRHT algorithms satisfy $\|\mathbf{M}-\mathbf{YY}^\dagger\mathbf{M}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ and $\|\mathbf{M}-\mathbf{QX}_{opt}\|_{\mathrm{F}} \le (1+\varepsilon)\|\mathbf{M}-\mathbf{M}_k\|_{\mathrm{F}}$ (for $\mathbf{M}=\mathbf{A},\mathbf{B},\mathbf{C}$).*

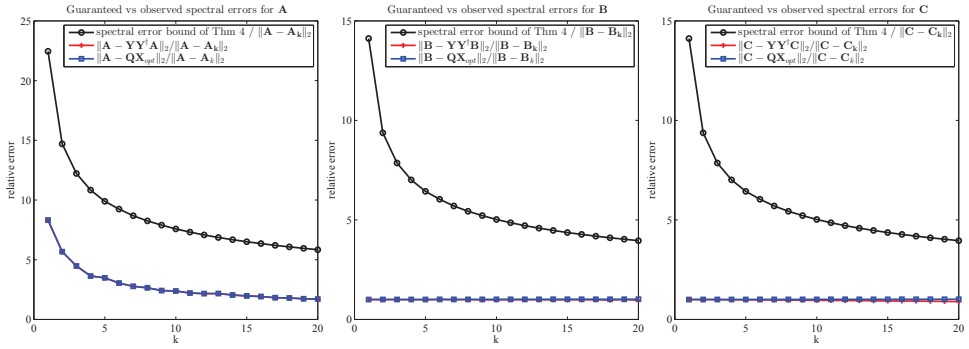

FIG. 6.4. *The empirical spectral norm residual errors relative to those of the optimal rank-k approximants* ($\|\mathbf{M} - \mathbf{YY}^\dagger\mathbf{M}\|_2/\|\mathbf{M} - \mathbf{M}_k\|_2$ *and* $\|\mathbf{M} - \mathbf{QX}_{opt}\|_2/\|\mathbf{M} - \mathbf{M}_k\|_2$*) plotted alongside the same ratio for the bound given in Theorem* 2.1*, when* $r = \lceil 2[\sqrt{k} + \sqrt{\ln(2n)}]^2 \ln(2k)\rceil$ *(for* $\mathbf{M} = \mathbf{A}, \mathbf{B}, \mathbf{C}$*).*

spectral norm residual errors of the rank-restricted and non-rank-restricted SRHT approximations are essentially the same.

Judging from Figures 6.3 and 6.4, even when we assume the constants present can be optimized away, the bounds given in Theorem 2.1 are pessimistic: it seems that in fact approximations with Frobenius norm residual error within $(1+\varepsilon)$ of the error of the optimal rank-k approximation can be obtained with r 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 2.1, Theorem 2.1—especially the spectral norm bounds—represents a significant improvement over prior efforts.

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 [21]. Regardless, we do not advocate using these approximation schemes for applications in which highly accurate low-rank approximations are needed. Rather, Theorem 2.1 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.

Acknowledgments. We would like to thank Joel Tropp and Mark Tygert for the initial suggestion that we attempt to sharpen the analysis of the SHRT low-rank approximation algorithm and for fruitful conversations on our approach. We are also grateful to an anonymous reviewer for pointing out the value in interpreting Lemma 4.11 as a relative error bound, and to Malik Magdon-Ismail for providing the proof of Lemma 5.3.

REFERENCES

[1] N. AILON AND B. CHAZELLE, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*, in Proceedings of the ACM Symposium on Theory of Computing (STOC), 2006.

[2] N. AILON AND E. LIBERTY, *Fast dimension reduction using Rademacher series on dual BCH codes*, in Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2008, pp. 1–9.

[3] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging LAPACK's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.
[4] C. Boutsidis, *Topics in Matrix Sampling Algorithms*, Ph.D. thesis, Rensselaer Polytechnic Institute, 2011.
[5] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near optimal column based matrix reconstruction*, in Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS), 2011.
[6] C. Boutsidis and P. Drineas, *Random projections for the nonnegative least-squares problem*, Linear Algebra Appl., 431 (2009), pp. 760–771.
[7] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Rich Coresets for Unconstrained Linear Regression*, preprint, arXiv:1202.3505, 2012.
[8] C. Boutsidis, A. Zouzias, and P. Drineas, *Random projections for k-means clustering*, in Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2010.
[9] C. Boutsidis, M. W. Mahoney, and P. Drineas, *An improved approximation algorithm for the column subset selection problem*, in Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2009, pp. 968–977.
[10] E. Candés and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.
[11] K. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Men, and D. Woodruff, *The fast Cauchy transform and faster robust linear regression*, in Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2013, pp. 466–477.
[12] K. Clarkson and D. Woodruff, *Numerical linear algebra in the streaming model*, in Proceedings of the ACM Symposium on Theory of Computing (STOC), 2009.
[13] K. Clarkson and D. Woodruff, *Low Rank Approximation and Regression in Input Sparsity Time*, preprint, arXiv:1207.6365, 2012.
[14] P. Drineas, *Randomized Algorithms for Matrix Operations*, Ph.D. thesis, Yale University, 2002.
[15] P. Drineas and R. Kannan, *Fast Monte-Carlo algorithms for approximate matrix multiplication*, in Proceedings of the Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2001.
[16] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo algorithms for matrices* I: *Approximating matrix multiplication*, SIAM J. Comput., 36 (2006), pp. 132–157.
[17] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo algorithms for matrices* II: *Computing a low-rank approximation to a matrix*, SIAM J. Comput., 36 (2006), pp. 158–183.
[18] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo algorithms for matrices* III: *Computing a compressed approximate matrix decomposition*, SIAM J. Comput., 36 (2006), pp. 184–206.
[19] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 217–249.
[20] A. Frieze, R. Kannan, and S. Vempala, *Fast Monte-Carlo algorithms for finding low-rank approximations*, in Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS), 1998.
[21] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.
[22] A. Gittens, *The Spectral Norm Error of the Naive Nystrom Extension*, preprint, arXiv: 1110.5305, 2011.
[23] A. Gittens and J. Tropp, *Tail Bounds for All Eigenvalues of a Sum of Random Matrices*, preprint, arXiv:1104.4513, 2011.
[24] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory, 57 (2011), pp. 1548–1566.
[25] D. Gross and V. Nesme, *Note on Sampling without Replacing from a Finite Collection of Matrices*, preprint, arXiv:1001.2738, 2010.
[26] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
[27] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.
[28] I. Ipsen and T. Wentworth, *The Effect of Coherence on Sampling from Matrices with Orthonormal Columns, and Preconditioned Least Squares Problems*, preprint, arXiv: 1203.4809, 2012.
[29] M. Ledoux, *On Talagrand's deviation inequalities for product measures*, ESAIM Probab. Statist., 1 (1996), pp. 63–87.

[30] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert, *Randomized algorithms for the low-rank approximation of matrices*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 20167–20172.

[31] M. Mahoney, *Algorithmic and Statistical Perspectives on Large-Scale Data Analysis*, Chapman & Hall/CRC Comput. Sci. Ser., CRC Press, Boca Raton, FL, 2011.

[32] M. Mahoney, L. Lim, and G. Carlsson, *Algorithmic and statistical challenges in modern largescale data analysis are the focus of MMDS* 2008, ACM SIGKDD Explorations Newsletter, 10 (2008), pp. 57–60.

[33] A. Magen and A. Zouzias, *Low rank matrix-valued Chernoff bounds and approximate matrix multiplication*, in Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2011, pp. 1422–1436.

[34] P.-G. Martinsson, V. Rokhlin, and M. Tygert, *A randomized algorithm for the decomposition of matrices*, Appl. Comput. Harmonic Anal., 30 (2010), pp. 47–68.

[35] N. H. Nguyen, T. T. Do, and T. D. Tran, *A fast and efficient algorithm for low-rank approximation of a matrix*, in Proceedings of the ACM Symposium on Theory of Computing (STOC), 2009.

[36] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.

[37] V. Rokhlin, A. Szlam, and M. Tygert, *A randomized algorithm for principal component analysis*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1100–1124.

[38] V. Rokhlin and M. Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 13212–13217.

[39] A. Talwalkar and A. Rostamizadeh, *Matrix coherence and the Nyström method*, in Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), 2010.

[40] T. Sarlós, *Improved approximation algorithms for large matrices via random projections*, in Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS), 2006.

[41] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adaptive Data Anal., 3 (2011), pp. 115–126.

[42] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.

[43] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices, preliminary report*, Appl. Comput. Harmonic Anal., 25 (2008), pp. 335–366.