SIAM J. Matrix Anal. Appl.
Vol. 36, No. 1, pp. 110–137

RANDOMIZED APPROXIMATION OF THE GRAM MATRIX: EXACT COMPUTATION AND PROBABILISTIC BOUNDS[*]

JOHN T. HOLODNAK[†] AND ILSE C. F. IPSEN[‡]

Abstract. Given a real matrix $\mathbf{A}$ with n columns, the problem is to approximate the Gram product $\mathbf{A}\mathbf{A}^T$ by $c \ll n$ weighted outer products of columns of $\mathbf{A}$. Necessary and sufficient conditions for the exact computation of $\mathbf{A}\mathbf{A}^T$ (in exact arithmetic) from $c \geq \operatorname{rank}(\mathbf{A})$ columns depend on the right singular vector matrix of $\mathbf{A}$. For a Monte Carlo matrix multiplication algorithm by Drineas et al. that samples outer products, we present probabilistic bounds for the two-norm relative error due to randomization. The bounds depend on the stable rank or the rank of $\mathbf{A}$, but not on the matrix dimensions. Numerical experiments illustrate that the bounds are informative, even for stringent success probabilities and matrices of small dimension. We also derive bounds for the smallest singular value and the condition number of matrices obtained by sampling rows from orthonormal matrices.

Key words. leverage scores, singular value decomposition, stable rank, coherence, matrix concentration inequalities, unbiased estimator

AMS subject classifications. 68W20, 65C05, 15A18, 65F20, 65F35

DOI. 10.1137/130940116

1. Introduction.

Given a real matrix $\mathbf{A} = \begin{pmatrix} A_1 & \dots & A_n \end{pmatrix}$ with n columns A_j, can one approximate the Gram matrix $\mathbf{A}\mathbf{A}^T$ from just a *few* columns? We answer this question by presenting deterministic conditions for the exact[1] computation of $\mathbf{A}\mathbf{A}^T$ from a few columns, and probabilistic error bounds for approximations.

Our motivation (section 1.1) is followed by an overview of the results (section 1.2), and a literature survey (section 1.3). Those not familiar with established notation can find a review in section 1.4.

1.1. Motivation.

The objective is the analysis of a randomized algorithm for approximating $\mathbf{A}\mathbf{A}^T$. Specifically, it is a Monte Carlo algorithm for sampling outer products and represents a special case of the groundbreaking work on randomized matrix multiplication by Drineas and Kannan [16] and Drineas, Kannan, and Mahoney [17].

The basic idea is to represent $\mathbf{A}\mathbf{A}^T$ as a sum of outer products of columns,

$$\mathbf{A}\mathbf{A}^T = A_1 A_1^T + \cdots + A_n A_n^T.$$

The Monte Carlo algorithm [16, 17], when provided with a user-specified positive integer c, samples c columns $A_{t_1}, \dots, A_{t_c}$ according to probabilities p_j, $1 \leq j \leq n$,

[*] Received by the editors October 7, 2013; accepted for publication (in revised form) by J. Tropp August 13, 2014; published electronically February 5, 2015. The research of the authors was partially supported by the National Science Foundation under grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute.

http://www.siam.org/journals/simax/36-1/94011.html

[†] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (jtholodn@ncsu.edu, http://www4.ncsu.edu/∼jtholodn/). The research of this author was supported in part by Department of Education grant P200A090081.

[‡] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (ipsen@ncsu.edu, http://www4.ncsu.edu/∼ipsen/). The research of this author was supported in part by NSF grant CCF-1145383, and also the XDATA Program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323 FA8750-12-C-0323.

[1] We assume infinite precision, and no round-off errors.

and then approximates $\mathbf{AA}^T$ by a weighted sum of c outer products

$$\mathbf{X} = w_1 A_{t_1} A_{t_1}^T + \cdots + w_c A_{t_c} A_{t_c}^T.$$

The weights are set to $w_j = 1/(cp_{t_j})$ so that $\mathbf{X}$ is an unbiased estimator, $\mathbb{E}[\mathbf{X}] = \mathbf{AA}^T$. Intuitively, one would expect the algorithm to do well for matrices of low rank.

The intuition is based on the singular value decomposition. Given left singular vectors U_j associated with the $k \equiv \text{rank}(\mathbf{A})$ nonzero singular values σ_j of $\mathbf{A}$, one can represent $\mathbf{AA}^T$ as a sum of k outer products,

$$\mathbf{AA}^T = \sigma_1^2 \, U_1 U_1^T + \cdots + \sigma_k^2 \, U_k U_k^T.$$

Hence for matrices $\mathbf{A}$ of low rank, a few left singular vectors and singular values suffice to reproduce $\mathbf{AA}^T$ exactly. Thus, if $\mathbf{A}$ has columns that "resemble" its left singular vectors, the Monte Carlo algorithm should have a chance to perform well.

1.2. Contributions and overview.

We sketch the main contributions of this paper. All proofs are relegated to section 7.

1.2.1. Deterministic conditions for exact computation (section 2).

To calibrate the potential of the Monte Carlo algorithm [16, 17] and establish connections to existing work in linear algebra, we first derive deterministic conditions that characterize when $\mathbf{AA}^T$ can be computed *exactly* from a few columns of $\mathbf{A}$. Specifically:

- We present necessary and sufficient conditions (Theorem 2.2) for computing $\mathbf{AA}^T$ exactly from $c \geq \text{rank}(\mathbf{A})$ columns $A_{t_1}, \ldots, A_{t_c}$ of $\mathbf{A}$,

$$\mathbf{AA}^T = w_1 \, A_{t_1} A_{t_1}^T + \cdots + w_c \, A_{t_c} A_{t_c}^T.$$

  The conditions and weights w_j depend on the right singular vector matrix $\mathbf{V}$ associated with the nonzero singular values of $\mathbf{A}$.
- For matrices with $\text{rank}(\mathbf{A}) = 1$, this is always possible (Corollary 2.4).
- In the special case where $c = \text{rank}(\mathbf{A})$ (Theorem 2.7), the weights are equal to inverse leverage scores, $w_j = 1/\|\mathbf{V}^T e_{t_j}\|_2^2$. However, they do not necessarily correspond to the largest leverage scores.

1.2.2. Sampling probabilities for the Monte Carlo algorithm (section 3).

Given an approximation $\mathbf{X}$ from the Monte Carlo algorithm [16, 17], we are interested in the two-norm relative error due to randomization, $\|\mathbf{X} - \mathbf{AA}^T\|_2/\|\mathbf{AA}^T\|_2$. Numerical experiments compare two types of sampling probabilities:

- "optimal" probabilities $p_j^{opt} = \|A_j\|_2^2/\|\mathbf{A}\|_F^2$ [17], and
- leverage score probabilities $p_j^{lev} = \|\mathbf{V}^T e_j\|_2^2/k$ [7, 9].

The experiments illustrate that sampling columns of $\mathbf{X}$ with the "optimal" probabilities produces a smaller error than sampling with leverage score probabilities. This was not obvious a priori, because the "optimal" probabilities are designed to minimize the expected value of the Frobenius norm absolute error, $\mathbb{E}[\|\mathbf{X} - \mathbf{AA}^T\|_F^2]$. Furthermore, corresponding probabilites p_j^{opt} and p_j^{lev} can differ by orders of magnitude.

For matrices $\mathbf{A}$ of rank one though, we show (Theorem 3.1) that the probabilities are identical, $p_j^{opt} = p_j^{lev}$ for $1 \leq j \leq n$, and that the Monte Carlo algorithm always produces the exact result, $\mathbf{X} = \mathbf{AA}^T$, when it samples with these probabilities.

1.2.3. Probabilistic bounds (sections 4 and 5).

We present probabilistic bounds for $\|\mathbf{X} - \mathbf{AA}^T\|_2/\|\mathbf{AA}^T\|_2$ when the Monte Carlo algorithm samples with two types of sampling probabilities.

- Sampling with "nearly optimal" probabilities $p_j^{\beta} \geq \beta\, p_j^{opt}$, where $\beta \leq 1$ (Theorems 4.1 and 4.2). We show that
$$\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_2 / \|\mathbf{A}\mathbf{A}^T\|_2 \leq \epsilon \quad \text{with probability at least } 1-\delta,$$
provided the number of sampled columns is at least
$$c \geq c_0(\epsilon)\, \frac{\ln(\rho(\mathbf{A})/\delta)}{\beta\epsilon^2} \mathsf{sr}(\mathbf{A}), \qquad \text{where} \quad 2 \leq c_0(\epsilon) \leq 2.7.$$
Here $\rho(\mathbf{A}) = \operatorname{rank}(\mathbf{A})$ or $\rho(\mathbf{A}) = 4\,\mathsf{sr}(\mathbf{A})$, where $\mathsf{sr}(\mathbf{A})$ is the stable rank of $\mathbf{A}$. The bound containing $\operatorname{rank}(\mathbf{A})$ is tighter for matrices with $\operatorname{rank}(\mathbf{A}) \leq 4\,\mathsf{sr}(\mathbf{A})$. Note that the amount of sampling depends on the rank or the stable rank, but not on the dimensions of $\mathbf{A}$. Numerical experiments (section 4.4) illustrate that the bounds are informative, even for stringent success probabilities and matrices of small dimension.
- Sampling with leverage score probabilities p_j^{lev} (Theorem 5.1). The bound corroborates the numerical experiments in section 3.2.3, but is not as tight as the bounds for "nearly optimal" probabilities, since it depends only on $\operatorname{rank}(\mathbf{A})$, and $\operatorname{rank}(\mathbf{A}) \geq \mathsf{sr}(\mathbf{A})$.

1.2.4. Singular value bounds (section 6). Given an $m \times n$ matrix $\mathbf{Q}$ with orthonormal rows, $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_m$, the Monte Carlo algorithm computes $\mathbf{QS}$ by sampling $c \geq m$ columns from $\mathbf{Q}$ with the "optimal" probabilities. The goal is to derive a positive lower bound for the smallest singular value $\sigma_m(\mathbf{QS})$, as well as an upper bound for the two-norm condition number with respect to left inversion $\kappa(\mathbf{QS}) \equiv \sigma_1(\mathbf{QS})/\sigma_m(\mathbf{QS})$.

Surprisingly, Theorem 4.1 leads to bounds (Theorems 6.1 and 6.3) that are not always as tight as the ones below. These bounds are based on a Chernoff inequality and represent a slight improvement over existing results.

- Bound for the smallest singular value (Theorem 6.2). We show that
$$\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon} \quad \text{with probability at least } 1-\delta,$$
provided the number of sampled columns is at least
$$c \geq c_1(\epsilon)\, m\, \frac{\ln(m/\delta)}{\epsilon^2}, \qquad \text{where} \quad 1 \leq c_1(\epsilon) \leq 2.$$
- Condition number bound (Theorem 6.4). We show that
$$\kappa(\mathbf{QS}) \leq \frac{\sqrt{1+\epsilon}}{\sqrt{1-\epsilon}} \quad \text{with probability at least } 1-\delta,$$
provided the number of sampled columns is at least
$$c \geq c_2(\epsilon)\, m\, \frac{\ln(2m/\delta)}{\epsilon^2}, \qquad \text{where} \quad 2 \leq c_2(\epsilon) \leq 2.6.$$

In addition, we derive corresponding bounds for uniform sampling with and without replacement (Theorems 6.2 and 6.4).

1.3. Literature review. We review bounds for the relative error due to randomization of general Gram matrix approximations $\mathbf{A}\mathbf{A}^T$, and also for the smallest singular value and condition number of sampled matrices $\mathbf{QS}$ when $\mathbf{Q}$ has orthonormal rows.

In addition to [16, 17], several other randomized matrix multiplication algorithms have been proposed [5, 13, 14, 40, 46, 48]. Sarlós's algorithms [48] are based on matrix transformations. Cohen and Lewis [13, 14] approximate large elements of a matrix product with a random walk algorithm. The algorithm by Belabbas and Wolfe [5] is related to the Monte Carlo algorithm [16, 17], but with different sampling methods and weights. A second algorithm by Drineas, Kannan, and Mahoney [17] relies on matrix sparsification, and a third algorithm [16] estimates each matrix element independently. Pagh [46] targets sparse matrices, while Liberty [40] estimates the Gram matrix $\mathbf{AA}^T$ by iteratively removing "unimportant" columns from $\mathbf{A}$.

Eriksson-Bique et al. [22] derive an importance sampling strategy that minimizes the variance of the inner products computed by the Monte Carlo method. Madrid, Guerra, and Rojas [41] present experimental comparisons of different sampling strategies for specific classes of matrices.

Excellent surveys of randomized matrix algorithms in general are given by Halko, Martinsson, and Tropp [32], and by Mahoney [45].

1.3.1. Gram matrix approximations. We review existing bounds for the error due to randomization of the Monte Carlo algorithm [16, 17] for approximating $\mathbf{AA}^T$, where $\mathbf{A}$ is a real $m \times n$ matrix. Relative error bounds $\|\mathbf{X} - \mathbf{AA}^T\|/\|\mathbf{AA}^T\|$ in the Frobenius norm and the two-norm are summarized in Tables 1 and 2.

Table 1 shows probabilistic lower bounds for the number of sampled columns so that the Frobenius norm relative error $\left\|\mathbf{X} - \mathbf{AA}^T\right\|_F / \left\|\mathbf{AA}^T\right\|_F \leq \epsilon$. Not listed is a bound for uniform sampling without replacement [38, Corollary 1], because it cannot easily be converted to the format of the other bounds, and a bound for a greedy sampling strategy [5, p. 5].

Table 2 shows probabilistic lower bounds for the number of sampled columns so that the two-norm relative error $\left\|\mathbf{X} - \mathbf{AA}^T\right\|_2 / \left\|\mathbf{AA}^T\right\|_2 \leq \epsilon$. These bounds imply, roughly, that the number of sampled columns should be at least $\Omega(\mathsf{sr}(\mathbf{A})\ \ln(\mathsf{sr}(\mathbf{A}))$ or $\Omega(\mathsf{sr}(\mathbf{A})\ \ln(m))$.

1.3.2. Singular value bounds. We review existing bounds for the smallest singular value of a sampled matrix $\mathbf{QS}$, where $\mathbf{Q}$ is $m \times n$ with orthonormal rows.

Table 1

Frobenius-norm error due to randomization: Lower bounds on the number c of sampled columns in $\mathbf{X}$, so that $\|\mathbf{X} - \mathbf{AA}^T\|_F/\|\mathbf{AA}^T\|_F \leq \epsilon$ with probability at least $1-\delta$. The second column specifies the sampling strategy: "opt" for sampling with "optimal" probabilities, and "u-wor" for uniform sampling without replacement. The last two bounds are special cases of bounds for general matrix products $\mathbf{AB}$.

Bound for # samples	Sampling	Reference
$\frac{(1+\sqrt{8\ln(1/\delta)})^2}{\epsilon^2}\frac{\|A\|_F^4}{\left\|AA^T\right\|_F^2}$	opt	[17, Theorem 2]
$\frac{1}{\epsilon^2\delta}\frac{\|A\|_F^4}{\left\|AA^T\right\|_F^2}$	opt	[24, Lemma 1], [25, Lemma 2]
$\frac{n^2}{(n-1)\delta\epsilon^2}\frac{\sum_{j=1}^n \left\|A_j\right\|_2^4}{\left\|AA^T\right\|_F^2}$	u-wor	[16, Lemma 7]
$\frac{36n\ln(1/\delta)}{\epsilon^2}\frac{\sum_{j=1}^n \|A_i\|_2^4}{\left\|AA^T\right\|_F^2}$	u-wor	[8, Lemma 4.13], [27, Lemma 4.3]

Table 2

Two-norm error due to randomization, for sampling with "optimal" probabilities: Lower bounds on the number c of sampled columns in $\mathbf{X}$, so that $\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\|_2/\|\mathbf{A}\mathbf{A}^T\|_2 \leq \epsilon$ with probability at least $1-\delta$ for all bounds but the first. The first bound contains an unspecified constant C and holds with probability at least $1-2\exp(\tilde{C}/\delta)$, where $\tilde{C}$ is another unspecified constant (our ϵ corresponds to $\epsilon^2/2$ in [47, *Theorem 1.1*]*). The penultimate bound is a special case of a bound for general matrix products $\mathbf{AB}$, while the last bound applies only to matrices with orthonormal rows.*

Bound for # samples	Reference
$C\frac{\mathsf{sr}(A)}{\epsilon^2\delta}\ln(\mathsf{sr}(A)/(\epsilon^2\delta))$	[47, Theorems 1.1 and 3.1, and their proofs]
$\frac{4\mathsf{sr}(A)}{\epsilon^2}\ln(2m/\delta)$	[43, Theorem 17], [42, Theorem 20]
$\frac{96\mathsf{sr}(A)}{\epsilon^2}\ln\left(\frac{96\mathsf{sr}(A)}{\epsilon^2\sqrt{\delta}}\right)$	[21, Theorem 4]
$\frac{20\mathsf{sr}(A)}{\epsilon^2}\ln(16\mathsf{sr}(A)/\delta)$	[44, Theorem 3.1], [55, Theorem 2.1]
$\frac{21(1+\mathsf{sr}(A))}{4\epsilon^2}\ln(4\mathsf{sr}(A)/\delta)$	[36, Example 4.3]
$\frac{8m}{\epsilon^2}\ln(m/\delta)$	[49, Theorem 3.9]

Table 3

Smallest singular value of a matrix $\mathbf{QS}$ whose columns are sampled from an $m\times n$ matrix $\mathbf{Q}$ with orthonormal rows: Lower bounds on the number c of sampled columns, so that $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ with probability at least $1-\delta$. The second column specifies the sampling strategy: "opt" for sampling with "optimal" probabilities, "u-wr" for uniform sampling with replacement, and "u-wor" for uniform sampling without replacement.

Bound for # samples	Sampling	Reference
$\frac{6n\mu}{\epsilon^2}\ln(m/\delta)$	u-wor	[8, Lemma 4.3]
$\frac{4m}{\epsilon^2}\ln(2m/\delta)$	opt	[6, Lemma 13]
$\frac{3n\mu}{\epsilon^2}\ln(m/\delta)$	u-wr, u-wor	[37, Corollary 4.2]
$\frac{8n\mu}{3\epsilon^2}\ln(m/\delta)$	u-wr	[8, Lemma 4.4]
$\frac{2n\mu}{\epsilon^2}\ln(m/\delta)$	u-wor	[26, Lemma 1]

Table 3 shows probabilistic lower bounds for the number of sampled columns so that the smallest singular value $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$. All bounds but one contain the coherence μ. Not listed is a bound [21, Lemma 4] that requires specific choices of ϵ, δ, and μ.

1.3.3. Condition number bounds. We are aware of only two existing bounds for the two-norm condition number $\kappa(\mathbf{QS})$ of a matrix $\mathbf{QS}$ whose columns are sampled from an $m\times n$ matrix $\mathbf{Q}$ with orthonormal rows. The first bound [1, Theorem 3.2] lacks explicit constants, while the second one [37, Corollary 4.2] applies to uniform sampling with and without replacement. It ensures $\kappa(\mathbf{QS}) \leq \frac{\sqrt{1+\epsilon}}{\sqrt{1-\epsilon}}$ with probability at least $1-\delta$, provided the number of sampled columns in $\mathbf{QS}$ is at least $c \geq 3n\mu \ln(2m/\delta)/\epsilon^2$.

1.3.4. Relation to subset selection. The Monte Carlo algorithm selects outer products from $\mathbf{AA}^T$, which is equivalent to selecting columns from $\mathbf{A}$, hence it can be viewed as a form of randomized column subset selection.

The traditional deterministic subset selection methods select exactly the required number of columns, by means of rank-revealing QR decompositions or SVDs [11, 28, 29, 31, 34]. In contrast, more recent methods are motivated by applications to graph sparsification [3, 4, 49]. They oversample columns from a matrix $\mathbf{Q}$ with orthonormal rows, by relying on a *barrier sampling* strategy.[2] The accuracy of the selected columns $\mathbf{QS}$ is determined by bounding the *reconstruction error*, which views $(\mathbf{QS})\,(\mathbf{QS})^T$ as an approximation to $\mathbf{QQ}^T = I$ [3, Theorem 3.1], [4, Theorem 3.1], [49, Theorem 3.2].

Boutsidis [6] extends this work to general Gram matrices $\mathbf{AA}^T$. Following [29], he selects columns from the right singular vector matrix $\mathbf{V}^T$ of $\mathbf{A}$, and applies barrier sampling simultaneously to the dominant and subdominant subspaces of $\mathbf{V}^T$.

In terms of randomized algorithms for subset selection, the two-stage algorithm by Boutsidis, Mahoney, and Drineas [9] samples columns in the first stage, and performs a deterministic subset selection on the sampled columns in the second stage. Other approaches include volume sampling [24, 25], and CUR decompositions [20].

1.3.5. Leverage scores. In the late seventies, statisticians introduced leverage scores for outlier detection in regression problems [12, 33, 53]. More recently, Drineas, Mahoney et al. have pioneered the use of leverage scores for importance sampling in randomized algorithms, such as CUR decompositions [20], least squares problems [19], and column subset selection [9], see also the perspectives on statistical leverage [45, section 6]. Fast approximation algorithms are being designed to make the computation of leverage scores more affordable [18, 39, 42].

1.4. Notation. All matrices are real. Matrices that can have more than one column are indicated in bold face, and column vectors and scalars in italics. The columns of the $m \times n$ matrix $\mathbf{A}$ are denoted by $\mathbf{A} = \begin{pmatrix} A_1 & \cdots & A_n \end{pmatrix}$. The $n \times n$ identity matrix is $\mathbf{I}_n \equiv \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix}$, whose columns are the canonical vectors e_j.

The thin singular value decomposition (SVD) of an $m\times n$ matrix $\mathbf{A}$ with $\operatorname{rank}(\mathbf{A}) = k$ is $\mathbf{A} = \mathbf{U\Sigma V}^T$, where the $m \times k$ matrix $\mathbf{U}$ and the $n \times k$ matrix $\mathbf{V}$ have orthonormal columns, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k = \mathbf{V}^T\mathbf{V}$, and the $k \times k$ diagonal matrix of singular values is $\mathbf{\Sigma} = \operatorname{diag}\begin{pmatrix} \sigma_1 & \ldots & \sigma_k \end{pmatrix}$, with $\sigma_1 \geq \cdots \geq \sigma_k > 0$. The Moore–Penrose inverse of $\mathbf{A}$ is $\mathbf{A}^\dagger \equiv \mathbf{V\Sigma}^{-1}\mathbf{U}^T$. The unique symmetric positive semidefinite square root of a symmetric positive semidefinite matrix $\mathbf{W}$ is denoted by $\mathbf{W}^{1/2}$.

The norms in this paper are the two-norm $\|\mathbf{A}\|_2 \equiv \sigma_1$ and the Frobenius norm

$$\|\mathbf{A}\|_F \equiv \sqrt{\sum_{j=1}^{n} \|A_j\|_2^2} = \sqrt{\sigma_1^2 + \cdots + \sigma_k^2}.$$

The *stable rank* of a nonzero matrix $\mathbf{A}$ is $\mathsf{sr}(\mathbf{A}) \equiv \|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2$, where $1 \leq \mathsf{sr}(\mathbf{A}) \leq \operatorname{rank}(\mathbf{A})$.

[2] The name comes about as follows: Adding a column q to $\mathbf{QS}$ amounts to a rank-one update qq^T for the Gram matrix $(\mathbf{QS})\,(\mathbf{QS})^T$. The eigenvalues of this matrix, due to interlacing, form "barriers" for the eigenvalues of the updated matrix $(\mathbf{QS})\,(\mathbf{QS})^T + qq^T$.

Given an $m \times n$ matrix $\mathbf{Q} = \begin{pmatrix} Q_1 & \cdots & Q_n \end{pmatrix}$ with orthonormal rows, $\mathbf{QQ}^T = \mathbf{I}_m$, the *two-norm condition number* with regard to left inversion is $\kappa(\mathbf{Q}) \equiv \sigma_1(\mathbf{Q})/\sigma_m(\mathbf{Q})$; the *leverage scores* [19, 20, 45] are the squared columns norms $\|Q_j\|_2^2$, $1 \leq j \leq m$; and the *coherence* [1, 10] is the largest leverage score,

$$\mu \equiv \max_{1 \leq j \leq m} \|Q_j\|_2^2.$$

The expected value of a scalar or a matrix-valued random random variable $\mathbf{X}$ is $\mathbb{E}[\mathbf{X}]$; and the probability of an event $\mathcal{X}$ is $\mathbb{P}[\mathcal{X}]$.

2. Deterministic conditions for exact computation.

To gauge the potential of the Monte Carlo algorithm, and to establish a connection to existing work in linear algebra, we first consider the best case: The *exact* computation of $\mathbf{AA}^T$ from a few columns. That is, given c not necessarily distinct columns $A_{t_1}, \ldots, A_{t_c}$, under which conditions is $w_1 A_{t_1} A_{t_1}^T + \cdots + w_c A_{t_c} A_{t_c}^T = \mathbf{AA}^T$?

Since a column can be selected more than once, and therefore the selected columns may not form a submatrix of $\mathbf{A}$, we express the c selected columns as $\mathbf{AS}$, where $\mathbf{S}$ is an $n \times c$ sampling matrix with

$$\mathbf{S} = \begin{pmatrix} e_{t_1} & \ldots & e_{t_c} \end{pmatrix}, \qquad 1 \leq t_1 \leq \cdots \leq t_c \leq n.$$

Then one can write

$$w_1 A_{t_1} A_{t_1}^T + \cdots + w_c A_{t_c} A_{t_c}^T = (\mathbf{AS})\mathbf{W}(\mathbf{AS})^T,$$

where $\mathbf{W} = \operatorname{diag}\begin{pmatrix} w_1 & \cdots & w_c \end{pmatrix}$ is a diagonal weighting matrix. We answer two questions in this section:

1. Given a set of c columns $\mathbf{AS}$ of $\mathbf{A}$, when is $\mathbf{AA}^T = (\mathbf{AS})\,\mathbf{W}\,(\mathbf{AS})^T$ *without any constraints* on $\mathbf{W}$? The answer is an expression for a matrix $\mathbf{W}$ with minimal Frobenius norm (section 2.1).
2. Given a set of c columns $\mathbf{AS}$ of $\mathbf{A}$, what are necessary and sufficient conditions under which $(\mathbf{AS})\mathbf{W}(\mathbf{AS})^T = \mathbf{AA}^T$ for a *diagonal matrix* $\mathbf{W}$? The answer depends on the right singular vector matrix of $\mathbf{A}$ (section 2.2).

2.1. Optimal approximation (no constraints on W).

For a given set of c columns $\mathbf{AS}$ of $\mathbf{A}$, we determine a matrix $\mathbf{W}$ of minimal Frobenius norm that minimizes the absolute error of $(\mathbf{AS})\mathbf{W}(\mathbf{AS})^T$ in the Frobenius norm.

The following is a special case of [23, Theorem 2.1], without any constraints on the number of columns in $\mathbf{AS}$. The idea is to represent $\mathbf{AS}$ in terms of the thin SVD of $\mathbf{A}$ as $\mathbf{AS} = \mathbf{U\Sigma}(\mathbf{V}^T\mathbf{S})$.

THEOREM 2.1. *Given c columns $\mathbf{AS}$ of $\mathbf{A}$, not necessarily distinct, the unique solution of*

$$\min_{\mathbf{W}} \|\mathbf{AA}^T - (\mathbf{AS})\mathbf{W}(\mathbf{AS})^T\|_F$$

with minimal Frobenius norm is $\mathbf{W}_{opt} = (\mathbf{AS})^\dagger\,\mathbf{AA}^T\,((\mathbf{AS})^\dagger)^T$.

If, in addition, $\operatorname{rank}(\mathbf{AS}) = \operatorname{rank}(\mathbf{A})$*, then*

$$(\mathbf{AS})\mathbf{W}_{opt}(\mathbf{AS})^T = \mathbf{AA}^T \qquad \text{and} \qquad \mathbf{W}_{opt} = (\mathbf{V}^T\mathbf{S})^\dagger((\mathbf{V}^T\mathbf{S})^\dagger)^T.$$

If also $c = \operatorname{rank}(\mathbf{AS}) = \operatorname{rank}(\mathbf{A})$*, then*

$$(\mathbf{AS})\mathbf{W}_{opt}(\mathbf{AS})^T = \mathbf{AA}^T \qquad \text{and} \qquad \mathbf{W}_{opt} = (\mathbf{V}^T\mathbf{S})^{-1}(\mathbf{V}^T\mathbf{S})^{-T}.$$

Proof. See section 7.1. □

Theorem 2.1 implies that if $\mathbf{AS}$ has maximal rank, then the solution $\mathbf{W}_{opt}$ of minimal Frobenius norm depends only on the right singular vector matrix of $\mathbf{A}$ and in particular only on those columns $\mathbf{V}^T\mathbf{S}$ that correspond to the columns in $\mathbf{AS}$.

2.2. Exact computation with outer products (diagonal W). We present necessary and sufficient conditions under which $(\mathbf{AS})\mathbf{W}(\mathbf{AS})^T = \mathbf{AA}^T$ for a nonnegative diagonal matrix $\mathbf{W}$, that is $w_1 A_{t_1} A_{t_1}^T + \cdots + w_c A_{t_c} A_{t_c}^T = \mathbf{AA}^T$.

THEOREM 2.2. *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix, and let* $c \geq k \equiv \operatorname{rank}(\mathbf{A})$. *Then*

$$\sum_{j=1}^{c} w_j \, A_{t_j} A_{t_j}^T = \mathbf{AA}^T$$

for weights $w_j \geq 0$ *if and only if the* $c \times k$ *matrix* $\mathbf{V}^T \begin{pmatrix} \sqrt{w_1}\, e_{t_1} & \cdots & \sqrt{w_c}\, e_{t_c} \end{pmatrix}$ *has orthonormal rows.*

Proof. See section 7.2. □

REMARK 2.3 (comparison with barrier sampling method). *Our results differ from those in* [3, 4, 49] *in that we present conditions for* $\mathbf{A}$ *and the weights for exact computation of* $\mathbf{AA}^T$, *while* [3, 4, 49] *present an algorithm that can produce an arbitrarily good approximation for any matrix* $\mathbf{A}$.

If $\mathbf{A}$ has rank one, then any c nonzero columns of $\mathbf{A}$ will do for representing $\mathbf{AA}^T$, and explicit expressions for the weights can be derived.

COROLLARY 2.4. *If* $\operatorname{rank}(\mathbf{A}) = 1$ *then for any* c *columns* $A_{t_j} \neq 0$,

$$\sum_{j=1}^{c} w_j \, A_{t_j} A_{t_j}^T = \mathbf{AA}^T, \qquad \text{where} \quad w_j = \frac{1}{c \, \|\mathbf{V}^T e_{t_j}\|_2^2} = \frac{\|\mathbf{A}\|_F^2}{\|A_{t_j}\|_2^2}, \quad 1 \leq j \leq c.$$

Proof. See section 7.3. □

Hence, in the special case of rank-one matrices, the weights are inverse leverage scores of $\mathbf{V}^T$ as well as inverse normalized column norms of $\mathbf{A}$. Furthermore, in the special case $c = 1$, Corollary 2.4 implies that any nonzero column of $\mathbf{A}$ can be chosen. In particular, choosing the column A_l of largest norm yields a weight $w_1 = 1/\|\mathbf{V}^T e_l\|_2^2$ of minimal value, where $\|\mathbf{V}^T e_l\|_2^2$ is the coherence of $\mathbf{V}^T$.

In the following, we look at Theorem 2.2 in more detail, and distinguish the two cases when the number of selected columns is greater than $\operatorname{rank}(\mathbf{A})$, and when it is equal to $\operatorname{rank}(\mathbf{A})$.

2.2.1. Number of selected columns greater than rank(A). We illustrate the conditions of Theorem 2.2 when $c > \operatorname{rank}(\mathbf{A})$. In this case, indices do not necessarily have to be distinct, and a column can occur repeatedly.

EXAMPLE 2.5. *Let*

$$\mathbf{V}^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

so that $\operatorname{rank}(\mathbf{A}) = 2$. *Also let* $c = 3$, *and select the first column twice,* $t_1 = t_2 = 1$ *and* $t_3 = 2$, *so that*

$$\mathbf{V}^T \begin{pmatrix} e_1 & e_1 & e_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The weights $w_1 = w_2 = 1/2$ *and* $w_3 = 1$ *give a matrix*

$$\mathbf{V}^T \begin{pmatrix} 2^{-1/2} e_1 & 2^{-1/2} e_1 & e_2 \end{pmatrix} = \begin{pmatrix} 2^{-1/2} & 2^{-1/2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

with orthonormal rows. Thus, an exact representation does not require distinct indices.

However, although the above weights yield an exact representation, the corresponding weight matrix does not have minimal Frobenius norm.

Remark 2.6 (connection to Theorem 2.1). *If $c > k \equiv \operatorname{rank}(\mathbf{A})$ in Theorem 2.2, then no diagonal weight matrix $\mathbf{W} = \operatorname{diag}\begin{pmatrix} w_1 & \cdots & w_c \end{pmatrix}$ can be a minimal norm solution $\mathbf{W}_{opt}$ in Theorem 2.1.*

To see this, note that for $c > k$, the columns $A_{t_1}, \ldots, A_{t_c}$ are linearly dependent. Hence the $c \times c$ minimal Frobenius norm solution $\mathbf{W}_{opt}$ has rank equal to $k < c$. If $\mathbf{W}_{opt}$ were to be diagonal, it could have only k nonzero diagonal elements, hence the number of outer products would be $k < c$, a contradiction.

To illustrate this, let

$$\mathbf{V}^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

so that $\operatorname{rank}(\mathbf{A}) = 2$. *Also, let $c = 3$, and select columns $t_1 = 1$, $t_2 = 2$, and $t_3 = 3$, so that*

$$\mathbf{V}^T\mathbf{S} \equiv \mathbf{V}^T \begin{pmatrix} e_1 & e_2 & e_3 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Theorem 2.1 *implies that the solution with minimal Frobenius norm is*

$$\mathbf{W}_{opt} = (\mathbf{V}^T\mathbf{S})^\dagger((\mathbf{V}\mathbf{S}^T)^\dagger) = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix},$$

which is not diagonal.

However $\mathbf{W} = \operatorname{diag}\begin{pmatrix} 1 & 2 & 1 \end{pmatrix}$ is also a solution since $\mathbf{V}^T\mathbf{S}\mathbf{W}^{1/2}$ has orthonormal rows. But $\mathbf{W}$ does not have minimal Frobenius norm since $\|\mathbf{W}\|_F^2 = 6$, while $\|\mathbf{W}_{opt}\|_F^2 = 5$.

2.2.2. Number of selected columns equal to rank(A).

If $c = \operatorname{rank}(\mathbf{A})$, then no column of $\mathbf{A}$ can be selected more than once, hence the selected columns form a submatrix of $\mathbf{A}$. In this case Theorem 2.2 can be strengthened: As for the rank-one case in Corollary 2.4, an explicit expression for the weights in terms of leverage scores can be derived.

Theorem 2.7. *Let $\mathbf{A}$ be an $m \times n$ matrix with $k \equiv \operatorname{rank}(\mathbf{A})$. In addition to the conclusions of Theorem 2.2 the following also holds: If*

$$\mathbf{V}^T \begin{pmatrix} \sqrt{w_1} e_{t_1} & \cdots & \sqrt{w_k} e_{t_k} \end{pmatrix}$$

has orthonormal rows, then it is an orthogonal matrix, and $w_j = 1/\|\mathbf{V}^T e_{t_j}\|_2^2$, $1 \le j \le k$.

Proof. See section 7.4. □

Note that the columns selected from $\mathbf{V}^T$ do not necessarily correspond to the largest leverage scores. The following example illustrates that the conditions in Theorem 2.7 are nontrivial.

Example 2.8. *In Theorem* 2.7 *it is not always possible to find k columns from* $\mathbf{V}^T$ *that yield an orthogonal matrix.*

For instance, let

$$\mathbf{V}^T = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -1/\sqrt{14} & -2/\sqrt{14} & 3/\sqrt{14} & 0 \end{pmatrix},$$

and $c = \operatorname{rank}(\mathbf{V}) = 2$. *Since no two columns of* $\mathbf{V}^T$ *are orthogonal, no two columns can be scaled to be orthonormal. Thus no* 2×2 *matrix submatrix of* $\mathbf{V}^T$ *can give rise to an orthogonal matrix.*

However, for $c = 3$ *it is possible to construct a* 2×3 *matrix with orthonormal rows. Selecting columns* $t_1 = 1$, $t_2 = 2$, *and* $t_3 = 3$ *from* $\mathbf{V}^T$, *and weights* $w_1 = \sqrt{5/2}$, $w_2 = \sqrt{2/5}$, *and* $w_3 = \sqrt{11/10}$ *yields a matrix*

$$\mathbf{V}^T \begin{pmatrix} \sqrt{\frac{5}{2}}e_1 & \sqrt{\frac{2}{5}}e_2 & \sqrt{\frac{11}{10}}e_3 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{5}{8}} & \sqrt{\frac{1}{10}} & \sqrt{\frac{11}{40}} \\ -\sqrt{\frac{5}{28}} & -\sqrt{\frac{4}{35}} & \sqrt{\frac{99}{140}} \end{pmatrix}$$

that has orthonormal rows.

Remark 2.9 (connection to Theorem 2.1). *In Theorem* 2.7 *the condition* $c = k$ *implies that the* $k \times k$ *matrix*

$$\mathbf{V}^T \begin{pmatrix} e_{t_1} & \dots & e_{t_k} \end{pmatrix} = \mathbf{V}^T \mathbf{S}$$

is nonsingular. From Theorem 2.1 *it follows that* $\mathbf{W}_{opt} = (\mathbf{V}^T\mathbf{S})^{-1}(\mathbf{V}^T\mathbf{S})^{-T}$ *is the unique minimal Frobenius norm solution for* $\mathbf{A}\mathbf{A}^T = (\mathbf{A}\mathbf{S})\mathbf{W}(\mathbf{A}\mathbf{S})^T$.

If, in addition, the rows of $\mathbf{V}^T\mathbf{S}\mathbf{W}_{opt}^{1/2}$ *are orthonormal, then the minimal norm solution* $\mathbf{W}_{opt}$ *is a diagonal matrix,*

$$\mathbf{W}_{opt} = (\mathbf{V}^T\mathbf{S})^{-1}(\mathbf{V}^T\mathbf{S})^{-T} = \operatorname{diag}\begin{pmatrix} \frac{1}{\|\mathbf{V}^T e_{t_1}\|_2^2} & \cdots & \frac{1}{\|\mathbf{V}^T e_{t_k}\|_2^2} \end{pmatrix}.$$

3. Monte Carlo algorithm for Gram matrix approximation.

We review the randomized algorithm to approximate the Gram matrix (section 3.1), and discuss and compare two different types of sampling probabilities (section 3.2).

3.1. The algorithm.

The randomized algorithm for approximating $\mathbf{A}\mathbf{A}^T$, presented as Algorithm 3.1, is a special case of the BasicMatrixMultiplication Algorithm [17, Figure 2] which samples according to the Exactly(c) algorithm [21, Algorithm 3], that is, independently and with replacement. This means a column can be sampled more than once.

A conceptual version of the randomized algorithm is presented as Algorithm 3.1. Given a user-specified number of samples c, and a set of probabilities p_j, this version assembles columns of the sampling matrix $\mathbf{S}$, then applies $\mathbf{S}$ to $\mathbf{A}$, and finally computes the product

$$\mathbf{X} = (\mathbf{A}\mathbf{S})\,(\mathbf{A}\mathbf{S})^T = \sum_{j=1}^{c} \frac{1}{c p_{t_j}}\, A_{t_j} A_{t_j}^T.$$

The choice of weights $1/(cp_{t_j})$ makes $\mathbf{X}$ an unbiased estimator, $\mathbb{E}[\mathbf{X}] = \mathbf{A}\mathbf{A}^T$ [17, Lemma 3].

Discounting the cost of sampling, Algorithm 3.1 requires $\mathcal{O}(m^2c)$ flops to compute an approximation to $\mathbf{A}\mathbf{A}^T$. Note that Algorithm 3.1 allows zero probabilities. Since an index corresponding to $p_j = 0$ can never be selected, division by zero does not occur in the computation of $\mathbf{S}$. Implementations of sampling with replacement are discussed in [22, section 2.1]. For matrices of small dimension, one can simply use the Matlab function `randsample`.

Algorithm 3.1. Conceptual version of randomized matrix multiplication [17, 21]

```
Input: $m \times n$ matrix $\mathbf{A}$, number of samples $1 \le c \le n$
       Probabilities $p_j$, $1 \le j \le n$, with $p_j \ge 0$ and $\sum_{j=1}^n p_j = 1$

Output: Approximation $\mathbf{X} = (\mathbf{AS})\,(\mathbf{AS})^T$, where $\mathbf{S}$ is $n \times c$ with $\mathbb{E}[\mathbf{S}\,\mathbf{S}^T] = \mathbf{I}_n$

  $\mathbf{S} = \mathbf{0}_{n \times c}$
  for $j = 1 : c$ do
    Sample $t_j$ from $\{1, \ldots, n\}$ with probability $p_{t_j}$
    independently and with replacement
    $S_j = e_{t_j}/\sqrt{c p_{t_j}}$
  end for
  $\mathbf{X} = (\mathbf{AS})\,(\mathbf{AS})^T$
```

3.2. Sampling probabilities. We consider two types of probabilities, the "optimal" probabilities from [17] (section 3.2.1), and leverage score probabilities (section 3.2.2) motivated by Corollary 2.4 and Theorem 2.7, and their use in other randomized algorithms [9, 19, 20]. We show (Theorem 3.1) that for rank-one matrices, Algorithm 3.1 with "optimal" probabilities produces the exact result with a single sample. Numerical experiments (section 3.2.3) illustrate that sampling with "optimal" probabilities results in smaller two-norm relative errors than sampling with leverage score probabilities, and that the two types of probabilities can differ significantly.

3.2.1. "Optimal" probabilities [17]. They are defined by

$$p_j^{opt} = \frac{\|A_j\|_2^2}{\|\mathbf{A}\|_F^2}, \qquad 1 \le j \le n, \tag{3.1}$$

and are called "optimal" because they minimize $\mathbb{E}[\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_F^2]$ [17, Lemma 4]. The "optimal" probabilities can be computed in $\mathcal{O}(mn)$ flops.

The analyses in [17, section 4.4] apply to the more general "nearly optimal" probabilities p_j^β, which satisfy $\sum_{j=1}^n p_j^\beta = 1$ and are constrained by

$$p_j^\beta \ge \beta\, p_j^{opt}, \qquad 1 \le j \le n, \tag{3.2}$$

where $0 < \beta \le 1$ is a scalar. In the special case $\beta = 1$, they revert to the "optimal" probabilites, $p_j^\beta = p_j^{opt}$, $1 \le j \le n$. Hence β can be viewed as the deviation of the probabilities p_j^β from the "optimal" probabilities p_j^{opt}.

3.2.2. Leverage score probabilities [7, 9]. The exact representation in Theorem 2.7 suggests probabilities based on the leverage scores of $\mathbf{V}^T$,

$$p_j^{lev} = \frac{\|\mathbf{V}^T e_j\|_2^2}{\|\mathbf{V}\|_F^2} = \frac{\|\mathbf{V}^T e_j\|_2^2}{k}, \qquad 1 \leq j \leq n, \tag{3.3}$$

where $k = \text{rank}(\mathbf{A})$.

Since the leverage score probabilities are proportional to the squared column norms of $\mathbf{V}^T$, they are the "optimal" probabilities for approximating $\mathbf{V}^T\mathbf{V} = \mathbf{I}_k$. Exact computation of leverage score probabilities, via SVD or QR decomposition, requires $\mathcal{O}(m^2 n)$ flops; thus, it is more expensive than the computation of the "optimal" probabilities.

In the special case of rank-one matrices, the "optimal" and leverage score probabilities are identical, and Algorithm 3.1 with "optimal" probabilities computes the exact result with any number of samples, and in particular a single sample. This follows directly from Corollary 2.4.

THEOREM 3.1. *If* rank$(\mathbf{A}) = 1$, *then* $p_j^{lev} = p_j^{opt}$, $1 \leq j \leq n$.

If $\mathbf{X}$ *is computed by Algorithm* 3.1 *with any* $c \geq 1$ *and probabilities* p_j^{opt}*, then* $\mathbf{X} = \mathbf{A}\mathbf{A}^T$*.*

3.2.3. Comparison of sampling probabilities. We compare the normwise relative errors due to randomization of Algorithm 3.1 when it samples with "optimal" probabilites and leverage score probabilities.

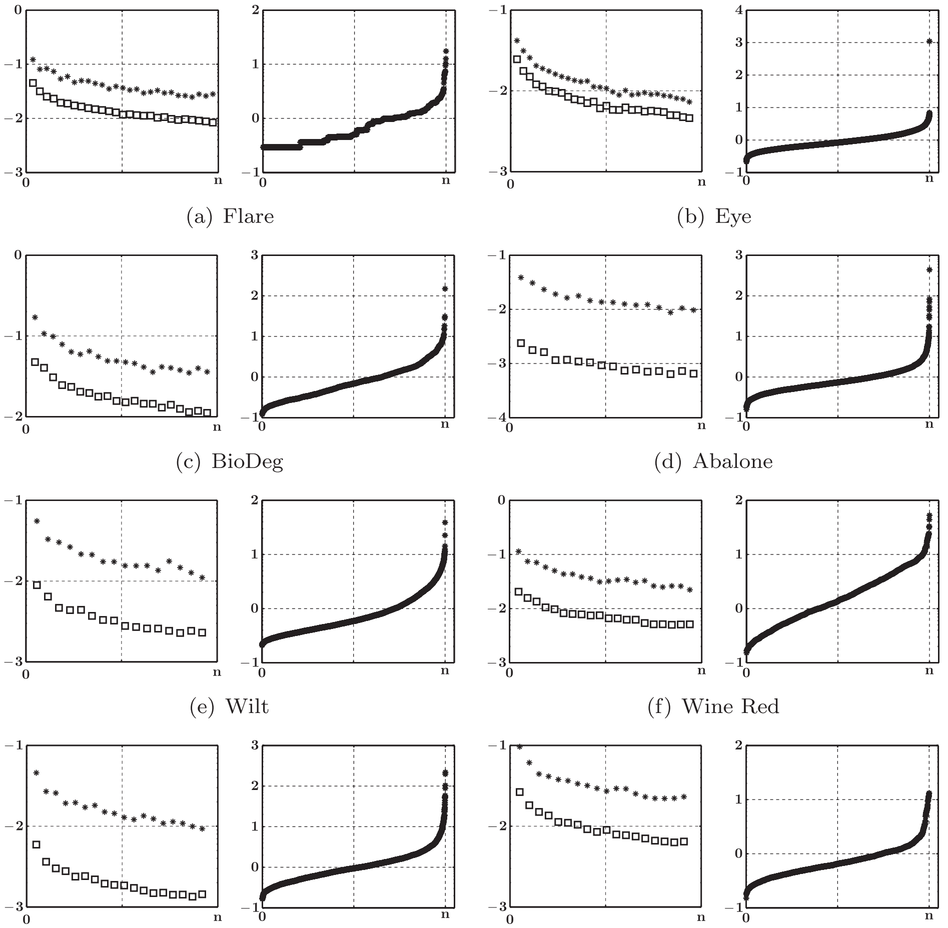
Experimental set up. We present experiments with eight representative matrices, described in Table 4, from the UCI Machine Learning Repository [2].

For each matrix, we ran Algorithm 3.1 twice: once sampling with "optimal" probabilities p_j^{opt} and once sampling with leverage score probabilities p_j^{lev}. The sampling amounts c range from 1 to n, with 100 runs for each value of c.

Figure 1 contains two plots for each matrix: The left plot shows the two-norm relative errors due to randomization, $\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_2 / \|\mathbf{A}\mathbf{A}^T\|_2$, averaged over 100 runs, versus the sampling amount c. The right plot shows the ratios of leverage score over "optimal" probabilities p_j^{lev}/p_j^{opt}, $1 \leq j \leq n$.

TABLE 4
Eight datasets from [2], *and the dimensions, rank, and stable rank of the associated matrices* $\mathbf{A}$.

Dataset	$m \times n$	rank$(\mathbf{A})$	sr$(\mathbf{A})$
Solar Flare	10×1389	10	1.10
EEG Eye State	15×14980	15	1.31
QSAR biodegradation	41×1055	41	1.13
Abalone	8×4177	8	1.002
Wilt	5×4399	5	1.03
Wine Quality—Red	12×1599	12	1.03
Wine Quality—White	12×4898	12	1.01
Yeast	8×1484	8	1.05



(a) Flare

(b) Eye

(c) BioDeg

(d) Abalone

(e) Wilt

(f) Wine Red

(g) Wine White

(h) Yeast

FIG. 1. *Relative errors due to randomization, and ratios of leverage score over "optimal" probabilities for the matrices in Table* 4. *Plots in columns* 1 *and* 3: *The average over* 100 *runs of* $\left\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\right\|_2/\left\|\mathbf{A}\mathbf{A}^T\right\|_2$ *when Algorithm* 3.1 *samples with "optimal" probabilities* (□) *and with leverage score probabilities* ($*$) *versus the number* c *of sampled columns in* $\mathbf{X}$. *The vertical axes are logarithmic, and the labels correspond to powers of* 10. *Plots in columns* 2 *and* 4: *Ratios* p_j^{lev}/p_j^{opt}, $1 \le j \le n$, *sorted in increasing magnitude from left to right.*

Conclusions. Sampling with "optimal" probabilities produces average errors that are lower, by as much as a factor of 10, than those from sampling with leverage score probabilities, for all sampling amounts c. Furthermore, corresponding leverage score and "optimal" probabilities tend to differ by several orders of magnitude.

4. Error due to randomization, for sampling with "nearly optimal" probabilities. We present two new probabilistic bounds (sections 4.1 and 4.2) for the two-norm relative error due to randomization, when Algorithm 3.1 samples with the "nearly optimal" probabilities in (3.2). The bounds depend on the stable rank or the rank of $\mathbf{A}$, but not on the matrix dimensions. Neither bound is always better than the other (section 4.3). The numerical experiments (section 4.4) illustrate that the bounds are informative, even for stringent success probabilities and matrices of small dimension.

4.1. First bound. The first bound depends on the stable rank of $\mathbf{A}$ and also, weakly, on the rank.

THEOREM 4.1. *Let $\mathbf{A} \neq \mathbf{0}$ be an $m \times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm* 3.1 *with the "nearly optimal" probabilities p_j^β in* (3.2).

Given $0 < \delta < 1$ and $0 < \epsilon \leq 1$, if the number of columns sampled by Algorithm 3.1 *is at least*

$$c \geq c_0(\epsilon)\, \mathsf{sr}(\mathbf{A})\, \frac{\ln(\operatorname{rank}(\mathbf{A})/\delta)}{\beta\, \epsilon^2}, \qquad \textit{where} \quad c_0(\epsilon) \equiv 2 + \frac{2\epsilon}{3},$$

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

$$\frac{\left\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \leq \epsilon.$$

Proof. See section 7.5. □

As the required error ϵ becomes smaller, so does the constant $c_0(\epsilon)$ in the lower bound for the number of samples, that is, $c_0(\epsilon) \to 2$ as $\epsilon \to 0$.

4.2. Second bound. This bound depends only on the stable rank of $\mathbf{A}$.

THEOREM 4.2. *Let $\mathbf{A} \neq \mathbf{0}$ be an $m \times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm* 3.1 *with the "nearly optimal" probabilities p_j^β in* (3.2).

Given $0 < \delta < 1$ and $0 < \epsilon \leq 1$, if the number of columns sampled by Algorithm 3.1 *is at least*

$$c \geq c_0(\epsilon)\, \mathsf{sr}(\mathbf{A})\, \frac{\ln(4\mathsf{sr}(\mathbf{A})/\delta)}{\beta\, \epsilon^2}, \qquad \textit{where} \quad c_0(\epsilon) \equiv 2 + \frac{2\epsilon}{3},$$

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

$$\frac{\left\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \leq \epsilon.$$

Proof. See section 7.6. □

4.3. Comparison. The bounds in Theorems 4.1 and 4.2 differ only in the arguments of the logarithms.

On the one hand, Theorem 4.2 is tighter than Theorem 4.1 if $4\,\mathsf{sr}(\mathbf{A}) < \operatorname{rank}(\mathbf{A})$. On the other hand, Theorem 4.1 is tighter for matrices with large stable rank, and in particular for matrices $\mathbf{A}$ with orthonormal rows where $\mathsf{sr}(\mathbf{A}) = \operatorname{rank}(\mathbf{A})$.

In general, Theorem 4.2 is tighter than all the bounds in Table 2, that is, to our knowledge, all published bounds.

4.4. Numerical experiments. We compare the bounds in Theorems 4.1 and 4.2 to the errors of Algorithm 3.1 for sampling with "optimal" probabilities.

Experimental set up. We present experiments with two matrices from the University of Florida Sparse Matrix Collection [15]. The matrices have the same dimension, and similar high ranks and low stable ranks; see Table 5. Note that only for low stable ranks can Algorithm 3.1 achieve any accuracy.

The sampling amounts c range from 1 to n, the number of columns, with 100 runs for each value of c. From the 100 errors $\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_2/\|\mathbf{A}\mathbf{A}^T\|_2$ for each c value, we plot the smallest, largest, and average.

TABLE 5
Matrices from [15], *their dimensions, rank and stable rank, and key quantities from* (4.1) *and* (4.2).

Matrix	$m \times n$	rank($\mathbf{A}$)	$\mathsf{sr}(\mathbf{A})$	$c\,\gamma_1$	$c\,\gamma_2$
us04	163×28016	115	5.27	16.43	13.44
bibd_16_8	163×28016	120	4.29	13.43	10.65

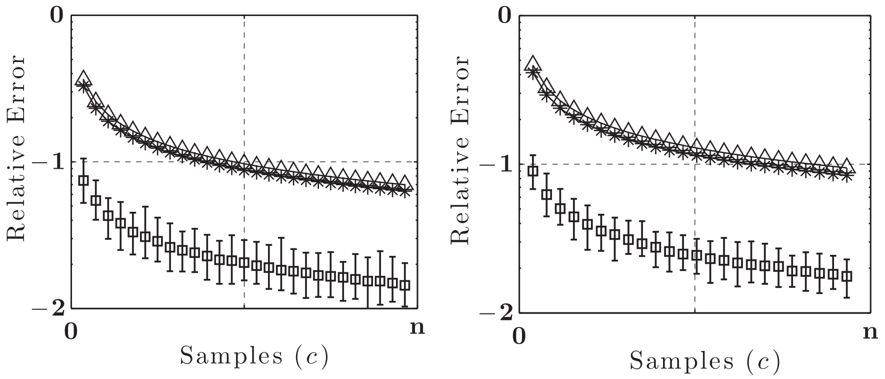


FIG. 2. *Relative errors due to randomization from Algorithm* 3.1, *and bounds* (4.1) *and* (4.2) *versus sampling amount c, for matrices us04 (left) and bidb_16_8 (right). Error bars represent the maximum and minimum of the errors* $\left\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\right\|_2 / \left\|\mathbf{A}\mathbf{A}^T\right\|_2$ *from Algorithm* 3.1 *over* 100 *runs, while the squares represent the average. The triangles* ($\triangle$) *represent the bound* (4.1), *while the stars* ($*$) *represent* (4.2). *The vertical axes are logarithmic, and the labels correspond to powers of* 10.

In Theorems 4.1 and 4.2, the success probability is 99 percent, that is, a failure probability of $\delta = .01$. The error bounds are plotted as a function of c. That is, for Theorem 4.1 we plot (see Theorem 7.6)

$$\frac{\left\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \leq \gamma_1 + \sqrt{\gamma_1\,(6+\gamma_1)}, \qquad \gamma_1 \equiv \mathsf{sr}(\mathbf{A})\,\frac{\ln\left(\mathrm{rank}(\mathbf{A})/.01\right)}{3\,c}, \tag{4.1}$$

while for Theorem 4.2 we plot (see Theorem 7.8)

$$\frac{\left\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \leq \gamma_2 + \sqrt{\gamma_2\,(6+\gamma_2)}, \qquad \gamma_2 \equiv \ \mathsf{sr}(\mathbf{A})\,\frac{\ln\left(4\mathsf{sr}(\mathbf{A})/.01\right)}{3\,c}. \tag{4.2}$$

The key quantities $c\,\gamma_1$ and $c\,\gamma_2$ are shown for both matrices in Table 5.

Figure 2 contains two plots, the left one for matrix us04 and the right one for matrix bibd_16_8. The plots show the relative errors $\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\|_2/\|\mathbf{A}\mathbf{A}^T\|_2$ and the bounds (4.1) and (4.2) versus the sampling amount c.

Conclusions. In both plots, the bounds corresponding to Theorems 4.1 and 4.2 are virtually indistinguishable, as was already predicted by the key quantities $c\gamma_1$ and $c\,\gamma_2$ in Table 5. The bounds overestimate the worst case error from Algorithm 3.1 by a factor of at most 10. Hence they are informative, even for matrices of small dimension and a stringent success probability.

5. Error due to randomization, for sampling with leverage score probabilities. For completeness, we present a normwise relative bound for the error due to randomization, when Algorithm 3.1 samples with leverage score probabilities (3.3). The bound corroborates the numerical experiments in section 3.2.3, and suggests that sampling with leverage score probabilities produces a larger error due to randomization than sampling with "nearly optimal" probabilities.

THEOREM 5.1. *Let $\mathbf{A} \neq \mathbf{0}$ be an $m \times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm* 3.1 *with the leverage score probabilites p_j^{lev} in* (3.3).

Given $0 < \delta < 1$ and $0 < \epsilon \leq 1$, if the number of columns sampled by Algorithm 3.1 *is at least*

$$c \geq c_0(\epsilon)\ \mathrm{rank}(\mathbf{A})\ \frac{\ln(\mathrm{rank}(\mathbf{A})/\delta)}{\epsilon^2}, \qquad \textit{where} \quad c_0(\epsilon) = 2 + \frac{2\epsilon}{3},$$

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

$$\frac{\left\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\right\|_2}{\left\|\mathbf{A}\mathbf{A}^T\right\|_2} \leq \epsilon.$$

Proof. See section 7.7. □

In the special case when $\mathbf{A}$ has orthonormal columns, the leverage score probabilities p_j^{lev} are equal to the "optimal" probabilities p_j^{opt} in (3.1). Furthermore, $\mathrm{rank}(\mathbf{A}) = \mathsf{sr}(\mathbf{A})$, so that Theorem 5.1 is equal to Theorem 4.1. For general matrices $\mathbf{A}$, though, $\mathrm{rank}(\mathbf{A}) \geq \mathsf{sr}(\mathbf{A})$, and Theorem 5.1 is not as tight as Theorem 4.1.

6. Singular value and condition number bounds.

As in [21], we apply the bounds for the Gram matrix approximation to a matrix with orthonormal rows, and derive bounds for the smallest singular value (section 6.1) and condition number (section 6.2) of a sampled matrix.

Specifically, let $\mathbf{Q}$ be a real $m \times n$ matrix with orthonormal rows, $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_m$. Then, as discussed in section 3.2.1, the "optimal" probabilities (3.1) for $\mathbf{Q}$ are equal to the leverage score probabilities (3.3),

$$p_j^{opt} = \frac{\|Q_j\|_2^2}{\|\mathbf{Q}\|_F^2} = \frac{\|Q_j\|_2^2}{m} = p_j^{lev}, \qquad 1 \leq j \leq m.$$

The connection between Gram matrix approximations $(\mathbf{QS})(\mathbf{QS})^T$ and singular values of the sampled matrix $\mathbf{QS}$ comes from the well-conditioning of singular values [30, Corollary 2.4.4],

$$\begin{aligned} \left|1 - \sigma_j(\mathbf{QS})^2\right| &= \left|\sigma_j\left(\mathbf{Q}\mathbf{Q}^T\right) - \sigma_j\left((\mathbf{QS})(\mathbf{QS})^T\right)\right| \\ &\leq \left\|\mathbf{Q}\mathbf{Q}^T - (\mathbf{QS})(\mathbf{QS})^T\right\|_2, \qquad 1 \leq j \leq m. \end{aligned} \tag{6.1}$$

6.1. Singular value bounds.

We present two bounds for the smallest singular value of a sampled matrix, for sampling with the "nearly optimal" probabilities (3.2), and for uniform sampling with and without replacement.

The first bound is based on the Gram matrix approximation in Theorem 4.1.

THEOREM 6.1. *Let $\mathbf{Q}$ be an $m \times n$ matrix with orthonormal rows and coherence μ, and let $\mathbf{QS}$ be computed by Algorithm* 3.1. *Given $0 < \epsilon < 1$ and $0 < \delta < 1$, we have $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ with probability at least $1 - \delta$, if Algorithm* 3.1

- *either samples with the "nearly optimal" probabilities p_j^β, and*

$$c \geq c_0(\epsilon)\, m\, \frac{\ln(m/\delta)}{\beta\epsilon^2},$$

- *or samples with uniform probabilities* $1/n$*, and*

$$c \geq c_0(\epsilon)\, n\, \mu\, \frac{\ln(m/\delta)}{\epsilon^2}.$$

Here $c_0(\epsilon) \equiv 2 + \frac{2}{3}\,\epsilon$*.*

Proof. See section 7.8. □

Since $c_0(\epsilon) \geq 2$, the above bound for uniform sampling is slightly less tight than the last bound in Table 3, i.e., [26, Lemma 1]. Although that bound technically holds only for uniform sampling *without* replacement, the same proof gives the same bound for uniform sampling *with* replacement.

This inspired us to derive a second bound, by modifying the argument in [26, Lemma 1], to obtain a slightly tighter constant. This is done with a direct application of a Chernoff bound (Theorem 7.9). The only difference between the next and the previous result is the smaller constant $c_1(\epsilon)$, and the added application to sampling without replacement.

THEOREM 6.2. *Let* $\mathbf{Q}$ *be an* $m\times n$ *matrix with orthonormal rows and coherence* μ*, and let* $\mathbf{QS}$ *be computed by Algorithm* 3.1*. Given* $0 < \epsilon < 1$ *and* $0 < \delta < 1$*, we have* $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ *with probability at least* $1-\delta$*, if Algorithm* 3.1

- *either samples with the "nearly optimal" probabilities* p_j^β*, and*

$$c \geq c_1(\epsilon)\, m\, \frac{\ln(m/\delta)}{\beta\epsilon^2},$$

- *or samples with uniform probabilities* $1/n$*, with or without replacement, and*

$$c \geq c_1(\epsilon)\, n\, \mu\, \frac{\ln(m/\delta)}{\epsilon^2}.$$

Here $c_1(\epsilon) \equiv \frac{\epsilon^2}{(1-\epsilon)\ln(1-\epsilon)+\epsilon}$ *and* $1 \leq c_1(\epsilon) \leq 2$*.*

Proof. See section 7.9. □

The constant $c_1(\epsilon)$ is slightly smaller than the constant 2 in [26, Lemma 1], which is the last bound in Table 3.

6.2. Condition number bounds. We present two bounds for the condition number $\kappa(\mathbf{QS}) \equiv \sigma_1(\mathbf{QS})/\sigma_m(\mathbf{QS})$ of a sampled matrix $\mathbf{QS}$ with full row rank.

The first condition number bound is based on a Gram matrix approximation, and is analogous to Theorem 6.1.

THEOREM 6.3. *Let* $\mathbf{Q}$ *be an* $m\times n$ *matrix with orthonormal rows and coherence* μ*, and let* $\mathbf{QS}$ *be computed by Algorithm* 3.1*. Given* $0 < \epsilon < 1$ *and* $0 < \delta < 1$*, we have* $\kappa(\mathbf{QS}) \leq \frac{\sqrt{1+\epsilon}}{\sqrt{1-\epsilon}}$ *with probability at least* $1-\delta$*, if Algorithm* 3.1

- *either samples with the "nearly optimal" probabilities* p_j^β*, and*

$$c \geq c_0(\epsilon)\, m\, \frac{\ln(m/\delta)}{\beta\epsilon^2},$$

- *or samples with uniform probabilities* $1/n$*, and*

$$c \geq c_0(\epsilon)\, n\, \mu\, \frac{\ln(m/\delta)}{\epsilon^2}.$$

Here $c_0(\epsilon) \equiv 2 + \frac{2}{3}\,\epsilon$*.*

Proof. See section 7.10. □

The second condition number bound is based on a Chernoff inequality, and is analogous to Theorem 6.2, but with a different constant, and an additional factor of two in the logarithm.

THEOREM 6.4. *Let $\mathbf{Q}$ be an $m \times n$ matrix with orthonormal rows and coherence μ, and let $\mathbf{QS}$ be computed by Algorithm 3.1. Given $0 < \epsilon < 1$ and $0 < \delta < 1$, we have $\kappa(\mathbf{QS}) \leq \frac{\sqrt{1+\epsilon}}{\sqrt{1-\epsilon}}$ with probability at least $1 - \delta$, if Algorithm 3.1*

- *either samples with the "nearly optimal" probabilities p_j^{β}, and*

$$c \geq c_2(\epsilon)\, m\, \frac{\ln(2m/\delta)}{\beta \epsilon^2},$$

- *or samples with uniform probabilities $1/n$, with or without replacement, and*

$$c \geq c_2(\epsilon)\, n\, \mu\, \frac{\ln(2m/\delta)}{\epsilon^2}.$$

Here $c_2(\epsilon) \equiv \frac{\epsilon^2}{(1+\epsilon)\ln(1+\epsilon)-\epsilon}$ and $2 \leq c_2(\epsilon) \leq 2.6$.

Proof. See section 7.11. □

It is difficult to compare the two condition number bounds, and neither bound is always tighter than the other. On the one hand, Theorem 6.4 has a smaller constant than Theorem 6.3 since $c_2(\epsilon) \leq c_1(\epsilon)$. On the other hand, though, Theorem 6.3 has an additional factor of two in the logarithm. For very large m/δ, the additional factor of 2 in the logarithm does not matter much and Theorem 6.4 is tighter.

In general, Theorem 6.4 is not always tighter than Theorem 6.3. For example, if $m = 100$, $\delta = 0.01$, $\epsilon = 0.1$, $\beta = 1$, and Algorithm 3.1 samples with "nearly optimal" probabilities, then Theorem 6.4 requires $1.57 \cdot 10^5$ samples, while Theorem 6.3 requires only $1.43 \cdot 10^5$; hence, it is tighter.

7. Proofs.

We present proofs for the results in sections 2–6.

7.1. Proof of Theorem 2.1.

We will use the two lemmas below. The first one is a special case of [23, Theorem 2.1] where the rank of the approximation is not restricted.

LEMMA 7.1. *Let $\mathbf{H}$ be $m \times n$, $\mathbf{B}$ be $m \times p$, and $\mathbf{C}$ be $q \times n$ matrices, and let $\mathbf{P_B}$ be the orthogonal projector onto $\mathsf{range}(\mathbf{B})$, and $\mathbf{P}_{\mathbf{C}^T}$ the orthogonal projector onto $\mathsf{range}(\mathbf{C}^T)$. Then the solution of*

$$\min_{\mathbf{W}} \|\mathbf{H} - \mathbf{B}\,\mathbf{W}\,\mathbf{C}\|_F$$

with minimal Frobenius norm is

$$\mathbf{W} = \mathbf{B}^{\dagger}\,\mathbf{P_B}\,\mathbf{H}\,\mathbf{P}_{\mathbf{C}^T}\,\mathbf{C}^{\dagger}.$$

LEMMA 7.2. *If $\mathbf{B}$ is $m \times p$ and $\mathbf{C}$ is $p \times n$, with $\operatorname{rank}(\mathbf{B}) = p = \operatorname{rank}(\mathbf{C})$, then $(\mathbf{BC})^{\dagger} = \mathbf{C}^{\dagger}\mathbf{B}^{\dagger}$.*

Proof. Set $\mathbf{Y} \equiv \mathbf{BC}$, and use $\mathbf{B}^{\dagger}\mathbf{B} = \mathbf{I}_p = \mathbf{CC}^{\dagger}$ to verify that $\mathbf{Z} \equiv \mathbf{C}^{\dagger}\mathbf{B}^{\dagger}$ satisfies the four conditions defining the Moore–Penrose inverse

$$\mathbf{YZY} = \mathbf{Y}, \quad \mathbf{ZYZ} = \mathbf{Z}, \quad (\mathbf{YZ})^T = \mathbf{YZ}, \quad (\mathbf{ZY})^T = \mathbf{ZY}. \tag{7.1}$$ □

Proof of Theorem 2.1. Abbreviate $\mathbf{A}_1 \equiv \mathbf{AS}$ and $\mathbf{V}_1^T \equiv \mathbf{V}^T\mathbf{S}$.

In Lemma 7.1, set $\mathbf{H} = \mathbf{AA}^T$, $\mathbf{B} = \mathbf{A}_1$, and $\mathbf{C} = \mathbf{A}_1^T$. Then $\mathbf{P_B} = \mathbf{A}_1\mathbf{A}_1^\dagger = \mathbf{P}_{\mathbf{C}^T}$, and

$$\mathbf{W}_{opt} = \mathbf{A}_1^\dagger\, \mathbf{A}_1\mathbf{A}_1^\dagger\, \mathbf{AA}^T\, \mathbf{A}_1\mathbf{A}_1^\dagger\, (\mathbf{A}_1^\dagger)^T.$$

The conditions for the Moore–Penrose inverse (7.1) imply $\mathbf{A}_1^\dagger\mathbf{A}_1\mathbf{A}_1^\dagger = \mathbf{A}_1^\dagger$, and

$$\mathbf{A}_1\mathbf{A}_1^\dagger\, (\mathbf{A}_1^\dagger)^T = \left(\mathbf{A}_1\mathbf{A}_1^\dagger\right)^T\, (\mathbf{A}_1^\dagger)^T = (\mathbf{A}_1^\dagger)^T\, \mathbf{A}_1^T\, (\mathbf{A}_1^\dagger)^T = (\mathbf{A}_1^\dagger)^T.$$

Hence $\mathbf{W}_{opt} = \mathbf{A}_1^\dagger\, \mathbf{AA}^T\, (\mathbf{A}_1^\dagger)^T$.

Special case rank$(\mathbf{A}_1)$ = rank$(\mathbf{A})$. This means the number of columns c in $\mathbf{A}_1 = \mathbf{U\Sigma V}_1^T$ is at least as large as $k \equiv \text{rank}(\mathbf{A})$. Hence $\mathbf{V}_1^T$ is $k \times c$ with $c \geq k$, and $\text{rank}(\mathbf{V}_1^T) = k = \text{rank}(\mathbf{U\Sigma})$. From Lemma 7.2 it follows that $\mathbf{A}_1^\dagger = (\mathbf{V}_1^\dagger)^T\, \mathbf{\Sigma}^{-1}\mathbf{U}^T$. Hence

$$\mathbf{W}_{opt} = (\mathbf{V}_1^\dagger)^T\, \mathbf{V}^T\mathbf{V}\, \mathbf{V}_1^\dagger = (\mathbf{V}_1^\dagger)^T\, \mathbf{V}_1^\dagger.$$

Furthermore rank$(\mathbf{A}_1)$ = rank$(\mathbf{A})$ implies that $\mathbf{A}_1$ has the same column space as $\mathbf{A}$. Hence the residual in Theorem 2.1 is zero, and $\mathbf{A}_1\mathbf{W}_{opt}\mathbf{A}_1^T = \mathbf{AA}^T$.

Special case $c = \text{rank}(\mathbf{A}_1) = \text{rank}(\mathbf{A})$. This means $c = k$, so that $\mathbf{V}_1$ is a $k \times k$ matrix. From $\text{rank}(\mathbf{A}) = k$ it follows that $\text{rank}(\mathbf{V}_1) = k$, so that $\mathbf{V}_1$ is nonsingular and $\mathbf{V}_1^\dagger = \mathbf{V}_1^{-1}$.

7.2. Proof of Theorem 2.2. Abbreviate

$$\mathbf{A}_1 \equiv \begin{pmatrix} A_{t_1} & \cdots & A_{t_c} \end{pmatrix}, \qquad \mathbf{V}_1^T \equiv \mathbf{V}^T \begin{pmatrix} e_{t_1} & \cdots & e_{t_c} \end{pmatrix},$$

so that the sum of outer products can be written as $\sum_{j=1}^c w_j\, A_{t_j} A_{t_j}^T = \mathbf{A}_1\mathbf{W}\mathbf{A}_1^T$, where $\mathbf{W} \equiv \text{diag}\begin{pmatrix} w_1 & \cdots & w_c \end{pmatrix}$.

1. *Show: If* $\mathbf{A}_1\mathbf{W}\mathbf{A}_1^T = \mathbf{AA}^T$ *for a diagonal* $\mathbf{W}$ *with non-negative diagonal, then* $\mathbf{V}_1^T\mathbf{W}^{1/2}$ *has orthonormal rows.* From $\mathbf{AA}^T = \mathbf{A}_1\mathbf{W}\mathbf{A}_1^T$ it follows that

$$\mathbf{U\Sigma}^2\mathbf{U}^T = \mathbf{AA}^T = \mathbf{A}_1\mathbf{W}\mathbf{A}_1^T = \mathbf{U\Sigma}\, \mathbf{V}_1^T\, \mathbf{W}\, \mathbf{V}_1\, \mathbf{\Sigma U}^T. \tag{7.2}$$

Multiplying by $\mathbf{\Sigma}^{-1}\mathbf{U}^T$ on the left and by $\mathbf{U\Sigma}^{-1}$ on the right gives $\mathbf{I}_k = \mathbf{V}_1^T\, \mathbf{W}\, \mathbf{V}_1$. Since $\mathbf{W}$ is positive semidefinite, it has a symmetric positive semidefinite square root $\mathbf{W}^{1/2}$. Hence $\mathbf{I}_k = \mathbf{V}_1^T\, \mathbf{W}\, \mathbf{V}_1 = (\mathbf{V}_1^T\mathbf{W}^{1/2})\, (\mathbf{V}_1^T\mathbf{W}^{1/2})^T$, and $\mathbf{V}_1^T\mathbf{W}^{1/2}$ has orthonormal rows.

2. *Show: If* $\mathbf{V}_1^T\mathbf{W}^{1/2}$ *has orthonormal rows, then* $\mathbf{A}_1\mathbf{W}\mathbf{A}_1^T = \mathbf{AA}^T$. Inserting $\mathbf{I}_k = (\mathbf{V}_1^T\mathbf{W}^{1/2})\, (\mathbf{V}_1^T\mathbf{W}^{1/2})^T = \mathbf{V}_1^T\mathbf{W}\mathbf{V}_1$ into $\mathbf{A}_1\mathbf{W}\mathbf{A}_1^T$ gives

$$\mathbf{A}_1\mathbf{W}\mathbf{A}_1^T = \mathbf{U\Sigma}\, \left(\mathbf{V}_1^T\, \mathbf{W}\, \mathbf{V}_1\right)\, \mathbf{\Sigma U}^T = \mathbf{U\Sigma}^2\mathbf{U}^T = \mathbf{AA}^T.$$

7.3. Proof of Corollary 2.4. Since rank$(\mathbf{A}) = 1$, the right singular vector matrix $\mathbf{V} = \begin{pmatrix} v_1 & \ldots & v_n \end{pmatrix}^T$ is an $n \times 1$ vector. Since $\mathbf{A}$ has only a single nonzero singular value, $\|A_j\|_2 = \|\mathbf{U\Sigma}\, v_j\|_2 = \|\mathbf{A}\|_F v_j$. Clearly $A_j \neq 0$ if and only $v_j \neq 0$, and $\|\mathbf{V}^T e_j\|_2^2 = v_j^2 = \|A_j\|_2^2/\|\mathbf{A}\|_F^2$. Let A_{t_j} be any c nonzero columns of $\mathbf{A}$. Then

$$\sum_{j=1}^c w_j A_{t_j} A_{t_j}^T = \mathbf{U\Sigma}\, \left(\sum_{j=1}^c w_j v_{t_j}^2\right)\, \mathbf{\Sigma U}^T = \mathbf{U\Sigma}^2\mathbf{U}^T = \mathbf{AA}^T$$

if and only if $\sum_{j=1}^c w_j v_{t_j}^2 = 1$. This is true if $w_j = 1/(cv_{t_j}^2)$, $1 \leq j \leq c$.

7.4. Proof of Theorem 2.7. Since Theorem 2.7 is a special case of Theorem 2.2, we only need to derive the expression for the weights. From $c = k$ it follows that $\mathbf{V}_1^T\mathbf{W}^{1/2}$ is $k \times k$ with orthonormal rows. Hence $\mathbf{V}_1^T\mathbf{W}^{1/2}$ is an orthogonal matrix, and must have orthonormal columns as well, $(\mathbf{W}^{1/2}\mathbf{V}_1)\,(\mathbf{W}^{1/2}\mathbf{V}_1)^T = \mathbf{I}_k$. Thus

$$\mathbf{V}_1\mathbf{V}_1^T = \operatorname{diag}\left(\left\|\mathbf{V}^T e_{t_1}\right\|_2^2 \quad \cdots \quad \left\|\mathbf{V}^T e_{t_c}\right\|_2^2\right) = \mathbf{W}^{-1}.$$

This and $\mathbf{W}^{1/2}$ being diagonal imply $w_j = 1/\|\mathbf{V}^T e_{t_j}\|_2^2$.

7.5. Proof of Theorem 4.1. We present two auxiliary results, a matrix Bernstein concentration inequality (Theorem 7.3) and a bound for the singular values of a difference of positive semidefinite matrices (Theorem 7.4), before deriving a probabilistic bound (Theorem 7.5). The subsequent combination of Theorem 7.5 and the invariance of the two-norm under unitary transformations yields Theorem 7.6 which, at last, leads to a proof for the desired Theorem 4.1.

THEOREM 7.3 (Theorem 1.4 in [51]). *Let $\mathbf{X}_j$ be c independent real symmetric random $m \times m$ matrices. Assume that, with probability one, $\mathbb{E}[\mathbf{X}_j] = \mathbf{0}$, $1 \le j \le c$, and $\max_{1\le j\le c} \|\mathbf{X}_j\|_2 \le \rho_1$. Let $\left\|\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2]\right\|_2 \le \rho_2$.*

Then for any $\epsilon \ge 0$

$$\mathbb{P}\left[\left\|\sum_{j=1}^{c} \mathbf{X}_j\right\|_2 \ge \epsilon\right] \le m\,\exp\left(-\frac{\epsilon^2/2}{\rho_2 + \rho_1\epsilon/3}\right).$$

THEOREM 7.4 (Theorem 2.1 in [54]). *If $\mathbf{B}$ and $\mathbf{C}$ are $m \times m$ real symmetric positive semidefinite matrices, with singular values $\sigma_1(\mathbf{B}) \ge \cdots \ge \sigma_m(\mathbf{B})$ and $\sigma_1(\mathbf{C}) \ge \cdots \ge \sigma_m(\mathbf{C})$, then the singular values of the difference are bounded by*

$$\sigma_j(\mathbf{B} - \mathbf{C}) \le \sigma_j\begin{pmatrix}\mathbf{B} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{pmatrix}, \qquad 1 \le j \le m.$$

In particular, $\|\mathbf{B} - \mathbf{C}\|_2 \le \max\{\|\mathbf{B}\|_2,\ \|\mathbf{C}\|_2\}$.

THEOREM 7.5. *Let $\mathbf{A} \ne \mathbf{0}$ be an $m \times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm 3.1 with the "nearly optimal" probabilites p_j^β in (3.2).*

For any $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{\left\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \le \gamma_0 + \sqrt{\gamma_0\,(6+\gamma_0)}, \qquad \textit{where} \quad \gamma_0 \equiv \mathsf{sr}(\mathbf{A})\,\frac{\ln(m/\delta)}{3\,\beta\,c}.$$

Proof. In order to apply Theorem 7.3, we need to change variables, and check that the assumptions are satisfied.

1. *Change of variables.* Define the $m\times m$ real symmetric matrix random variables $\mathbf{Y}_j \equiv \frac{1}{c\,p_{t_j}}\,A_{t_j}A_{t_j}^T$, and write the output of Algorithm 3.1 as

$$\mathbf{X} = (\mathbf{AS})\,(\mathbf{AS})^T = \mathbf{Y}_1 + \cdots + \mathbf{Y}_c.$$

Since $\mathbb{E}[\mathbf{Y}_j] = \mathbf{A}\mathbf{A}^T/c$, but Theorem 7.3 requires random variables with zero mean, set $\mathbf{X}_j \equiv \mathbf{Y}_j - \frac{1}{c}\mathbf{A}\mathbf{A}^T$. Then

$$\mathbf{X} - \mathbf{A}\mathbf{A}^T = (\mathbf{AS})\,(\mathbf{AS})^T - \mathbf{A}\mathbf{A}^T = \sum_{j=1}^{c}\left(\mathbf{Y}_j - \frac{1}{c}\,\mathbf{A}\mathbf{A}^T\right) = \sum_{j=1}^{c}\mathbf{X}_j.$$

Hence, we show $\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_2 \le \epsilon$ by showing $\left\|\sum_{j=1}^c \mathbf{X}_j\right\|_2 \le \epsilon$.

Next we have to check that the assumptions of Theorem 7.3 are satisfied. In order to derive bounds for $\max_{1\leq j\leq c}\|\mathbf{X}_j\|_2$ and $\left\|\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2]\right\|_2$, we assume general nonzero probabilities p_j for the moment, that is, $p_j>0$, $1\leq j\leq n$.

2. *Bound for* $\max_{1\leq j\leq c}\|\mathbf{X}_j\|_2$. Since $\mathbf{X}_j$ is a difference of positive semidefinite matrices, apply Theorem 7.4 to obtain

$$\|\mathbf{X}_j\|_2 \leq \max\left\{\|\mathbf{Y}_j\|_2,\, \tfrac{1}{c}\left\|\mathbf{A}\mathbf{A}^T\right\|_2\right\} \leq \frac{\hat{\rho}_1}{c}, \qquad \hat{\rho}_1 \equiv \max_{1\leq i\leq n}\left\{\frac{\|A_i\|_2^2}{p_i},\, \|\mathbf{A}\|_2^2\right\}.$$

3. *Bound for* $\left\|\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2]\right\|_2$. To determine the expected value of

$$\mathbf{X}_j^2 = \mathbf{Y}_j^2 - \tfrac{1}{c}\,\mathbf{A}\mathbf{A}^T\,\mathbf{Y}_j - \tfrac{1}{c}\mathbf{Y}_j\,\mathbf{A}\mathbf{A}^T + \tfrac{1}{c^2}(\mathbf{A}\mathbf{A}^T)^2$$

use the linearity of the expected value and $\mathbb{E}[\mathbf{Y}_j]=\mathbf{A}\mathbf{A}^T/c$ to obtain

$$\mathbb{E}[\mathbf{X}_j^2] = \mathbb{E}[\mathbf{Y}_j^2] - \frac{1}{c^2}\,(\mathbf{A}\mathbf{A}^T)^2.$$

Applying the definition of expected value again yields

$$\mathbb{E}[\mathbf{Y}_j^2] = \frac{1}{c^2}\,\sum_{i=1}^n p_i\,\frac{(A_iA_i^T)^2}{p_i^2} = \frac{1}{c^2}\,\sum_{i=1}^n \frac{(A_iA_i^T)^2}{p_i}.$$

Hence

$$\begin{aligned}\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2] &= \frac{1}{c}\,\left(\sum_{i=1}^n \frac{(A_iA_i^T)^2}{p_i} - (\mathbf{A}\mathbf{A}^T)^2\right) = \frac{1}{c}\mathbf{A}\,\left(\sum_{i=1}^n e_i\frac{\|A_i\|_2^2}{p_i}e_i^T - \mathbf{A}^T\mathbf{A}\right)\,\mathbf{A}^T\\ &= \frac{1}{c}\mathbf{A}\,(\mathbf{L}-\mathbf{A}^T\mathbf{A})\,\mathbf{A}^T,\end{aligned}$$

where $\mathbf{L}\equiv\operatorname{diag}\begin{pmatrix}\|A_1\|_2^2/p_1 & \dots & \|A_n\|_2^2/p_n\end{pmatrix}$. Taking norms and applying Theorem 7.4 to $\|\mathbf{L}-\mathbf{A}^T\mathbf{A}\|_2$ gives

$$\left\|\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2]\right\|_2 \leq \frac{\|\mathbf{A}\|_2^2}{c}\,\max\left\{\|\mathbf{L}\|_2,\,\|\mathbf{A}\|_2^2\right\} = \frac{\|\mathbf{A}\|_2^2}{c}\,\hat{\rho}_1.$$

4. *Application of Theorem* 7.3. The required upper bounds for Theorem 7.3 are

$$\|\mathbf{X}_j\|_2 \leq \rho_1 \equiv \frac{\hat{\rho}_1}{c} \qquad \textit{and} \qquad \left\|\sum_{j=1}^c \mathbb{E}[\mathbf{X}_j^2]\right\|_2 \leq \rho_2 \equiv \frac{\|\mathbf{A}\|_2^2}{c}\,\hat{\rho}_1.$$

Inserting these bounds into Theorem 7.3 gives

$$\mathbb{P}\left[\left\|\sum_{j=1}^c \mathbf{X}_j\right\|_2 > \epsilon\right] \leq m\,\exp\left(\frac{-c\epsilon^2}{2\hat{\rho}_1\,(\|\mathbf{A}\|_2^2+\epsilon/3)}\right).$$

Hence $\left\|\sum_{j=1}^c \mathbf{X}_j\right\|_2 \leq \epsilon$ with probability at least $1-\delta$, where

$$\delta \equiv m\,\exp\left(\frac{-c\epsilon^2}{2\hat{\rho}_1\,(\|\mathbf{A}\|_2^2+\epsilon/3)}\right).$$

Solving for ϵ gives

$$\epsilon = \tau_1\,\hat{\rho}_1 + \sqrt{\tau_1\,\hat{\rho}_1\,\;(6\|\mathbf{A}\|_2^2+\tau_1\,\hat{\rho}_1)}, \qquad \tau_1 \equiv \frac{\ln(m/\delta)}{3c}.$$

5. *Specialization to "nearly optimal" probabilities.* We remove zero columns from the matrix. This does not change the norm or the stable rank. The "nearly optimal" probabilities for the resulting submatrix are $p_j^\beta = \beta\|A_j\|_2^2/\|\mathbf{A}\|_F^2$, with $p_j > 0$ for all j. Now replace p_j^β by their lower bounds (3.2). This gives $\hat{\rho}_1 \leq \|\mathbf{A}\|_2^2\,\tau_2$, where $\tau_2 \equiv \mathsf{sr}(\mathbf{A})/\beta \geq 1$, and

$$\epsilon \leq \|\mathbf{A}\|_2^2 \left(\tau_1\tau_2 + \sqrt{\tau_1\tau_2\ (6+\tau_1\tau_2)}\right).$$

Finally observe that $\gamma_0 = \tau_1\tau_2$ and divide by $\|\mathbf{A}\|_2^2 = \|\mathbf{A}\mathbf{A}^T\|_2$. □

We make Theorem 7.5 tighter and replace the dimension m by rank$(\mathbf{A})$. The idea is to apply Theorem 7.5 to the $k \times k$ matrix $(\mathbf{\Sigma}\mathbf{V}^T)\,(\mathbf{\Sigma}\mathbf{V}^T)^T$ instead of the $m \times m$ matrix $\mathbf{A}\mathbf{A}^T$.

THEOREM 7.6. *Let $\mathbf{A} \neq \mathbf{0}$ be an $m \times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm* 3.1 *with the "nearly optimal" probabilites p_j^β in* (3.2).

For any $\delta > 0$, with probability at least $1-\delta$,

$$\frac{\left\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\right\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2} \leq \gamma_1 + \sqrt{\gamma_1\,(6+\gamma_1)}, \qquad \textit{where} \quad \gamma_1 \equiv \mathsf{sr}(\mathbf{A})\,\frac{\ln(\mathrm{rank}(\mathbf{A})/\delta)}{3\,\beta\,c}.$$

Proof. The invariance of the two-norm under unitary transformations implies

$$\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\|_2 = \left\|(\mathbf{\Sigma}\mathbf{V}^T\mathbf{S})\,(\mathbf{\Sigma}\mathbf{V}^T\mathbf{S})^T - (\mathbf{\Sigma}\mathbf{V}^T)\,(\mathbf{\Sigma}\mathbf{V}^T)^T\right\|_2.$$

Apply Theorem 7.5 to the $k \times n$ matrix $B \equiv \mathbf{\Sigma}\mathbf{V}^T$ with probabilities

$$p_j^\beta \geq \beta\,\frac{\|A_j\|_2^2}{\|\mathbf{A}\|_F^2} = \beta\,\frac{\|B_j\|_2^2}{\|\mathbf{B}\|_F^2}. \qquad \square$$

Note that Algorithm 3.1 is still applied to the original matrix $\mathbf{A}$, with probabilities (3.2) computed from $\mathbf{A}$. It is only the bound that has changed.

Proof of Theorem 4.1. At last, we set $\gamma_1 + \sqrt{\gamma_1\,(6+\gamma_1)} \leq \epsilon$ and solve for c as follows. In $\gamma_1 + \sqrt{\gamma_1\,(6+\gamma_1)}$, write

$$\gamma_1 = \tfrac{\ln(\mathrm{rank}(\mathbf{A})/\delta)}{3\,\beta\,c}\,\mathsf{sr}(\mathbf{A}) = \tfrac{t}{3c}, \qquad \text{where} \quad t \equiv \frac{\ln(\mathrm{rank}(\mathbf{A})/\delta)\,\mathsf{sr}(\mathbf{A})}{\beta}.$$

We want to determine $\alpha > 0$ so that $c = \alpha t/\epsilon^2$ satisfies

$$\gamma_1 + \sqrt{\gamma_1\,(6+\gamma_1)} = \frac{t}{3c} + \sqrt{\frac{t}{3c}\left(6+\frac{t}{3c}\right)} \leq \epsilon.$$

Solving for α gives $\alpha \geq 2 + 2\epsilon/3 = c_0(\epsilon)$.

7.6. Proof of Theorem 4.2. To start with, we need a matrix Bernstein concentration inequality, along with the the Löwner partial ordering [35, section 7.7] and the instrinsic dimension [52, section 7].

If $\mathbf{A}_1$ and $\mathbf{A}_2$ are $m \times m$ real symmetric matrices, then $\mathbf{A}_1 \preceq \mathbf{A}_2$ means that $\mathbf{A}_2 - \mathbf{A}_1$ is positive semidefinite [35, Definition 7.7.1]. The *intrinsic dimension* of an $m \times m$ symmetric positive semidefinite matrix $\mathbf{A}$ is [52, Definition 7.1.1]:

$$\mathsf{intdim}(\mathbf{A}) \equiv \mathsf{trace}(\mathbf{A})/\,\|\mathbf{A}\|_2\,,$$

where $1 \leq \mathsf{intdim}(\mathbf{A}) \leq \mathrm{rank}(\mathbf{A}) \leq m$.

Theorem 7.7 (Theorem 7.3.1 and (7.3.2) in [52]). *Let $\mathbf{X}_j$ be c independent real symmetric random matrices, with $\mathbb{E}[\mathbf{X}_j]=\mathbf{0}$, $1\le j\le c$. Let $\max_{1\le j\le c}\|\mathbf{X}_j\|_2\le\rho_1$, and let $\mathbf{P}$ be a symmetric positive semidefinite matrix so that $\sum_{j=1}^c\mathbb{E}[\mathbf{X}_j^2]\preceq\mathbf{P}$. Then for any $\epsilon\ge\|\mathbf{P}\|_2^{1/2}+\rho_1/3$*

$$\mathbb{P}\left[\left\|\sum_{j=1}^c\mathbf{X}_j\right\|_2\ge\epsilon\right]\le 4\,\mathsf{intdim}(\mathbf{P})\,\exp\left(\frac{-\epsilon^2/2}{\|\mathbf{P}\|_2+\rho_1\epsilon/3}\right).$$

Now we apply the above theorem to sampling with "nearly optimal" probabilities.

Theorem 7.8. *Let $\mathbf{A}\neq\mathbf{0}$ be an $m\times n$ matrix, and let $\mathbf{X}$ be computed by Algorithm 3.1 with the "nearly optimal" probabilities p_j^β in (3.2).*

For any $0<\delta<1$, with probability at least $1-\delta$,

$$\frac{\|\mathbf{X}-\mathbf{A}\mathbf{A}^T\|_2}{\|\mathbf{A}\mathbf{A}^T\|_2}\le\gamma_2+\sqrt{\gamma_2\,(6+\gamma_2)},\qquad \textit{where}\quad \gamma_2\equiv\mathsf{sr}(\mathbf{A})\,\frac{\ln\left(4\mathsf{sr}(\mathbf{A})/\delta\right)}{3\,\beta\,c}.$$

Proof. In order to apply Theorem 7.7, we need to change variables and check that the assumptions are satisfied.

1. *Change of variables.* As in item 1 of the proof of Theorem 7.5, we define the real symmetric matrix random variables $\mathbf{Y}_j\equiv\frac{1}{c\,p_{t_j}}\,A_{t_j}A_{t_j}^T$, and write the output of Algorithm 3.1 as

$$\mathbf{X}=(\mathbf{A}\mathbf{S})\,(\mathbf{A}\mathbf{S})^T=\mathbf{Y}_1+\cdots+\mathbf{Y}_c.$$

The zero mean versions are $\mathbf{X}_j\equiv\mathbf{Y}_j-\frac{1}{c}\mathbf{A}\mathbf{A}^T$, so that $\mathbf{X}-\mathbf{A}\mathbf{A}^T=\sum_{j=1}^c\mathbf{X}_j$.

Next we have to check that the assumptions of Theorem 7.7 are satisfied, for the "nearly optimal" probabilities $p_j^\beta=\beta\|A_j\|_2^2/\|\mathbf{A}\|_F^2$. Since Theorem 7.7 does not depend on the matrix dimensions, we can assume that all zero columns of $\mathbf{A}$ have been removed, so that all $p_j^\beta>0$.

2. *Bound for* $\max_{1\le j\le c}\|\mathbf{X}_j\|_2$. From item 2 in the proof of Theorem 7.5 it follows that $\|\mathbf{X}_j\|_2\le\rho_1$, where

$$\rho_1=\frac{1}{c}\max_{1\le j\le n}\left\{\frac{\|A_j\|_2^2}{p_j^\beta},\,\|\mathbf{A}\|_2^2\right\}\le\frac{\|\mathbf{A}\|_F^2}{\beta c}.$$

3. *The matrix* $\mathbf{P}$. From item 3 in the proof of Theorem 7.5 it follows that

$$\sum_{j=1}^c\mathbb{E}[\mathbf{X}_j^2]=\tfrac{1}{c}\mathbf{A}\mathbf{L}\mathbf{A}^T-\tfrac{1}{c}\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T,$$

where $\mathbf{L}\equiv\operatorname{diag}\begin{pmatrix}\|A_1\|_2^2/p_1^\beta & \cdots & \|A_n\|_2^2/p_n^\beta\end{pmatrix}\preceq\left(\|\mathbf{A}\|_F^2/\beta\right)\mathbf{I}_n$. Since $\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T$ is positive semidefinite, so is

$$\tfrac{1}{c}\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T=\tfrac{1}{c}\mathbf{A}\mathbf{L}\mathbf{A}^T-\tfrac{1}{c}\left(\mathbf{A}\mathbf{L}\mathbf{A}^T-\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{A}^T\right)=\tfrac{1}{c}\mathbf{A}\mathbf{L}\mathbf{A}^T-\sum_{j=1}^c\mathbb{E}[\mathbf{X}_j^2].$$

Thus, $\sum_{j=1}^c\mathbb{E}[\mathbf{X}_j^2]\preceq\frac{1}{c}\mathbf{A}\mathbf{L}\mathbf{A}^T\preceq\frac{\|\mathbf{A}\|_F^2}{\beta c}\mathbf{A}\mathbf{A}^T$, where the the second inequality follows from [35, Theorem 7.7.2(a)]. Set $\mathbf{P}\equiv\frac{\|\mathbf{A}\|_F^2}{\beta c}\mathbf{A}\mathbf{A}^T$. Then

$$\|\mathbf{P}\|_2=\frac{\|\mathbf{A}\|_2^2\|\mathbf{A}\|_F^2}{\beta c}\qquad\textit{and}\qquad\mathsf{intdim}(\mathbf{P})=\frac{\|\mathbf{A}\|_F^4}{\|\mathbf{A}\|_F^2\,\|\mathbf{A}\|_2^2}=\mathsf{sr}(\mathbf{A}).$$

4. *Application of Theorem* 7.7. Substituting the above expressions for $\|\mathbf{P}\|_2$, $\mathsf{intdim}(\mathbf{P})$, and $\rho_1 = \frac{\|\mathbf{A}\|_F^2}{\beta\, c}$ into Theorem 7.7 gives

$$\mathbb{P}\left[\left\|\sum_{j=1}^{c} \mathbf{X}_j\right\|_2 \geq \epsilon\right] \leq 4\,\mathsf{sr}(\mathbf{A})\ \exp\left(\frac{-\epsilon^2\beta c}{2\,\|\mathbf{A}\|_F^2\,\left(\|\mathbf{A}\|_2^2 + \epsilon/3\right)}\right).$$

Hence $\left\|\sum_{j=1}^{c} \mathbf{X}_j\right\|_2 \leq \epsilon$ with probability at least $1-\delta$, where

$$\delta \equiv 4\,\mathsf{sr}(\mathbf{A})\ \exp\left(\frac{-\epsilon^2\beta c}{2\,\|\mathbf{A}\|_F^2\,\left(\|\mathbf{A}\|_2^2 + \epsilon/3\right)}\right).$$

Solving for ϵ gives

$$\epsilon = \hat{\gamma}_2 + \sqrt{\hat{\gamma}_2\,(6\,\|\mathbf{A}\|_2^2 + \hat{\gamma}_2)}, \qquad \text{where} \quad \hat{\gamma}_2 \equiv \|\mathbf{A}\|_F^2\,\frac{\ln(4\,\mathsf{sr}(\mathbf{A})/\delta)}{3\beta c} = \|\mathbf{A}\|_2^2\,\gamma_2.$$

It remains to show the last requirement of Theorem 7.7, that is, $\epsilon \geq \|\mathbf{P}\|_2^{1/2} + \rho_1/3$. Replacing ϵ by its above expression in terms of $\hat{\gamma}_2$ shows that the requirement is true if $\hat{\gamma}_2 \geq \rho_1/3$ and $\sqrt{6\|\mathbf{A}\|_2^2\,\hat{\gamma}_2} \geq \|\mathbf{P}\|_2^{1/2}$. This is the case if $\ln(4\,\mathsf{sr}(\mathbf{A})/\delta) > 1$. Since $\mathsf{sr}(\mathbf{A}) \geq 1$, this is definitely true if $\delta < 4/e$. Since we assumed $\delta < 1$ from the start, the requirement is fulfilled automatically.

At last, divide both sides of $\left\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\right\|_2 \leq \hat{\gamma}_2 + \sqrt{\hat{\gamma}_2\,(6\,\|\mathbf{A}\|_2^2 + \hat{\gamma}_2)}$ by $\left\|\mathbf{A}\mathbf{A}^T\right\|_2 = \|\mathbf{A}\|_2^2$. □

Proof of Theorem 4.2. As in the proof of Theorem 4.1, solve for c in $\gamma_2 + \sqrt{\gamma_2\,(6+\gamma_2)} \leq \epsilon$.

7.7. Proof of Theorem 5.1. To get a relative error bound, substitute the thin SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ into

$$\begin{aligned}\|\mathbf{X} - \mathbf{A}\mathbf{A}^T\|_2 &= \|(\mathbf{A}\mathbf{S})\,(\mathbf{A}\mathbf{S})^T - \mathbf{A}\mathbf{A}^T\|_2 = \|(\mathbf{\Sigma}\mathbf{V}^T\mathbf{S})\,(\mathbf{\Sigma}\mathbf{V}^T\mathbf{S})^T - \mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\|_2\\ &\leq \|\mathbf{\Sigma}\|_2^2\,\|(\mathbf{V}^T\mathbf{S})\,(\mathbf{V}^T\mathbf{S})^T - \mathbf{V}^T\mathbf{V}\|_2\\ &= \|\mathbf{A}\mathbf{A}^T\|_2\,\|(\mathbf{V}^T\mathbf{S})\,(\mathbf{V}^T\mathbf{S})^T - \mathbf{V}^T\mathbf{V}\|_2.\end{aligned}$$

The last term can be viewed as sampling columns from $\mathbf{V}^T$ to approximate the product $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$. Now apply Theorem 4.1, where $\|\mathbf{V}\|_F^2 = k = \text{rank}(\mathbf{A})$ and $\|\mathbf{V}\|_2^2 = 1$, so that $\mathsf{sr}(\mathbf{V}) = k = \text{rank}(\mathbf{A})$.

7.8. Proof of Theorem 6.1. We present separate proofs for the two types of sampling probabilities.

Sampling with "nearly optimal" probabilities. Applying Theorem 4.1 shows that $\|\mathbf{Q}\mathbf{Q}^T - (\mathbf{Q}\mathbf{S})\,(\mathbf{Q}\mathbf{S})^T\|_2 \leq \epsilon$ with probability at least $1-\delta$, if $c \geq c_0(\epsilon)\,\frac{m}{\beta\epsilon^2}\,\ln(m/\delta)$.

Sampling with uniform probabilities. Use the β factor to express the uniform probabilities as "nearly optimal" probabilities,

$$\frac{1}{n} = \frac{m}{n\,\mu}\,\frac{\mu}{m} \geq \frac{m}{n\,\mu}\,\frac{\|Q_j\|_2^2}{\|\mathbf{Q}\|_F^2} = \beta\,\frac{\|Q_j\|_2^2}{\|\mathbf{Q}\|_F^2} = \beta\,p_j^{opt}, \qquad 1 \leq j \leq n.$$

Now apply Theorem 4.1 with $\beta = m/(n\mu)$.

For both sampling methods, the connection (6.1) implies that $\sigma_m(\mathbf{Q}\mathbf{S}) \geq \sqrt{1-\epsilon}$ with probability at least $1-\delta$.

7.9. Proof of Theorem 6.2. First we present the concentration inequality on which the proof is based. Below $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ denote the smallest and largest eigenvalues, respectively, of the symmetric positive semidefinite matrix $\mathbf{X}$.

THEOREM 7.9 (Theorem 5.1.1 in [52]). *Let $\mathbf{X}_j$ be c independent $m \times m$ real symmetric positive semidefinite random matrices, with $\max_{1\leq j\leq c} \|\mathbf{X}_j\|_2 \leq \rho$. Define*

$$\rho_{\max} \equiv \lambda_{\max}\left(\mathbb{E}\left[\sum_{j=1}^{c} X_j\right]\right), \qquad \rho_{\min} \equiv \lambda_{\min}\left(\mathbb{E}\left[\sum_{j=1}^{c} X_j\right]\right),$$

and $f(x) \equiv e^x/(1+x)^{1+x}$. Then, for any $0 < \epsilon < 1$

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{j=1}^{c} X_j\right) \leq (1-\epsilon)\rho_{\min}\right] \leq m\, f(-\epsilon)^{\rho_{\min}/\rho}$$

and

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{j=1}^{c} X_j\right) \geq (1+\epsilon)\rho_{\max}\right] \leq m\, f(\epsilon)^{\rho_{\max}/\rho}.$$

Proof of Theorem 6.2. Write $(\mathbf{QS})\,(\mathbf{QS})^T = \sum_{j=1}^{c} X_j$, where $\mathbf{X}_j \equiv \frac{Q_{t_j} Q_{t_j}^T}{c\, p_{t_j}}$. To apply Theorem 7.9 we need to compute ρ, $\rho_{\min}$, and $\rho_{\max}$.

Sampling with "nearly optimal" probabilities. The definition of "nearly optimal" probabilities (3.2) and the fact that $\|\mathbf{Q}\|_F^2 = m$ imply $\|\mathbf{X}_j\|_2 = \frac{\|Q_{t_j}\|_2^2}{c p_{t_j}^{\beta}} \leq \frac{m}{c\,\beta}$. Hence we can set $\rho \equiv \frac{m}{c\,\beta}$. The definition of $\mathbf{X}_j$ implies

$$\mathbb{E}\left[\sum_{j=1}^{c} X_{t_j}\right] = \frac{1}{c}\sum_{j=1}^{c}\sum_{i=1}^{n} Q_i Q_i^T = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_m,$$

so that $\rho_{\min} = 1$. Now apply Theorem 7.9 to conclude

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{j=1}^{c} X_j\right) \leq (1-\epsilon)\right] \leq m f(-\epsilon)^{c\beta/m}.$$

Setting the right-hand side equal to δ and solving for c gives

$$c = \frac{m}{\beta}\,\frac{\ln(\delta/m)}{\ln f(-\epsilon)} = c_1(\epsilon)\, m\, \frac{\ln(m/\delta)}{\beta\epsilon^2},$$

where the second equality follows from $\ln f(x) = x - (1+x)\ln(1+x)$. The function $c_1(x)$ is decreasing in $[0,1]$, and L'Hôpital's rule implies that $c_1(\epsilon) \to 2$ as $\epsilon \to 0$ and $c_1(\epsilon) \to 1$ as $\epsilon \to 1$.

Sampling with uniform probabilities. An analogous proof with $p_j = 1/n$ shows that $\|\mathbf{X}_j\|_2 \leq \rho \equiv n\mu/c$.

Uniform sampling without replacement. Theorem 7.9 also holds when the matrices $\mathbf{X}_j$ are sampled uniformly without replacement [50, Theorem 2.2].

For all three sampling methods, the connection (6.1) implies that $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ with probability at least $1-\delta$.

7.10. Proof of Theorem 6.3. The proof follows from Theorem 6.1, and the connection (6.1), since $|1-\sigma_j^2(\mathbf{QS})| \leq \epsilon$, $1 \leq j \leq m$, implies that both $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ and $\sigma_1(\mathbf{QS}) \leq \sqrt{1+\epsilon}$.

7.11. Proof of Theorem 6.4. We derive separate bounds for the smallest and largest singular values of $\mathbf{QS}$.

Sampling with "nearly optimal" probabilities. The proof of Theorem 6.2 implies that

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{j=1}^{c} X_j\right) \leq (1-\epsilon)\right] \leq m f(-\epsilon)^{c\beta/m}.$$

Similarly, we can apply Theorem 7.9 with $\rho_{\max}=1$ to conclude

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{j=1}^{c} X_j\right) \geq (1+\epsilon)\right] \leq m f(\epsilon)^{c\beta/m}.$$

Since $f(-\epsilon) \leq f(\epsilon)$, Boole's inequality implies

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{j=1}^{c} X_j\right) \leq (1-\epsilon) \text{ and } \lambda_{\max}\left(\sum_{j=1}^{c} X_j\right) \geq (1+\epsilon)\right] \leq 2m f(\epsilon)^{c\beta/m}.$$

Hence, $\sigma_m(\mathbf{QS}) \geq \sqrt{1-\epsilon}$ and $\sigma_1(\mathbf{QS}) \leq \sqrt{1+\epsilon}$ hold simultaneously with probability at least $1-\delta$, if

$$c \geq c_2(\epsilon)\, m\, \frac{\ln(2m/\delta)}{\beta\epsilon^2}.$$

This bound for c also ensures that $\kappa(\mathbf{QS}) \leq \frac{\sqrt{1+\epsilon}}{\sqrt{1-\epsilon}}$ with probability at least $1-\delta$. The function $c_2(x)$ is increasing in $[0,1]$, and L'Hôpital's rule implies that $c_2(\epsilon) \to 2$ as $\epsilon \to 0$ and $c_2(\epsilon) \to 1/(2\ln(2)-1) \leq 2.6$ as $\epsilon \to 1$.

Uniform sampling, with or without replacement. The proof is analogous to the corresponding part of the proof of Theorem 6.2.

Acknowledgments. We thank Petros Drineas and Michael Mahoney for useful discussions, and the four anonymous reviewers whose suggestions helped us to improve the quality of the paper.

REFERENCES

[1] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging LAPACK's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.

[2] K. Bache and M. Lichman, *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml (2013).

[3] J. D. Batson, D. A. Spielman, and N. Srivastava, *Twice-Ramanujan sparsifiers*, in Proceedings of the 2009 ACM International Symposium on Theory of Computing (STOC'09), ACM, New York, 2009, pp. 255–262.

[4] J. Batson, D. A. Spielman, and N. Srivastava, *Twice-Ramanujan sparsifiers*, SIAM J. Comput., 41 (2012), pp. 1704–1721.

[5] M.-A. Belabbas and P. J. Wolfe, *On sparse representations of linear operators and the approximation of matrix products*, in Proceedings of the 42nd Annual Conference on Information Sciences and Systems, IEEE, Piscataway, NJ, 2008, pp. 258–263.

[6] C. Boutsidis, *Topics in Matrix Sampling Algorithms*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 2011.

[7] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near-optimal column-based matrix reconstruction*, in Proceedings of the IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 2011, pp. 305–314.

[8] C. Boutsidis and A. Gittens, *Improved matrix algorithms via the subsampled randomized Hadamard transform*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1301–1340.

[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, SIAM, Philadelphia, 2009. pp. 968–977.

[10] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[11] S. Chandrasekaran and I. C. F. Ipsen, *On rank-revealing QR factorisations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.

[12] S. Chatterjee and A. S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Statist. Sci., 1 (1986), pp. 379–393.

[13] E. Cohen and D. D. Lewis, *Approximating matrix multiplication for pattern recognition tasks*, in Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 1997, pp. 682–691.

[14] E. Cohen and D. D. Lewis, *Approximating matrix multiplication for pattern recognition tasks*, J. Algorithms, 30 (1999), pp. 211–252.

[15] T. Davis and Y. Hu, *The University of Florida Sparse Matrix Collection*, ACM Trans. Math. Software, 38 (2011), 1.

[16] P. Drineas and R. Kannan, *Fast Monte-Carlo algrithms for approximate matrix multiplication*, in Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 2001, pp. 452–459.

[17] 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.

[18] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, *Fast approximation of matrix coherence and statistical leverage*, J. Mach. Learn. Res., 13 (2012), pp. 3475–3506.

[19] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for l_2 regression and applications*, in Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2006, pp. 1127–1136.

[20] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Relative-error CUR matrix decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 844–881.

[21] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2010), pp. 219–249.

[22] S. Eriksson-Bique, M. Solbrig, M. Stefanelli, S. Warkentin, R. Abbey, and I. C. F. Ipsen, *Importance sampling for a Monte Carlo matrix multiplication algorithm, with application to information retrieval*, SIAM J. Sci. Comput., 33 (2011), pp. 1689–1706.

[23] S. Friedland and A. Torokhti, *Generalized rank-constrained matrix approximations*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 656–659.

[24] A. Frieze, R. Kannan, and S. Vempala, *Fast Monte-Carlo algorithms for finding low-rank approximations*, in Proceedings of the 39th Annual Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 1998, pp. 370–378.

[25] A. Frieze, R. Kannan, and S. Vempala, *Fast Monte-Carlo algorithms for finding low-rank approximations*, J. ACM, 51 (2004), pp. 1025–1041.

[26] A. Gittens, *The Spectral Norm Error of the Naïve Nyström Extension*, preprint, arXiv: 1110.5305v1, 2011.

[27] A. Gittens, *Topics in Randomized Numerical Linear Algebra*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2013.

[28] G. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[29] G. Golub, V. Klema, and G. Stewart, *Rank Degeneracy and Least Squares Problems*, Technical report STAN-CS-76-559, Computer Science Department, Stanford University, Stanford, CA, 1976.

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, MD, 2013.

[31] 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.

[32] 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.

[33] D. C. Hoaglin and R. E. Welsch, *The Hat matrix in regression and ANOVA*, Amer. Statist., 32 (1978), pp. 17–22.

[34] H. Hong and C. Pan, *The rank-revealing QR decomposition and SVD*, Math. Comp., 58 (1992), pp. 213–32.

[35] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, 2013.

[36] D. Hsu, S. M. Kakade, and T. Zhang, *Tail inequalities for sums of random matrices that depend on the intrinsic matrix dimension*, Electron. Comm. Probab., 17 (2012), 14.

[37] I. C. F. Ipsen and T. Wentworth, *The Effect of Coherence on Sampling from Matrices with Orthonormal Columns, and Preconditioned Least Squares Problems*, preprint, arXiv: 1203.4809v2, 2012.

[38] S. Kumar, M. Mohri, and A. Talwalkar, *Sampling techniques for the Nyström method*, in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, Clearwater Beach, FL, Vol. 5, 2009, pp. 304–311.

[39] M. Li, G. L. Miller, and R. Peng, *Iterative row sampling*, in Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos, CA, 2013, pp. 127–136.

[40] E. Liberty, *Simple and deterministic matrix sketching*, in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), ACM, New York, 2013, pp. 581–588.

[41] H. Madrid, V. Guerra, and M. Rojas, *Sampling techniques for Monte Carlo matrix multiplication with applications to image processing*, in Proceedings of the 4th Mexican Conference on Pattern Recognition, Springer-Verlag, New York, 2012, pp. 45–54.

[42] M. Magdon-Ismail, *Row Sampling for Matrix Algorithms via a Non-Commutative Bernstein Bound*, preprint, arXiv:1008.0587, 2010.

[43] M. Magdon-Ismail, *Using a Non-Commutative Bernstein Bound to Approximate Some Matrix Algorithms in the Spectral Norm*, preprint, arXiv1103.5453v1, 2011.

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

[45] M. W. Mahoney, *Randomized algorithms for matrices and data*, Found. Trends Mach. Learn., 3 (2011), pp. 123–224.

[46] R. Pagh, *Compressed matrix multiplication*, ACM Trans. Comput. Theory, 5 (2013) 9.

[47] M. Rudelson and R. Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, J. ACM, 54 (2007), 21.

[48] T. Sarlós, *Improved approximation for large matrices via random projections*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 2006, pp. 143–152.

[49] N. Srivastava, *Spectral Sparsification and Restricted Invertibility*, Ph.D. thesis, Yale University, New Haven, CT, 2010.

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

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

[52] J. A. Tropp, *User-Friendly Tools for Random Matrices: An Introduction*, http://users.cms.caltech.edu/∼jtropp/pubs.html (2012).

[53] P. F. Velleman and R. E. Welsch, *Efficient computing of regression diagnostics*, Amer. Statist., 35 (1981), pp. 234–242.

[54] X. Zhan, *Singular values of differences of positive semidefinite matrices*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 819–823.

[55] A. Zouzias, *Randomized Primitives for Linear Algebra and Applications*, Ph.D. thesis, University of Toronto, Toronto, Canada, 2013.