Near-Optimal Column-Based Matrix Reconstruction

Christos Boutsidis
Mathematical Sciences Department
IBM T.J. Watson Research Center
Yorktown Heights, New York
cboutsi@us.ibm.com

Petros Drineas and Malik Magdon-Ismail
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York
drinep@cs.rpi.edu, magdon@cs.rpi.edu

Abstract— We consider low-rank reconstruction of a matrix using a subset of its columns and we present asymptotically optimal algorithms for both spectral norm and Frobenius norm reconstruction. The main tools we introduce to obtain our results are: (i) the use of fast approximate SVD-like decompositions for column-based matrix reconstruction, and (ii) two deterministic algorithms for selecting rows from matrices with orthonormal columns, building upon the sparse representation theorem for decompositions of the identity that appeared in [1].

Keywords-low-rank matrix approximation; subset selection; SVD; approximate SVD; spectral sparsification

1. INTRODUCTION

The best rank k approximation to a matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ is $\mathrm{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$ are the top k singular values of A, with associated left and right singular vectors $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_i \in \mathbb{R}^n$ respectively. (See Section 1.1 for notation.) The singular values and singular vectors of A can be computed via the Singular Value Decomposition (SVD) of A in $O(mn\min\{m,n\})$ time. There is considerable interest (e.g. [3], [5], [7], [8], [10], [14], [18], [19], [20]) in determining a minimum set of $r \ll n$ columns of A which is approximately as good as A_k at reconstructing A. Such columns are important for interpretting data [20], building robust machine learning algorithms [3], etc.

Let $\mathrm{A} \in \mathbb{R}^{m\times n}$ and let $\mathrm{C} \in \mathbb{R}^{m\times r}$ consist of r columns of A for some $k \leq r < n$. We are interested in the reconstruction errors (see Section 1.1 for notation)

$$\|\mathrm{A} - \mathrm{C}\mathrm{C}^{+}\mathrm{A}\|_{\xi} \qquad \text{and} \qquad \|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi},$$

for $\xi = 2, F$ (see Section 1.1 for notation). The former is the reconstruction error for A using the columns in C; the latter is the error from the best rank k reconstruction of A (under the appropriate norm) within the column space of C. For fixed A, k, and r, we would like these errors to be as close to

$$\|\mathrm{A} - \mathrm{A}_k\|_{\xi}$$

as possible. We present polynomial-time near-optimal constructions for arbitrary $r > k$, settling important open questions regarding column-based matrix reconstruction.

- **Spectral norm:** What is the best reconstruction error with $r > k$ columns? We present polynomial-time (deterministic and randomized) algorithms with approximation error asymptotically matching a lower bound proven in this work. Prior work had focused on the $r = k$ case and presented near-optimal polynomial-time algorithms [5], [16].
- **Frobenius norm:** How many columns are needed for relative error approximation, i.e. a reconstruction error of $(1+\epsilon)\|\mathrm{A} - \mathrm{A}_k\|_F$, for $\epsilon > 0$? We show that $O(k/\epsilon)$ columns contain a rank-k subspace which reconstructs A to relative error, and we present the first sub-SVD-time (randomized) algorithm to identify these columns. This matches the $\Omega(k/\epsilon)$ lower bound in [7] and improves the best known upper bound of $O(k\log k + k/\epsilon)$ [5], [7], [11], [22].

1.1. Notation

$\mathrm{A}, \mathrm{B}, \ldots$ are matrices; $\mathbf{a}, \mathbf{b}, \ldots$ are column vectors. I_n is the $n\times n$ identity matrix; $\mathbf{0}_{m\times n}$ is the $m\times n$ matrix of zeros; $\mathbf{1}_n$ is the $n\times 1$ vector of ones; $\boldsymbol{e}_i$ is the standard basis (whose dimensionality will be clear from the context); $\mathrm{rank}(\mathrm{A})$ is the rank of A. The Frobenius and the spectral matrix-norms are: $\|\mathrm{A}\|_F^2 = \sum_{i,j} \mathrm{A}_{ij}^2$ and $\|\mathrm{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathrm{A}\mathbf{x}\|_2$; $\|\mathrm{A}\|_{\xi}$ is used if a result holds for both norms $\xi = 2$ and $\xi = F$. The Singular Value Decomposition (SVD) of A, with $\mathrm{rank}(\mathrm{A}) = \rho$ is

$$\mathrm{A} = \underbrace{\begin{pmatrix} \mathrm{U}_k & \mathrm{U}_{\rho-k} \end{pmatrix}}_{\mathrm{U}_{\mathrm{A}}\in\mathbb{R}^{m\times\rho}} \underbrace{\begin{pmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\rho-k} \end{pmatrix}}_{\Sigma_{\mathrm{A}}\in\mathbb{R}^{\rho\times\rho}} \underbrace{\begin{pmatrix} \mathrm{V}_k^{\mathrm{T}} \\ \mathrm{V}_{\rho-k}^{\mathrm{T}} \end{pmatrix}}_{\mathrm{V}_{\mathrm{A}}^{\mathrm{T}}\in\mathbb{R}^{\rho\times n}},$$

with singular values $\sigma_1 \geq \ldots \sigma_k \geq \sigma_{k+1} \geq \ldots \geq \sigma_\rho > 0$. We will use $\sigma_i(\mathrm{A})$ to denote the i-th singular value of A when the matrix is not clear from the context. The matrices $\mathrm{U}_k \in \mathbb{R}^{m\times k}$ and $\mathrm{U}_{\rho-k} \in \mathbb{R}^{m\times(\rho-k)}$ contain the left singular vectors of A; and, similarly, the matrices $\mathrm{V}_k \in \mathbb{R}^{n\times k}$ and $\mathrm{V}_{\rho-k} \in \mathbb{R}^{n\times(\rho-k)}$ contain the right singular vectors of A. It is well-known that $\mathrm{A}_k = \mathrm{U}_k\Sigma_k\mathrm{V}_k^{\mathrm{T}}$ minimizes $\|\mathrm{A} - \mathrm{X}\|_{\xi}$ over all matrices $\mathrm{X} \in \mathbb{R}^{m\times n}$ of rank at most k. We use $\mathrm{A}_{\rho-k}$ to denote the matrix $\mathrm{A} - \mathrm{A}_k = \mathrm{U}_{\rho-k}\Sigma_{\rho-k}\mathrm{V}_{\rho-k}^{\mathrm{T}}$. Also, $\mathrm{A}^{+} = \mathrm{V}_{\mathrm{A}}\Sigma_{\mathrm{A}}^{-1}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}$ denotes the Moore-Penrose pseudo-inverse of A. For a symmetric positive definite matrix $\mathrm{A} = \mathrm{B}\mathrm{B}^{\mathrm{T}}$, $\lambda_i(\mathrm{A}) = \sigma_i^2(\mathrm{B})$ denotes the i-th eigenvalue of A.

Finally, given a matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ and a matrix $\mathrm{C} \in \mathbb{R}^{m\times r}$ with $r > k$, we formally define the matrix $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A}) \in \mathbb{R}^{m\times n}$ as the best approximation to A within the column space of C that has rank at most k; $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$ minimizes the residual $\|\mathrm{A} - \hat{\mathrm{A}}\|_{\xi}$, over all $\hat{\mathrm{A}}$ in the column space of C that have rank at most k (one can write $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A}) = \mathrm{CX}$ where $\mathrm{X} \in \mathbb{R}^{r\times n}$ has rank at most k). In general, $\Pi^{2}_{\mathrm{C},k}(\mathrm{A}) \neq \Pi^{F}_{\mathrm{C},k}(\mathrm{A})$; Section 4.2 discusses the computation of $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$.

1.2. Our main results

Since $\|\mathrm{A} - \mathrm{CC}^{+}\mathrm{A}\|_{\xi} \leq \|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi}$, we will state all our bounds in terms of the latter quantity. Note that we chose to state our Frobenius norm bounds in terms of the *square* of the Frobenius norm; this choice facilitates comparisons with prior work and simplifies our proofs.

Theorem 1 (Deterministic spectral norm reconstruction)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ *and a target rank* $k < \rho$*, there exists a deterministic polynomial-time algorithm to select* $r > k$ *columns of* A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that*

$$\begin{aligned}\|\mathrm{A} - \Pi^{2}_{\mathrm{C},k}(\mathrm{A})\|_2 &\leq \left(1 + \frac{1+\sqrt{(\rho-k)/r}}{1-\sqrt{k/r}}\right)\|\mathrm{A} - \mathrm{A}_k\|_2 \\ &= O\left(\sqrt{\rho/r}\right)\|\mathrm{A} - \mathrm{A}_k\|_2.\end{aligned}$$

The matrix C *can be computed in* $T_{SVD} + O(rn(k^2 + (\rho-k)^2))$ *time, where* T_{SVD} *is the time needed to compute all* ρ *right singular vectors of* A*.*

Our algorithm uses the matrices V_k and $\mathrm{V}_{\rho-k}$ of the right singular vectors of A. These matrices can be computed in $O(mn\min\{m,n\})$ time via the SVD. The asymptotic multiplicative error of the above theorem matches a lower bound that we prove in Section 5. This is the first spectral reconstruction algorithm with asymptotically optimal guarantees for arbitrary $r > k$. Previous work presented near-optimal algorithms for $r = k$ [16]. We note that in Section 3 we will present a result that achieves a slightly worse error bound (essentially replacing ρ by n in the accuracy guarantee), but only uses the top k right singular vectors of A (i.e., the matrix V_k).

Theorem 2 (Deterministic Frobenius norm reconstruction)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ *and a target rank* $k < \rho$*, there exists a deterministic polynomial-time algorithm to select* $r > k$ *columns of* A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that*

$$\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2 \leq \left(1 + \frac{1}{(1-\sqrt{k/r})^2}\right)\|\mathrm{A} - \mathrm{A}_k\|_F^2.$$

The matrix C *can be computed in* $T_{\mathrm{V}_k} + O\left(mn + nrk^2\right)$ *time, where* T_{V_k} *is the time needed to compute the top* k *right singular vectors of* A*.*

Our bound implies a constant-factor approximation. Previous work presents deterministic near-optimal algorithms for $r = k$ [5]; we are unaware of any deterministic algorithms for $r > k$.

The next two theorems guarantee (up to small constant factors) the same bounds as Theorems 1 and 2, but the proposed algorithms are considerably more efficient. In particular, there is no need to exactly compute the right singular vectors of A, because approximations suffice.

Theorem 3 (Fast spectral norm reconstruction)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ*, a target rank* $2 \leq k < \rho$*, and* $0 < \epsilon < 1$*, there exists a randomized algorithm to select* $r > k$ *columns of* A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that*

$$\begin{aligned}\mathbf{E}\left[\|\mathrm{A} - \Pi^{2}_{\mathrm{C},k}(\mathrm{A})\|_2\right] &\leq \left(\sqrt{2}+\epsilon\right)\left(1 + \frac{1+\sqrt{n/r}}{1-\sqrt{k/r}}\right)\|\mathrm{A} - \mathrm{A}_k\|_2 \\ &= O\left(\sqrt{n/r}\right)\|\mathrm{A} - \mathrm{A}_k\|_2.\end{aligned}$$

The matrix C *can be computed in* $O\left(mnk\epsilon^{-1}\log\left(k^{-1}\min\{m,n\}\right) + nrk^2\right)$ *time.*

Theorem 4 (Fast Frobenius norm reconstruction)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ*, a target rank* $2 \leq k < \rho$*, and* $0 < \epsilon < 1$*, there exists a randomized algorithm to select* $r > k$ *columns of* A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that*

$$\mathbf{E}\left[\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2\right] \leq (1+\epsilon)\left(1 + \frac{1}{(1-\sqrt{k/r})^2}\right)\|\mathrm{A}-\mathrm{A}_k\|_F^2.$$

The matrix C *can be computed in* $O\left(mnk\epsilon^{-1} + nrk^2\right)$ *time.*

Our last, yet perhaps most interesting result, guarantees relative-error Frobenius norm approximation by combining the algorithm of Theorem 4 with one round of adaptive sampling [7], [8]. This is the first relative-error approximation for Frobenius norm reconstruction that uses a linear number of columns in k (the target rank). Previous work [11], [22], [7], [5] achieves relative error with $O(k\log k + k/\epsilon)$ columns. Our result asymptotically matches the $\Omega(k/\epsilon)$ lower bound in [7].

Theorem 5 (Fast relative-error Frobenius norm reconstruction)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ*, a target rank* $2 \leq k < \rho$*, and* $0 < \epsilon < 1$*, there exists a randomized algorithm to select at most*

$$r = \frac{2k}{\epsilon}\left(1 + o(1)\right)$$

columns of A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that,*

$$\mathbf{E}\left[\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2\right] \leq (1+\epsilon)\|\mathrm{A} - \mathrm{A}_k\|_F^2.$$

The matrix C *can be computed in* $O((mnk + nk^3)\epsilon^{-2/3})$ *time.*

Running times: Our running times are in terms of the number of operations needed to compute the matrix C, and for simplicity we assume that A is dense; if A is sparse, additional savings might be possible. Our accuracy guarantees are in terms of the optimal matrix $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$, which would require additional time to compute. For the Frobenius norm, computing $\Pi^{F}_{\mathrm{C},k}(\mathrm{A})$ is straightforward, and only requires an additional $O\left(mnr + (m+n)\,r^2\right)$ time (see the discussion in Section 4.2). For the spectral norm, we are not aware of any algorithm to compute $\Pi^{2}_{\mathrm{C},k}(\mathrm{A})$ exactly. In Section 4.2 we present a simple approach that computes $\hat{\Pi}^{2}_{\mathrm{C},k}(\mathrm{A})$, a constant-factor approximation to $\Pi^{2}_{\mathrm{C},k}(\mathrm{A})$, in $O\left(mnr + (m+n)\,r^2\right)$ time. Our bounds in Theorems 1 and 3 can be restated in terms of the error $\left\|\mathrm{A} - \hat{\Pi}^{2}_{\mathrm{C},k}(\mathrm{A})\right\|_2$; the accuracy guarantees only weaken by small constant factors.

1.3. Prior results on column-based matrix reconstructions

There is a long literature on algorithms for column-based matrix reconstruction using $r \geq k$ columns. The first result goes back to [15], with the most recent one being, to the best of our knowledge, the work in [5]. Table I provides a summary on lower bounds for the ratio

$$\frac{\|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|^2_{\xi}}{\|\mathrm{A} - \mathrm{A}_k\|^2_{\xi}},$$

where C is a matrix consisting of r columns of A, with $r \geq k$. Our Theorem 17 in the Appendix contributes a new lower bound for the spectral norm case when $r > k$. (Note that any lower bound for the ratio $\|\mathrm{A}-\mathrm{CC}^{+}\mathrm{A}\|^2_{\xi}/\|\mathrm{A}-\mathrm{A}_k\|^2_{\xi}$ implies a lower bound for $\|\mathrm{A}-\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|^2_{\xi}/\|\mathrm{A}-\mathrm{A}_k\|^2_{\xi}$; the converse, however, is not true.)

1.3.1. The Frobenius norm case: We present known guarantees for the approximation ratio

$$\frac{\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|^2_{F}}{\|\mathrm{A} - \mathrm{A}_k\|^2_{F}}.$$

When $r = k$, [5] gives a $(k+1)$ approximation running in $O(knm^3 \log m)$ time; this approximation ratio matches a lower bound in [8]. [5] also presented a faster randomized algorithm achieving an expected $(1+\epsilon)(k+1)$ approximation, running in $O(mn \log nk^2\epsilon^{-2} + n\log^3 n \cdot k^7\epsilon^{-6}\log\left(k\epsilon^{-1}\log n\right))$ time.

When $r = \Omega(k \log k)$, relative-error approximations are known. [11] presented the first result that achieved such a bound, using random sampling of the columns of A according to the Euclidean norms of the rows of V_k. More specifically, a $(1+\epsilon)$-approximation was proven using $r = \Omega\left(k\epsilon^{-2}\log\left(k\epsilon^{-1}\right)\right)$ columns in $T_{\mathrm{V}_k} + O(kn + r\log r)$ time. [22] argued that the same technique gives a $(1+\epsilon)$-approximation using $r = \Omega\left(k\log k + k\epsilon^{-1}\right)$ columns and showed how to improve the running time to $T_{\tilde{\mathrm{V}}_k} + O(kn + r\log r)$, where $\tilde{\mathrm{V}}_k \in \mathbb{R}^{n\times k}$ contains the right singular vectors of an approximation to A_k and can be computed in $o(mn\min\{m,n\})$ time (sub-SVD). In [7], the authors leveraged volume sampling and presented an approach that achieves a relative error approximation using $O(k^2\log k + k\epsilon^{-1})$ columns in $O(mnk^2\log k)$ time. Also, it is possible to combine the fast volume sampling approach in [5] (setting, for example, $\epsilon = 1/2$) with $O(\log k)$ rounds of adaptive sampling as described in [7] to achieve a relative error approximation using $O\left(k\log k + k\epsilon^{-1}\right)$ columns. The running time of this combined algorithm is $O\left(mnk^2\log n + nk^7\log^3 n\log\left(k\log n\right)\right)$. The techniques in [11] do not apply to general $r > k$, since $\Omega(k\log k)$ columns must be sampled in order to preserve rank with random sampling.

A related line of work (including [6], [12], [13], [23]) has focused on the construction of coresets and sketches for high dimensional subspace approximation with respect to general ℓ_p norms. In our setting, $p = 2$ corresponds to Frobenius norm matrix reconstruction, and Theorem 1.3 of [23] presents an exponential in k/ϵ algorithm to select $O\left(k^2/\epsilon\log\left(k/\epsilon\right)\right)$ columns that guarantee relative error approximation. It would be interesting to understand if the techniques of [6], [12], [13], [23] can be extended to match our results here in the special case of $p = 2$.

1.3.2. The spectral norm case: We present known guarantees for the approximation ratio

$$\frac{\|\mathrm{A} - \Pi^{2}_{\mathrm{C},k}(\mathrm{A})\|^2_{2}}{\|\mathrm{A} - \mathrm{A}_k\|^2_{2}}.$$

In general, results for spectral norm have been sparse. When $r = k$, the strongest bound emerges from Strong Rank Revealing QR (RRQR) [16] (specifically Algorithm 4 in [16]), which, for $f > 1$, runs in $O(mnk\log_f n)$ time and guarantees an $f^2k(n-k)+1$ approximation. For $r > k$, to the best of our knowledge, there is no easy way to extend the RRQR guarantees. In fact, we are not aware of any bound applicable to this domain other than those obtained by trivially extending the Frobenius norm bounds, because any α-approximation in the Frobenius norm gives an $\alpha(\rho-k)$-approximation in the spectral norm:

$$\begin{aligned}\|\mathrm{A} - \Pi^{2}_{\mathrm{C},k}(\mathrm{A})\|^2_2 &\leq \|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|^2_2 \leq \|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|^2_F \\ &\leq \alpha\|\mathrm{A} - \mathrm{A}_k\|^2_F \leq \alpha(\rho-k)\|\mathrm{A} - \mathrm{A}_k\|^2_2.\end{aligned}$$

2. Main Tools

Our two main tools are the use of matrix factorizations for column-based low-rank matrix reconstruction, and two deterministic sparsification lemmas which extend the work of [1].

2.1. Matrix factorizations

Our first tool suggests how to use matrix factorizations to reconstruct a matrix from a subset of its columns: Lemmas

r	Spectral norm ($\xi = 2$)	Frobenius norm ($\xi = F$)
$r = k$	n/k [5]	$k+1$ [8]
$r > k$	n/r (Section 5)	$1 + k/r$ [7] (also see Section 5)

Table I
LOWER BOUNDS FOR THE APPROXIMATION RATIO $\|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi}^2 / \|\mathrm{A} - \mathrm{A}_k\|_{\xi}^2$.

6, 8, and 9. Lemmas 8 and 9 present factorizations of the matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ of the form

$$\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E},$$

where $\mathrm{B} \in \mathbb{R}^{m\times k}$, $\mathrm{Z} \in \mathbb{R}^{n\times k}$, $\mathrm{E} \in \mathbb{R}^{m\times n}$, and Z consists of orthonormal columns. Lemma 6 shows how to apply these factorizations by drawing a connection between matrix factorizations and column selection. Lemma 6 is the starting point of all our column reconstruction results.

Lemma 6. *Let* $\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}$, *with* $\mathrm{E}\mathrm{Z} = \mathbf{0}_{m\times k}$ *and* $\mathrm{Z}^{\mathrm{T}}\mathrm{Z} = \mathrm{I}_k$. *Let* $\mathrm{S} \in \mathbb{R}^{n\times r}$ *be any matrix such that* $rank(\mathrm{Z}^{\mathrm{T}}\mathrm{S}) = rank(\mathrm{Z}) = k$. *Let* $\mathrm{C} = \mathrm{A}\mathrm{S} \in \mathbb{R}^{m\times r}$. *Then,*

$$\|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi}^2 \leq \|\mathrm{E}\|_{\xi}^2 + \|\mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\|_{\xi}^2.$$

Proof: The optimality of $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$ implies that $\|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi}^2 \leq \|\mathrm{A} - \mathrm{X}\|_{\xi}^2$ over all matrices $\mathrm{X} \in \mathbb{R}^{m\times n}$ of rank at most k in the column space of C. Consider the matrix $\mathrm{X} = \mathrm{C}\left(\mathrm{Z}^{\mathrm{T}}\mathrm{S}\right)^{+}\mathrm{Z}^{\mathrm{T}}$ (clearly X is in the column space of C and $\mathrm{rank}(\mathrm{X}) \leq k$ because $\mathrm{Z} \in \mathbb{R}^{n\times k}$):

$$\|\mathrm{A} - \mathrm{C}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\|_{\xi}^2 =$$

$$\begin{aligned}
&= \|\underbrace{\mathrm{B}\mathrm{Z}^{\mathrm{T}} + \left(\mathrm{A} - \mathrm{B}\mathrm{Z}^{\mathrm{T}}\right)}_{\mathrm{A}} - \underbrace{\left(\mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}\right)\mathrm{S}}_{\mathrm{C}=\mathrm{A}\mathrm{S}}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\|_{\xi}^2 \\
&= \|\mathrm{B}\mathrm{Z}^{\mathrm{T}} - \mathrm{B}\mathrm{Z}^{\mathrm{T}}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}} + \mathrm{E} + \mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\|_{\xi}^2 \\
&\overset{(a)}{=} \|\mathrm{E} + \mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\|_{\xi}^2 \\
&\overset{(b)}{\leq} \|\mathrm{E}\|_{\xi}^2 + \|\mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\|_{\xi}^2.
\end{aligned}$$

(a) follows because, by assumption, $\mathrm{rank}(\mathrm{Z}^{\mathrm{T}}\mathrm{S}) = k$, and thus $(\mathrm{Z}^{\mathrm{T}}\mathrm{S})(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+} = \mathrm{I}_k$ which implies $\mathrm{B}\mathrm{Z}^{\mathrm{T}} - \mathrm{B}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}} = \mathbf{0}_{m\times n}$. (b) follows by matrix-Pythagoras because $\mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\mathrm{Z}^{\mathrm{T}}\mathrm{E}^{T} = \mathbf{0}_{m\times n}$ (recall that $\mathrm{E} = \mathrm{A} - \mathrm{B}\mathrm{Z}^{T}$ and $\mathrm{E}\mathrm{Z} = \mathbf{0}_{m\times k}$ by assumption). The lemma follows by strong submultiplicativity because Z has orthonormal columns, hence $\|\mathrm{Z}\|_2 = 1$. ∎

In this work, we view C as a dimensionally-reduced or sampled sketch of A; S is the dimension-reduction or sampling matrix. In words, Lemma 6 argues that if the matrix S preserves the rank of an approximate factorization of the original matrix A, then the reconstruction of A from $\mathrm{C} = \mathrm{A}\mathrm{S}$ has an error that is essentially proportional to the error of the approximate factorization. The importance of this lemma is that it indicates an algorithm for matrix reconstruction using a subset of the columns of A: first, compute *any* factorization of the form $\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}$ satisfying the assumptions of the lemma; then, compute a sampling matrix S which satisfies the rank assumption and controls the error $\|\mathrm{E}\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\|_{\xi}$.

An immediate corollary of Lemma 6 emerges by considering the SVD of A. More specifically, consider the following factorization of A: $\mathrm{A} = \mathrm{A}\mathrm{V}_k\mathrm{V}_k^{\mathrm{T}} + (\mathrm{A} - \mathrm{A}_k)$, where V_k is the matrix of the top k right singular vectors of A. In the parlance of Lemma 6, $\mathrm{Z} = \mathrm{V}_k$, $\mathrm{B} = \mathrm{A}\mathrm{V}_k$, $\mathrm{E} = \mathrm{A} - \mathrm{A}_k$, and clearly $\mathrm{E}\mathrm{Z} = \mathbf{0}_{m\times k}$.

Lemma 7. *Let* $\mathrm{S} \in \mathbb{R}^{n\times r}$ *be a matrix such that* $rank(\mathrm{V}_k^{\mathrm{T}}\mathrm{S}) = k$. *Let* $\mathrm{C} = \mathrm{A}\mathrm{S}$*; then,*

$$\|\mathrm{A} - \Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})\|_{\xi}^2 \leq \|\mathrm{A} - \mathrm{A}_k\|_{\xi}^2 + \|(\mathrm{A} - \mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^{+}\|_{\xi}^2.$$

The above lemma will be useful for designing the deterministic (spectral norm and Frobenius norm) column-reconstruction algorithms of Theorems 1 and 2. However, computing the SVD is costly and thus we would like to design a factorization of the form $\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}$ that is as good as the SVD, but can be computed in $O(mnk)$ time. The next two lemmas achieve this goal by extending the algorithms in [18], [21] (see [2] for their proofs). We will use these factorizations to design fast column reconstruction algorithms in Theorems 3, 4, and 5.

Lemma 8 (Randomized fast spectral norm SVD)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ, *a target rank* $2 \leq k < \rho$, *and* $0 < \epsilon < 1$, *there exists an algorithm that computes a factorization* $\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}$, *with* $\mathrm{B} = \mathrm{A}\mathrm{Z}$, $\mathrm{Z}^{\mathrm{T}}\mathrm{Z} = \mathrm{I}_k$, *and* $\mathrm{E}\mathrm{Z} = \mathbf{0}_{m\times k}$ *such that*

$$\mathbf{E}\left[\|\mathrm{E}\|_2\right] \leq \left(\sqrt{2} + \epsilon\right)\|\mathrm{A} - \mathrm{A}_k\|_2.$$

The proposed algorithm runs in $O\left(mnk\epsilon^{-1}\log\left(k^{-1}\min\{m,n\}\right)\right)$ *time.*

Lemma 9 (Randomized fast Frobenius norm SVD)**.** *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ, *a target rank* $2 \leq k < \rho$, *and* $0 < \epsilon < 1$, *there exists an algorithm that computes a factorization* $\mathrm{A} = \mathrm{B}\mathrm{Z}^{\mathrm{T}} + \mathrm{E}$, *with* $\mathrm{B} = \mathrm{A}\mathrm{Z}$, $\mathrm{Z}^{\mathrm{T}}\mathrm{Z} = \mathrm{I}_k$, *and* $\mathrm{E}\mathrm{Z} = \mathbf{0}_{m\times k}$ *such that*

$$\mathbf{E}\left[\|\mathrm{E}\|_F^2\right] \leq (1+\epsilon)\|\mathrm{A} - \mathrm{A}_k\|_F^2.$$

The proposed algorithm runs in $O\left(mnk\epsilon^{-1}\right)$ *time.*

2.2. Sparse approximate decompositions of the identity

Lemmas 6, 8 and 9 argue that, in order to achieve almost optimal column-based matrix reconstruction, we need a sampling matrix S that preserves the rank of Z and controls the error $\|\mathrm{ES}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^{+}\|_{\xi}$. We present algorithms to compute such a matrix S in Lemmas 10 and 11. These lemmas were motivated by an important linear-algebraic result for a decomposition of the identity presented by Batson *et al.* [1]. It is worth emphasizing that the result of [1] can not be directly applied to the column reconstruction problem. Indeed, in our setting, it is necessary to control properties related to *both* matrices Z and $\mathrm{E} = \mathrm{A} - \mathrm{BZ}^{\mathrm{T}}$ *simultaneously*. In the spectral-norm reconstruction case, we need to control the singular values of the two matrices; in the Frobenius-norm reconstruction case, we need to control singular values and Frobenius norms of two matrices.

Lemma 10 (Dual Set Spectral Sparsification.)**.** *Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be two equal cardinality decompositions of the identity, where $\mathbf{v}_i \in \mathbb{R}^k$ ($k < n$), $\mathbf{u}_i \in \mathbb{R}^\ell$ ($\ell \leq n$), $\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}} = \mathrm{I}_k$, and $\sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \mathrm{I}_\ell$. Given an integer r with $k < r \leq n$, there exists a set of weights $s_i \geq 0$ ($i = 1, \ldots, n$) at most r of which are non-zero, such that*

$$\begin{aligned} \lambda_k\left(\sum_{i=1}^n s_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right) &\geq \left(1 - \sqrt{\frac{k}{r}}\right)^2 \qquad \textit{and} \\ \lambda_1\left(\sum_{i=1}^n s_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right) &\leq \left(1 + \sqrt{\frac{\ell}{r}}\right)^2. \end{aligned}$$

The weights s_i can be computed deterministically in $O\left(rn\left(k^2 + \ell^2\right)\right)$ time.

Proof Sketch. The main insight is to decouple the analysis of the lower bound on λ_k and the upper bound on λ_1 in [1]. Once this is done, one can accomodate two *different* sets of vectors, and the rest of the analysis follows a similar line of reasoning as the original single set analysis of [1]. The details are in [2]. ∎

In matrix notation, let U and V be the matrices whose *rows* are the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ respectively. We can now construct the sampling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ as follows: for $i = 1, \ldots, n$, if s_i is non-zero then include $\sqrt{s_i}\mathbf{e}_i$ as a column of S; here $\mathbf{e}_i$ is the i-th standard basis vector [1]. Using this matrix notation, the above lemma guarantees that $\sigma_k\left(\mathrm{V}^{\mathrm{T}}\mathrm{S}\right) \geq 1 - \sqrt{k/r}$ and $\sigma_1\left(\mathrm{U}^{\mathrm{T}}\mathrm{S}\right) \leq 1 + \sqrt{\ell/r}$. Clearly, S may be viewed as a matrix that samples and rescales r *rows* of U and V (columns of U^{T} and V^{T}), namely the rows that correspond to non-zero weights s_i.

[1]Note that we slightly abused notation: indeed, the number of columns of S is less than or equal to r, since at most r of the weights are non-zero. Here, we use r to also denote the actual number of non-zero weights, which is equal to the number of columns of the matrix S.

Lemma 11 (Dual Set Spectral-Frobenius Sparsification.)**.** *Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a decomposition of the identity, where $\mathbf{v}_i \in \mathbb{R}^k$ ($k < n$) and $\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}} = \mathrm{I}_k$; let $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ be an arbitrary set of vectors, where $\mathbf{a}_i \in \mathbb{R}^\ell$. Then, given an integer r such that $k < r \leq n$, there exists a set of weights $s_i \geq 0$ ($i = 1 \ldots n$), at most r of which are non-zero, such that*

$$\begin{aligned} \lambda_k\left(\sum_{i=1}^n s_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right) &\geq \left(1 - \sqrt{\frac{k}{r}}\right)^2 \qquad \textit{and} \\ \mathrm{Tr}\left(\sum_{i=1}^n s_i\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\right) &\leq \mathrm{Tr}\left(\sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\right) = \sum_{i=1}^n \|\mathbf{a}_i\|_2^2. \end{aligned}$$

The weights s_i can be computed deterministically in $O\left(rnk^2 + n\ell\right)$ time.

Proof Sketch. After decoupling the analysis as in Lemma 10, the main insight is to introduce a new potential function which controls the Frobenius norm of the sparsified second set of vectors. This new potential function turns out to be the trace. The two set analysis for two *different* potential functions then follows a similar line as Lemma 10. Again, the details are in [2]. ∎

In matrix notation (here A denotes the matrix whose *rows* are the vectors $\mathbf{a}_i$), the above lemma guarantees that $\sigma_k\left(\mathrm{V}^{\mathrm{T}}\mathrm{S}\right) \geq 1 - \sqrt{k/r}$ and $\|\mathrm{A}^{\mathrm{T}}\mathrm{S}\|_F^2 \leq \|\mathrm{A}\|_F^2$.

3. Proofs of our Main Results

In this section, we leverage the main tools described in Section 2 in order to prove the results of Section 1.2 (Theorems 1 through 5). We start with a proof of Theorem 1, using Lemmas 7 and 10.

Proof of Theorem 1. Apply the algorithm of Lemma 10 on the following two sets of vectors: the n rows of the matrix V_k and the n rows of the matrix $\mathrm{V}_{\rho-k}$. The output of the algorithm is a sampling and rescaling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ (see discussion after Lemma 10 in Section 2.2). Let $\mathrm{C} = \mathrm{AS}$ and note that C consists of a subset of r *rescaled* columns of A. Lemma 10 guarantees that $\sigma_k(\mathrm{V}_k^{\mathrm{T}}\mathrm{S}) \geq 1 - \sqrt{k/r} > 0$ (assuming $r > k$), and so $\mathrm{rank}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S}) = k$. Also, $\sigma_1(\mathrm{V}_{\rho-k}^{\mathrm{T}}\mathrm{S}) = \|\mathrm{V}_{\rho-k}^{\mathrm{T}}\mathrm{S}\|_2 \leq 1 + \sqrt{(\rho-k)/r}$. Applying Lemma 7, we get $\|\mathrm{A} - \Pi_{\mathrm{C},k}^2(\mathrm{A})\|_2^2 \leq$

$$\begin{aligned} &\leq \|\mathrm{A} - \mathrm{A}_k\|_2^2 + \|(\mathrm{A} - \mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^{+}\|_2^2 \\ &\leq \|\mathrm{A} - \mathrm{A}_k\|_2^2 + \|(\mathrm{A} - \mathrm{A}_k)\mathrm{S}\|_2^2\|(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^{+}\|_2^2 \\ &= \|\mathrm{A} - \mathrm{A}_k\|_2^2 + \|\mathrm{U}_{\rho-k}\Sigma_{\rho-k}\mathrm{V}_{\rho-k}^{\mathrm{T}}\mathrm{S}\|_2^2\|(V_k^{\mathrm{T}}\mathrm{S})^{+}\|_2^2 \\ &\leq \|\mathrm{A} - \mathrm{A}_k\|_2^2 + \|\Sigma_{\rho-k}\|_2^2\|\mathrm{V}_{\rho-k}^{\mathrm{T}}\mathrm{S}\|_2^2\|(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^{+}\|_2^2 \\ &\leq \|\mathrm{A} - \mathrm{A}_k\|_2^2\left(1 + \frac{(1 + \sqrt{(\rho-k)/r})^2}{(1 - \sqrt{k/r})^2}\right), \end{aligned}$$

where the last inequality follows because $\|\Sigma_{\rho-k}\|_2 = \|\mathrm{A} - \mathrm{A}_k\|_2$ and $\|(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^{+}\|_2 = 1/\sigma_k(\mathrm{V}_k^{\mathrm{T}}\mathrm{S}) \leq 1/(1 - \sqrt{k/r})$. Theorem 1 now follows by taking square roots of both sides

and using $\sqrt{1+x^2} \leq 1+x$. The running time is equal to the time needed to compute V_k and $\mathrm{V}_{\rho-k}$ plus the running time of the algorithm in Lemma 10. Finally, we note that the rescaling of the columns of C does not change the span of its columns and thus is irrelevant in the construction of $\Pi^2_{\mathrm{C},k}(\mathrm{A})$. ∎

Our next theorem describes a deterministic algorithm for spectral norm reconstruction that only needs to compute V_k and will serve as a prequel to the proof of Theorem 3. The accuracy guarantee of this theorem is essentially identical to the one in Theorem 1, with $\rho - k$ being replaced by n.

Theorem 12. *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ *and a target rank* $k < \rho$*, there exists a deterministic polynomial-time algorithm to select* $r > k$ *columns of* A *and form a matrix* $\mathrm{C} \in \mathbb{R}^{m\times r}$ *such that*

$$\|\mathrm{A}-\Pi^2_{\mathrm{C},k}(\mathrm{A})\|_2 \leq \|\mathrm{A}-\mathrm{A}_k\|_2 + \left(\frac{1+\sqrt{n/r}}{1-\sqrt{k/r}}\right)\|\mathrm{A}-\mathrm{A}_k\|_2.$$

The matrix C *can be computed in* $T_{\mathrm{V}_k} + O(nrk^2)$ *time, where* T_{V_k} *is the time needed to compute the top* k *right singular vectors of* A.

Proof: The proof is very similar to the proof of Theorem 1, so we only highlight the differences. First, apply the algorithm of Lemma 10 on the following two sets of vectors: the n rows of the matrix V_k and the n rows of the matrix I_n. The output of the algorithm is a sampling and rescaling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ (see discussion after Lemma 10 in Section 2.2). Let $\mathrm{C} = \mathrm{AS}$ and note that C consists of a subset of r *rescaled* columns of A. Lemma 10 guarantees that $\|\mathrm{I}_n\mathrm{S}\|_2 \leq 1+\sqrt{n/r}$. We now replicate the proof of Theorem 1 up to the point where $\|(\mathrm{A}-\mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_2^2$ is bounded. We continue as follows:

$$\begin{aligned}\|(\mathrm{A}-\mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_2^2 &= \|(\mathrm{A}-\mathrm{A}_k)\mathrm{I}_n\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_2^2 \\ &\leq \|(\mathrm{A}-\mathrm{A}_k)\|_2^2\|\mathrm{I}_n\mathrm{S}\|_2^2\|(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_2^2.\end{aligned}$$

The remainder of the proof now follows the same line as in Theorem 1. Again, the rescaling of the columns of C is irrelevant to the construction of $\Pi^2_{\mathrm{C},k}(\mathrm{A})$. To analyze the running time of the proposed algorithm, we need to look more closely at Lemma 10 and the related algorithm. The details are in [2], where we argue that the algorithm of Lemma 10 can be implemented in $O(nrk^2)$ time. The total running time is the time needed to compute V_k plus $O(nrk^2)$. ∎

Proof of Theorem 3. In order to prove Theorem 3 we will follow the proof of Theorem 1 using Lemma 8 (a fast matrix factorization) instead of Lemma 7 (the exact SVD of A). More specifically, instead of using the top k right singular vectors of A (the matrix V_k), we use the matrix $\mathrm{Z} \in \mathbb{R}^{n\times k}$ of Lemma 8. We now apply the algorithm of Lemma 10 on the following two sets of vectors: the n rows of the matrix Z and the n rows of the matrix I_n. The output of the algorithm is a sampling and rescaling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ (see discussion after Lemma 10 in ection 2.2). Let $\mathrm{C} = \mathrm{AS}$ and note that C consists of a subset of r *rescaled* columns of A. The proof of Theorem 3 is now identical to the proof of Theorem 12, except for using Lemma 6 instead of Lemma 7 in the first step of the proof:

$$\begin{aligned}\|\mathrm{A}-\Pi^2_{\mathrm{C},k}(\mathrm{A})\|_2^2 &\leq \|\mathrm{E}\|_2^2 + \|\mathrm{ES}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2^2 \\ &= \|\mathrm{E}\|_2^2 + \|\mathrm{EI}_n\mathrm{S}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2^2 \\ &\leq \|\mathrm{E}\|_2^2\left(1+\|\mathrm{I}_n\mathrm{S}\|_2^2\|(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2^2\right),\end{aligned}$$

where E is the residual error from the matrix factorization of Lemma 8. Taking square roots (using $\sqrt{1+x^2} \leq 1+x$) and using the bounds guaranteed by Lemma 10 for $\|\mathrm{I}_n\mathrm{S}\|_2$ and $\|(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2$, we obtain a bound in terms of $\|\mathrm{E}\|_2$. Finally, since E is a random variable, taking expectations and applying the bound of Lemma 8 concludes the proof of the theorem. Again, the rescaling of the columns of C is irrelevant to the construction of $\Pi^2_{\mathrm{C},k}(\mathrm{A})$. The running time is the time needed to compute the matrix Z from Lemma 8 plus an additional $O(nrk^2)$ time as in Theorem 12. ∎

Proof of Theorem 2. First, apply the algorithm of Lemma 11 on the following two sets of vectors: the n rows of the matrix V_k and the n rows of the matrix $(\mathrm{A}-\mathrm{A}_k)^{\mathrm{T}}$. The output of the algorithm is a sampling and rescaling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ (see discussion after Lemma 10 in Section 2.2). Let $\mathrm{C} = \mathrm{AS}$ and note that C consists of a subset of r *rescaled* columns of A. We follow the proof of Theorem 1 in the previous section up to the point where we need to bound the term $\|(\mathrm{A}-\mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_F^2$. By strong submultiplicativity,

$$\|(\mathrm{A}-\mathrm{A}_k)\mathrm{S}(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_F^2 \leq \|(\mathrm{A}-\mathrm{A}_k)\mathrm{S}\|_F^2\|(\mathrm{V}_k^{\mathrm{T}}\mathrm{S})^+\|_2^2.$$

To conclude, we apply Lemma 11 to bound the two terms in the right-hand side of the above inequality. Again, the rescaling of the columns of C is irrelevant to the construction of $\Pi^F_{\mathrm{C},k}(\mathrm{A})$. The running time of the proposed algorithm is equal to the time needed to compute V_k plus the time needed to compute $\mathrm{A}-\mathrm{A}_k$ (which is equal to $O(mnk)$ given V_k) plus the time needed to run the algorithm of Lemma 11, which is equal to $O\left(nrk^2+nm\right)$. ∎

Proof of Theorem 4. We will follow the proof of Theorem 2, but, as with the proof of Theorem 3, instead of using the top k left singular vectors of A (the matrix V_k), we will use the matrix Z of Lemma 9 that is computed via a fast, approximate matrix factorization. More specifically, let Z be the matrix of Lemma 9 and run the algorithm of Lemma 11 on the following two sets of vectors: the n rows of the matrix Z and the n rows of the matrix E^{T}. The output of the algorithm is a sampling and rescaling matrix $\mathrm{S} \in \mathbb{R}^{n\times r}$ (see discussion after Lemma 10 in Section 2.2). Let $\mathrm{C} = \mathrm{AS}$ and note that C consists of a subset of r *rescaled* columns of

A. The proof of Theorem 4 is now identical to the proof of Theorem 2, except for using Lemma 6 instead of Lemma 7. Ultimately, we obtain

$$\begin{aligned}\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\|_F^2 &\leq \|\mathrm{E}\|_F^2+\|\mathrm{ES}(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2^2\\ &\leq \|\mathrm{E}\|_F^2+\|\mathrm{ES}\|_F^2\|(\mathrm{Z}^{\mathrm{T}}\mathrm{S})^+\|_2^2\\ &\leq \left(1+\left(1-\sqrt{k/r}\right)^{-2}\right)\|\mathrm{E}\|_F^2.\end{aligned}$$

The last inequality follows from the bounds of Lemma 11. The theorem now follows by taking the expectation of both sides and using Lemma 9 to bound $\mathbf{E}[\|\mathrm{E}\|_F^2]$. Again, the rescaling of the columns of C is irrelevant to the construction of $\Pi^F_{\mathrm{C},k}(\mathrm{A})$. The overall running time is derived by replacing the time needed to compute V_k in Theorem 2 with the time needed to compute the fast approximate factorization of Lemma 9. ∎

Proof of Theorem 5. Finally, we will prove Theorem 5 by combining the results of Theorem 4 (a constant factor approximation algorithm) with one round of adaptive sampling. We first recall the following lemma, which has appeared in prior work [9], [14].

Lemma 13. *Given a matrix* $\mathrm{A}\in\mathbb{R}^{m\times n}$*, a target rank* k*, and an integer* r*, there exists an algorithm to select* r *columns from* A *to form the matrix* $\mathrm{C}\in\mathbb{R}^{m\times r}$ *such that*

$$\mathbf{E}\left[\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\|_F^2\right]\leq\|\mathrm{A}-\mathrm{A}_k\|_F^2+\frac{k}{r}\|\mathrm{A}\|_F^2.$$

The matrix C *can be computed in* $O(mn+r\log r)$ *time.*

Algorithms for the above lemma choose r columns of A in r independent identically distributed (i.i.d.) trials, where in each trial a column of A is sampled with probability proportional to its norm-squared (importance sampling). We now state Theorem 2.1 of [8], which builds upon Lemma 13 to provide an adaptive sampling procedure that improves the accuracy guarantees of Lemma 13.

Lemma 14. *Given a matrix* $\mathrm{A}\in\mathbb{R}^{m\times n}$*, let* $\mathrm{C}_1\in\mathbb{R}^{m\times r}$ *consist of* r *columns of* A*, and define the residual* $\mathrm{B}=\mathrm{A}-\mathrm{C}_1\mathrm{C}_1^+\mathrm{A}\in\mathbb{R}^{m\times n}$*. For* $i=1,\ldots,n$*, let*

$$p_i=\|\mathbf{b}_i\|_2^2/\|\mathrm{B}\|_F^2,$$

where $\mathbf{b}_i$ *is the* i*-th column of the matrix* B*. Sample a further* s *columns from* A *in* s *i.i.d. trials, where in each trial the* i*-th column is chosen with probability* p_i*. Let* $\mathrm{C}_2\in\mathbb{R}^{m\times s}$ *contain the* s *sampled columns and let* $\mathrm{C}=[\mathrm{C}_1\ \ \mathrm{C}_2]\in\mathbb{R}^{m\times(r+s)}$ *contain the columns of both* C_1 *and* C_2*, all of which are columns of* A*. Then, for any integer* $k>0$*,*

$$\mathbf{E}\left[\left\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\right\|_F^2\right]\leq\|\mathrm{A}-\mathrm{A}_k\|_F^2+\frac{k}{s}\|\mathrm{B}\|_F^2.$$

Note that Lemma 14 is an extension of Lemma 13; one can obtain Lemma 13 by setting C_1 to be empty in Lemma 14. We are now ready to prove Theorem 5. First, fix $d>1$ and define $c_0=(1+\epsilon_0)(1+1/(1-\sqrt{k/\hat{r}})^2)$, where $\hat{r}=\lceil dk\rceil$. (We will choose d and ϵ_0 later.) Now run the algorithm of Theorem 4 to sample $\hat{r}=\lceil dk\rceil$ columns of A and form the matrix C_1. Then, run the adaptive sampling algorithm of Lemma 14 with $\mathrm{B}=\mathrm{A}-\mathrm{C}_1\mathrm{C}_1^+\mathrm{A}$ and sample a further $s=\lceil c_0k/\epsilon\rceil$ columns of A to form the matrix C_2. Let $\mathrm{C}=[\mathrm{C}_1\ \ \mathrm{C}_2]\in\mathbb{R}^{n\times(\hat{r}+s)}$ contain all the sampled columns. We will analyze the expectation $\mathbf{E}\left[\left\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\right\|_F^2\right]$. Using the bound of Lemma 14, we first compute the expectation with respect to C_2 conditioned on C_1:

$$\mathbf{E}_{\mathrm{C}_2}\left[\left\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\right\|_F^2\middle|\,\mathrm{C}_1\right]\leq\|\mathrm{A}-\mathrm{A}_k\|_F^2+\frac{k}{s}\|\mathrm{B}\|_F^2.$$

We now compute the expectation with respect to C_1 (only B depends on C_1):

$$\mathbf{E}_{\mathrm{C}_1}\left[\mathbf{E}_{\mathrm{C}_2}\left[\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\|_F^2\middle|\,\mathrm{C}_1\right]\right]\leq$$

$$\|\mathrm{A}-\mathrm{A}_k\|_F^2+\frac{k}{s}\mathbf{E}_{\mathrm{C}_1}\left[\|\mathrm{A}-\mathrm{C}_1\mathrm{C}_1^+\mathrm{A}\|_F^2\right]. \qquad (1)$$

By the law of iterated expectation, the left hand side is exactly equal to $\mathbf{E}\left[\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\|_F^2\right]$. We now use the accuracy guarantee of Theorem 4 and our definition of c_0 to bound

$$\begin{aligned}\mathbf{E}_{\mathrm{C}_1}\left[\|\mathrm{A}-\mathrm{C}_1\mathrm{C}_1^+\mathrm{A}\|_F^2\right]&\leq\mathbf{E}_{\mathrm{C}_1}\left[\|\mathrm{A}-\Pi^F_{C_1,k}(\mathrm{A})\|_F^2\right]\\ &\leq c_0\|\mathrm{A}-\mathrm{A}_k\|_F^2.\end{aligned}$$

Using the bound in (1), we obtain

$$\mathbf{E}\left[\|\mathrm{A}-\Pi^F_{\mathrm{C},k}(\mathrm{A})\|_F^2\right]\leq\|\mathrm{A}-\mathrm{A}_k\|_F^2\left(1+c_0k/s\right).$$

Finally, recall that for our choice of s, $s\geq c_0k/\epsilon$, and so we obtain the relative error bound. The number of columns needed is $r=\hat{r}+s=dk+c_0k/\epsilon$. Set $d=(1+\alpha)^2$, where $\alpha=\sqrt[3]{(1+\epsilon_0)/\epsilon}$. After some algebra, this yields $r=k(\alpha^3+(1+\alpha)^3)=\frac{2k}{\epsilon}(1+O(\epsilon_0+\epsilon^{1/3}))$ sampled columns. The time needed to compute the matrix C is the sum of three terms: the running time of Theorem 4 (which is $O(mnk\epsilon_0^{-1}+n\hat{r}k^2)$), plus the time needed to compute $\mathrm{A}-\mathrm{C}_1\mathrm{C}_1^+\mathrm{A}$ (which is $O(mn\hat{r})$), plus the time needed to run the algorithm of Lemma 14 (which is $O(mn+s\log s)$). Assume $r<n$ (otherwise the problem is trivial), set $\epsilon_0=\epsilon^{2/3}$ and use $d=O(\epsilon^{-2/3})$ to get the final asymptotic run time. ∎

Comments. The number of columns required for relative error approximation is approximately $\frac{2k}{\epsilon}$, a 2-factor from optimal, since $\frac{k}{\epsilon}$ are needed ([7] and Section 5). We get a much better running time of $O(mnk+nk^3+n\log\epsilon^{-1})$ using just a constant factor more columns by setting d and ϵ_0 in the proof to constants (for example setting $d=100$; $\epsilon_0=\frac{62}{181}\approx\frac{1}{3}$ results in $\frac{3k}{\epsilon}(1+o(1))$ columns).

4. Matrix Pythagoras and the computation of $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$

4.1. Matrix norm properties

Recall notation from Section 1.1; for any matrix A of rank at most ρ, it is well-known that $\|\mathrm{A}\|_F^2 = \sum_{i=1}^{\rho} \sigma_i^2(\mathrm{A})$ and $\|\mathrm{A}\|_2 = \sigma_1(\mathrm{A})$. Also, the best rank k approximation to A satisfies $\|\mathrm{A} - \mathrm{A}_k\|_2 = \sigma_{k+1}(\mathrm{A})$ and $\|\mathrm{A} - \mathrm{A}_k\|_F^2 = \sum_{i=k+1}^{\rho} \sigma_i^2(\mathrm{A})$. For any two matrices A and B of appropriate dimensions, $\|\mathrm{A}\|_2 \leq \|\mathrm{A}\|_F \leq \sqrt{\rho}\|\mathrm{A}\|_2$, $\|\mathrm{AB}\|_F \leq \|\mathrm{A}\|_F\|\mathrm{B}\|_2$, and $\|\mathrm{AB}\|_F \leq \|\mathrm{A}\|_2\|\mathrm{B}\|_F$. The latter two properties are stronger versions of the standard submultiplicativity property.

We refer to the next lemma as matrix-Pythogoras:

Lemma 15. *If* $\mathrm{X}, \mathrm{Y} \in \mathbb{R}^{m\times n}$ *and* $\mathrm{XY}^{\mathrm{T}} = \mathbf{0}_{m\times m}$ *or* $\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \mathbf{0}_{n\times n}$*, then*

$$\|\mathrm{X}+\mathrm{Y}\|_F^2 = \|\mathrm{X}\|_F^2 + \|\mathrm{Y}\|_F^2,$$
$$\max\{\|\mathrm{X}\|_2^2, \|\mathrm{Y}\|_2^2\} \leq \|\mathrm{X}+\mathrm{Y}\|_2^2 \leq \|\mathrm{X}\|_2^2 + \|\mathrm{Y}\|_2^2.$$

Proof: Since $\mathrm{XY}^{\mathrm{T}} = \mathbf{0}_{m\times m}$, $(\mathrm{X}+\mathrm{Y})(\mathrm{X}+\mathrm{Y})^{\mathrm{T}} = \mathrm{XX}^{\mathrm{T}} + \mathrm{YY}^{\mathrm{T}}$. For $\xi = F$,

$$\|\mathrm{X}+\mathrm{Y}\|_F^2 = \mathrm{Tr}\left((\mathrm{X}+\mathrm{Y})(\mathrm{X}+\mathrm{Y})^{\mathrm{T}}\right) = \mathrm{Tr}\left(\mathrm{XX}^{\mathrm{T}} + \mathrm{YY}^{\mathrm{T}}\right) =$$
$$= \|\mathrm{X}\|_F^2 + \|\mathrm{Y}\|_F^2.$$

Let $\mathbf{z}$ be any vector in $\mathbb{R}^m$. For $\xi = 2$,

$$\|\mathrm{X}+\mathrm{Y}\|_2^2 = \max_{\|\mathbf{z}\|_2=1} \mathbf{z}^{\mathrm{T}}(\mathrm{X}+\mathrm{Y})(\mathrm{X}+\mathrm{Y})^{\mathrm{T}}\mathbf{z} =$$
$$= \max_{\|\mathbf{z}\|_2=1} \left(\mathbf{z}^{\mathrm{T}}\mathrm{XX}^{\mathrm{T}}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\mathbf{z}\right).$$

We have that $\max_{\|\mathbf{z}\|_2=1}\left(\mathbf{z}^{\mathrm{T}}\mathrm{XX}^{\mathrm{T}}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\mathbf{z}\right)$ is at most

$$\max_{\|\mathbf{z}\|_2=1} \mathbf{z}^{\mathrm{T}}\mathrm{XX}^{\mathrm{T}}\mathbf{z} + \max_{\|\mathbf{z}\|_2=1} \mathbf{z}^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\mathbf{z} = \|\mathrm{X}\|_2^2 + \|\mathrm{Y}\|_2^2$$

and that

$$\max_{\|\mathbf{z}\|_2=1} (\mathbf{z}^{\mathrm{T}}\mathrm{XX}^{\mathrm{T}}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\mathbf{z}) \geq \max_{\|\mathbf{z}\|_2=1} \mathbf{z}^{\mathrm{T}}\mathrm{XX}^{\mathrm{T}}\mathbf{z} = \|\mathrm{X}\|_2^2,$$

since $\mathbf{z}^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\mathbf{z}$ is non-negative for any vector $\mathbf{z}$. We get the same lower bound with $\|\mathrm{Y}\|_2^2$ instead, which means we can lower bound with $\max\{\|\mathrm{X}\|_2^2, \|\mathrm{Y}\|_2^2\}$. The case with $\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \mathbf{0}_{n\times n}$ can be proven similarly. ∎

4.2. Computing the best rank k approximation $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$

Let $\mathrm{A} \in \mathbb{R}^{m\times n}$, let $k < n$ be an integer, and let $\mathrm{C} \in \mathbb{R}^{m\times r}$ with $r > k$. Recall that $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A}) \in \mathbb{R}^{m\times n}$ is the best rank k approximation to A in the column space of C: We can write $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A}) = \mathrm{CX}^{\xi}$, where

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

In order to compute (or approximate) $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$ given A, C, and k, we will use the following algorithm:

1: Orthonormalize the columns of C in $O(mr^2)$ time to construct the matrix $\mathrm{Q} \in \mathbb{R}^{m\times r}$.
2: Compute $\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k \in \mathbb{R}^{r\times n}$ via SVD in $O(mnr + nr^2)$ – the best rank-k approximation of $\mathrm{Q}^{\mathrm{T}}\mathrm{A}$.
3: Return $\mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k \in \mathbb{R}^{m\times n}$ in $O(mnk)$ time.

Clearly, $\mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k$ is a rank k matrix that lies in the column span of C. Note that though $\Pi^{\xi}_{\mathrm{C},k}(\mathrm{A})$ can depend on ξ, our algorithm computes the same matrix, independent of ξ. The next lemma, which is essentially Lemma 4.3 in [4] together with a slight improvment of Theorem 9.3 in [18], proves that this algorithm computes $\Pi^{F}_{\mathrm{C},k}(\mathrm{A})$ and a constant factor approximation to $\Pi^{2}_{\mathrm{C},k}(\mathrm{A})$.

Lemma 16. *Given* $\mathrm{A} \in \mathbb{R}^{m\times n}$, $\mathrm{C} \in \mathbb{R}^{m\times r}$ *and an integer* k*, the matrix* $\mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k \in \mathbb{R}^{m\times n}$ *described above (where* Q *is an orthonormal basis for the columns of* C*) can be computed in* $O\left(mnr + (m+n)r^2\right)$ *time and satisfies:*

$$\|\mathrm{A} - \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k\|_F^2 = \|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2,$$
$$\|\mathrm{A} - \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k\|_2^2 \leq 2\|\mathrm{A} - \Pi^{2}_{\mathrm{C},k}(\mathrm{A})\|_2^2.$$

Proof: Our proof for the Frobenius norm case is a mild modification of the proof of Lemma 4.3 [4]. First, note that $\Pi^{F}_{\mathrm{C},k}(\mathrm{A}) = \Pi^{F}_{\mathrm{Q},k}(\mathrm{A})$, because $\mathrm{Q} \in \mathbb{R}^{m\times r}$ is an orthonormal basis for the column space of C. Thus,

$$\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2 = \|\mathrm{A} - \Pi^{F}_{\mathrm{Q},k}(\mathrm{A})\|_F^2 =$$
$$\min_{\Psi:\mathrm{rank}(\Psi)\leq k} \|\mathrm{A} - \mathrm{Q}\Psi\|_F^2.$$

Now, using matrix-Pythagoras and the orthonormality of Q,

$$\|\mathrm{A} - \mathrm{Q}\Psi\|_F^2 = \|\mathrm{A} - \mathrm{QQ}^{\mathrm{T}}\mathrm{A} + \mathrm{Q}(\mathrm{Q}^{\mathrm{T}}\mathrm{A} - \Psi)\|_F^2 =$$
$$\|\mathrm{A} - \mathrm{QQ}^{\mathrm{T}}\mathrm{A}\|_F^2 + \|\mathrm{Q}^{\mathrm{T}}\mathrm{A} - \Psi\|_F^2.$$

Setting $\Psi = (\mathrm{Q}^{\mathrm{T}}\mathrm{A})_k$ minimizes the above quantity over all rank-k matrices Ψ. Thus, combining the above results, $\|\mathrm{A} - \Pi^{F}_{\mathrm{C},k}(\mathrm{A})\|_F^2 = \|\mathrm{A} - \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k\|_F^2$.

We now proceed to the spectral-norm part of the proof, which combines ideas from Theorem 9.3 [18] and matrix-Pythagoras. We first manipulate $\|\mathrm{A} - \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k\|_2^2 =$

$$\begin{aligned} &= \|\mathrm{A} - \mathrm{QQ}^{\mathrm{T}}\mathrm{A} + \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A} - (\mathrm{Q}^{\mathrm{T}}\mathrm{A})_k\right)\|_2^2 \\ &\leq \|\mathrm{A} - \mathrm{QQ}^{\mathrm{T}}\mathrm{A}\|_2^2 + \|\mathrm{QQ}^{\mathrm{T}}\mathrm{A} - (\mathrm{QQ}^{\mathrm{T}}\mathrm{A})_k\|_2^2 \\ &\overset{(a)}{\leq} \|\mathrm{A} - \Pi^{2}_{\mathrm{Q},k}(\mathrm{A})\|_2^2 + \|\mathrm{A} - \mathrm{A}_k\|_2^2 \\ &\leq 2\|\mathrm{A} - \Pi^{2}_{\mathrm{Q},k}(\mathrm{A})\|_2^2. \end{aligned}$$

The first inequality follows from the simple fact that $\left(\mathrm{QQ}^{\mathrm{T}}\mathrm{A}\right)_k = \mathrm{Q}\left(\mathrm{Q}^{\mathrm{T}}\mathrm{A}\right)_k$ and matrix-Pythagoras; the first term in (a) follows because $\mathrm{QQ}^{\mathrm{T}}\mathrm{A}$ is the (unconstrained, not necessarily of rank at most k) best approximation to A in the column space of Q; the second term in (a) follows because QQ^{T} is a projector matrix and thus

$$\|\mathrm{QQ}^{\mathrm{T}}\mathrm{A} - (\mathrm{QQ}^{\mathrm{T}}\mathrm{A})_k\|_2^2 = \sigma_{k+1}^2(\mathrm{QQ}^{\mathrm{T}}\mathrm{A}) \leq \sigma_{k+1}^2(\mathrm{A}) = \|\mathrm{A} - \mathrm{A}_k\|_2^2.$$

The last inequality follows because

$$\|\mathrm{A}-\mathrm{A}_k\|_2^2 \le \|\mathrm{A}-\Pi_{\mathrm{Q},k}^2(\mathrm{A})\|_2^2.$$

■

5. Lower bounds

Theorem 17. *For any $\alpha > 0$, any $k \ge 1$, and any $r \ge 1$, there exists a matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ for which*

$$\frac{\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_2^2}{\|\mathrm{A}-\mathrm{A}_k\|_2^2} \ge \frac{n+\alpha^2}{r+\alpha^2}.$$

Here C is any matrix that consists of r columns of A. As $\alpha \to 0$, the lower bound is n/r for the approximation ratio of spectral norm column-based matrix reconstruction.

Proof: We extend the lower bound in [5] to arbitrary $r > k$. Consider the matrix

$$\mathrm{A} = [\mathbf{e}_1 + \alpha\mathbf{e}_2, \mathbf{e}_1 + \alpha\mathbf{e}_3, \ldots, \mathbf{e}_1 + \alpha\mathbf{e}_{n+1}] \in \mathbb{R}^{(n+1)\times n},$$

where $\mathbf{e}_i \in \mathbb{R}^{n+1}$ are the standard basis vectors. Then,

$$\mathrm{A}^{\mathrm{T}}\mathrm{A} = \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + \alpha^2\mathrm{I}_n, \qquad \sigma_1^2(\mathrm{A}) = n+\alpha^2, \qquad \text{and}$$
$$\sigma_i^2(\mathrm{A}) = \alpha^2 \quad \text{for} \quad i > 1.$$

Thus, for all $k \ge 1$, $\|\mathrm{A}-\mathrm{A}_k\|_2^2 = \alpha^2$. Intuitively, as $\alpha \to 0$, A is a rank-one matrix. Consider any r columns of A and note that, up to row permutations, all sets of r columns of A are equivalent. So, without loss of generality, let C consist of the first r columns of A. We now compute the optimal reconstruction of A from C as follows: let $\mathbf{a}_j$ be the j-th column of A. In order to reconstruct $\mathbf{a}_j$, we minimize $\|\mathbf{a}_j - \mathrm{C}\mathbf{x}\|_2^2$ over all vectors $\mathbf{x} \in \mathbb{R}^r$. Note that if $j \le r$ then the reconstruction error is zero. For $j > r$, $\mathbf{a}_j = \mathbf{e}_1 + \alpha\mathbf{e}_{j+1}$,

$$\mathrm{C}\mathbf{x} = \mathbf{e}_1\sum_{i=1}^r x_i + \alpha\sum_{i=1}^r x_i\mathbf{e}_{i+1}.$$

Then,

$$\begin{aligned}\|\mathbf{a}_j - \mathrm{C}\mathbf{x}\|_2^2 &= \|\mathbf{e}_1\left(\sum_{i=1}^r x_i - 1\right) + \alpha\sum_{i=1}^r x_i\mathbf{e}_{i+1} - e_{j+1}\|_2^2 \\ &= \left(\sum_{i=1}^r x_i - 1\right)^2 + \alpha^2\sum_{i=1}^r x_i^2 + 1.\end{aligned}$$

The above quadratic form in $\mathbf{x}$ is minimized when $x_i = \left(r+\alpha^2\right)^{-1}$ for all $i = 1, \ldots, r$. Let $\hat{\mathrm{A}} = \mathrm{A} - \mathrm{CC}^+\mathrm{A}$ and let the j-th column of $\hat{\mathrm{A}}$ be $\hat{\mathbf{a}}_j$. Then, for $j \le r$, $\hat{\mathbf{a}}_j$ is an all-zeros vector; for $j > r$, $\hat{\mathbf{a}}_j = \alpha\mathbf{e}_{j+1} - \frac{\alpha}{r+\alpha^2}\sum_{i=1}^r \mathbf{e}_{i+1}$. Thus,

$$\hat{\mathrm{A}}^{\mathrm{T}}\hat{\mathrm{A}} = \begin{bmatrix}\mathbf{0}_{r\times r} & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathrm{Z}\end{bmatrix},$$

where

$$\mathrm{Z} = \frac{\alpha^2}{r+\alpha^2}\mathbf{1}_{n-r}\mathbf{1}_{n-r}^{\mathrm{T}} + \alpha^2\mathrm{I}_{n-r}.$$

This immediately implies that

$$\begin{aligned}\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_2^2 &= \|\hat{\mathrm{A}}\|_2^2 = \|\hat{\mathrm{A}}^{\mathrm{T}}\hat{\mathrm{A}}\|_2 = \|\mathrm{Z}\|_2^2 \\ &= \frac{(n-r)\alpha^2}{r+\alpha^2} + \alpha^2 = \frac{n+\alpha^2}{r+\alpha^2}\alpha^2.\end{aligned}$$

This concludes our proof, because

$$\alpha^2 = \|A - A_k\|_2^2.$$

■

5.1. Frobenius norm approximation

Note that a lower bound for the ratio

$$\|\mathrm{A}-\Pi_{\mathrm{C},k}^{\xi}(\mathrm{A})\|_{\xi}^2/\|\mathrm{A}-\mathrm{A}_k\|_{\xi}^2,$$

does not imply a lower bound for the ratio

$$\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_{\xi}^2/\|\mathrm{A}-\mathrm{A}_k\|_{\xi}^2,$$

because

$$\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_{\xi}^2/\|\mathrm{A}-\mathrm{A}_k\|_{\xi}^2 \le \|\mathrm{A}-\Pi_{\mathrm{C},k}^{\xi}(\mathrm{A})\|_{\xi}^2/\|\mathrm{A}-\mathrm{A}_k\|_{\xi}^2.$$

Also, notice that Proposition 4 in [7] shows a lower bound $1 + k/2r$ for the ratio $\|\mathrm{A}-\Pi_{\mathrm{C},k}^{F}(\mathrm{A})\|_F^2/\|\mathrm{A}-\mathrm{A}_k\|_F^2$. For completeness, we extend the bound of [7] for the ratio $\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_F^2/\|\mathrm{A}-\mathrm{A}_k\|_F^2$; in fact, we obtain a lower bound which is asymptotically $1 + k/r$.

The matrix Z constructed in the previous proof is all we need. The trace of Z is the Frobenius norm of the residual error matrix in approximating A using any r columns. This gives the following lemma.

Lemma 18. *For any $\alpha > 0$ and $r \ge 1$, there exists a matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ for which*

$$\frac{\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_F^2}{\|\mathrm{A}-\mathrm{A}_1\|_F^2} \ge \frac{n-r}{n-1}\left(1 + \frac{1}{r+\alpha^2}\right).$$

Proof: In the proof of Theorem 17,

$$\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_F^2 = \mathrm{Tr}(\mathrm{Z}) = \alpha^2(n-r)(1 + \frac{1}{r+\alpha^2}),$$

and $\|\mathrm{A}-\mathrm{A}_1\|_F^2 = (n-1)\alpha^2$. ■

Now, construct a matrix with k copies of A along the diagonal. The size of each block is $\frac{n}{k}$. We sample r columns in total, with r_i from each block. Lemma 18 holds in each block, with n and r replaced by $\frac{n}{k}$ and r_i.

Theorem 19. *For any $\alpha > 0$, any $k \ge 1$, and any $r \ge 1$, there exists a matrix $\mathrm{A} \in \mathbb{R}^{m\times n}$ for which*

$$\frac{\|\mathrm{A}-\mathrm{CC}^+\mathrm{A}\|_F^2}{\|\mathrm{A}-\mathrm{A}_k\|_F^2} \ge \frac{n-r}{n-k}\left(1 + \frac{k}{r+\alpha^2}\right).$$

Here C is any matrix that consists of r columns of A. As $\alpha \to 0$ and $n \to \infty$ the lower bound is $1 + k/r$ for the approximation ratio of Frobenius norm column-based matrix reconstruction.

Proof: Let B be the block diagonal matrix with k copies of A along the diagonal (A is the matrix defined in the proof of Theorem 17). Let r_i be the number of columns selected in each block, $\sum_{i=1}^{k} r_i = r$. We can treat the Frobenius error in each block independently. Let Z_i be the error matrix in each block, as in the proof of Theorem 17. Then, using Lemma 18, the approximation error is

$$\begin{aligned} \|\mathrm{A} - \mathrm{CC}^{+}\mathrm{A}\|_F^2 &= \sum_{i=1}^{k} \mathrm{Tr}(\mathrm{Z}_i) \\ &= \alpha^2 \sum_{i=1}^{k} \left(\frac{n}{k} - r_i\right)\left(1 + \frac{1}{r_i + \alpha^2}\right). \end{aligned}$$

Minimizing this expression subject to the constraint that $\sum_{i=1}^{k} r_i = r$ gives $r_i = r/k$. The result follows after a little algebra using $\|\mathrm{A} - \mathrm{A}_k\|_F^2 = (n-k)\alpha^2$. ∎

6. Open Problems

Several interesting questions remain unanswered; we highlight two. First, is it possible to improve the running time of the deterministic algorithms of Lemmas 10 and 11? Recently, Zouzias [24] made progress in improving the running time of the spectral sparsification result of [1]; can we get a similar improvement for the 2-set algorithms presented here? Second, in the parlance of Theorem 5, is there a *deterministic* algorithm that selects $O(k/\epsilon)$ columns from A and guarantees relative-error accuracy for the error $\|\mathrm{A} - \Pi_{\mathrm{C},k}^{F}(\mathrm{A})\|_F^2$? In a very recent development, [17] partially answers this question by extending the volume sampling approach of [5] to deterministically select $\frac{k}{\epsilon}(1+o(1))$ columns and obtain a relative error bound for the term $\|\mathrm{A} - \mathrm{CC}^{+}\mathrm{A}\|_F^2$. Notice that it is not obvious if [17] implies a similar deterministic bound for the error $\|\mathrm{A} - \Pi_{\mathrm{C},k}^{F}(\mathrm{A})\|_F^2$.

Acknowledgment

We would like to thank A. Deshpande, D. Feldman, K. Varadarajan and J. Tropp for useful discussions. We also thank D. Feldman and K. Varadarajan for pointing out the connections between the subspace approximation line of research [6], [12], [13], [23] and ours. This work has been supported by two NSF CCF grants to Petros Drineas and Malik Magdon-Ismail.

References

[1] J.D. Batson, D.A. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proc. 41st STOC*, pages 255–262, 2009.

[2] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal Column-based Matrix Reconstruction. arXiv report: arxiv.org/abs/1103.0995, 2011.

[3] T.F. Chan and P.C. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13:727–741, 1992.

[4] K.L. Clarkson and D.P. Woodruff. Numerical linear algebra in the streaming model. In *Proc. 41st STOC*, pages 205–214, 2009.

[5] A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proc 42th STOC*, pages 329–338, 2010.

[6] A. Deshpande and K. R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proc. 39th STOC*, pages 641–650, 2007.

[7] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *RANDOM - APPROX*, 2006.

[8] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(12):225–247, 2006.

[9] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proc. 10th SODA*, pages 291–299, 1999.

[10] P. Drineas, I. Kerenidis, and P. Raghavan. Competitive recommendation systems. In *Proc 34th STOC*, pages 82–90, 2002.

[11] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Polynomial time algorithm for column-row based relative-error low-rank matrix approximation. Technical Report 2006-04, DIMACS, March 2006.

[12] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proc. 43rd STOC*, 2011.

[13] D. Feldman, M. Monemizadeh, C. Sohler, and D. Woodruf. Coresets and sketches for high dimensional subspace approximation problems. In *Proc. 21st SODA*, pages 630–649, 2010.

[14] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Proc. 39th FOCS*, pages 370–378, 1998.

[15] G. Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7:206–216, 1965.

[16] M. Gu and S.C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17:848–869, 1996.

[17] V. Guruswami and A. K. Sinop. Optimal column-based low-rank matrix reconstruction. arXiv report: arXiv:1104.1732v1, http://arxiv.org/abs/1104.1732, April 09 2011.

[18] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 2011.

[19] E. Liberty, F. Woolfe, P.G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

[20] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. In *Proc. PNAS,*, 106:697–702, 2009.

[21] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

[22] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th FOCS*, pages 143–152, 2006.

[23] N. D. Shyamalkumar and K. R. Varadarajan. Efficient subspace approximation algorithms. In *Proc. 18th SODA*, pages 532–540, 2007.

[24] A. Zouzias. arXiv report: arXiv:1103.2793v1, http://arxiv.org/abs/1103.2793.