Published in final edited form as:

Proc IEEE Int Congr Big Data. 2019 December ; 2019: 26–35. doi:10.1109/bigdata47090.2019.9006345.

SketchyCoreSVD: SketchySVD from Random Subsampling of the Data Matrix

Chandrajit Bajaj, Yi Wang, Tianming Wang

The University of Texas at Austin, Austin, TX 78712, USA

Abstract

We present a method called SketchyCoreSVD to compute the near-optimal rank r SVD of a data matrix by building random sketches only from its subsampled columns and rows. We provide theoretical guarantees under incoherence assumptions, and validate the performance of our SketchyCoreSVD method on various large static and time-varying datasets.

Keywords

Low rank; SVD; sketching; sampling

I. INTRODUCTION

Data matrices in practice often have fast decaying spectrum. Finding low rank approximation to a data matrix is a fundamental task in numerous applications. Due to the massive size of data in modern applications, it may not be possible to store the full data matrix, causing problems for classic solvers.

It turns out that randomized sketches of the data matrix, built from dimension reduction maps, suffice for low-rank approximation. The first one-pass algorithm based on random sketches appear in [1]. Such idea is further advocated in the survey article [2]. Ever since, quite a few algorithms [3]–[6] based on random sketches have been developed. A tabular summary description of prior sketch building algorithms is provided in Appendix A.

The most recent paper called SketchySVD [7] is shown to provide consistent and practical approximations compared to the predecessors. For a data matrix A , SketchySVD finds the basis for the column/row space by QR decomposition of the right/left sketch while the first k singular values are well preserved in the core sketch. Theoretical guarantees have been derived for sketches built from Gaussian maps. The more practical dimension reduction maps such as Scrambled Subsampled Randomized Fourier Transform (SSRFT) [1] and sparse sign matrices [8], [9] are shown to exhibit similar performances in practice.

Building upon SketchySVD, we would like to claim that there is redundancy in the data matrix that can be further exploited when constructing sketches, thereby reducing the computation cost without sacrificing approximation accuracy. Such redundancy is

characterized by incoherence (see Definition 1 in Section II-B), which is widely used in compressed sensing and matrix recovery. Several papers [10]–[12] also establish guarantees for Nyström method [13] based on the incoherence assumption. Overall, our contributions are the following: 1) Propose SketchyCoreSVD, a method to compute SketchySVD from subsampled columns and rows of the data matrix; 2) Establish theoretical guarantees based on incoherence assumptions of the data matrix; 3) Verify through experiments that SketchyCoreSVD is able to reduce the computation cost of SketchySVD without sacrificing approximation accuracy.

II. ALGORITHM

A. SketchyCoreSVD

Suppose the data matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is huge in size, and its spectrum is decaying fast enough for compression. Let r be a suitable choice of rank. The main goal is to find a near-optimal rank r approximation to $\mathbf{A}$, via dimension reduction maps applied on subsampled columns and rows of $\mathbf{A}$. Choose sketch sizes k and s such that $r \leq k \leq s \leq \min\{m, n, m', n'\}$, where m, n, m' and n' satisfy

$$\frac{m}{M} = \frac{n}{N} = p < 1, \quad \frac{m'}{M} = \frac{n'}{N} = q < 1, \quad \text{and} \quad p \leq q.$$

Inspired by SketchySVD [7], the steps of our method, termed SketchyCoreSVD, are the following.

Step 1: Building sketches.

- 1) Uniformly sample (without replacement) m rows of $\mathbf{A}$. Denote the indices of the sampled rows to be Δ . The map $\mathbf{\Gamma} \in \mathbb{R}^{k \times m}$ is applied to $\mathbf{A}^{(\Delta, :)}$ from the left,

$$\mathbf{X} = \mathbf{\Gamma} \mathbf{A}^{(\Delta, :)} \in \mathbb{R}^{k \times N};$$

- 2) Uniformly sample (without replacement) n columns of $\mathbf{A}$. Denote the indices of the sampled columns to be Θ . The map $\mathbf{\Omega} \in \mathbb{R}^{k \times n}$ is applied to $\mathbf{X}^{(:, \Theta)}$ from the right,

$$\mathbf{Y} = \mathbf{X}^{(:, \Theta)} \mathbf{\Omega}^* \in \mathbb{R}^{M \times k};$$

- 3) Uniformly sample (without replacement) m' row indices and n' column indices, denoted by Δ' and Θ' , respectively. Apply random matrix maps $\mathbf{\Phi} \in \mathbb{R}^{s \times m'}$ and $\mathbf{\Psi} \in \mathbb{R}^{s \times n'}$ to the intersection of $\mathbf{Y}^{(\Delta', \Theta')} \in \mathbb{R}^{m' \times n'}$,

$$\mathbf{Z} = \mathbf{\Phi} \mathbf{Y}^{(\Delta', \Theta')} \mathbf{\Psi}^* \in \mathbb{R}^{s \times s}.$$

Step 2: Computations.

- 1) Compute the QR decomposition of $\mathbf{Z}^*$,

$$X^* = PR_1,$$

where $P \in \mathbb{R}^{N \times k}$, and $R_1 \in \mathbb{R}^{k \times k}$,

- 2) Compute the QR decomposition of Y ,

$$Y = QR_2,$$

where $Q \in \mathbb{R}^{M \times k}$, and $R_2 \in \mathbb{R}^{k \times k}$,

- 3) Compute the core approximation,

$$C = (\Phi Q^{(\Delta', :)})^\dagger \cdot Z \cdot ((\Psi P^{(\Theta', :)})^\dagger)^* \in \mathbb{R}^{k \times k},$$

- 4) Denote $\hat{A} = QCP^*$ as the initial approximation. Let $[[C]]_r$ be the best rank r approximation to C . The final near-optimal rank r approximation to A is $[[\hat{A}]]_r$, which is equal to

$$Q[[C]]_r P^*.$$

While the storage cost is the same, SketchyCoreSVD has reduced computational cost. For Gaussian maps, the (dominated) flops count of the two methods are shown in Table I. If the same k and s are used, SketchyCoreSVD has lower complexity. The cost is reduced mainly when constructing the sketches, depending on the sampling ratios p and q .

B. Theoretical Guarantees

We generally follow the outline of [7], with substantial differences when dealing with subsamples of the data matrix. The first step is to prove that Q and P capture the range and co-range of A , i.e.,

$$A \approx QQ^*A, \quad A \approx APP^*.$$

These are proved in [2] for Q and P computed from the sketches of the full matrix. For our case, only the randomly selected subsets of the columns/rows are used. Thus we must impose certain conditions on the columns/rows of A .

The intuition is that the columns/rows are “more or less the same”. Such a property can be characterized by incoherence. Suppose $A \in \mathbb{R}^{M \times N}$ has an SVD of the following form

$$U\Sigma V^* = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^*,$$

where $[[A]]_r = U_1 \Sigma_1 V_1^*$ is the best rank r approximation of A .

Definition 1.— $[[A]]_r \in \mathbb{R}^{M \times N}$ is (μ, ν) -incoherent if

$$\max_i \left\{ \|\mathbf{U}_1^{(i, :)}\|_2 \right\} \leq \sqrt{\frac{\mu r}{M}}, \quad \max_j \left\{ \|\mathbf{V}_1^{(:, j)}\|_2 \right\} \leq \sqrt{\frac{\nu r}{N}}.$$

Based on the incoherence assumption, we provide the following error bound.

Theorem 1.—Suppose $\mathbf{A} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* + \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2^*$, where $[[\mathbf{A}]]_r = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$, the best rank r approximation of $\mathbf{A}$, is (μ, ν) -incoherent. Then for our SketchyCoreSVD algorithm with $k \geq r + 4$, $m \geq 8\mu r \log r$, and $n \geq 8\nu r \log r$,

$$\begin{aligned} & \max \{ \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F, \|\mathbf{A} - \mathbf{A}\mathbf{P}\mathbf{P}^*\|_F \} \\ & \leq (C_1(p, k, r) + 1) \cdot \|\mathbf{\Sigma}_2\|_F + C_2(p, k, r) \cdot \|\mathbf{\Sigma}_2\|_2 \end{aligned}$$

with probability at least $1 - \frac{4}{r^3} - \frac{4}{k^3}$, where

$$C_1(p, k, r) = \sqrt{\frac{6e^2}{p}} \cdot \frac{k}{k-r+1} \cdot \frac{3}{k-r+1},$$

$$C_2(p, k, r) = \sqrt{\frac{36e^2}{p}} \cdot \frac{\sqrt{k \log k}}{k-r+1} \cdot \frac{3}{k-r+1}.$$

Remark.—Note that $C_1(p, k, r)$ and $C_2(p, k, r)$ both decrease as p or k increases. Thus we are advised to use a not too small sampling ratio p , and a bigger sketch size k whenever possible.

The probability of success depends on r and k . Such dependency is in part due to the need to apply Lemma 1 to bound the smallest singular values of some subsampled orthonormal matrices with small row norms.

Let $\mathbf{Q}$ and $\mathbf{P}$ be the basis computed by SketchyCoreSVD for the columns and rows, respectively. There exist $\mu' \in \left[1, \frac{M}{k}\right]$ and $\nu' \in \left[1, \frac{N}{k}\right]$ such that

$$\max_i \left\{ \|\mathbf{Q}^{(i, :)}\|_2 \right\} \leq \sqrt{\frac{\mu' k}{M}}, \quad \max_j \left\{ \|\mathbf{P}^{(:, j)}\|_2 \right\} \leq \sqrt{\frac{\nu' k}{N}}.$$

The existence of μ' and ν' can be easily shown. For example, consider μ' . Since $\max_i \left\{ \|\mathbf{Q}^{(i, :)}\|_2^2 \right\} \leq 1$, it follows that $\mu' \leq \frac{M}{k}$. If $\mu' < 1$, then $\|\mathbf{Q}\|_F^2 < M \cdot \frac{k}{M} = k$, contradicting to the fact that $\|\mathbf{Q}\|_F^2 = k$.

We derive the following guarantee for the rank r approximation $[[\hat{\mathbf{A}}]]_r$ computed by SketchyCoreSVD, provided that m' and n' are greater than some multiples of $\mu' k \log k$ and $\nu' k \log k$, respectively.

Theorem 2.—*Condition on the success of Theorem 1, for our SketchyCoreSVD algorithm with $s = k + 4$, $m' = 8\mu' k \log k$, and $n' = 8\nu' k \log k$, the initial approximation $\hat{\mathbf{A}}$ satisfies*

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_F \leq C_3(p, q, s, k, r) \cdot \|\Sigma_2\|_F + C_4(p, q, s, k, r) \cdot \|\Sigma_2\|_2$$

with probability at least $1 - \frac{4}{k^3} - \frac{6}{s^3}$, where

$$\begin{aligned} C_3(p, q, s, k, r) &= C_1(p, k, r) \cdot (\sqrt{3}C(q, s, k) + \sqrt{2}), \\ C_4(p, q, s, k, r) &= C_2(p, k, r) \cdot (\sqrt{3}C(q, s, k) + \sqrt{2}), \end{aligned}$$

and $C(q, s, k)$ is defined as

$$\frac{6e^2}{q} \cdot \frac{s^1 + 6/(s - k + 1)}{(s - k + 1)^2} \cdot (\sqrt{s} + \sqrt{6 \log s})^2.$$

Moreover, the final rank r approximation satisfies

$$\|\mathbf{A} - [[\hat{\mathbf{A}}]]_r\|_F \leq (2C_3(p, q, s, k, r) + 1) \cdot \|\Sigma_2\|_F + (2C_4(p, q, s, k, r)) \cdot \|\Sigma_2\|_2.$$

Remark.—*Note that $C(q, s, k)$ decreases as q or s increases. Thus we are advised to use a not too small sampling ratio q , and a bigger sketch size s whenever possible.*

We admit the dependence on μ' and ν' is the deficiency of our current theoretical guarantees. Ideally, one should derive bounds for μ' and ν' based on μ and ν . One way to get around this is to slightly modify the algorithm: after building the left and right sketches, calculate the QR factorizations, and sample m' row indices and n' column indices based on the actual incoherence parameters estimated from the row norms of $\mathbf{Q}$ and $\mathbf{P}$, respectively.

Nevertheless, we observe that $\mu' = O(\mu)$, $\nu' = O(\nu)$ in practice. See Table VII for some empirical evidences. In the experiments shown in Section III, we simply choose $m' = m$ and $n' = n$, i.e., $q = p$.

III. NUMERICAL EXPERIMENTS

If full spectrum can be obtained, the guide to choose the proper rank is to find the “knee” in the scree plot. At rank r ,

$$screer(r) = \frac{1}{\|\mathbf{A}\|_F^2} \sum_{i=r+1}^n \sigma_i^2(\mathbf{A}).$$

For the rank r approximation $[[\hat{\mathbf{A}}]]_r$ to $\mathbf{A}$, we measure its approximation error as

$$err = \frac{\|[[\hat{\mathbf{A}}]]_r - \mathbf{A}\|_F^2}{\|\mathbf{A}\|_F^2}.$$

We compare our SketchyCoreSVD method to SketchySVD [7], which suggests one to choose $s = 2k + 1$ and $k = \mathcal{O}(r)$. The dimension reduction maps used are Gaussian maps. For our SketchyCoreSVD, we choose the sampling ratio p such that $p \cdot \min\{M, N\} = s$, and $q = p$.

The experiments are executed from MATLAB R2018a on a 64-bit Linux machine with 8 Intel i7-7700 CPUs at 3.60 GHz and 32 GB of RAM. The reported errors are averaged over 20 trials. In the experiments, we first load the data into memory and then build the sketches. The computation time reported does not include computing the final approximation $[[\hat{\mathbf{A}}]]_r$ since in practice it is usually stored in the factorized form for subsequent tasks. The need to compute $[[\hat{\mathbf{A}}]]_r$ is only for the error metric. We also provide visual comparisons on the first few left singular vectors.

Yale Face Dataset.

The data is originally from the Yale Face Database B [14]. For this dataset, $\mathbf{A} \in \mathbb{R}^{2500 \times 640}$ since there are 640 face images of size 50×50 . Based on the scree plot in Figure 1, we choose $r = 20$. The optimal err is 0.033.

We choose $k = 4r + 1 = 81$ and $s = 2k + 1 = 163$ for both methods. In Table II, we see that the error of SketchySVD is about twice the optimal error. Our SketchyCoreSVD can achieve about the same error bound with less time. As the sampling ratio p increases from 30% to 40%, the err decreases while computation time gradually increases. The visual comparison in Figure 2 shows that for SketchyCoreSVD, the singular vector(s) can be estimated accurately with sampling ratio $p = 40\%$.

Cardiac Magnetic Resonance Imaging Dataset.

We have a collection of time-varying, 2D slice stacks of cardiac magnetic resonance images (Cardiac MRI). We select a 2D slice and consider a time-sequence of this in the form of a data matrix $\mathbf{A} \in \mathbb{R}^{45056 \times 160}$ (i.e., 160 time snapshots, each of size 256×176). Based on the scree plot in Figure 3, we choose $r = 5$. The optimal err is 0.0011.

We choose $k = 4r + 1 = 21$ and $s = 2k + 1 = 43$ for both methods. In Table III, we see that the error of SketchySVD is less than twice the optimal error. Our SketchyCoreSVD can achieve the same error bound in less time. As sampling ratio p increases from 30% to 40%, err decreases while computation time gradually increases. The visual comparison in Figure 4

shows that for SketchyCoreSVD, the singular vectors can be estimated accurately with sampling ratio $p = 40\%$.

The BR1003 Breast Cancer Dataset.

The data is extracted from a collection of Fourier-transform infrared spectroscopy (FTIR) spectral signatures of breast tissues [15]. $\mathbf{A} \in \mathbb{R}^{1506 \times 783090}$ contains 789030 spectral signatures, each of length 1506. Based on the scree plot in Figure 5, we choose $r = 6$. The optimal err is 0.002.

We choose $k = 4r + 1 = 25$ and $s = 2k + 1 = 51$ for both methods. In Table IV, we see that the error of SketchySVD is almost the same as the optimal error. Our SketchyCoreSVD can achieve about the same error bound in less than half of the time. As the sampling ratio p increases from 4% to 8%, the err decreases while the computation time gradually increases. The visual comparison in Figure 6 and the SNR values in Table V show that for SketchyCoreSVD, the singular vectors can be estimated accurately with sampling ratio $p = 8\%$.

Video Dataset.

This dataset is a color video of size $1080 \times 1920 \times 3 \times 2498$. It was originally used by [16] to test low rank tensor approximations. We convert the video into to grayscale, reduce the spatial size by a factor of 2, and discard the first 100 and last 198 frames due to camera movement. The resulting data matrix $\mathbf{A} \in \mathbb{R}^{518400 \times 2200}$. Based on the scree plot in Figure 7, we choose $r = 25$. The optimal err is 0.0066.

We choose $k = 4r + 1 = 101$ and $s = 2k + 1 = 203$ for both methods. In Table VI, we see that the error of SketchySVD is about twice the optimal error. SketchyCoreSVD can achieve about the same error bound in less than half of the time. As the sampling ratio p increases from 10% to 20%, the err decreases while the computation time gradually increases. The visual comparison in Figure 8 shows that for SketchyCoreSVD, the singular vectors can be estimated accurately with sampling ratio $p = 20\%$.

IV. PROOFS

A. Proof of Theorem 1

The following proofs are for the approximation of the column space of $\mathbf{A}$, i.e., $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\|_F$. Similar argument holds for $\|\mathbf{A} - \mathbf{A}\mathbf{P}\mathbf{P}^*\|_F$. We first state a few auxiliary lemmas to facilitate the main proof.

Lemma 1, adapted from [17, Lemma 3.4], is essential for our proofs since it provides a lower bound on the singular values of a submatrix from randomly sampled rows of an “incoherent” orthonormal matrix.

Lemma 1.—*Suppose $\max_j \{\|\mathbf{V}_1^{(j, \cdot)}\|_2\} \leq \sqrt{\frac{\mu r}{N}}$. For $\alpha > 0$, select the sample size*

$$n \geq 8\mu r \log r.$$

Draw a random subset Θ from $\{1, \dots, N\}$ by sampling n coordinates without replacement. Then with probability at least $1 - \frac{2}{r^3}$,

$$\sqrt{\frac{n}{6N}} \leq \sigma_r(V_1^{(\Theta, :)}) \quad \text{and} \quad \sigma_1(V_1^{(\Theta, :)}) \leq \sqrt{\frac{13n}{6N}}.$$

If we further apply Gaussian maps to the subsampled rows, Lemma 2 ensures that the full rankness is preserved almost surely.

Lemma 2.—Suppose $V_1^{(\Theta, :)} \in \mathbb{R}^{n \times r}$ is of full column rank, and $\Omega \in \mathbb{R}^{k \times n}$ is standard normal Gaussian. Then $\Omega V_1^{(\Theta, :)} \in \mathbb{R}^{k \times r}$ is of full column rank almost surely.

Proof: Denote the truncated SVD of $V_1^{(\Theta, :)}$ by $W_1 \Lambda W_2^*$, where

$W_1 \in \mathbb{R}^{n \times r}$, $\Lambda \in \mathbb{R}^{r \times r}$, and $W_2 \in \mathbb{R}^{r \times r}$. W_1 is orthonormal, $\Omega W_1 \in \mathbb{R}^{k \times r}$ is standard Gaussian. Thus with probability one, ΩW_1 is of full column rank. Consequently $\Omega V_1^{(\Theta, :)} = (\Omega W_1) \Lambda W_2^*$ is also of full column rank. ■

From Lemma 2, we know that $(V_1^*)^{(:, \Theta)} \Omega^* \in \mathbb{R}^{r \times k}$ is of full row rank. Define $\Omega_1 = (V_1^*)^{(:, \Theta)} \Omega^*$, and $\Omega_2 = (V_2^*)^{(:, \Theta)} \Omega^*$. Lemma 3 below provides a deterministic bound of $\|A - QQ^*A\|_F^2$. It can be proved similarly as [2, Theorem 9.1]. We include its proof for completeness, and defer it to Section IV-B.

Lemma 3.—Assuming that $\Omega_1 \in \mathbb{R}^{r \times k}$ has full row rank, and $Q \in \mathbb{R}^{M \times k}$ is from the QR decomposition of $A^{(:, \Theta)} \Omega$. Then

$$\|A - QQ^*A\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2, \quad (1)$$

where Σ_2 is the diagonal matrix containing the $(r+1)$ to $\min\{M, N\}$ singular values of A .

For the bound in (1), Ω_2 and Ω_1 are not independent in our case, contrary to what [7] deals with. Our strategy is to separately bound $\Sigma_2 \Omega_2$ and $\Omega_1^\dagger$. For $\Omega_1^\dagger$, we need to deal with the operator norm of the pseudo-inverse of random Gaussian at some point. Lemma 4, adapted from [2, Proposition 10.3], provides such a bound.

Lemma 4.—For $G \in \mathbb{R}^{k \times r}$ ($k \geq r+4$) being standard normal Gaussian,

$$\mathbb{P}\left\{\|G^\dagger\|_2 \geq \frac{e\sqrt{k}}{k-r+1} \cdot t\right\} \leq t^{-(k-r+1)}, \quad \forall t \geq 1.$$

Taking $t = \frac{3}{k-k-r+1}$, then

$$\|\mathbf{G}^\dagger\|_2 \leq \frac{e\sqrt{k}}{k-r+1} \cdot \frac{3}{k-k-r+1}$$

with probability at least $1 - \frac{1}{k^3}$.

We bound $\|\Sigma_2 \Omega_2\|_F$ with the help of Lemma 5 and Lemma 6.

Lemma 5.—([2, Proposition 10.1]) Fix matrices $\mathbf{S}$, $\mathbf{T}$, and draw a standard Gaussian matrix $\mathbf{G}$. Then

$$\mathbb{E}(\|\mathbf{SGT}\|_F^2) = \|\mathbf{S}\|_F^2 \|\mathbf{T}\|_F^2.$$

Lemma 6.—([2, Proposition 10.3]) Suppose h is a Lipschitz function on matrices:

$$|h(\mathbf{X}) - h(\mathbf{Y})| \leq L \|\mathbf{X} - \mathbf{Y}\|_F, \quad \forall \mathbf{X}, \mathbf{Y}.$$

Draw a standard Gaussian matrix $\mathbf{G}$. Then

$$\mathbb{P}\left\{h(\mathbf{G}) \geq \mathbb{E}(h(\mathbf{G})) + Lt\right\} \leq e^{-t^2/2}.$$

We are now ready to present the main proof of Theorem 1.

$$\begin{aligned} & \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F^2 \\ (a) & \leq \|\Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2 \\ & = \|\Sigma_2\|_F^2 + \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^* \Omega_1^\dagger\|_F^2 \\ & = \|\Sigma_2\|_F^2 + \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^* ((\mathbf{V}_1^*)^{(:,\Theta)} \Omega^*)^\dagger\|_F^2 \\ (b) & = \|\Sigma_2\|_F^2 + \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^* (\mathbf{W}_2 \Lambda \mathbf{W}_1^* \Omega^*)^\dagger\|_F^2 \\ (c) & = \|\Sigma_2\|_F^2 + \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^* (\mathbf{W}_1^* \Omega^*)^\dagger \Lambda^\dagger \mathbf{W}_2^\dagger\|_F^2 \\ & \leq \|\Sigma_2\|_F^2 + \|(\mathbf{W}_1^* \Omega^*)^\dagger \Lambda^\dagger \mathbf{W}_2^\dagger\|_2^2 \cdot \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^*\|_F^2 \\ & \leq \|\Sigma_2\|_F^2 + \|\Lambda^\dagger\|_2^2 \cdot \|(\mathbf{W}_1^* \Omega^*)^\dagger\|_2^2 \cdot \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^*\|_F^2 \\ (d) & \leq \|\Sigma_2\|_F^2 + \frac{6N}{n} \cdot \|(\mathbf{W}_1^* \Omega^*)^\dagger\|_2^2 \cdot \|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^*\|_F^2 \\ (e) & \leq \|\Sigma_2\|_F^2 + \frac{6N}{n} \cdot \frac{e^{2k} + 6/(k-r+1)}{(k-r+1)^2} \cdot \underbrace{\|\Sigma_2(\mathbf{V}_2^*)^{(:,\Theta)} \Omega^*\|_F^2}_{T_1}. \end{aligned}$$

where (a) holds with probability at least $1 - \frac{2}{k^3}$, by applying Lemma 1, Lemma 2, and

Lemma 3; in (b), we assume $\mathbf{V}_1^{(\Theta, :)} = \mathbf{W}_1 \Lambda \mathbf{W}_2^*$ is the SVD (of rank r); (c) is due to the

property of pseudo-inverse; (d) is due to the lower bound of the singular value in Lemma 1; (e) holds with probability at least $1 - \frac{1}{k^3}$, by applying Lemma 4 to $\mathbf{\Omega W}_1$.

Bound for T_1 . By Lemma 5,

$$\begin{aligned} & \mathbb{E}(\|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{\Omega}^* \|_F^2) \\ &= \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)}\|_F^2 \cdot \|\mathbf{I}_k\|_F^2 \leq k \|\mathbf{\Sigma}_2\|_F^2, \end{aligned}$$

where $\mathbf{I}_k$ is the identity matrix of size k , and the last inequality is due to $\|(\mathbf{V}_2^*)^{(\cdot, \Theta)}\|_2 = \|\mathbf{V}_2^{(\Theta, \cdot)}\|_2 \leq 1$. Consider the function $h(\mathbf{X}) = \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{X}\|_F$,

$$\begin{aligned} & |h(\mathbf{X}) - h(\mathbf{Y})| \\ &= \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{X}\|_F - \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{Y}\|_F \\ &\leq \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)}(\mathbf{X} - \mathbf{Y})\|_F \\ &\leq \|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)}\|_2 \cdot \|\mathbf{X} - \mathbf{Y}\|_F \leq \|\mathbf{\Sigma}_2\|_2 \cdot \|\mathbf{X} - \mathbf{Y}\|_F. \end{aligned}$$

Thus the Lipschitz constant is $\|\mathbf{\Sigma}_2\|_2$. Now by Lemma 6, with probability at least $1 - e^{-t^2/2}$,

$$\|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{\Omega}^* \|_F \leq \sqrt{k} \|\mathbf{\Sigma}_2\|_F + \|\mathbf{\Sigma}_2\|_2 \cdot t.$$

A particular choice of $t = \sqrt{6 \log k}$ gives

$$\|\mathbf{\Sigma}_2(\mathbf{V}_2^*)^{(\cdot, \Theta)} \mathbf{\Omega}^* \|_F \leq \sqrt{k} \|\mathbf{\Sigma}_2\|_F + \sqrt{6 \log k} \|\mathbf{\Sigma}_2\|_2$$

with probability at least $1 - \frac{1}{k^3}$.

Based on the derived bound for T_1 , and the inequality $\sqrt{a^2 + b^2} \leq a + b$ for $a, b \geq 0$,

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q Q}^* \mathbf{A}\|_F &\leq (C_1(p, k, r) + 1) \cdot \|\mathbf{\Sigma}_2\|_F \\ &\quad + C_2(p, k, r) \cdot \|\mathbf{\Sigma}_2\|_2, \end{aligned}$$

where

$$C_1(p, k, r) = \sqrt{\frac{6e^2}{p}} \cdot \frac{k}{k-r+1} \cdot k \frac{3}{k-r+1},$$

$$C_2(p, k, r) = \sqrt{\frac{36e^2}{p}} \cdot \frac{\sqrt{k \log k}}{k-r+1} \cdot k \frac{3}{k-r+1}.$$

B. Proof of Lemma 3

First, the left unitary factor U plays no essential role. To see this, define

$$\tilde{A} = U^* A = \begin{bmatrix} \Sigma_1 V_1^* \\ \Sigma_2 V_2^* \end{bmatrix}, \quad \tilde{Y} = \tilde{A}(:, \Theta) \Omega^* = \begin{bmatrix} \Sigma_1 \Omega_1 \\ \Sigma_2 \Omega_2 \end{bmatrix}.$$

Denote P_Y as the projection to the column space of Y . We have

$$\begin{aligned} \|(I - P_Y)A\|_F &= \|(I - QQ^*)A\|_F \\ &= \|U^*(I - QQ^*)U\tilde{A}\|_F \\ &= \|(I - (U^*Q)(U^*Q)^*)\tilde{A}\|_F \\ &= \|(I - P_{\tilde{Y}})\tilde{A}\|_F. \end{aligned}$$

It suffices to prove that

$$\|(I - P_{\tilde{Y}})\tilde{A}\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2.$$

Second, we assume that Σ_2 is not zero matrix. Otherwise,

$$\begin{aligned} \text{range}(\tilde{A}) &= \text{range}\left(\begin{bmatrix} \Sigma_1 V_1^* \\ 0 \end{bmatrix}\right) \\ &= \text{range}\left(\begin{bmatrix} \Sigma_1 \Omega_1 \\ 0 \end{bmatrix}\right) = \text{range}(\tilde{Y}), \end{aligned}$$

where the second equality holds since both V_1^* and Ω_1 have full row rank. As the result,

$$\|(I - P_{\tilde{Y}})\tilde{A}\|_F = 0,$$

and the conclusion follows.

Next, we flatten out the top block of $\tilde{Y}$ to obtain

$$Z = \tilde{Y} \cdot \Omega_1^\dagger \Sigma_1^{-1} = \begin{bmatrix} I \\ F \end{bmatrix}, \quad \text{where } F = \Sigma_2 \Omega_2 \Omega_1^\dagger \Sigma_1^{-1}.$$

Since $\text{range}(Z) \subset \text{range}(\tilde{Y})$,

$$\|(I - P_{\tilde{Y}})\tilde{A}\|_F \leq \|(I - P_Z)\tilde{A}\|_F$$

by applying [2, Proposition 8.5]. Taking squares,

$$\begin{aligned}
\|(I - P_{\tilde{Y}})\tilde{A}\|_F^2 &\leq \|(I - P_Z)\tilde{A}\|_F^2 \\
&= \text{Tr}(\tilde{A}^*(I - P_Z)\tilde{A}) \\
&= \text{Tr}(\Sigma(I - P_Z)\Sigma).
\end{aligned}$$

Note that Z has full column rank,

$$P_Z = Z(Z^*Z)^{-1}Z^* = \begin{bmatrix} I \\ F \end{bmatrix} (I + F^*F)^{-1} \begin{bmatrix} I \\ F \end{bmatrix}^*,$$

and $I - P_Z$ is equal to

$$\begin{bmatrix} I - (I + F^*F)^{-1} & -(I + F^*F)^{-1}F^* \\ -F(I + F^*F)^{-1} & I - F(I + F^*F)^{-1}F^* \end{bmatrix}.$$

The top left block satisfies

$$I - (I + F^*F)^{-1} \leq F^*F$$

by applying [2, Proposition 8.2], and the bottom right block satisfies

$$I - F(I + F^*F)^{-1}F^* \leq I$$

since $F(I + F^*F)^{-1}F^* \geq 0$. Let $B = -(I + F^*F)^{-1}F^*$,

$$I - P_Z \leq \begin{bmatrix} F^*F & B \\ B^* & I \end{bmatrix},$$

and consequently,

$$\Sigma(I - P_Z)\Sigma \leq \begin{bmatrix} \Sigma_1 F^*F \Sigma_1 & \Sigma_1 B \Sigma_2 \\ \Sigma_2 B^* \Sigma_1 & \Sigma_2^2 \end{bmatrix}.$$

The last step is to notice that

$$\begin{aligned}
\text{Tr}(\Sigma(I - P_Z)\Sigma) &\leq \text{Tr} \left(\begin{bmatrix} \Sigma_1 F^*F \Sigma_1 & \Sigma_1 B \Sigma_2 \\ \Sigma_2 B^* \Sigma_1 & \Sigma_2^2 \end{bmatrix} \right) \\
&= \text{Tr}(\Sigma_1 F^*F \Sigma_1) + \text{Tr}(\Sigma_2^2) \\
&= \|\Sigma_1\|_F^2 + \|\Sigma_2\|_F^2 \\
&= \|\Sigma_2\|_F^2 + \|\Sigma_1\|_F^2.
\end{aligned}$$

C. Proof of Theorem 2

The final approximation $[[\hat{\mathbf{A}}]]_r$ is the best rank r approximation of $\hat{\mathbf{A}}$. Note that

$$\begin{aligned}\|\mathbf{A} - [[\hat{\mathbf{A}}]]_r\|_F &\leq \|\mathbf{A} - \hat{\mathbf{A}}\|_F + \|\hat{\mathbf{A}} - [[\hat{\mathbf{A}}]]_r\|_F \\ &\leq \|\mathbf{A} - \hat{\mathbf{A}}\|_F + \|\hat{\mathbf{A}} - [[\mathbf{A}]]_r\|_F \\ &\leq 2\|\mathbf{A} - \hat{\mathbf{A}}\|_F + \|\mathbf{A} - [[\mathbf{A}]]_r\|_F \\ &= 2\|\mathbf{A} - \hat{\mathbf{A}}\|_F + \|\Sigma_2\|_F,\end{aligned}$$

where the second inequality holds since $[[\hat{\mathbf{A}}]]_r$ is the best rank r approximation. We could bound $\|\mathbf{A} - [[\hat{\mathbf{A}}]]_r\|_F$ by bounding the initial approximation error $\|\mathbf{A} - \hat{\mathbf{A}}\|_F$.

$$\begin{aligned}\|\mathbf{A} - \hat{\mathbf{A}}\|_F^2 &= \|\mathbf{A} - \mathbf{Q}\mathbf{C}\mathbf{P}^*\|_F^2 \\ &= \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{P}\mathbf{P}^* + \mathbf{Q}(\mathbf{Q}^*\mathbf{A}\mathbf{P} - \mathbf{C})\mathbf{P}^*\|_F^2 \\ &= \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{P}\mathbf{P}^*\|_F^2 + \|\mathbf{Q}(\mathbf{C} - \mathbf{Q}^*\mathbf{A}\mathbf{P})\mathbf{P}^*\|_F^2 \\ &= \underbrace{\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{P}\mathbf{P}^*\|_F^2}_{T_2} + \underbrace{\|\mathbf{C} - \mathbf{Q}^*\mathbf{A}\mathbf{P}\|_F^2}_{T_3}\end{aligned}$$

where the third equality is due to

$$\begin{aligned}\langle \mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{P}\mathbf{P}^*, \mathbf{Q}(\mathbf{C} - \mathbf{Q}^*\mathbf{A}\mathbf{P})\mathbf{P}^* \rangle \\ = \langle \mathbf{Q}^*(\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{P}\mathbf{P}^*)\mathbf{P}, \mathbf{C} - \mathbf{Q}^*\mathbf{A}\mathbf{P} \rangle = 0.\end{aligned}$$

Bound for T_2 . Note that

$$\begin{aligned}T_2 &= \|\mathbf{A}(\mathbf{I} - \mathbf{P}\mathbf{P}^*) + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\mathbf{P}^*\|_F^2 \\ &= \|\mathbf{A}(\mathbf{I} - \mathbf{P}\mathbf{P}^*)\|_F^2 + \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\mathbf{P}^*\|_F^2 \\ &= \|\mathbf{A}(\mathbf{I} - \mathbf{P}\mathbf{P}^*)\|_F^2 + \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\|_F^2 \\ &= \|(U_2\Sigma_2V_2^*)(\mathbf{I} - \mathbf{P}\mathbf{P}^*)\|_F^2 \\ &\quad + \|(U_1\Sigma_1V_1^*)(\mathbf{I} - \mathbf{P}\mathbf{P}^*)\|_F^2 + \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\|_F^2 \\ &\leq \|\Sigma_2\|_F^2 + \|(U_1\Sigma_1V_1^*)(\mathbf{I} - \mathbf{P}\mathbf{P}^*)\|_F^2 \\ &\quad + \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\|_F^2,\end{aligned}$$

where the second equality is due to

$$\langle \mathbf{A}(\mathbf{I} - \mathbf{P}\mathbf{P}^*), (\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{A}\mathbf{P}\mathbf{P}^* \rangle = 0,$$

and the fourth equality is due to

$$\langle (U_2\Sigma_2V_2^*)(\mathbf{I} - \mathbf{P}\mathbf{P}^*), (U_1\Sigma_1V_1^*)(\mathbf{I} - \mathbf{P}\mathbf{P}^*) \rangle = 0.$$

Denote $\Gamma_1 = \Gamma U_1^{(\Delta, \cdot)}$, and $\Gamma_2 = \Gamma U_2^{(\Delta, \cdot)}$. Following a similar argument as the proof of Lemma 3,

$$\|(U_1 \Sigma_1 V_1^*)(I - PP^*)\|_F^2 \leq \|\Gamma_1^\dagger \Gamma_2 \Sigma_2\|_F^2.$$

Therefore, we arrive at

$$\begin{aligned} T_2 &\leq \|\Sigma_2\|_F^2 + \|\Gamma_1^\dagger \Gamma_2 \Sigma_2\|_F^2 + \|(I - QQ^*)AP\|_F^2 \\ &\leq \|\Sigma_2\|_F^2 + \|\Gamma_1^\dagger \Gamma_2 \Sigma_2\|_F^2 + \|(I - QQ^*)A\|_F^2 \\ &\leq \|\Sigma_2\|_F^2 + \|\Gamma_1^\dagger \Gamma_2 \Sigma_2\|_F^2 + \|\Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2 \\ &= 2\|\Sigma_2\|_F^2 + \|\Gamma_1^\dagger \Gamma_2 \Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2. \end{aligned}$$

Bound for T_3 . We decompose $C - Q^*AP$ into several parts, and bound them separately. Denote the complement of Q and P to be $Q_\perp$ and $P_\perp$, respectively. Define

$$\begin{aligned} \Phi_1 &= \Phi Q^{(\Delta', :)} \quad \text{and} \quad \Phi_2 = \Phi Q_\perp^{(\Delta', :)}, \\ \Psi_1 &= \Psi P^{(\Theta', :)} \quad \text{and} \quad \Psi_2 = \Psi P_\perp^{(\Theta', :)}. \end{aligned}$$

Lemma 7.—Assume that Φ_1 and Ψ_1 are both of full column rank. Then

$$\begin{aligned} C - Q^*AP &= \Phi_1^\dagger \Phi_2 (Q_\perp^* AP) + (Q^* AP_\perp) \Psi_2^* (\Psi_1^\dagger)^* \\ &\quad + \Phi_1^\dagger \Phi_2 (Q_\perp^* AP_\perp) \Psi_2^* (\Psi_1^\dagger)^*. \end{aligned}$$

The proof of Lemma 7 is the same as [7, Lemma A.3]. We include its proof for completeness, and defer it to Section IV-D.

Remark.—To ensure that Φ_1 and Ψ_1 are of full column rank, one way is to argue that Q and P are also incoherent, i.e.,

$$\begin{aligned} \max_i \{ \|Q^{(i, :)}\|_2 \} &\leq \sqrt{\frac{\mu' k}{M}}, \\ \max_j \{ \|P^{(j, :)}\|_2 \} &\leq \sqrt{\frac{\nu' k}{N}}, \end{aligned} \tag{2}$$

for some $\mu' \ll M$, $\nu' \ll N$, and apply Lemma 1. As shown in Table VII, we empirically observe that (2) holds, and the incoherence parameters $\mu' = O(\mu)$, $\nu' = O(\nu)$.

However, we find that the bounds of (μ', ν') in (2) are difficult to obtain. As mentioned at the end of Section II, one way to get around this is to sample $m' \approx O(\mu' k \log k)$ row indices and $n' \approx O(\nu' k \log k)$ column indices based on the actual incoherence parameters estimated from the row norms of Q and P , respectively.

Based on Lemma 7,

$$\begin{aligned}
T_3 &= \|\Phi_1^\dagger \Phi_2(Q_\perp^* \mathbf{A} \mathbf{P}) + (Q^* \mathbf{A} \mathbf{P}_\perp) \Psi_2^* (\Psi_1^\dagger)^*\|_F^2 \\
&\quad + \|\Phi_1^\dagger \Phi_2(Q_\perp^* \mathbf{A} \mathbf{P}_\perp) \Psi_2^* (\Psi_1^\dagger)^*\|_F^2 \\
&\leq 3 \cdot \underbrace{\left(\|\Phi_1^\dagger \Phi_2(Q_\perp^* \mathbf{A} \mathbf{P})\|_F^2 \right)}_{T_{31}} + \underbrace{\left(\|(Q^* \mathbf{A} \mathbf{P}_\perp) \Psi_2^* (\Psi_1^\dagger)^*\|_F^2 \right)}_{T_{32}} \\
&\quad + \underbrace{\|\Phi_1^\dagger \Phi_2(Q_\perp^* \mathbf{A} \mathbf{P}_\perp) \Psi_2^* (\Psi_1^\dagger)^*\|_F^2}_{T_{33}}.
\end{aligned}$$

We bound the three terms separately. The bounds all follow from a similar argument as deriving the bound for $\|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2$ in the proof of Theorem 1. Suppose $s \geq k + 4$.

$$\begin{aligned}
T_{31} &\leq \frac{6e^2}{q} \cdot \frac{s^1 + 6/(s - k + 1)}{(s - k + 1)^2} \cdot (\sqrt{s} \|Q_\perp^* \mathbf{A} \mathbf{P}\|_F + \sqrt{6 \log s} \|Q_\perp^* \mathbf{A} \mathbf{P}\|_2)^2 \\
&\leq C(q, s, k) \cdot \|Q_\perp^* \mathbf{A} \mathbf{P}\|_F^2
\end{aligned}$$

with probability at least $1 - \frac{2}{s^3}$, where

$$C(q, s, k) = \frac{6e^2}{q} \cdot \frac{s^1 + 6/(s - k + 1)}{(s - k + 1)^2} \cdot (\sqrt{s} + \sqrt{6 \log s})^2.$$

Similarly, with probability at least $1 - \frac{2}{s^3}$,

$$T_{32} \leq C(q, s, k) \cdot \|Q^* \mathbf{A} \mathbf{P}_\perp\|_F^2.$$

Apply the argument twice,

$$\begin{aligned}
T_{33} &\leq C(q, s, k) \cdot \|(Q_\perp^* \mathbf{A} \mathbf{P}_\perp) \Psi_2^* (\Psi_1^\dagger)^*\|_F^2 \\
&\leq (C(q, s, k))^2 \cdot \|Q_\perp^* \mathbf{A} \mathbf{P}_\perp\|_F^2
\end{aligned}$$

with probability at least $1 - \frac{2}{s^3}$. Combining the three estimates,

$$\begin{aligned}
\frac{T_3}{3} &\leq C(q, s, k) \cdot \left(\|Q_{\perp}^* A P\|_F^2 + \|Q^* A P_{\perp}\|_F^2 \right. \\
&\quad \left. + C(q, s, k) \cdot \|Q_{\perp}^* A P_{\perp}\|_F^2 \right) \\
&= C(q, s, k) \cdot \left(\|Q_{\perp}^* A P\|_F^2 + \|Q^* A P_{\perp}\|_F^2 \right. \\
&\quad \left. + \|Q_{\perp}^* A P_{\perp}\|_F^2 + (C(q, s, k) - 1) \cdot \|Q_{\perp}^* A P_{\perp}\|_F^2 \right) \\
&= C(q, s, k) \cdot \left(\|A - Q Q^* A P P^*\|_F^2 \right. \\
&\quad \left. + (C(q, s, k) - 1) \cdot \|Q_{\perp}^* A P_{\perp}\|_F^2 \right) \\
&= C(q, s, k) \cdot \|A - Q Q^* A P P^*\|_F^2 \\
&\quad + C(q, s, k)(C(q, s, k) - 1) \cdot \|Q_{\perp}^* A P_{\perp}\|_F^2 \\
&\leq C(q, s, k) \cdot \|A - Q Q^* A P P^*\|_F^2 \\
&\quad + C(q, s, k)(C(q, s, k) - 1) \cdot \|(\mathbf{I} - Q Q^*) A\|_F^2.
\end{aligned}$$

Thus for the square of the initial approximation error,

$$\begin{aligned}
&\|A - \hat{A}\|_F^2 \\
&\leq (3C(q, s, k) + 1) \cdot \|A - Q Q^* A P P^*\|_F^2 \\
&\quad + 3C(q, s, k)(C(q, s, k) - 1) \cdot \|(\mathbf{I} - Q Q^*) A P\|_F^2 \\
&\leq (3C(q, s, k) + 1)(2\|\Sigma_2\|_F^2 + \|\Gamma_1^{\dagger} \Gamma_2 \Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^{\dagger}\|_F^2) \\
&\quad + 3C(q, s, k)(C(q, s, k) - 1)(\|\Sigma_2\|_F^2 + \|\Sigma_2 \Omega_2 \Omega_1^{\dagger}\|_F^2).
\end{aligned}$$

Taking the square root,

$$\begin{aligned}
&\|A - \hat{A}\|_F \\
&\leq \sqrt{3(C(q, s, k))^2 + 3C(q, s, k) + 2} \cdot \\
&\quad (C_1(p, k, r)\|\Sigma_2\|_F + C_2(p, k, r)\|\Sigma_2\|_2) \\
&\leq (\sqrt{3}C(q, s, k) + \sqrt{2}) \cdot \\
&\quad (C_1(p, k, r)\|\Sigma_2\|_F + C_2(p, k, r)\|\Sigma_2\|_2) \\
&= C_3(p, q, s, k, r) \cdot \|\Sigma_2\|_F + C_4(p, q, s, k, r) \cdot \|\Sigma_2\|_2,
\end{aligned}$$

where

$$\begin{aligned}
C_3(p, q, s, k, r) &= C_1(p, k, r) \cdot (\sqrt{3}C(q, s, k) + \sqrt{2}), \\
C_4(p, q, s, k, r) &= C_2(p, k, r) \cdot (\sqrt{3}C(q, s, k) + \sqrt{2}).
\end{aligned}$$

D. Proof of Lemma 7

Define $S_{\Delta'} \in \mathbb{R}^{m' \times M}$ such that when left multiplied to a matrix, the result is equal to the rows with indices Δ' of that matrix. Define $S_{\Theta'}$ in the same way. The core sketch can be written as

$$\begin{aligned}
Z &= \Phi A^{(\Delta', \Theta')} \Psi^* = \Phi S_{\Delta'} A S_{\Theta}^* \Psi^* \\
&= \Phi S_{\Delta'} (A - QQ^* A P P^*) S_{\Theta}^* \Psi^* \\
&\quad + (\Phi S_{\Delta'}, Q) Q^* A P (P^* S_{\Theta}^* \Psi^*) \\
&= \Phi S_{\Delta'} (A - QQ^* A P P^*) S_{\Theta}^* \Psi^* \\
&\quad + (\Phi Q^{(\Delta', :)}) Q^* A P (\Psi P^{(\Theta', :)})^*.
\end{aligned}$$

Left multiply by $\Phi_1^\dagger$ and right-multiply by $(\Psi_1^\dagger)^*$,

$$C = \Phi_1^\dagger \Phi S_{\Delta'} (A - QQ^* A P P^*) S_{\Theta}^* \Psi^* (\Psi_1^\dagger)^* + Q^* A P.$$

Notice that

$$\begin{aligned}
\Phi_1^\dagger \Phi S_{\Delta'} &= \Phi_1^\dagger \Phi S_{\Delta'} Q Q^* + \Phi_1^\dagger \Phi S_{\Delta'} Q_\perp Q_\perp^* \\
&= Q^* + \Phi_1^\dagger \Phi_2 Q_\perp^*,
\end{aligned}$$

$$\begin{aligned}
S_{\Theta}^* \Psi^* (\Psi_1^\dagger)^* &= P P^* S_{\Theta}^* \Psi^* (\Psi_1^\dagger)^* + P_\perp P_\perp^* S_{\Theta}^* \Psi^* (\Psi_1^\dagger)^* \\
&= P + P_\perp \Psi_2^* (\Psi_1^\dagger)^*.
\end{aligned}$$

Combining all the pieces,

$$\begin{aligned}
C - Q^* A P &= (Q^* + \Phi_1^\dagger \Phi_2 Q_\perp^*) (A - QQ^* A P P^*) (P + P_\perp \Psi_2^* (\Psi_1^\dagger)^*) \\
&= \Phi_1^\dagger \Phi_2 (Q_\perp^* A P) + (Q^* A P_\perp) \Psi_2^* (\Psi_1^\dagger)^* \\
&\quad + \Phi_1^\dagger \Phi_2 (Q_\perp^* A P_\perp) \Psi_2^* (\Psi_1^\dagger)^*.
\end{aligned}$$

V. CONCLUSION

Our SketchyCoreSVD method is demonstrably useful in many applications. For example, the procedure to obtain a reduced model of fluid Navier Stokes flow dynamics is usually combined with a dimension reduction step. As a proof of concept, we conduct a Navier Stokes fluid flow simulation, and show the performance of SketchyCoreSVD in Appendix B.

For future work, we would like to derive tighter bounds for the incoherence parameters of the column and row spaces constructed from the sketches. It is also interesting to extend the algorithm and theory to compute low-rank tensor decompositions, and demonstrate its efficacy in several appropriate applications.

Acknowledgments

This research was supported in part by a grant from NIH - R01GM117594.

APPENDIX A: ALGORITHMS BASED ON SKETCHES

We provide a brief comparison of prior algorithms based on sketches.

[2] builds sketches

$$X = \Gamma A \quad \text{and} \quad Y = A \Omega^* .$$

Compute $(P, \sim, \sim) = \text{svds}(X^*, r)$, $(Q, \sim, \sim) = \text{svds}(Y, r)$; $C_1 = Q^* Y ((\Omega P)^\dagger)^*$, $C_2 = (\Gamma Q)^\dagger X P$;
 $(U, \Sigma, V) = \text{svd}((C_1 + C_2)/2)$;

$$\hat{A} = (QU)\Sigma(PV)^* .$$

[3] builds sketches

$$X = \Gamma A \quad \text{and} \quad Y = A \Omega^* .$$

Compute $Y = QR$; $\Gamma Q = UT$;

$$\hat{A} = QT^\dagger [[U^* Y]]_r .$$

[4] builds sketches

$$X = \Gamma A \quad \text{and} \quad Y = A \Omega^* .$$

Compute $(V, \sim, \sim) = \text{svds}(Y, r)$; $\Gamma V = UT$;

$$\hat{A} = VT^\dagger [[U^* Y]]_r .$$

[5] with simplification suggested by [18]. Build sketches

$$X = \Gamma A, \quad Y = A \Omega^*, \quad \text{and} \quad Z = \Phi A \Psi^* .$$

Compute $X^* = PR_1$, $Y = QR_2$; $\Phi Q = U_1 \Sigma_1 V_1^*$, $\Psi P = U_2 \Sigma_2 V_2^*$;

$$\hat{A} = QV_1 \Sigma_1^\dagger [[U_1^* Z U_2]]_r \Sigma_2^\dagger V_2^* P^* .$$

[6] builds sketches

$$X = \Gamma A, \quad \text{and} \quad Y = A \Omega^* ,$$

Compute $Y = QR$;

$$\hat{A} = Q[[(\Gamma Q)^\dagger X]]_r .$$

APPENDIX B: MORE NUMERICS

We use a Navier Stokes simulated flow system which generates vorticity patterns for an incompressible fluid under certain initial and boundary conditions. For each point of the 100×50 grids, fluid velocity values in x and y directions are recorded for 200 time instances. As an example, we demonstrate our result on the y component of the fluid velocity, captured in

the data matrix $A \in \mathbb{R}^{5000 \times 200}$. Based on the scree plot in Figure 9, we choose $r = 7$. The optimal err is 0.0016.

Fig. 9.
Scree plot for Navier Stokes dataset.

TABLE VIII

PERFORMANCE COMPARISONS FOR NAVIER STOKES DATASET.

	SketchySVD	SketchyCoreSVD
p	-	0.3
err	0.0016	0.0016
time (sec)	0.0097	0.0062

Fig. 10.

Visual comparison for the first six left singular vectors computed on the Navier Stokes dataset. Ground truth (left column), SketchySVD (middle column), SketchyCoreSVD (right column, $p = 0.3$). For each subfigure, $x \in [1, 8]$ (horizontal), and $y \in [-2, 2]$ (vertical).

TABLE IX

COMPARISON OF PSNR VALUES FOR THE FIRST SIX LEFT SINGULAR VECTORS COMPUTED ON THE NAVIER STOKES DATASET, WHERE $p = 0.3$ FOR OUR SKETCHYCORESVD METHOD.

	SketchySVD	SketchyCoreSVD
1	115.3884	107.3097
2	116.2469	107.2117
3	112.3364	98.2154
4	96.8527	84.5132
5	97.7459	83.6819
6	88.7866	85.8143

We choose $k = 4r + 1 = 29$ and $s = 2k + 1 = 59$ for both methods. In Table VIII, we see that the error of SketchySVD is the same as the optimal error. SketchyCoreSVD can achieve the same error bound in less time. The visual comparison in Figure 10 and PSNR values in Table IX show that for SketchyCoreSVD, the singular vectors can be estimated accurately with sampling ratio $p = 30\%$.

REFERENCES

- [1]. Woolfe F, Liberty E, Rokhlin V, and Tygert M, "A fast randomized algorithm for the approximation of matrices," *Applied and Computational Harmonic Analysis*, vol. 25, no. 3, pp. 335–366, 2008.
- [2]. Halko N, Martinsson P-G, and Tropp JA, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.
- [3]. Woodruff DP et al., "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.
- [4]. Cohen MB, Elder S, Musco C, Musco C, and Persu M, "Dimensionality reduction for k-means clustering and low rank approximation," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing ACM*, 2015, pp. 163–172.
- [5]. Boutsidis C, Woodruff DP, and Zhong P, "Optimal principal component analysis in distributed and streaming models," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing ACM*, 2016, pp. 236–249.
- [6]. Tropp JA, Yurtsever A, Udell M, and Cevher V, "Practical sketching algorithms for low-rank matrix approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1454–1485, 2017.
- [7]. Tropp JA "Streaming low-rank matrix approximation with an application to scientific simulation," *arXiv preprint arXiv:1902.08651*, 2019.
- [8]. Achlioptas D, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.
- [9]. Li P, Hastie TJ, and Church KW, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining ACM*, 2006, pp. 287–296.
- [10]. Talwalkar A and Rostamizadeh A, "Matrix coherence and the Nystrom method," *arXiv preprint arXiv:1408.2044*, 2014.
- [11]. Gittens A, "The spectral norm error of the naive nystrom extension," *arXiv preprint arXiv:1110.5305*, 2011.
- [12]. Chiu J and Demanet L, "Sublinear randomized algorithms for skeleton decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1361–1383, 2013.
- [13]. Williams C and Seeger M, "Using the Nyström method to speed up kernel machines," in *Advances in neural information processing systems*, 2001, pp. 682–688.
- [14]. Georgiades A, Belhumeur P, and Kriegman D, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 6 2001 [Online]. Available: 10.1109/34.927464
- [15]. Gupta S, Mittal S, Kajdacsy-Balla A, Bhargava R, and Bajaj C, "A fully automated, faster noise rejection approach to increasing the analytical capability of chemical imaging for digital histopathology," *PLOS ONE*, vol. 14, no. 4, pp. 1–15, 04 2019 [Online]. Available: 10.1371/journal.pone.0205219
- [16]. Malik OA and Becker S, "Low-rank Tucker decomposition of large tensors using TensorSketch," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 096–10 106.
- [17]. Tropp JA, "Improved analysis of the subsampled randomized Hadamard transform," *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.
- [18]. Upadhyay J, "The price of privacy for low-rank factorization," in *Advances in Neural Information Processing Systems*, 2018, pp. 4176–4187.

Fig. 1.
Scree plot for Yale Face dataset.

Fig. 2.

Visual comparison for the first left singular vector computed on the Yale Face dataset.

Ground truth (left column), SketchySVD (middle column), SketchyCoreSVD (right column, $p = 0.4$). The peak signal to noise ratio (PSNR) is 44.0320 for SketchySVD, and 44.2639 for SketchyCoreSVD.

Fig. 3.
Scree plot for Cardiac MRI dataset.

Fig. 4.

Visual comparison for the first two left singular vectors computed on the Cardiac MRI dataset. Ground truth (left column), SketchySVD (middle column), SketchyCoreSVD (right column, $p = 0.4$). The PSNR values are $\{68.5232, 59.1094\}$ for SketchySVD, and $\{69.5770, 58.3338\}$ for SketchyCoreSVD.

Fig. 5.
Scree plot for BR1003 dataset.

Fig. 6. Visual comparison for the first six left singular vectors computed on the BR1003 dataset, where $p = 0.08$ for our SketchyCoreSVD method.

Fig. 7.
Scree plot for Video dataset.

Fig. 8.

Visual comparison for the first two left singular vectors computed on the Video dataset. Ground truth (left column), SketchySVD (middle column), SketchyCoreSVD (right column, $p = 0.2$). The PSNR values are $\{55.5701, 50.2379\}$ for SketchySVD, and $\{55.8322, 49.2919\}$ for SketchyCoreSVD.

TABLE I

COMPARISON OF COMPUTATIONAL COST.

	SketchySVD	SketchyCoreSVD
X	$O(kMN)$	$O(kpMN)$
Y	$O(kMN)$	$O(kpMN)$
Z	$O(sMN)$	$O(sq^2MN)$
QR of X	$O(k^2N)$	$O(k^2N)$
QR of Y	$O(k^2M)$	$O(k^2M)$
C	$O(ks(M+N))$	$O(ksq(M+N))$
$[[C]]_r$	$O(k^3)$	$O(k^3)$
$[[\hat{A}]]_r$	$O(rMN)$	$O(rMN)$

TABLE II

PERFORMANCE COMPARISONS FOR YALE FACE DATASET.

	SketchySVD	SketchyCoreSVD		
p	-	0.3	0.35	0.4
err	0.066	0.0765	0.0737	0.0717
time (sec)	0.0239	0.0134	0.0146	0.0166

TABLE III

PERFORMANCE COMPARISONS FOR CARDIAC MRI DATASET.

	SketchySVD	SketchyCoreSVD		
p	-	0.3	0.35	0.4
err	0.0019	0.0021	0.0021	0.0019
time (sec)	0.0567	0.0316	0.038	0.0396

TABLE IV

PERFORMANCE COMPARISONS FOR BR1003 DATASET.

	SketchySVD	SketchyCoreSVD		
p	-	0.04	0.06	0.08
err	0.0025	0.0031	0.0029	0.0027
time (sec)	3.2049	0.816	1.1643	1.3614

TABLE V

COMPARISON OF SIGNAL TO NOISE RATIO (SNR) VALUES FOR THE FIRST SIX LEFT SINGULAR VECTORS COMPUTED ON THE BR1003 DATASET, WHERE $p = 0.08$ FOR OUR SKETCHYCORESVD METHOD.

	SketchySVD	SketchyCoreSVD
1	56.6010	54.2577
2	41.9409	39.0797
3	37.9187	34.9353
4	32.6139	31.8365
5	25.3890	23.0102
6	18.6244	15.6122

TABLE VI

PERFORMANCE COMPARISONS FOR VIDEO DATASET.

	SketchySVD	SketchyCoreSVD		
p	-	0.1	0.15	0.2
err	0.0148	0.0213	0.0177	0.0165
time (sec)	8.0062	2.3224	2.6266	3.5619

TABLE VII

COMPARISON OF (μ, ν) AND (μ', ν') ON DIFFERENT DATASETS. THE PARAMETERS (r, k, s, p) FOR EACH DATASET ARE CHOSEN AS THE SAME VALUES AS USED TO SHOW THE VISUAL COMPARISON.

	Yale Face	Cardiac MRI	BR1003	Video
μ	4.1137	127.5935	5.4270	20.3505
μ'	5.9454	159.8982	5.2750	22.8109
ν	2.7068	2.1507	32.8387	14.0194
ν'	4.1355	2.2260	57.4357	6.7429