Sketched Column-based Matrix Approximation

Jeongmin Chae†, Selin Bac††, Usama Saleem*, Shaama Mallikarjun Sharada* and Urbashi Mitra†

***Abstract*—A new, practical algorithm, fast Sketched Column-based Matrix Approximation (fSCMA), is proposed for low-rank matrix approximation. fSCMA leverages randomly, but fully sampled columns combined with structural side information, to achieve efficient and accurate approximations. The algorithm leverages both matrix sketching and side information to reduce complexity. A theoretical spectral bound on the reconstruction error is derived, improving the error bound by a factor of n (in terms of key parameters) compared to state-of-the-art algorithms (SoTA). Experimental results on synthetic data demonstrate that fSCMA achieves competitive performance relative to SoTA, validating theoretical bounds, while significantly reducing computational complexity. Additionally, fSCMA shows strong improvement over prior methods when applied to real data.**

I. Introduction

Many computational science applications employ low-rank matrix approximation [1]–[3]. For matrices of large dimension, these low-rank approximations provide efficiency in storage and processing [4]–[6]. Motivated by quantum chemistry applications, we previously proposed low-rank approximation methods [7], when only a subset of columns of the matrix are available. In [8], Quasi-Polynomial Matrix Approximation (QPMA) recovered a ground truth matrix that admits the following structure: $\mathbf{M} = \mathbf{QS} + \mathbf{E}$, where $\mathbf{S}$ captures the prior quasi-polynomial side information. The matrix $\mathbf{S}$ is *known* and $\mathbf{Q}$ is an *unknown* coefficient matrix; $\mathbf{E}$ represents an *unknown* perturbation from the polynomial structure. In fact, the matrix $\mathbf{S}$ can be any full rank matrix. With this side information in hand and only column samples, QPMA outperforms state-of-the-art methods [9], [10]. In [11], we proposed sCMA, a computationally efficient strategy for addressing the recovery of $\mathbf{M}$. sCMA performs sketching on $\hat{\mathbf{Q}}\mathbf{S}$, where $\hat{\mathbf{Q}}$ is an estimated coefficient matrix, and restricts access to only a limited number of its rows. While sCMA offers a notable reduction in complexity relative to QPMA, it still inherits the computational burden associated with estimating $\mathbf{Q}$. In contrast, the proposed algorithm circumvents this bottleneck by eliminating the need to solve for $\mathbf{Q}$ altogether.

Herein, we further improve upon QPMA and sCMA, via a novel, low complexity method called fast Sketched Column-based Matrix Approximation (fSCMA). fSCMA also employs a modest number of fully sampled columns, however, the sketching operation enables a significant reduction in complexity by obviating the need for a gradient search to solve for the coefficient matrix $\mathbf{Q}$. In particular, column space estimation is enabled by the sampled columns and its sketch and row space estimation is facilitated through the known $\mathbf{S}$ matrix. In contrast to some prior approaches, fSCMA does not require a singular value decomposition on the $\mathbf{QS}$ and thus scales well with problem dimension. Specifically, sCMA sketches the rows of $\hat{\mathbf{Q}}\mathbf{S}$ while fSCMA sketches the sampled columns.

While [12], [13] investigate low-rank matrix approximation with side information, both works rely on the assumption that the column and row spaces of the target matrix are fully known or can be accurately inferred from the side information. For example, [13] assumes that the true column and row spaces are perfectly embedded within the subspaces defined by the side information, along with access to additional randomly sampled entries. In contrast, our proposed method operates under a setting where the side information is only approximately informative.

As the name implies, fSCMA applies a randomized sketching matrix to the set of sampled columns. Matrix sketching has been studied extensively in the context of matrix approximation [6], [14]–[17]. Sketching provides matrix approximation via the projection onto a lower dimensional subspace while minimizing information loss. In [6], it is assumed that the full ground truth matrix is known and sketching is employed to create a low rank approximation. In [16], a low-rank matrix approximation is created from two noisy sketched matrices. Both [6] and [16] assume the ground truth matrix is low-rank. In contrast, herein, we assume access to only a collection of columns of the high-rank ground truth matrix, as well as structural side information.

Given the unique nature of our problem statement, comparison methods are challenging to find. To this end, the CUR approximation [9], [18] bears some similarities to our approach; therein, ground truth matrix is approximated as $\mathbf{M} \approx \mathbf{CUR}$. The matrices $\mathbf{C}$ and $\mathbf{R}$ are comprised of true columns and rows, respectively and be viewed as sketched versions of the row and columns of $\mathbf{M}$. In contrast, we only have access to sketched columns and the side information and we do not know anything else about the ground truth matrix. Recently, CUR+ [9] was introduced to address missing values within the CUR decomposition framework. This method constructs a low-rank approximation by utilizing a small number of fully sampled true rows and columns, along with additional random collected samples from the original matrix. Although CUR+ is relevant to our missing-value scenario, it relies on access to more information than our proposed framework. However, we do compare CUR+ to fSCMA and characterize the informativeness of our side information and the computational efficiency of both methods. Table I compares the information assumed to be available for key methods that we have presented.

The key contributions of this work are as follows.

- We propose a practical low-rank matrix approximation method that uses randomly sampled columns and struc-

† J. Chae and U.Mitra are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA. E-mail: {chaej, ubli}@usc.edu.

†† S. Bac is with the Department of Chemical Engineering, University of California, Santa Barbara, USA.

* U.Saleem and S. Mallikarjun Sharada are with the Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA. E-mail: {usaleem, ssharada}@usc.edu.

Algorithm	column-space	row-space
fSCMA	A sketch of $\mathbf{A}$	right singular vectors of $\mathbf{S}$
QPMA	$\mathbf{A}$	right singular vectors of $\hat{\mathbf{Q}}\mathbf{S}$
CUR+	$\mathbf{A}$	A few rows of $\mathbf{M}$
sCMA	$\mathbf{A}$	A few rows of $\hat{\mathbf{Q}}\mathbf{S}$

TABLE I
COMPARISON OF INFORMATION AVAILABLE FOR COLUMN-SPACE AND ROW-SPACE ESTIMATION TO EACH METHOD

tural side information, eliminating the need for coefficient matrix estimation.

- A theoretical bound on the reconstruction error achieved by fSCMA is derived, characterized by the relationship between the row-space of $\mathbf{E}$ and $\mathbf{S}$, as well as the spectral properties of the ground truth matrix and the sketch. This bound shows an improvement over QPMA by a factor related to the dimension of the ground truth matrix in terms of the order of key parameters. [1]
- fSCMA is compared numerically with QPMA, CUR+, and sCMA on synthetic data, demonstrating competitive NMSE performance while achieving up to 50 times greater computational efficiency compared to QPMA. Additionally, fSCMA is validated using real data from the original quantum chemistry application.

II. Problem formulation

We present the concrete problem formulation and introduce key definitions.

A. Signal model

Let $\mathbf{M} \in \mathbb{R}^{n\times m}$ be a matrix whose actual rank is k; this is the matrix we wish to approximate. Herein, we consider the problem of obtaining a low rank matrix approximation of $\mathbf{M}$, $\hat{\mathbf{M}}$, from d randomly sampled columns, we assume that $d \leq k$. We further assume that prior side information is captured by a *known* matrix $\mathbf{S}$. Thus, the true matrix $\mathbf{M}$ is modeled as

$$\mathbf{M} = \mathbf{Q}\mathbf{S} + \mathbf{E}, \tag{1}$$

where, $\mathbf{Q} \in \mathbb{R}^{n\times l}$ is an *unknown* coefficient matrix with respect to $\mathbf{S}$. The structural side information matrix, $\mathbf{S} \in \mathbb{R}^{l\times m}$, encodes the row-space prior knowledge of $\mathbf{M}$ and $\mathbf{E}$ is the perturbation/noise matrix. We note that the side information matrix $\mathbf{S}$ can be arbitrary, but should be tailored to the corresponding application. For example, in image processing, $\mathbf{S}$ can be drawn from the Discrete Cosine Transform (DCT), where each column of the matrix represents a cosine function with a different frequency [19] or a function of Legendre polynomials [20]. In our prior work [7], [8], [21], we have employed polynomial side information [2].

[1] In contrast to sCMA [11], we can show that the sketch of sampled columns is sufficient to represent the column-space of $\mathbf{M}$.

[2] In particular, to predict chemical reaction rate coefficients [22], [23] given information at specific points along the reaction coordinate, $\mathbf{s} = [s_1, \ldots, s_m]$ and polynomial order l, we assume that the side information $\mathbf{S}$ has the following polynomial structure. In addition, when $m = 4$, $\mathcal{C} = \{1, 3, 4\}$, i.e., $d = 3$, the sampling operation $\mathbf{\Psi}$ is defined as

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ s_1 & s_2 & \ldots & s_m \\ s_1^2 & s_2^2 & \ldots & s_m^2 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^{l-1} & s_2^{l-1} & \ldots & s_m^{l-1} \end{bmatrix} \text{ and } \mathbf{\Psi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

We set the target rank of $\hat{\mathbf{M}}$ as l. In particular, we consider the following regime,

$$l \leq d \leq k. \tag{3}$$

Since the structure of $\mathbf{E}$ is unknown, there exist two possible cases with respect to the row-space of $\mathbf{E}$: (1) the row-space of $\mathbf{E}$ fully coincides with that of $\mathbf{S}$, i.e., perfect side information, then $\mathbf{E} = \mathbf{0}$, and (2) the row-space of $\mathbf{E}$ is not fully contained in the row-space of $\mathbf{S}$. We focus primarily on the second case. If the row-space of $\mathbf{E}$ completely coincides with that of $\mathbf{S}$, the rank of $\mathbf{M}$ becomes l, resulting in $l = d = k$. In this case, the problem becomes trivial, as it corresponds to sampling all columns. Otherwise, the relationship $l < k$ always holds. We assume that $\mathbf{S}$ is full row rank, i.e., $\text{rank}(\mathbf{S}) = l$.

As mentioned previously, we observe a subset of the columns of $\mathbf{M}$. This column sampling operation $\mathbf{\Psi}$ is defined as follows. Let $\mathcal{C} = \{c_1, \ldots, c_d\} \subset [m]$ denote the set of sampled column indices. Clearly $|\mathcal{C}| = d$. Then, $\mathbf{\Psi} \in \{0, 1\}^{m\times d}$ is $\mathbf{\Psi} \doteq \mathbf{I}_{\mathcal{C}}$, where $\mathbf{I}$ is the identity matrix of dimension m and the notation $\mathbf{I}_{\mathcal{C}}$ means that we consider the sub-matrix of $\mathbf{I}$ formed by its columns indexed by entries in the set $\mathcal{C}$[1]. Thus, the observed matrix of the sampled columns, $\mathbf{A}$, can be equivalently expressed as

$$\mathbf{A} = \mathbf{M}\mathbf{\Psi}. \tag{4}$$

B. Notation and Key Definitions

We next define the following necessary quantities. We denote the Singular Value Decomposition (SVD) of a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ as

$$\mathbf{X} \overset{\text{SVD}}{=} \underbrace{\mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\mathsf{T}}_{\text{rank}-r \text{ approximation}} + \underbrace{\mathbf{U}_{X,\perp}\mathbf{\Sigma}_{X,\perp}\mathbf{V}_{X,\perp}^\mathsf{T}}_{\text{reminder}}, \tag{5}$$

where $r \leq \min\{n, m\}$, and $\mathbf{U}_X \in \mathbb{R}^{n\times r}$, $\mathbf{\Sigma}_X \in \mathbb{R}^{r\times r}$ and $\mathbf{V}_X^\mathsf{T} \in \mathbb{R}^{r\times m}$ (the matrix formed by the left and right singular vectors corresponding to the top-r singular values of $\mathbf{X}$). Additionally, we denote the rank-r SVD of $\mathbf{X}$ as $\mathbf{X} \overset{r-\text{SVD}}{=} \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\mathsf{T}$. Then, the SVD of the side information matrix $\mathbf{S} \in \mathbb{R}^{l\times m}$ is given by

$$\mathbf{S} \overset{\text{SVD}}{=} \mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^\mathsf{T}, \tag{6}$$

where the dimension of each component is $\mathbf{U}_S \in \mathbb{R}^{l\times l}$, $\mathbf{\Sigma}_S \in \mathbb{R}^{l\times l}$ and $\mathbf{V}_S \in \mathbb{R}^{m\times l}$. Note that $\text{rank}(\mathbf{S}) = l << \{n, m\}$. Similarly, we denote the SVD of $\mathbf{M}$ as $\mathbf{M} \overset{\text{SVD}}{=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, where the dimension of each component is $\mathbf{U} \in \mathbb{R}^{n\times m}$, $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{m\times m}$.

Next, we consider the following standard definition of the Johnson-Lindenstrauss Transform (JLT) [24].

Definition 1. *(Johnson-Lindenstrauss Transform). A random matrix $\mathbf{R} \in \mathbb{R}^{n\times r}$ forms a Johnson-Lindenstrauss transform if, for any (row) vector $\mathbf{x} \in \mathbb{R}^n$ and any $\epsilon > 0$,*

$$\mathbb{P}\left[(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{x}\mathbf{R}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2\right] \geq 1 - e^{-C\epsilon^2 r},$$

where $C > 0$ is a positive constant.

The Johnson-Lindenstrauss Transform condition is a necessary assumption in our work to map high-dimensional data

from the matrix of sampled columns onto a lower-dimensional subspace while approximately preserving the column subspace, all with high probability. We invoke the Johnson-Lindenstrauss transform to establish error guarantees for matrix sketching applied to the column-space approximation of $\mathbf{M}$ (See the proof of Lemma 2 in Appendix E).

In addition, consider the following standard definition of an orthogonal projector [25], and matrix incoherence [26].

Definition 2 (Orthogonal projectors)**.** *Let $\mathbf{X}$ be a $n \times m$ matrix. When $\mathbf{X}$ has full column rank, the orthogonal projector is defined as*

$$\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}. \tag{7}$$

We write $\mathbf{P}_X$ for the unique orthogonal projector with range$(\mathbf{P}_X) =$ range$(\mathbf{X})$.

Definition 3 (Incoherence)**.** *Let $\mathbf{X}$ be a $n \times m$ matrix of rank l and $\mathbf{X} \overset{l\text{-SVD}}{=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, where the dimension is $\mathbf{U} \in \mathbb{R}^{n\times l}$, $\mathbf{\Sigma} \in \mathbb{R}^{l\times l}$ and $\mathbf{V}^\mathsf{T} \in \mathbb{R}^{l\times m}$. Let $\mathbf{u}_i$ be the i-th row of $\mathbf{U}$ and $\mathbf{v}_j$ be the j-th row of $\mathbf{V}$. Then, the incoherence of $\mathbf{X}$ is given by $\mu(\mathbf{X}) = \max\left(\max_{i\in[n]} \frac{n}{l}\|\mathbf{u}_i\|_2^2, \max_{j\in[m]} \frac{m}{l}\|\mathbf{v}_j\|_2^2\right)$.*

In the analysis, we will use the shorthand notation, $\mu = \mu(\mathbf{M})$ and $\mu_s = \mu(\mathbf{S})$. Incoherence is a crucial assumption that ensures the "energy" of a matrix is distributed uniformly, facilitating its recovery from a limited number of randomly selected entries [26], [27]. Our setting differs from standard matrix completion, where individual entries are sampled randomly. In contrast, we observe only a few randomly chosen columns. However, the incorporation of side information allows us to leverage the conventional definition of incoherence. Incoherence guarantees that sketching is likely to catch enough information, because no single coordinate is too important.

III. Fast-Sketched Column-based Matrix Approximation

We now introduce the proposed algorithm, *fast-Sketched Column Matrix Approximation (fSCMA)*. As noted in the Introduction, when the ground truth matrix is available, low rank approximations can be formed by sampling true columns and rows, resulting in $\mathbf{M} \approx \mathbf{AUB}$ [9], [18]. If $\mathbf{B}$ were known, then a natural way to cast the optimization in our column-based setting is

$$\underset{\mathbf{Z}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \|\mathbf{M}\mathbf{\Psi} - \mathbf{AZB}\mathbf{\Psi}\|_F. \tag{8}$$

We observe that the optimization over $\mathbf{Z}$ depends on the number of sampled columns d. Moreover, in our column-sampled observation setting, we cannot collect row samples to form $\mathbf{B}$. Thus, we seek to determine an efficient way to estimate the column- and row-spaces given only a set of sampled columns and side information. Our proposed framework, fSCMA, consists of three stages: (i) the column-space approximation of $\mathbf{M}$ by randomly sketching a set of sampled columns $\mathbf{A}$; (ii) the row-space approximation of $\mathbf{M}$ by leveraging its side information structure $\mathbf{S}$; and (iii) performing the final low rank matrix approximation step, constrained by the previously obtained column and row-space approximations.

(i) Sketching-based column-space approximation. The goal of this step is to develop a practical approach for estimating the column-space of $\mathbf{M}$ given $\mathbf{A}$. We argue that as long as a sufficient number of independent columns are sampled (as will be shown in Lemma 1), the following optimization gives us a good low dimensional representation of $\mathbf{A}$

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times p}}{\operatorname{argmin}} \left\|\left(\mathbf{I} - \mathbf{U}_A\mathbf{U}_A^\mathsf{T}\right)\mathbf{X}\mathbf{X}^\mathsf{T}\right\|_F, \tag{9}$$

where $\mathbf{U}_A$ consists of the left singular vectors corresponding to the top-d singular values of $\mathbf{A}$. If $p = d$ in (9), the solution to the above is given by $\mathbf{A}$ from Eckart-Young-Mirsky theorem [28]. However, when $p \leq d$, we need a carefully constructed sketch of $\mathbf{A}$ using the sketching matrix $\mathbf{R}$ as shown in (10) using the Johnson Lindenstrauss Transform with the proper sample complexity (as will be shown in Theorem 1). This approach produces an $n \times p$ matrix that reliably estimates the column space of $\mathbf{A}$, and consequently, of $\mathbf{M}$.

We define a sketch of a set of observed columns $\mathbf{A}$, $\tilde{\mathbf{A}}$ as follows. Let $\mathbf{R} \in \mathbb{R}^{d\times p}$ be a Johnson-Lindenstrauss Transform such as a random Gaussian matrix or a sparse random matrix. We denote $\mathbf{R}$ a *sketching matrix*. Here, p is the sketching parameter. We shall focus on the case where $l \leq p \leq d$. The sketching parameter p controls the trade-off between the compression rate and the approximation accuracy. A larger p implies that more information is retained, and better a approximation of the original matrix $\mathbf{A}$ is achieved. A smaller p yields greater compression, but introduces more distortion in the singular values and subspace approximation of $\mathbf{A}$. We obtain a sketch $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times p}$ of $\mathbf{A} \in \mathbb{R}^{m\times d}$ as

$$\tilde{\mathbf{A}} = \mathbf{AR}. \tag{10}$$

In Theorem 1, we characterize the sampling complexity of d and p with respect to the reconstruction error. Empirically, we fixed d and determined a good value of p; our findings suggest that the relationship between p and d is approximately $p = \lceil 0.5 - 0.7 \times d \rceil$. The validation of this approximation is seen in Fig. 3. In addition, we numerically compare the performance of fSCMA versus p in Fig. 3. We observe that $\tilde{\mathbf{A}}$ serves as a good approximation of the column-space of $\mathbf{M}$, provided that a sufficient number of independent columns are sampled. Furthermore, when compared to our prior approach, QPMA [8], which estimates the column-space of $\mathbf{M}$ by performing an SVD on $\mathbf{A}$, the proposed algorithm, fSCMA, significantly reduces computational complexity by eliminating the SVD step, which costs $O(ndl)$, versus directly utilizing $\tilde{\mathbf{A}}$.

(ii) Row-space approximation. We next design a good estimate of the row space of $\mathbf{M}$. As previously noted, the row-space of the side structural information, $\mathbf{S}$ (see (1)), for $\mathbf{M}$ will be leveraged. By a standard linear algebra property [29], for any matrix $\mathbf{Q} \in \mathbb{R}^{n\times l}$ and $\mathbf{S} \in \mathbb{R}^{l\times m}$, the row space of $\mathbf{QS}$ is contained in (or equal to) the row space of $\mathbf{S}$ as left multiplication by $\mathbf{Q}$ forms linear combinations of the rows of $\mathbf{S}$. Therefore, it follows that rowspace$(\mathbf{QS}) \subseteq$ rowspace$(\mathbf{S})$. Based on the SVD representation of $\mathbf{S}$ in (6), $\mathbf{M}$ can be expressed as: $\mathbf{M} = \mathbf{Q}\mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^\mathsf{T} + \mathbf{E}$, which suggests that the row-space of $\mathbf{M}$ can be viewed as a perturbed version of the row-space of $\mathbf{S}$, influenced by the row-space of $\mathbf{E}$ [30].

However, the structure of $\mathbf{E}$ is completely unknown. Thus, we can only estimate the row space of $\mathbf{M}$ from $\mathbf{S}$ through the following minimization

$$\mathbf{V}_S = \underset{\bar{\mathbf{V}}\in\mathbb{R}^{m\times l}, \bar{\mathbf{V}}^\mathsf{T}\bar{\mathbf{V}}=\mathbf{I}}{\operatorname{argmin}} \left\| \left(\mathbf{I} - \bar{\mathbf{V}}\bar{\mathbf{V}}^\mathsf{T}\right) \mathbf{S}^\mathsf{T}\mathbf{S} \right\|_F. \quad (11)$$

The solution to the above is readily obtained through a rank-l SVD of $\mathbf{S}$. In Theorem 1, we characterize the error bound in terms of the relationship between the row spaces of $\mathbf{S}$ and $\mathbf{E}$. This approach eliminates the computational complexity associated with solving for $\mathbf{Q}$ in QPMA, which requires both gradient descent and SVD to estimate the row-space. In particular, fSCMA avoids the cost of gradient descent, $O(nd^2l \cdot T)$, where T is the number of iterations, and the cost of $O(nml)$ for the SVD. These operations scale poorly with the matrix dimensions n and m, making them impractical for large-scale problems.

(iii) Low rank matrix approximation. Lastly, we exploit the column sketch $\tilde{\mathbf{A}}$ and the row-space of $\mathbf{S}$ (i.e., $\mathbf{V}_S$) to obtain the desired low-rank approximation as follows,

$$\mathbf{Z} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\| \mathbf{M}\mathbf{\Psi} - \tilde{\mathbf{A}}\mathbf{X}\mathbf{V}_S^\mathsf{T}\mathbf{\Psi} \right\|_F^2. \quad (12)$$

This is a Frobenius norm regression problem where we seek the $\mathbf{X}$ that best satisfies the linear system in the least-squares sense. The closed form solution for $\mathbf{Z}$ [10], [18] is given by $\mathbf{Z} = \tilde{\mathbf{A}}^\dagger\mathbf{M}\mathbf{\Psi}\left(\mathbf{V}_S^\mathsf{T}\mathbf{\Psi}\right)^\dagger \in \mathbb{R}^{p\times l}$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse. Equation (12) solves for $\mathbf{Z}$ only using the sketch of $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$ and the row-space of $\mathbf{S}$, $\mathbf{V}_S$, without full knowledge of $\mathbf{M}$, but with knowledge of the column sketch and the side information $\mathbf{S}$. Finally, our low rank approximation of $\mathbf{M}$, $\hat{\mathbf{M}}$, is then obtained as

$$\hat{\mathbf{M}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\dagger\mathbf{A}\left(\mathbf{V}_S^\mathsf{T}\mathbf{\Psi}\right)^\dagger\mathbf{V}_S^\mathsf{T} = \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}.$$

The dimensions of $\mathbf{Z}$ and the computation of the pseudo-inverse of $\mathbf{V}_S^\mathsf{T}\mathbf{\Psi} \in \mathbb{R}^{l\times d}$ depend on l, p and d, we observe that these parameters do not scale at the same rate as the system parameters (m and n) which are much larger. The complete algorithm is summarized in Algorithm 1.

IV. Main theorem and analysis

In this section, we provide our main result and the proof sketch. The full proof can be found in the Appendix. To perform the analysis, we partition the singular value decomposition of $\mathbf{M}$ as follows. Recall that the target rank of the approximated matrix is l.

$$\mathbf{M} \overset{\text{SVD}}{=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U}\begin{bmatrix} \underbrace{\mathbf{\Sigma}_1}_{l} & \\ & \underbrace{\mathbf{\Sigma}_2}_{m-l} \end{bmatrix}\begin{bmatrix} \underbrace{\mathbf{V}_1^\mathsf{T}}_{l\times m} \\ \underbrace{\mathbf{V}_2^\mathsf{T}}_{(m-l)\times m} \end{bmatrix}. \quad (13)$$

The matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are square.

A. Main result

We now present our main result.

Theorem 1. *Assume that d columns are sampled uniformly at random from the underlying ground truth, $\mathbf{M}$. Let $\mathbf{R}$ denote the Johnson-Lindenstrauss Transform of size $d \times p$. Then, with probability $1 - \max\{e^{-c_1\epsilon^2 p}, c_2 l^{-3}\}$, where $\delta = \max\{e^{-c_1\epsilon^2 p}, c_2 l^{-3}\}$, if $d \geq \max\{c_3\mu l \ln l, c_4\mu_s l \ln l\}$ and $p = O\left(\log(1/\delta)/\epsilon^2\right)$, we have*

$$\begin{aligned} &\frac{\left\|\mathbf{M} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2}{\|\mathbf{M}\|_2^2} \\ &\leq 4\frac{\sigma_{l+1}^2(\mathbf{M})}{\sigma_1^2(\mathbf{M})}\left(5 + \frac{6n}{d}\right) + 8(1+\epsilon)(m-l)\cdot\gamma \\ &\quad + \frac{8\sigma_2^2(\mathbf{E})}{\sigma_1^2(\mathbf{M})}\left(2 + \frac{2m}{d}\right) + \frac{4\left\|\mathbf{E} - \mathbf{E}\mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right\|_2^2}{\sigma_1^2(\mathbf{M})}, \quad (14) \end{aligned}$$

where $\gamma \doteq \frac{\sigma_{l+1}^2(\mathbf{M})}{\sigma_1^2\left(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right)\cdot\sigma_1^2(\mathbf{M})}$. And $c_1, c_2, c_3, c_4, \epsilon > 0$ are numerical constants and $0 < \delta < 1$. ⊠

Algorithm 1 Proposed fSCMA algorithm

Input: A matrix of sampled columns $\mathbf{A} \in \mathbb{R}^{n\times d}$, Side information matrix $\mathbf{S} \in \mathbb{R}^{l\times m}$, Column sampling matrix $\mathbf{\Psi} \in \{0,1\}^{m\times d}$, A JLT sketching matrix $\mathbf{R} \in \mathbb{R}^{d\times p}$

Parameters: Degree of $\mathbf{S}$, l, Sketching matrix parameter, p, and the number of sampled columns d

Algorithm:

1. Sketching-based column-space approximation

- Obtain a sketch of $\mathbf{A}$ via random projection $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$

2. row-space approximation

- Obtain orthogonal basis of the row-space of side information $\mathbf{S}$ by means of SVD: $\mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^\mathsf{T} = \text{svd}(\mathbf{S})$

3. Low-rank matrix approximation: Using $\tilde{\mathbf{A}}$ and $\mathbf{V}_S$ obtained from the previous steps,

- Solve (12) and obtain $\mathbf{Z} = \tilde{\mathbf{A}}^\dagger\mathbf{A}\left(\mathbf{V}_S^\mathsf{T}\mathbf{\Psi}\right)^\dagger \in \mathbb{R}^{p\times l}$
- Construct $\hat{\mathbf{M}} = \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}$

Output: $\hat{\mathbf{M}} = \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}$

Proof. The full proof of Theorem 1 is provided in the Appendix. It is derived by leveraging large-deviation results from random matrix theory [31] to ensure that the estimated column-space remains well-behaved, given a sufficient number of sampled columns. This is followed by a careful application of matrix sketching techniques.

B. Discussion

Interpreting the error. Theorem 1 addresses three primary sources of error: (i) the unrecoverable energy resulting from the high rank nature of the original matrix $\mathbf{M}$, (ii) imperfections in the side information, and (iii) the loss due to using the sketched column information in $\mathbf{A}$ to estimate the column-space of $\mathbf{M}$. The first and second terms in (14) quantify the irreducible approximation error due to the high-rank components of the matrix and the additional distortion introduced by the sketching operator, respectively. We assume that both terms $\sigma_1^2(\mathbf{M})$ and $\sigma_l^2\left(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right)$ in (14) are constant, i.e., $\sigma_1^2(\mathbf{M}) = O(1)$ and $\sigma_l^2\left(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right) = O(1)$. Otherwise, we observe that fSCMA incurs a multiplicative factor of $O(n/d + (m-l))$ in the best rank-l approximation error, given by $\left\|\mathbf{M} - \hat{\mathbf{M}}\right\|_2^2 = \sigma_{l+1}^2(\mathbf{M})$. This arises because our column-based sampling strategy yields a more challenging problem

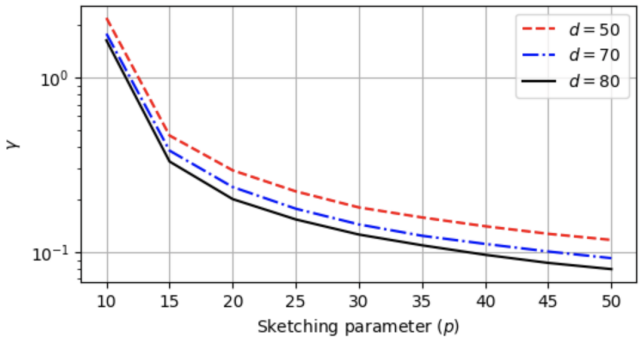

Fig. 1. Characterizing γ as a function of p and d. γ decreases as p increases, leading to a corresponding reduction in NMSE.

than classical rank-l approximation. Such a multiplicative factor is standard in the high-rank matrix approximation literature [18], [28], [32], [33].

Furthermore, we characterize the imperfections of the side information. Due to the unknown structure of $\mathbf{E}$, when $\mathbf{E} \neq \mathbf{0}$, the error arising from the structure of the row-space of $\mathbf{E}$ and the side information $\mathbf{S}$ contributes to the third and fourth terms in (14). First, if the row-space of $\mathbf{E}$ is a subspace of the row-space $\mathbf{S}$, i.e., co-range$(\mathbf{E}) \subset$ range$(\mathbf{V}_S)$, the row-space of $\mathbf{E}$ is a linear combination of the rows in $\mathbf{V}_S^\mathsf{T}$. This coincides with the case of the row-space of $\mathbf{M}$ being equal to that of $\mathbf{S}$, and then the rank$(\mathbf{M}) = l$. Then, it is straightforward to observe $\sigma_{l+1}(\mathbf{M}) = 0$, and the error becomes $\left\|\mathbf{M} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2 = 0$, as we sample all columns. Only in this case, can we perfectly recover $\mathbf{M}$. Otherwise, $\mathbf{M}$ has higher rank, $l < k$. Specifically, $\left\|\mathbf{E}\left(\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right)\right\|_2^2$ represents how informative the side information is. For example, the matrix $\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^\mathsf{T}$ serves as a projection onto the complement of the row-space of $\mathbf{S}$. If rowspace$(\mathbf{E}) \subset$ rowspace$(\mathbf{S})$, then $\left\|\mathbf{E}\left(\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right)\right\|_2 = 0$. Consequently, $\left\|\mathbf{E}\left(\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right)\right\|_2$ effectively measures how informative $\mathbf{S}$ is in describing the structure of $\mathbf{M}$; the smaller the value, the more informative the side information.

Interpreting γ. Starting from the SVD of $\mathbf{M}$ from (13), we can determine an alternative expression for $\mathbf{A}$ which is given by $\mathbf{A} = \mathbf{M}\mathbf{\Psi} = \mathbf{U}\mathbf{\Sigma}_1\mathbf{V}_1^\mathsf{T}\mathbf{\Psi} + \mathbf{U}\mathbf{\Sigma}_2\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}$. Here, the first term $\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}$ represents the sampled columns of a matrix composed of the top l singular vectors and values of $\mathbf{M}$. Consequently, $\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}$ can be interpreted as a sketch of the best rank-l approximation of $\mathbf{M}$. We then obtain the bound: $\left\|\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_2^2 \leq \|\mathbf{U}\|_2^2 \|\mathbf{\Sigma}_1\|_2^2 \left\|\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_2^2 = \sigma_1^2(\mathbf{M}) \cdot \sigma_1^2(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R})$, using classical matrix norm bounds. Thus, the term $\gamma \doteq \frac{\sigma_{l+1}^2(\mathbf{M})}{\sigma_1^2(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R})\cdot\sigma_1^2(\mathbf{M})}$ in Theorem 1 is an "effective eigenratio between $\mathbf{M}$ and $\tilde{\mathbf{A}}$". The term $\frac{\sigma_{l+1}^2(\mathbf{M})}{\sigma_1^2(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R})\cdot\sigma_1^2(\mathbf{M})}$ measures *how informative the sketch $\tilde{\mathbf{A}}$ is with respect to the original matrix* $\mathbf{M}$. We observe that the more informative $\tilde{\mathbf{A}}$, i.e., a better rank-l approximation of $\mathbf{M}$, this ratio becomes smaller, which makes the bound tighter. When $d << m$, the sketched $\mathbf{A}$ incurs a loss, impacting the bound. This observation is numerically validated in Fig. 1. Lastly, as the target rank l increases, we observe that the influence of the second term diminishes as the error between the high rank $\mathbf{M}$ and its estimate $\hat{\mathbf{M}}$ decreases.

Comparison with QPMA [8] and sCMA [11]. We compare our spectral error result with those of our prior column-based matrix approximation methods. We assume for the rest of the paper that the incoherences, are constant, i.e., $\mu = O(1)$ and $\mu_s = O(1)$. We denote the low-rank matrix approximation obtained by QPMA, sCMA and fSCMA as $\hat{\mathbf{M}}_{\text{QPMA}}$, $\hat{\mathbf{M}}_{\text{sCMA}}$ and $\hat{\mathbf{M}}_{\text{fSCMA}}$, respectively. Then, with high probability, QPMA and sCMA, converting the corresponding bounds [8], [11] to big O notation, we have,

$$\left\|\mathbf{M} - \hat{\mathbf{M}}_{\text{QPMA}}\right\|_2^2 \leq O\left(\frac{mn}{d^2}\sigma_{l+1}^2(\mathbf{M})\right) + O\left(\frac{m}{d}\|\mathbf{E}\|_F^2\right),$$

and

$$\left\|\mathbf{M} - \hat{\mathbf{M}}_{\text{sCMA}}\right\|_2^2 \leq O\left(\frac{n}{p}\sigma_{l+1}^2(\mathbf{M})\right) + O(1)\cdot\|\mathbf{M}\|_F^2.$$

In contrast, fSCMA shows the following error bound with $d = O(l\log l)$,

$$\left\|\mathbf{M} - \hat{\mathbf{M}}_{\text{fSCMA}}\right\|_2^2 \leq O\left(\frac{n}{d} + m - l\right)\sigma_{l+1}^2(\mathbf{M}) + O\left(\frac{m}{d}\|\mathbf{E}\|_2^2\right).$$

In comparison with QPMA which requires $d = O(l^2\log l)$ number of sampled columns, fSCMA achieves a sample complexity of $d = O(l\log l)$, which improves the sample complexity of QPMA by a factor of l. The additional complexity $O(l)$ of QPMA is due to the use of gradient descent to optimize the convex objective function to solve for $\mathbf{X}$ in (12). In contrast, fSCMA, has a closed form solution for $\mathbf{X}$, if $p < d$. Furthermore, QPMA has an additional multiplicative factor of $O(\frac{m}{d})$ associated with $\sigma_{l+1}^2(\mathbf{M})$ compared to the proposed method also due to the gradient descent step. This tighter error bound in Theorem 1 is achieved by eliminating the gradient descent step in QPMA.

Time complexity of fSCMA. We analyze the computational time complexity of fSCMA (Algorithm 1) versus other algorithms, emphasizing the dominant operations in each step, such as the SVD, pseudo-inverse computation, and gradient descent, versus matrix multiplication. Notably, the column-space approximation step in fSCMA does not involve any computationally dominant matrix operations. Next, the row-space $\mathbf{V}_S$ is obtained through a rank-l SVD on $\mathbf{S}$, and this takes $O(ml^2)$ time [34]. Next, the low rank matrix approximation step is performed by the pseudo-inverse of $\tilde{\mathbf{A}}$, and $\mathbf{V}_S^\mathsf{T}\mathbf{\Psi}$. The pseudo-inverse of $\tilde{\mathbf{A}}$ and $\mathbf{V}_S^\mathsf{T}\mathbf{\Psi}$ are $O(np^2)$ and $O(l^2d)$, respectively. Therefore, the overall complexity of fSCMA is $O(ml^2 + np^2 + l^2d)$. On the other hand, QPMA involves SVDs on $\mathbf{A}$ and $\hat{\mathbf{Q}}\mathbf{S}$ and gradient descent, with an overall cost is $O(ndl + nd^2l + nml + n^2dl\cdot T)$, where T is the number of iterations during the gradient descent step. Finally, sCMA requires the estimation of the coefficient matrix $\mathbf{Q}$ and the computation of the pseudo-inverse of $\mathbf{\Omega}\mathbf{Q}\mathbf{S}\mathbf{\Psi}$, where $\mathbf{\Omega}$ is the row-sampling operator, with an overall cost of $O(n^2dl + d^2p)$. We numerically compare the time complexity of QPMA, sCMA and fSCMA in Section VI. Due to the fact that $l \leq p \leq d << \min\{n, m\}$, the runtime complexity of fSCMA is significantly lower than that of QPMA.

V. Proof Sketch and Key Lemmas

We present the proof sketch of our key result along with the key lemmas required to establish Theorem 1. The full proofs

are provided in the Appendix. For Theorem 1, we start with bounding $\left\|\mathbf{M}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2$:

$$\begin{aligned}&\left\|\mathbf{M}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2\\&=\left\|\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}+\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2\\&\overset{(a)}{\leq}\underbrace{2\left\|\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\right\|_2^2}_{\otimes}+\underbrace{2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2}_{\oplus},\end{aligned}$$

where (a) is due to the triangle inequality and the fact that for $a,b\geq 0$, $(a+b)^2\leq 2(a^2+b^2)$. Note that $\mathbf{P}_{\mathbf{U}_A}=\mathbf{U}_A\mathbf{U}_A^{\mathsf{T}}$, where $\mathbf{U}_A$ is the left singular vectors of rank-d SVD of $\mathbf{A}$ and $\mathbf{P}_{\mathbf{V}_S}=\mathbf{V}_S\mathbf{V}_S^{\mathsf{T}}$, where $\mathbf{V}_S$ is defined in (6). Next, we bound each term denoted by $\otimes$ and $\oplus$ with high probability via Propositions 1 and 2. Note that $\otimes$ represents the energy of $\mathbf{M}$ that is orthogonal to the estimated (l-dimensional) row and column spaces.

Proposition 1. *If $d\geq a_1\mu l\ln l$, under the conditions of Theorem 1, we have that*

$$\begin{aligned}&\|\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2\\&\quad\leq 2\sigma_{l+1}^2(\mathbf{M})\left(1+2\frac{n}{d}\right)+2\left\|\mathbf{E}-\mathbf{E}\mathbf{V}_S\mathbf{V}_S^{\mathsf{T}}\right\|_2^2.\end{aligned}$$

with probability at least $1-a_2l^{-3}$, where a_1 and $a_2>0$ are constants.

Remark. The complete proof of Proposition 1 is provided in Appendix D. The proof follows from first invoking [35, Theorem 6] to bound the energy of $\mathbf{M}$ orthogonal to $\mathbf{U}_A$ and $\mathbf{V}_S$.

Proposition 1 accounts for errors arising from two factors: (i) the high rank of $\mathbf{M}$, where $\text{rank}(\mathbf{M})=k>>l$ and (ii) the sub-sampling of columns, both of which contribute to the first term. When $\mathbf{E}=\mathbf{0}$, the true rank of $\mathbf{M}$, k, is equal to l and the error becomes zero. This is because $\sigma_{l+1}^2(\mathbf{M})=0$ and the row-space of $\mathbf{E}$ fully aligns with the row-space of $\mathbf{S}$, leading to $\left\|\mathbf{E}-\mathbf{E}\mathbf{V}_S\mathbf{V}_S^{\mathsf{T}}\right\|_2^2=0$. The second term arises due to the unknown structure of $\mathbf{E}$.

Next, $\oplus$ represents the error arising from the use of the sketch $\tilde{\mathbf{A}}$ for estimating the column-space of $\mathbf{M}$, as well as the final matrix approximation step, which involves estimating $\mathbf{Z}$. This error is bounded using the Proposition 2 below.

Proposition 2. *If $d\geq\max\{b_1\mu l\ln l,b_2\mu_s l\ln l\}$ and $p=O(\log(1/\delta)/\epsilon^2)$, where $\delta\doteq\max\{e^{-b_3\epsilon^2 p},b_4l^{-3}\}$, then under the conditions of Theorem 1, we have that,*

$$\begin{aligned}&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2\\&\leq 4\sigma_{l+1}^2(\mathbf{M})\left(2+\frac{2n}{d}+\frac{(1+\epsilon)(m-l)}{\sigma_1^2\left(\mathbf{V}_1^{\mathsf{T}}\mathbf{\Psi}\mathbf{R}\right)}\right)\\&\qquad+4\sigma_2^2(\mathbf{E})\left(2+\frac{2m}{d}\right),\end{aligned}$$

with probability $1-\delta$, where b_1,b_2,b_3,b_4 and $\epsilon>0$ are constants.

Remark. The proof of Proposition 2 is provided in Appendix G. It follows by carefully applying the Johnson-Lindenstrauss transform (see Definition 1) to $\mathbf{R}$ with respect to $\mathbf{A}$. As will be seen in the Appendix, the impact of $\mathbf{A}$ is revealed in the columns of $\mathbf{V}_2$ in (13). We next analyze the error between the projection of $\mathbf{M}$ onto the column-space of $\mathbf{A}$ and that of $\tilde{\mathbf{A}}$. Combining Propositions 1 and 2 completes the proof of Theorem 1.

VI. Numerical Results

In this section, we evaluate the performance of fSCMA on both synthetic and real-world datasets. All experiments on synthetic data are averaged over 300 independent iterations. The code is available at https://github.com/JeongminChae/fSCMA.

Benchmark Algorithms. While most matrix approximation algorithms require access to the complete, true rows and columns of the matrix – making direct comparisons more challenging – recently proposed methods such as CUR+ [9], sCMA [11], and QPMA [8] operate using only a limited number of sampled columns (or rows) and are thus appropriate for comparison to fSCMA. These algorithms are assessed based on four criteria: (i) sampling strategies employed, (ii) the informativeness of the side information utilized, (iii) the effectiveness of sketching, and (iv) computational efficiency.

We briefly describe QPMA, sCMA and CUR+. Similar to fSCMA, both QPMA and sCMA compute a low-rank matrix approximation of $\mathbf{M}$ using only d sampled columns and the side information $\mathbf{S}$. However, both QPMA and sCMA require solving for $\mathbf{Q}$ in (1) as part of their process. QPMA consists of three-step process; (i) column-space estimation, (ii) row-space estimation, and (iii) low rank matrix approximation of $\mathbf{M}$. For step (i), QPMA performs and SVD on the sampled columns, $\mathbf{A}$, and derives its column-space. Next, it solves for $\mathbf{Q}$ using gradient descent, leveraging $\mathbf{A}$ and $\mathbf{S}$. With the estimated coefficient matrix, $\hat{\mathbf{Q}}$, QPMA estimates the row-space by performing an SVD on $\hat{\mathbf{Q}}\mathbf{S}$. Finally, QPMA solves for an intersection matrix using a gradient descent method that incorporates the estimated column-space and row-space. QPMA is implemented with a fixed step-size $\eta=0.01$ for solving $\mathbf{Q}$ via gradient descent, with a maximum number of iterations $T=1500$. sCMA [11] has a two-step process; first solving for $\mathbf{Q}$ employing least squares estimation, with the given sampled columns, $\mathbf{A}$, and $\mathbf{S}$. And, with the estimated coefficient matrix, $\hat{\mathbf{Q}}$, sCMA samples a few rows of $\hat{\mathbf{Q}}\mathbf{S}$. Let the sampled row matrix be $\mathbf{B}$. Then, sCMA computes its low-rank matrix approximation as $\mathbf{A}(\mathbf{B}\mathbf{\Psi})^{\dagger}\mathbf{B}$. Finally, CUR+ samples a subset of the full **true** columns, i.e., $\mathbf{A}$, and full rows of the **true** matrix, i.e., $\mathbf{B}$. Thus, for column- and row-space estimation, CUR+ performs an SVD on a set of true columns and rows of $\mathbf{M}$. Lastly, the low-rank matrix approximation step (corresponding line 9 in Algorithm 1 in [8]m) is performed with the low rank column-space, i.e., $\mathbf{U}_A$, and row-space, i.e., $\mathbf{V}_B$, obtained from SVD on the **true** columns and rows; $\text{argmin}_{\mathbf{X}}\|\hat{\mathbf{Q}}\mathbf{S}-\mathbf{U}_A\mathbf{X}\mathbf{V}_B^{\mathsf{T}}\|_F^2$. On the other hand, fSCMA estimates the column-space and the row-space through a sketch of $\mathbf{A}$ and the side information $\mathbf{S}$.

To compare the sampling complexity of each algorithm, fSCMA samples only d columns, resulting in access to a total

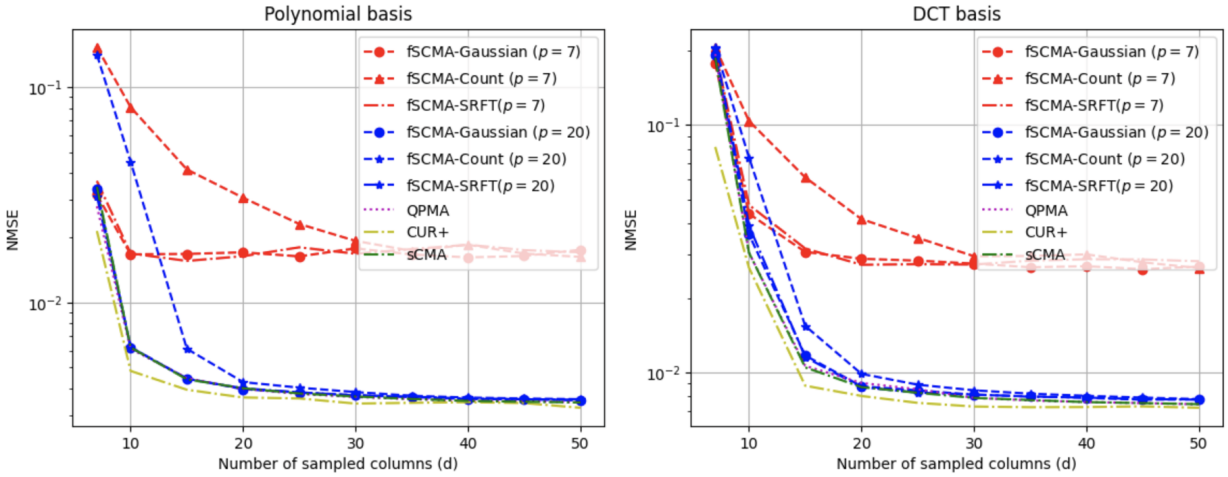

Fig. 2. The NMSE versus the number of true sampled columns, d when $\mathbf{S}$ is a polynomial basis matrix and a DCT basis. Here, $k = 100$ and $l = 7$.

TABLE II
COMPARISON OF SAMPLING NEEDS FOR BASELINE ALGORITHMS

Algorithm	# rows	# columns	Total samples
CUR+	d	d	$2nd - d^2$
QPMA	0	d	nd
sCMA	0	d	nd
fSCMA	0	d	nd

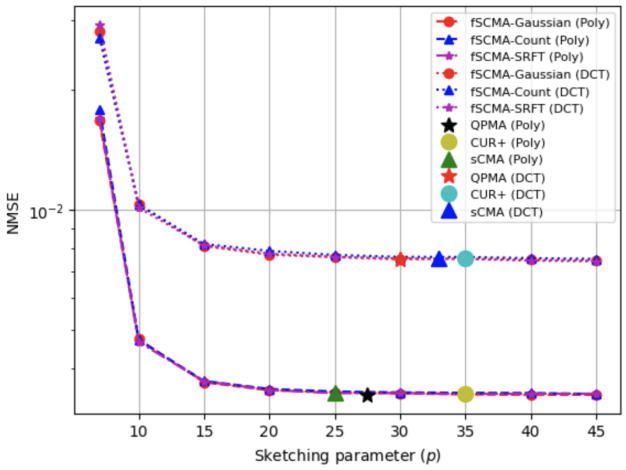

Fig. 3. Characterizing the informativeness of the sketch $\tilde{\mathbf{A}}$ through numerical performance varying p when $k = 100$ and $d = 50$.

of nd entries. Additionally, fSCMA estimates the column-space by projecting the d-dimensional space onto an p-dimensional subspace using random projections. In contrast, CUR+ requires a larger number of samples compared to fSCMA, utilizing d rows and d columns each. Although QPMA and sCMA also sample d columns similar to fSCMA, both algorithms benefit from additional information in the form of the estimated coefficient matrix, $\hat{\mathbf{Q}}$. Via the comparison with benchmark algorithms, we seek to evaluate the informativeness of the side information $\mathbf{S}$ in the absence of coefficient information $\mathbf{Q}$, as well as the effectiveness of sketching. Table II summarizes the sampling needs of each scheme.

A. Synthetic data

Data generation. We generate the data as follows. Matrix dimensions are $n = m = 100$ throughout. The entries of the coefficient matrix $\mathbf{Q} \in \mathbb{R}^{n\times l}$ are drawn identically and independently from $\mathcal{N}(0,1)$. For $\mathbf{S} \in \mathbb{R}^{l\times m}$, we consider the *discrete cosine transform (DCT)* [36] and a *polynomials* of the reaction coordinate values $\mathbf{s}$ as seen in (2). For the DCT matrix, we generate a $l \times m$ matrix $\mathbf{S}$, where each of the l rows corresponds to a different frequency and i-th column of $\mathbf{S}$ is defined as $\mathbf{S}_{\{j,i\}} = \alpha_j \cos\left(\frac{\pi j(2i+1))}{m}\right)$, where $j \in \{0,\dots,l-1\}$ is the index of frequency components, $i \in \{0,\dots,m-1\}$ is the index for sample points. The value α_j is a normalization factor, where $\alpha_j = \sqrt{\frac{1}{m}}$ if $j = 0$ and $\alpha_j = \sqrt{\frac{2}{m}}$ if $j > 0$. The polynomial matrix is constructed using arbitrarily sampled polynomial coordinate values [21], [23] as in Eq. (2). To control the rank of the underlying ground truth matrix, we generate the perturbation matrix $\mathbf{E}$ as $\mathbf{U}_{QS}\mathbf{R}_1\mathbf{V}_{QS}^\mathsf{T} + \mathbf{U}_{QS,\perp,[:,1:k-l]}\mathbf{R}_2\mathbf{V}_{QS,\perp,[:,1:k-l]}^\mathsf{T}$, where $\mathbf{QS} \overset{SVD}{=} \mathbf{U}_{QS}\mathbf{\Sigma}_{QS}\mathbf{V}_{QS}^\mathsf{T} + \mathbf{U}_{QS,\perp}\mathbf{\Sigma}_{QS,\perp}\mathbf{V}_{QS,\perp}^\mathsf{T}$. The entries of $\mathbf{R}_1 \in \mathbb{R}^{l\times l}$ and $\mathbf{R}_2 \in \mathbb{R}^{k-l\times k-l}$ are generated i.i.d. from $\mathcal{N}(0,\sigma)$ with $\sigma = 0.001$. With this approach, we obtain the rank-k ground truth matrix, $\mathbf{M} = \mathbf{QS} + \mathbf{E}$. We use the normalized mean squared error (NMSE), defined as $\text{NMSE}(\mathbf{M},\hat{\mathbf{M}}) = \frac{\|\mathbf{M}-\hat{\mathbf{M}}\|_F}{\|\mathbf{M}\|_F}$,where $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$, to measure the performance.

Sketching is done by post-multiplying $\mathbf{A}$ with an $d \times p$ JLT random matrix $\mathbf{R}$. For the sketching matrix $\mathbf{R}$, we investigate the numerical performance of fSCMA using the following three different random matrices drawn from different distributions: (i) random Gaussian matrices whose entries are i.i.d. Gaussian random variables with zero mean and unit variance [6], [37], [38], (ii) Count matrices [39] and (iii) the Subsampled Randomized Fourier Transform (SRFT) [17], [40]. We further explain how we generate $\mathbf{R}$ for the cases of Count matrices and SRFT.

- **Count sketching.** For any fixed matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and the sketching dimension p, we first hash each column with a discrete value which is uniformly sampled from $\{1,\dots,p\}$, then flip the sign of each column with prob-

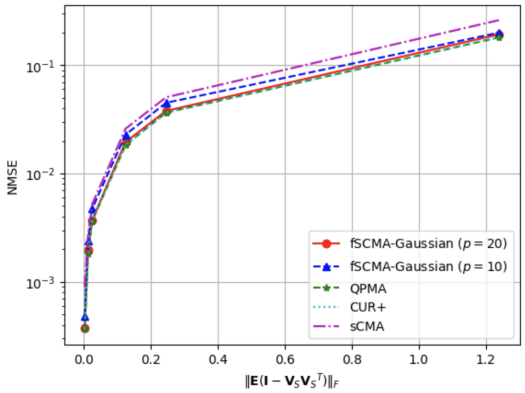

Fig. 4. The sensitivity of fSCMA to the row-space between $\mathbf{E}$ and $\mathbf{QS}$ when $d = 50$, $k = 100$, $l = 7$ and $p \in \{10, 20\}$.

ability $\frac{1}{2}$, and finally sum up the columns with the same hash value. The result is an $n \times p$ matrix $\tilde{\mathbf{A}} = \mathbf{AR}$. The matrix $\mathbf{R} \in \mathbb{R}^{d\times p}$ has exactly one nonzero entry in each row, and the entry can be either +1 or -1.

- **Subsampled Randomized Fourier Transform (SRFT).** SRFT sketching matrix $\mathbf{R}$ takes the form $\mathbf{R} = \mathbf{DFP}$, where $\mathbf{D}$ is a $d \times d$ diagonal matrix with independent Rademacher entries (± 1 with equal probability.); $\mathbf{F}$ is a $d \times d$ discrete Fourier transform matrices; $\mathbf{P}$ is a $d \times p$ restriction matrix onto p coordinates, chosen uniformly at random.

In this paper, we denote fSCMA with Gaussian, Count and SRFT sketching matrices as fSCMA-Gaussian, fSCMA-Count, and fSCMA-SRFT, respectively.

Varying d. We begin by evaluating the performance of all algorithms as a function of the number of sampled columns (d) in Fig. 2. We evaluate performance for when the true rank $k = 100$ and for two types of $\mathbf{S}$, while fixing the sketching dimension to $p = \{7, 20\}$ when the target rank $l = 7$. Recall the relationship between the parameters: $l \leq p \leq d \leq k$. When $d < 20$, we set $p = d$. The standard deviation of the components of $\mathbf{E}$ is set to $\sigma = 0.001$. It is important to note that the overall noise level is significantly higher, as numerical evaluations are based on the Frobenius norm.

We first observe that as d increases, the NMSE decreases since more information about the column-space is incorporated. As expected, the larger $p = 20$ shows lower NMSE compare to when $p = 7$, which implies that the sketch $\tilde{\mathbf{A}}$ contains more information about the column-space of $\mathbf{M}$. Furthermore, for both polynomial and DCT matrices, $\mathbf{S}$, the performances of all fSCMA algorithms are comparable to that of CUR+, QPMA and sCMA when $p = 20$ and $d \geq 25$. This implies that the sketch $\tilde{\mathbf{A}}$ can successfully represent the lower-dimensional column-space of $\mathbf{A}$ when the number of sampled columns $d \geq 25$ and $p = 20$. This is impressive because fSCMA with $p = 20$ shows near identical performance to the benchmark algorithms which sample $d = 50$ columns. On the other hand, when $p = 7$, we see that the lower-dimensional column-space of $\tilde{\mathbf{A}} \in \mathbb{R}^{100\times 7}$ fails to capture as much column information as when $p = 20$. We recall that QPMA, CUR+, and sCMA incorporate additional information via the

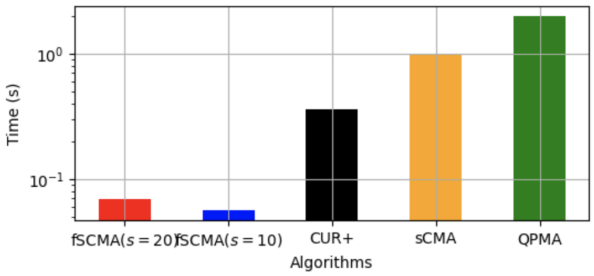

Fig. 5. Comparison of run time complexity (log scale) for all algorithms.

estimated of $\mathbf{Q}$. However, the fact that NMSE performance remains comparable when $d \geq 25$ and $p = 20$ numerically confirms our argument that the performance primarily depends on the row-space of the side information $\mathbf{S}$, rather than precisely solving for $\mathbf{Q}$, as long as a sufficient number of columns (d) are sampled. Additionally, this demonstrates that fSCMA effectively leverages the information in $\mathbf{S}$ to estimate the row-space of $\mathbf{M}$ without requiring the estimation of $\mathbf{Q}$.

Informativeness of the sketch $\tilde{\mathbf{A}}$. We next attempt to answer the following question: *how much column-space information of $\mathbf{M}$ is being captured by the sketching matrix $\tilde{\mathbf{A}}$?* To this end, for QPMA, CUR+, and sCMA, we fix the number of observed columns to $d = 50$ and numerically compute the p required for fSCMA to attain the same (fixed) numerical error as that of QPMA, CUR+ and sCMA. The results are shown in Fig 3. The data is generated following the same procedure as described in Section VI-A with $l = 7$, $k = 100$ and $d = 50$. We plot the NMSE of QPMA, CUR+ and sCMA by comparing the best NMSE performance of fSCMA with different sketching matrices. For the polynomial matrix $\mathbf{S}$, the NMSE of QPMA, CUR+ and sCMA when $d = 50$ correspond to that of fSCMA when p is 27, 35 and 25, respectively. Also, for the DCT matrix, the NMSE of QPMA, CUR+ and sCMA when $d = 50$ correspond to that of fSCMA when p is 30, 35, and 33 Empirically, we have determined that fSCMA has a sketching dimension that is generally some fraction of the dimension of $\mathbf{A}$, that is $p \approx \alpha d, \alpha \in [0.5, 0.7]$ to achieve the performance of the benchmark algorithms in case of the polynomial matrix, and $\alpha \in [0.6, 0.7]$ for the DCT matrix. These results also align well with those in Fig. 2, which shows that fSCMA achieves performance comparable to that of the benchmark algorithms when $d \geq 25$.

Varying p. Additionally, we analyze the effect of varying the sketching parameter, p, in Fig. 3. We first observe that the NMSE decreases as p increases. We see that a larger p can capture more information about the column-space of $\mathbf{A}$, thereby providing more information about $\mathbf{M}$ to $\mathbf{Z}$. For example, if $p = d$, i.e., $p = d = 50$, $\tilde{\mathbf{A}}$ will have the complete column-space as $\mathbf{A}$, that is, $\tilde{\mathbf{A}} = \mathbf{A}$. We also observe that the NMSE remains nearly constant in the $p \geq 20$ region (when $d = 50$), suggesting that the sketch $\tilde{\mathbf{A}}$ adequately captures the column-space of $\mathbf{A}$ with $p = 20$.

Sensitivity to the structure of $\mathbf{E}$. We next investigate the sensitivity of fSCMA to the structure of $\mathbf{E}$ in Fig. 4. We generate the data as previously with $k = 100, d = 50$ and $p \in \{10, 20\}$. Recall that $\mathbf{E}$ is generated as

$\mathbf{U}_{QS}\mathbf{R}_1\mathbf{V}_{QS}^{\mathsf{T}} + \mathbf{U}_{QS,\perp,[:,1:k-l]}\mathbf{R}_2\mathbf{V}_{QS,\perp,[:,1:k-l]}^{\mathsf{T}}$, where $\mathbf{R}_2$ is generated i.i.d from $\mathcal{N}(0,\sigma)$. We vary the standard deviation, $\sigma = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ to control the canonical angle [30] between $\mathbf{QS}$ and $\mathbf{E}$. We use $\|\mathbf{E}(\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^{\mathsf{T}})\|_F$ in Theorem 1 (see the fourth term in (14)) as a distance measure between the two row spaces of $\mathbf{E}$ and $\mathbf{V}_S$. $\mathbf{R}_2$ plays a role to control the distance between the row space of $\mathbf{E}$ and $\mathbf{S}$. A larger σ indicates greater misalignment and hence a larger distance. Only when $\mathbf{E}$ lies perfectly in the row-space of $\mathbf{S}$, i.e., $\|\mathbf{E}(\mathbf{I} - \mathbf{V}_S\mathbf{V}_S^{\mathsf{T}})\|_F = 0$, will rank($\mathbf{M}$) = l, and therefore $\sigma_{l+1}(\mathbf{M}) = 0$. The numerical NMSE goes to zero as well, in this case. As the Frobenius norm of the projection of $\mathbf{E}$ onto the row-space of $\mathbf{S}$ increases, we observe that the NMSE of all algorithms increase. For fSCMA in particular, this observation is consistent with Theorem 1.

Run time complexity and efficiency. We numerically compare the run-time complexity and efficiency of fSCMA in Fig. 5. Among all the benchmark algorithms, fSCMA clearly exhibits the shortest runtime. We recall that all benchmark algorithms require solving for the coefficient matrix $\mathbf{Q}$ using either gradient descent or least-squares, which is not necessary for fSCMA. Furthermore, QPMA estimates the column- and row-spaces via an SVD. The main source of runtime complexity for QPMA comes from the large number of iterations for gradient descent to solve for $\mathbf{Q}$. We see that fSCMA enjoys a significant reduction in computational run time – more than 50 times faster than QPMA and more than 20 times faster than CUR+.

B. Real data

We evaluate fSCMA on the the matrix of Hessian eigenvalues constituting the reaction path of a chemical reaction provided in [7], specifically the `Ir` reaction system, where $\mathbf{M} \in \mathbb{R}^{24\times 131}$. As l is unknown *a priori*, we perform a rank-d SVD on the randomly sampled columns $\mathbf{A} \in \mathbb{R}^{24\times d}$ for $d \in \{7, 10, 12, 14, 16, 18, 20\}$. Guided by the singular value gap of $\mathbf{A}$ and our assumption $l \leq d$, for the real data we find that $l = 7$ is suggested by this approach. In addition, the polynomial basis matrix $\mathbf{S}$ is determined for the reaction coordinate values $\mathbf{s} = [1 + 0.01 \cdot m]$, where $m = 131$, following the methodology proposed in [7]. For a more detailed description of the dataset, please refer to [7]. We observe that l is a hyper-parameter. Given that a sufficient number of columns are sampled, d (see Proposition 1), from the Eckart-Young-Mirsky theorem [28], we can obtain a good estimate of the target rank l by the rank-d SVD of $\mathbf{A}$.

To simulate a column-sampling-only setting while limiting access to true rows, we implement CUR+ by providing d estimated rows (via fSCMA) alongside $24 - d$ true rows; while this is not fully fair comparison, it enables CUR+ to be implemented without access to the true rows of $\mathbf{M}$. Additionally, we set the sketching parameter $p = 7$. We present the results in Fig. 6. We observe that over the entire sampling region, fSCMA exhibits performance nearly identical to QPMA; thus fSCMA sucessfully leverages the polynomial matrix $\mathbf{S}$ to estimate the row-space of $\mathbf{M}$. In addition, in the low-sample regime when $d \leq 11$, fSCMA outperforms CUR+.

As the sample size increases, CUR+ tends to yield better performance. However, the comparison to CUR+ is not fully fair, as we provide estimated rows (via fSCMA) to CUR+ which is not possible in real-world environments. Furthemore, the computational complexity of CUR+ becomes large when increasing the number of sampled columns. Thus, we see that fSCMA does, in fact, work well for the application that motivated its creation.

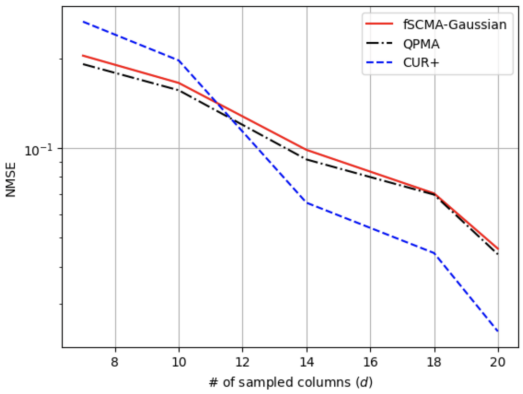

Fig. 6. NMSE versus the number of sampled columns d for Ir chemical system with $l = 7$, $p = 7$ and $k = 24$.

VII. Conclusions

We have formulated a new algorithm, fast Sketched Column-based Matrix Approximation (fSCMA) for low rank matrix approximation which leverages sketching and removes the need to estimate an interim variable yielding strongly reduced complexity over our prior methods [8]. Our approach is tailored to our constraints of access to only a few fully sampled columns and key structural side information. Sketching is performed on the sampled columns to estimate the column space. We provided a theoretical guarantee on the reconstruction error. This error bound is comprised of a component due to the deviation of the true matrix from the structural side information as well as a term due to the spectral properties of the ground truth matrix and the sketching matrices. The dependence on the sketched column-space of the target matrix is also characterized. Numerical results validate the theoretical guarantees and application of the proposed method to real data shows strong complexity improvements while maintaining performance.

Acknowledgements

This work is funded in part by one or more of the following grants: ONR N00014-22-1-2363, NSF CCF 2148313, NSF CCF-2200221, ARO W911NF2410094, NSF CCF 2311653, ARO W911NF1910269, and the NSF Center for Pandemic Insights DBI-2412522.

Appendix A
Technical preliminaries

Recall the SVD of $\mathbf{M}$ in (13). With this, a set of sampled column matrix $\mathbf{A}$ defined in (4) and its sketch $\tilde{\mathbf{A}}$ defined in (10) can be expressed as

$$\mathbf{A} = \mathbf{M}\mathbf{\Psi} = \mathbf{U}\begin{bmatrix}\underbrace{\mathbf{\Sigma}_1\mathbf{\Psi}_1}_{l\times d}\\ \underbrace{\mathbf{\Sigma}_2\mathbf{\Psi}_2}_{(m-l)\times d}\end{bmatrix} \text{ and } \tilde{\mathbf{A}} = \mathbf{A}\mathbf{R} = \mathbf{U}\begin{bmatrix}\underbrace{\mathbf{\Sigma}_1\mathbf{\Psi}_1\mathbf{R}}_{l\times p}\\ \underbrace{\mathbf{\Sigma}_2\mathbf{\Psi}_2\mathbf{R}}_{(m-l)\times p}\end{bmatrix},$$

TABLE III
SUMMARY OF KEY PARAMETERS

Parameter	Description	Parameter	Description
l	Target rank	n, m	Dimension of $\mathbf{M}$
d	# of sampled columns	μ	Incoherence of $\mathbf{M}$
p	Sketching parameter	μ_s	Incoherence of $\mathbf{S}$

where $\mathbf{\Psi}_1 = \mathbf{V}_1^\mathsf{T}\mathbf{\Psi}$ and $\mathbf{\Psi}_2 = \mathbf{V}_2^\mathsf{T}\mathbf{\Psi}$, where $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\mathbf{V}_1^\mathsf{T}$ and $\mathbf{V}_2^\mathsf{T}$ are defined in (13). Recall the definition of SVD of a matrix $\mathbf{X}$ in (5). We define rank-l SVD of $\mathbf{A} \in \mathbb{R}^{n\times d}$ as $\mathbf{A} \overset{l-\text{SVD}}{=} \mathbf{U}_A\mathbf{\Sigma}_A\mathbf{V}_A^\mathsf{T}$, where $l \le \min\{n, d\}$. Then, the rank-ρ SVD of $\tilde{\mathbf{A}}$ is given by

$$\tilde{\mathbf{A}} \overset{\rho-\text{SVD}}{=} \mathbf{U}_{\tilde{A}}\mathbf{\Sigma}_{\tilde{A}}\mathbf{V}_{\tilde{A}}^\mathsf{T}, \tag{15}$$

where ρ is the rank of $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times p}$, and $l \le \rho \le p$. Lastly, we recall the key parameters that are used throughout the Appendix.

APPENDIX B
PRIOR RESULTS

Our key proof relies on prior results from [25], characterizing projections (see (7)):

Proposition 3. *Let $\mathbf{P}_X$ be an orthogonal projector onto $\mathbf{X}$. Suppose $\mathbf{U}$ is unitary. Then $\mathbf{U}^\mathsf{T}\mathbf{P}_X\mathbf{U} = \mathbf{P}_{U^\mathsf{T}X}$.*

Proposition 4. *For a matrix $\mathbf{M}$ and $\mathbf{N}$, suppose range$(\mathbf{N}) \subset$ range$(\mathbf{M})$. Then, for each matrix $\mathbf{X}$, it holds that $\|\mathbf{P}_N\mathbf{X}\|_2 \le \|\mathbf{P}_M\mathbf{X}\|_2$ and that $\|(\mathbf{I}-\mathbf{P}_M)\mathbf{X}\|_2 \le \|(\mathbf{I}-\mathbf{P}_N)\mathbf{X}\|_2$.*

Proposition 5. *(Perturbation of Inverses). Suppose that a matrix $\mathbf{X}$ is a positive semidefinite matrix, i.e., $\mathbf{0} \preccurlyeq \mathbf{X}$. Then, $\mathbf{I} - (\mathbf{I}+\mathbf{X})^{-1} \preccurlyeq \mathbf{X}$.*

Proposition 6. *We have $\|\mathbf{X}\| \le \|\mathbf{Y}\| + \|\mathbf{T}\|$ for each partitioned positive semi definite matrix $\mathbf{X} = \begin{bmatrix}\mathbf{Y} & \mathbf{D}\\ \mathbf{D}^\mathsf{T} & \mathbf{T}\end{bmatrix}$.*

We also need the following results on bounding singular values of random matrices and large deviation events:

Lemma 1. *[9] With probability $1-c_1l^{-3}$, and if $d \ge c_2\mu l\ln l$, for constants $c_1 > 0$ and $c_2 > 0$, we have*

$$\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\|_2^2 \le \sigma_{l+1}^2(\mathbf{M})\left(1+2\frac{n}{d}\right). \quad \boxtimes$$

Next, we provide an additional necessary theorem from [31] which describes the large-deviation behavior of specific types of matrix random variables.

Theorem 2. *[31] Let $\mathcal{X}$ be a finite set of positive semi definite matrices of dimension k. If there exists a constant $B < \infty$ such that $\max_{\mathbf{X}\in\mathcal{X}}\lambda_{max}(\mathbf{X}) \le B$, and, if we sample $\{\mathbf{X}_1, \dots, \mathbf{X}_q\}$ uniformly at random from $\mathcal{X}$ without replacement, with $\mu_{max} := q\lambda_{max}(\mathbb{E}[\mathbf{X}_1])$ and $\mu_{min} := q\lambda_{min}(\mathbb{E}[\mathbf{X}_1])$. Then we have that,*

$$P\left[\lambda_{max}\left(\sum_{i=1}^q \mathbf{X}_i\right) \ge (1+\rho)\mu_{max}\right] \le k\cdot exp\frac{-\mu_{max}}{B}[(1+\rho)\ln(1+\rho)-\rho] \text{ for } \rho\in[0,1)$$

$$P\left[\lambda_{min}\left(\sum_{i=1}^q \mathbf{X}_i\right) \le (1-\rho)\mu_{min}\right] \le k\cdot exp\frac{-\mu_{min}}{B}[(1-\rho)\ln(1-\rho)+\rho] \text{ for } \rho\in[0,1).$$

APPENDIX C
PROOF OF THEOREM 1

In this section, we provide the proof of our main Theorem 1. We will see below that our term of interest can be bounded by the sum of two terms, we will then provide two propositions (Proposition 1 and Proposition 2) which provide bounds on these two individual terms. First, we bound $\left\|\mathbf{M} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2$,

$$\begin{aligned}&\left\|\mathbf{M} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2\\ &= \left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} + \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2\\ &\overset{(a)}{\le} \underbrace{2\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\right\|_2^2}_{\otimes} + \underbrace{2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^\mathsf{T}\right\|_2^2}_{\oplus},\end{aligned} \tag{16}$$

where (a) is due to the triangle inequality and the fact that for real value $a, b \ge 0$, $(a+b)^2 \le 2(a^2+b^2)$. Then, Proposition 1 and Proposition 2 bound $\otimes$ and $\oplus$ respectively with high probability.

APPENDIX D
PROOF OF PROPOSITION 1

Proposition 1 bounds the energy of $\mathbf{M}$ that is captured by the column-space of $\mathbf{A}$ and the row-space of $\mathbf{S}$.

Proposition 1. If $d \ge c_1\mu l\ln l$, under the condition of Theorem 1, with probability at least $1-c_2l^{-3}$, and for $c_1, c_2 > 0$, we have that

$$\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2 \le 2\sigma_{l+1}^2(\mathbf{M})\left(1+2\frac{n}{d}\right) + 2\left\|\mathbf{E} - \mathbf{E}\mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right\|_2^2. \quad \boxtimes$$

Proof:

$$\begin{aligned}&\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2\\ &= \|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} + \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2\\ &\overset{(a)}{\le} 2\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2 + 2\|\mathbf{P}_{\mathbf{U}_A}\|_2^2\|\mathbf{M} - \mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2\\ &\overset{(b)}{=} 2\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\|_2^2 + 2\|\mathbf{M} - \mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2,\end{aligned} \tag{17}$$

where (a) is due to the triangle inequality and matrix norm inequality and (b) is due to $\|\mathbf{P}_{\mathbf{U}_A}\|_2^2 = 1$. We bound $\|\mathbf{M} - \mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2$ as follows.

$$\begin{aligned}\|\mathbf{M} - \mathbf{M}\mathbf{P}_{\mathbf{V}_S}\|_2^2 &= \|\mathbf{Q}\mathbf{S} + \mathbf{E} - (\mathbf{Q}\mathbf{S}+\mathbf{E})\mathbf{P}_{\mathbf{V}_S}\|_2^2\\ &\overset{(a)}{=} \left\|\mathbf{Q}\mathbf{S} + \mathbf{E} - \left(\mathbf{Q}\mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^\mathsf{T} + \mathbf{E}\right)\mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right\|_2^2\\ &\overset{(b)}{=} \left\|\mathbf{Q}\mathbf{S} + \mathbf{E} - \left(\mathbf{Q}\mathbf{S} + \mathbf{E}\mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right)\right\|_2^2\\ &= \left\|\mathbf{E} - \mathbf{E}\mathbf{V}_S\mathbf{V}_S^\mathsf{T}\right\|_2^2,\end{aligned} \tag{18}$$

where (a) is the SVD of $\mathbf{S}$ defined in (6) and (b) is due to $\mathbf{V}_S^\mathsf{T}\mathbf{V}_S = \mathbf{I}$. By plugging Lemma 1 and (18) into (17), we complete the proof of Proposition 1.

APPENDIX E
LEMMA 2 AND PROOF OF LEMMA 2

Lemma 2 provides a bound on the gap between the energy of $\mathbf{M}$ captured by the column-space of $\mathbf{A}$ and that captured by its sketch $\tilde{\mathbf{A}}$.

Lemma 2. *If $d \geq c_1 \mu l \ln l$ and $p = O(\log(1/\delta)/\epsilon^2)$, where $\delta \doteq \max\{e^{-c_2\epsilon^2 p}, c_3 l^{-3}\}$, we have*

$$\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2 \leq 2\sigma_{l+1}^2(\mathbf{M})\left(2 + \frac{2n}{d} + (1+\epsilon)(m-l)\left\|\left(\mathbf{V}_1^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right)^\dagger\right\|_2^2\right),$$

with probability $1-\delta$ where c_1, c_2, c_3, and $\epsilon > 0$ are numerical constants.

Proof:

$$\begin{aligned}\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2 &= \left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{M} + \mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2 \\ &\overset{(a)}{\leq} \underbrace{2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{M}\right\|_2^2}_{\odot_{1,1}} + \underbrace{2\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2}_{\odot_{1,2}}, \quad (19)\end{aligned}$$

where (a) is due to the triangle inequality. $\odot_{1,1}$ is bounded by Lemma 1. Next, $\odot_{1,2}$ can be bounded as follows. We note that the left unitary factor $\mathbf{U}$ in (13). We execute the proof for an auxiliary ground truth matrix $\tilde{\mathbf{M}} = \mathbf{U}^\mathsf{T}\mathbf{M}$ and the associated matrix $\ddot{\mathbf{A}} = \tilde{\mathbf{M}}\mathbf{\Psi}\mathbf{R}$ defined by

$$\tilde{\mathbf{M}} = \begin{bmatrix}\mathbf{\Sigma}_1\mathbf{V}_1^\mathsf{T} \\ \mathbf{\Sigma}_2\mathbf{V}_2^\mathsf{T}\end{bmatrix} \text{ and } \ddot{\mathbf{A}} = \begin{bmatrix}\mathbf{\Sigma}_1\mathbf{\Psi}_1\mathbf{R} \\ \mathbf{\Sigma}_2\mathbf{\Psi}_2\mathbf{R}\end{bmatrix}. \quad (20)$$

Recall the definition of $\mathbf{M}$ in (13). Due to the unitary invariance of the spectral norm and Proposition 3 (see Appendix B), we have the following identity

$$\begin{aligned}\odot_{1,2} = \left\|\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\right)\mathbf{M}\right\|_2^2 &= \left\|\mathbf{U}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\right)\mathbf{U}\tilde{\mathbf{M}}\right\|_2^2 \\ &= \left\|\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^\mathsf{T}\mathbf{U}_{\tilde{A}}}\right)\tilde{\mathbf{M}}\right\|_2^2 \\ &\overset{(a)}{=} \left\|\left(\mathbf{I} - \mathbf{P}_{\ddot{A}}\right)\tilde{\mathbf{M}}\right\|_2^2, \quad (21)\end{aligned}$$

where (a) holds because $\ddot{\mathbf{A}} = \tilde{\mathbf{M}}\mathbf{\Psi}\mathbf{R} = \mathbf{U}^\mathsf{T}\mathbf{M}\mathbf{\Psi}\mathbf{R} = \mathbf{U}^\mathsf{T}\mathbf{A}\mathbf{R} = \mathbf{U}^\mathsf{T}\tilde{\mathbf{A}}$. And $\mathbf{P}_{\ddot{A}} = \mathbf{U}^\mathsf{T}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\dagger\mathbf{U} = \mathbf{U}^\mathsf{T}\mathbf{U}_{\tilde{A}}\mathbf{U}_{\tilde{A}}^\mathsf{T}\mathbf{U}$. In view of (21), it suffices to obtain an upper bound of $\left\|\left(\mathbf{I} - \mathbf{P}_{\ddot{A}}\right)\tilde{\mathbf{M}}\right\|_2^2$ to bound $\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2$. First, we whiten out the top block of $\ddot{\mathbf{A}}$ in (20) to obtain the matrix

$$\mathbf{W} = \ddot{\mathbf{A}} \cdot (\mathbf{\Psi}_1\mathbf{R})^\dagger \mathbf{\Sigma}_1^{-1} = \begin{bmatrix}\mathbf{I} \\ \mathbf{F}\end{bmatrix}, \quad \text{where } \mathbf{F} = \mathbf{\Sigma}_2\mathbf{\Psi}_2\mathbf{R}(\mathbf{\Psi}_1\mathbf{R})^\dagger\mathbf{\Sigma}_1^{-1}. \quad (22)$$

This construction ensures that range$(\mathbf{W}) \subset$ range$(\ddot{\mathbf{A}})$, so Proposition 4 implies that the error satisfies $\left\|\left(\mathbf{I} - \mathbf{P}_{\ddot{A}}\right)\tilde{\mathbf{M}}\right\|_2 \leq \left\|\left(\mathbf{I} - \mathbf{P}_W\right)\tilde{\mathbf{M}}\right\|_2$. Squaring each side, we obtain

$$\begin{aligned}\left\|\left(\mathbf{I} - \mathbf{P}_{\ddot{A}}\right)\tilde{\mathbf{M}}\right\|_2^2 &\leq \left\|\left(\mathbf{I} - \mathbf{P}_W\right)\tilde{\mathbf{M}}\right\|_2^2 \\ &= \left\|\tilde{\mathbf{M}}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\tilde{\mathbf{M}}\right\|_2 \overset{(a)}{=} \left\|\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\mathbf{\Sigma}\right\|_2\end{aligned}$$

where (a) follows from the definition $\tilde{\mathbf{M}} = \mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ and the unitary invariance of the spectral norm. We next bound $\left\|\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\mathbf{\Sigma}\right\|_2$. We provide a detailed representation of the projector $\mathbf{I} - \mathbf{P}_W$. The construction in (22) ensures that $\mathbf{W}$ has full column rank, so we can apply the formula (7) for an orthogonal projector to see that

$$\mathbf{P}_W = \mathbf{W}\left(\mathbf{W}^\mathsf{T}\mathbf{W}\right)^{-1}\mathbf{W}^\mathsf{T} = \begin{bmatrix}\mathbf{I} \\ \mathbf{F}\end{bmatrix}\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1}\begin{bmatrix}\mathbf{I} \\ \mathbf{F}\end{bmatrix}^\mathsf{T}.$$

Expanding the above expression, we have the following complementary projector with respect to $\mathbf{W}$,

$$\mathbf{I} - \mathbf{P}_W = \begin{bmatrix}\mathbf{I} - \left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1} & -\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)\mathbf{F}^\mathsf{T} \\ -\mathbf{F}\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1} & \mathbf{I} - \mathbf{F}\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1}\mathbf{F}^\mathsf{T}\end{bmatrix}.$$

The partitioning in above conforms the dimension of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in the partitioning of $\mathbf{\Sigma}$ in (13). Proposition 5 (see Appendix B) shows that the top-left block satisfies $\mathbf{I} - \left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1} \preccurlyeq \mathbf{F}^\mathsf{T}\mathbf{F}$. The bottom-right block satisfies $\mathbf{I} - \mathbf{F}\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1}\mathbf{F}^\mathsf{T} \preccurlyeq \mathbf{I}$, this is due to the conjugation rule that guarantees that $\mathbf{0} \preccurlyeq \mathbf{F}\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)^{-1}\mathbf{F}^\mathsf{T}$. Let us denote the off-diagonal blocks as $\mathbf{B} = -\left(\mathbf{I} + \mathbf{F}^\mathsf{T}\mathbf{F}\right)\mathbf{F}^\mathsf{T}$. Then, $\mathbf{I} - \mathbf{P}_W \preccurlyeq \begin{bmatrix}\mathbf{F}^\mathsf{T}\mathbf{F} & \mathbf{B} \\ \mathbf{B}^\mathsf{T} & \mathbf{I}\end{bmatrix}$. Now, we bound $\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\mathbf{\Sigma}$ as follows,

$$\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\mathbf{\Sigma} \preccurlyeq \begin{bmatrix}\mathbf{\Sigma}_1^\mathsf{T}\mathbf{F}^\mathsf{T}\mathbf{F}\mathbf{\Sigma}_1 & \mathbf{\Sigma}_1^\mathsf{T}\mathbf{B}\mathbf{\Sigma}_2 \\ \mathbf{\Sigma}_2^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{\Sigma}_1 & \mathbf{\Sigma}_2^\mathsf{T}\mathbf{\Sigma}_2\end{bmatrix}.$$

The conjugation rule shows that the matrix on the left-hand side is positive semi definite, so the matrix on the right-hand side is as well. Proposition 6 results from the norm bound

$$\begin{aligned}\left\|\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{I} - \mathbf{P}_W\right)\mathbf{\Sigma}\right\|_2 &\leq \left\|\mathbf{\Sigma}_1^\mathsf{T}\mathbf{F}^\mathsf{T}\mathbf{F}\mathbf{\Sigma}_1\right\|_2 + \left\|\mathbf{\Sigma}_2^\mathsf{T}\mathbf{\Sigma}_2\right\|_2 \\ &= \left\|\mathbf{F}\mathbf{\Sigma}_1\right\|_2^2 + \left\|\mathbf{\Sigma}_2\right\|_2^2. \quad (23)\end{aligned}$$

By plugging $\mathbf{F} = \mathbf{\Sigma}_2\mathbf{\Psi}_2\mathbf{R}(\mathbf{\Psi}_1\mathbf{R})^\dagger\mathbf{\Sigma}_1^{-1}$ in (23), we have

$$\begin{aligned}&\left\|\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\right)\mathbf{M}\right\|_2^2 \\ &\leq \left\|\mathbf{\Sigma}_2\mathbf{\Psi}_2\mathbf{R}(\mathbf{\Psi}_1\mathbf{R})^\dagger\right\|_2^2 + \left\|\mathbf{\Sigma}_2\right\|_2^2 \\ &\overset{(a)}{\leq} \sigma_{l+1}^2(\mathbf{M}) + \left\|\mathbf{\Sigma}_2\right\|_2^2\left\|\mathbf{\Psi}_2\mathbf{R}\right\|_2^2\left\|(\mathbf{\Psi}_1\mathbf{R})^\dagger\right\|_2^2 \\ &\leq \sigma_{l+1}^2(\mathbf{M}) + \sigma_{l+1}^2(\mathbf{M})\left\|\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_2^2\left\|(\mathbf{\Psi}_1\mathbf{R})^\dagger\right\|_2^2, \quad (24)\end{aligned}$$

where (a) follows the matrix norm inequality. We bound $\left\|\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_2^2$ as follows. Let $\mathbf{x}_i^\mathsf{T}$ be i th row of $\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}$ and $\mathbf{y}_i^\mathsf{T}$ be i th row of $\mathbf{V}_2^\mathsf{T}$, where $1 \leq i \leq m-l$.

$$\begin{aligned}\left\|\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_2^2 \leq \left\|\mathbf{V}_2^\mathsf{T}\mathbf{\Psi}\mathbf{R}\right\|_F^2 &\overset{(a)}{=} \sum_{i=1}^{m-l}\left\|\mathbf{x}_i^\mathsf{T}\mathbf{R}\right\|_2^2 \\ &\overset{(b)}{\leq} \sum_{i=1}^{m-l}(1+\epsilon)\left\|\mathbf{x}_i^\mathsf{T}\right\|_2^2 \\ &\overset{(c)}{\leq} \sum_{i=1}^{m-l}(1+\epsilon)\left\|\mathbf{y}_i^\mathsf{T}\right\|_2^2 \\ &\overset{(d)}{=} (1+\epsilon)(m-l), \quad (25)\end{aligned}$$

where (a) follows from the definition of the Frobenius norm. (b) follows Johnson-Lindenstrauss Transform in Definition 1.

And (c) is due to the column sampling operator $\mathbf{\Psi}$. Finally, (d) follows from $\left\|\mathbf{y}_i^{\mathsf{T}}\right\|_2^2 = 1$. Plugging (25) in (24) and (19), we complete the proof of Lemma 2. ⊠

APPENDIX F
LEMMA 3 AND PROOF OF LEMMA 3

We present Lemma 3 and its proof. Lemma 3 provides a bound on the gap between the row-space of $\mathbf{E}$ and that of $\mathbf{S}$.

Lemma 3. *With probability $1 - c_1 l^{-3}$, and if $d \geq c_2 \mu_s l \ln l$, for constants $c_1 > 0$ and $c_2 > 0$, we have*

$$\left\|\mathbf{E}\left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right)\right\|_2^2 \leq \sigma_2^2(\mathbf{E})\left(1 + \frac{2m}{d}\right)$$

Proof: ⊠

$$\begin{aligned}
&\left\|\mathbf{E}\left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right)\right\|_2^2 \\
&\overset{(a)}{\leq} \|\mathbf{E}\|_2^2 \left\|\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&\overset{(b)}{\leq} \|\mathbf{E}\|_2^2 \left(1 + \left\|\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2\right) \\
&\overset{(c)}{\leq} \|\mathbf{E}\|_2^2 \left(1 + \left\|\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\right\|_2^2\right), \qquad (26)
\end{aligned}$$

where (a) is due to matrix inequality and (b) and (c) are due to $\|\mathbf{I}\|_2^2 = 1$, $\|\mathbf{\Psi}\|_2^2 = 1$ and $\left\|\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 = 1$.

Now, we bound $\left\|\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\right\|_2^2$ in (26). Note that $\left\|\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\right\|_2^2 = 1/\sigma_{min}^2\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right) = 1/\lambda_{min}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S\right)$. Thus, we need to bound the minimum eigenvalue of $\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S$, where $\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi} \in \mathbb{R}^{l\times d}$, $l \leq d$, is full rank. We let $\bar{\mathbf{v}}_i$, where $i \in [m]$, be the ith column vector of $\mathbf{V}_S^{\mathsf{T}} \in \mathbb{R}^{l\times m}$. Then, we have $\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S = \sum_{j=1}^{d} \bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}$. It is straightforward to show that

$$\mathbb{E}\left[\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S\right] = \frac{d}{m}\mathbf{I}_l \text{ and } \mathbb{E}\left[\bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}\right] = \frac{1}{m}\mathbf{I}_l. \qquad (27)$$

To bound the minimum eigenvalue of $\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S$, we leverage Theorem 2 (see Appendix B) and Definition 3, where we first need to bound the maximum eigenvalue of $\bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}$, which is a rank-1 matrix, whose eigenvalue

$$\max_{1\leq i\leq m}\lambda_{max}\left(\bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}\right) = \max_{1\leq i\leq m}|\bar{\mathbf{v}}_i|^2 \leq \mu_s\frac{l}{m} \qquad (28)$$

and

$$\lambda_{max}\left(\mathbb{E}\left[\bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}\right]\right) = \lambda_{min}\left(\mathbb{E}\left[\bar{\mathbf{v}}_{i_j}\bar{\mathbf{v}}_{i_j}^{\mathsf{T}}\right]\right) = \frac{1}{m}. \qquad (29)$$

Thus, we have,

$$\begin{aligned}
&P\left[\lambda_{min}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S\right) \leq (1-\rho)\frac{d}{m}\right] \\
&\leq l\cdot\exp\frac{-d/m}{l\mu_s/m}[(1-\rho)\ln(1-\rho)+\rho] \text{ for } \rho\in[0,1) \qquad (30)\\
&= l\cdot\exp\frac{-d}{l\mu_s}[(1-\rho)\ln(1-\rho)+\rho]. \qquad (31)
\end{aligned}$$

By setting $\rho = 1/2$, we have $P\left[\lambda_{min}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S\right) \leq \frac{d}{2m}\right] \leq l\,\exp\frac{-d}{7l\mu_s} = l\,\exp^{-d/7l\mu_s}$. This expression can be algebraically manipulated, such that with a probability $1 - l^{-c_1}$, where $c_1 > 0$ is a constant, if $d \geq c_2\mu_s l\ln\, l$, for a constant $c_2 > 0$, we have

$$\lambda_{min}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\mathbf{V}_S\right) \geq \frac{d}{2m}. \qquad (32)$$

Finally, by plugging (32) into (26), we obtain Lemma 3. ⊠

APPENDIX G
PROOF OF PROPOSITION 2

Proposition 2. If $d \geq \max\{c_1\mu l\ln l, c_2\mu_s l\ln l\}$, and $p = O(\log(1/\delta)/\epsilon^2)$, where $\delta \doteq \max\{e^{-c_3\epsilon^2 p}, c_4 l^{-3}\}$, under the condition of Theorem 1, we have,

$$\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&\leq 4\sigma_{l+1}^2(\mathbf{M})\left(2 + \frac{2n}{d} + \frac{(1+\epsilon)(m-l)}{\sigma_1^2\left(\mathbf{V}_1^{\mathsf{T}}\mathbf{\Psi}\mathbf{R}\right)}\right) \\
&\quad + 4\sigma_2^2(\mathbf{E})\left(2 + \frac{2m}{d}\right) \qquad (33)
\end{aligned}$$

with probability $1 - \max\{e^{-c_3\epsilon^2 p}, c_4 l^{-3}\}$ where c_1, c_2, c_3, c_4 and $\epsilon > 0$ are constants. ⊠

Proof:

$$\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S} - \tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&= \left\|\mathbf{P}_{\mathbf{U}_A}(\mathbf{Q}\mathbf{S}+\mathbf{E})\mathbf{P}_{\mathbf{V}_S} - \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\dagger}\mathbf{A}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&\overset{(a)}{=} \left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{Q}\mathbf{S} + \mathbf{P}_{\mathbf{U}_A}\mathbf{E}\mathbf{P}_{\mathbf{V}_S} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{A}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&\overset{(b)}{=} \|\mathbf{P}_{\mathbf{U}_A}\mathbf{Q}\mathbf{S} + \mathbf{P}_{\mathbf{U}_A}\mathbf{E}\mathbf{P}_{\mathbf{V}_S} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{S}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}} \\
&\quad - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\|_2^2, \qquad (34)
\end{aligned}$$

where (a) is followed by the ρ-SVD of $\tilde{\mathbf{A}}$ defined in (15), i.e., $\mathbf{P}_{\mathbf{U}_{\tilde{A}}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\dagger} = \mathbf{U}_{\tilde{A}}\mathbf{U}_{\tilde{A}}^{\top}$. And (b) is due to $\mathbf{A} = \mathbf{M}\mathbf{\Psi} = (\mathbf{Q}\mathbf{S}+\mathbf{E})\mathbf{\Psi}$. Then, $\mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{S}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}$ in (34) can be expressed as

$$\begin{aligned}
\mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{S}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}} &= \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}} \\
&= \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{U}_S\mathbf{\Sigma}_S\mathbf{V}_S^{\mathsf{T}} = \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{Q}\mathbf{S}. \qquad (35)
\end{aligned}$$

We plug (35) in (34), and continue bounding (34) as

$$\begin{aligned}
&\left\|\left(\mathbf{P}_{\mathbf{U}_A} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\right)\mathbf{Q}\mathbf{S} + \mathbf{P}_{\mathbf{U}_A}\mathbf{E}\mathbf{P}_{\mathbf{V}_S} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2 \\
&= \|\left(\mathbf{P}_{\mathbf{U}_A} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\right)(\mathbf{M}-\mathbf{E}) + \mathbf{P}_{\mathbf{U}_A}\mathbf{E}\mathbf{P}_{\mathbf{V}_S} \\
&\qquad - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\|_2^2 \\
&= \|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{E}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_S}) \qquad (36)\\
&\qquad + \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right)\|_2^2 \\
&\overset{(a)}{\leq} 2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2 \\
&\qquad + 2\left\|\mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right) - \mathbf{P}_{\mathbf{U}_A}\mathbf{E}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_S})\right\|_2^2 \\
&\overset{(b)}{\leq} 2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2 + 4\left\|\mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{E}\left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right)\right\|_2^2 \\
&\qquad + 4\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{E}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_S})\right\|_2^2
\end{aligned}$$

$$\overset{(c)}{\leq} \underbrace{2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}-\mathbf{P}_{\mathbf{U}_{\tilde{A}}}\mathbf{M}\right\|_2^2}_{\odot_1}+\underbrace{4\left\|\mathbf{E}\left(\mathbf{I}-\mathbf{\Psi}\left(\mathbf{V}_S^{\mathsf{T}}\mathbf{\Psi}\right)^{\dagger}\mathbf{V}_S^{\mathsf{T}}\right)\right\|_2^2}_{\odot_2} + 4\left\|\mathbf{E}\right\|_2^2, \tag{37}$$

where (a) and (b) are due to matrix norm inequality and (c) is because of the matrix norm inequality and $\|\mathbf{P}_{\mathbf{U}_A}\|_2^2 = 1$ and $\|\mathbf{I}-\mathbf{P}_{\mathbf{V}_S}\|_2^2 = 1$. Using Lemma 2 and Lemma 3 to bound $\odot_1$ and $\odot_2$ each, we finally have,

$$\begin{aligned}&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\mathbf{V}_S}-\tilde{\mathbf{A}}\mathbf{Z}\mathbf{V}_S^{\mathsf{T}}\right\|_2^2\\ &\leq 4\sigma_{l+1}^2\left(\mathbf{M}\right)\left(2+\frac{2n}{d}+\left(1+\epsilon\right)\left(m-l\right)\left\|\left(\mathbf{V}_1^{\mathsf{T}}\mathbf{\Psi}\mathbf{R}\right)^{\dagger}\right\|_2^2\right)\\ &\quad+4\sigma_2^2\left(\mathbf{E}\right)\left(2+\frac{2m}{d}\right)\end{aligned} \tag{38}$$

REFERENCES

[1] I. Markovsky, "Structured low-rank approximation and its applications," *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

[2] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Streaming low-rank matrix approximation with an application to scientific simulation," *SIAM Journal on Scientific Computing*, vol. 41, no. 4, pp. A2430–A2463, 2019.

[3] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, 2018.

[4] 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*, vol. 104, no. 51, pp. 20 167–20 172, 2007.

[5] D. Achlioptas and F. McSherry, "Fast computation of low-rank matrix approximations," *Journal of the ACM (JACM)*, vol. 54, no. 2, pp. 9–es, 2007.

[6] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Practical sketching algorithms for low-rank matrix approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1454–1485, 2017.

[7] S. J. Quiton, J. Chae, S. Bac, K. Kron, U. Mitra, and S. M. Sharada, "Toward efficient direct dynamics studies of chemical reactions: A novel matrix completion algorithm," *Journal of Chemical Theory and Computation*, 2022.

[8] J. Chae, P. Narayanamurthy, S. Bac, S. M. Sharada, and U. Mitra, "Matrix approximation with side information: When column sampling is enough," *IEEE Transactions on Signal Processing*, vol. 72, pp. 2276–2291, 2024.

[9] M. Xu, R. Jin, and Z.-H. Zhou, "Cur algorithm for partially observed matrices," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1412–1421.

[10] K. Hamm and L. Huang, "Perturbations of cur decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 42, no. 1, pp. 351–375, 2021.

[11] J. Chae, P. Narayanamurthy, S. Bac, S. M. Sharada, and U. Mitra, "Sketched column-based matrix approximation with side information," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 9816–9820.

[12] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, "Matrix completion with noisy side information," *Advances in neural information processing systems*, vol. 28, 2015.

[13] M. Xu, R. Jin, and Z.-H. Zhou, "Speedup matrix completion with side information: Application to multi-label learning," *Advances in neural information processing systems*, vol. 26, 2013.

[14] A. Desai, M. Ghashami, and J. M. Phillips, "Improved practical matrix sketching with guarantees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1678–1690, 2016.

[15] F. Ban, D. Woodruff, and R. Zhang, "Regularized weighted low rank approximation," *Advances in neural information processing systems*, vol. 32, 2019.

[16] A. Ma, D. Stöger, and Y. Zhu, "Robust recovery of low-rank matrices and low-tubal-rank tensors from noisy sketches," *SIAM Journal on Matrix Analysis and Applications*, vol. 44, no. 4, pp. 1566–1588, 2023.

[17] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, "A fast randomized algorithm for the approximation of matrices," *Applied and Computational Harmonic Analysis*, vol. 25, no. 3, pp. 335–366, 2008.

[18] M. W. Mahoney and P. Drineas, "Cur matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[19] G. Strang, "The discrete cosine transform," *SIAM review*, vol. 41, no. 1, pp. 135–147, 1999.

[20] N. Kingsbury, "Image processing with complex wavelets," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1760, pp. 2543–2560, 1999.

[21] J. Chae, P. Narayanamurthy, S. Bac, S. M. Sharada, and U. Mitra, "Column-based matrix approximation with quasi-polynomial structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] S. Bac, S. J. D. Quiton, K. J. Kron, J. Chae, U. Mitra, and S. Mallikarjun Sharada, "A matrix completion algorithm for efficient calculation of quantum and variational effects in chemical reactions," *The Journal of Chemical Physics*, Apr. 2022, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/5.0091155

[23] S. J. Quiton, J. Chae, S. Bac, K. Kron, U. Mitra, and S. Mallikarjun Sharada, "PVMC," Mar. 2022, (accessed 2022-03-07). [Online]. Available: https://github.com/RateTheory/PVMC

[24] W. B. Johnson, J. Lindenstrauss *et al.*, "Extensions of lipschitz mappings into a hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[25] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[26] E. J. Candès and B. Recht, "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, Dec. 2009. [Online]. Available: http://link.springer.com/10.1007/s10208-009-9045-5

[27] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[28] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[29] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[30] G. W. Stewart, *Perturbation theory for the singular value decomposition*. Citeseer, 1998.

[31] J. A. Tropp, "Improved analysis of the subsampled randomized hadamard transform," *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.

[32] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Practical sketching algorithms for low-rank matrix approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1454–1485, 2017.

[33] M. Azizyan, A. Krishnamurthy, and A. Singh, "Extreme compressive sampling for covariance estimation," *IEEE Transactions on Information Theory*, vol. 64, no. 12, pp. 7613–7635, 2018.

[34] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear algebra and its applications*, vol. 415, no. 1, pp. 20–30, 2006.

[35] M. Xu, R. Jin, and Z.-H. Zhou, "Supplementary of cur algorithm for partially observed matrices," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1412–1421.

[36] J. Makhoul, "A fast cosine transform in one and two dimensions," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 27–34, 1980.

[37] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[38] P.-G. Martinsson, V. Rokhlin, and M. Tygert, "A randomized algorithm for the decomposition of matrices," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 47–68, 2011.

[39] K. L. Clarkson and D. P. Woodruff, "Low-rank approximation and regression in input sparsity time," *Journal of the ACM (JACM)*, vol. 63, no. 6, pp. 1–45, 2017.

[40] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast johnson-lindenstrauss transform," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 2006, pp. 557–563.